Poznan University of Technology
Faculty of Computing and Telecommunications

Doctoral dissertation

Computer-aided decision support methods resistant to imperfections of learning data

Grzegorz Miebs

Supervisor: Miłosz Kadziński, Ph.D., Habil.

Poznań, 2023

Abstract

Due to the rapid increase in data availability and ubiquitous computing, computer-based decision support systems are becoming more common. Unfortunately, together with their popularity, some improper usages arise. Imperfect learning data is one of the main reasons behind this problem. Those imperfections might originate from low-quality data, measurement errors, and human errors, such as a lack of knowledge about which method and how it should be used. Classical algorithms are not designed to work with such data. This doctoral thesis addressed this problem by introducing new computer-based decision support algorithms robust to imperfect learning data. The novel methods handled problems related to incomplete preference information expressed as example decisions. New methods for robust weight calculations and group decision-making were proposed as well. Additionally, two new architectures were designed to use neural networks in robust decision-making. Finally, a system recommending an adequate procedure for a given decision problem was introduced. All proposed algorithms were used to solve real-world decision problems, proving their practical usability.

List of publications

The dissertation consists of the introductory section and the following ten original publications:

[P1] M. Kadziński, L. Rocchi, G. Miebs, D. Grohmann, M. E. Menconi, and L. Paolotti. Multiple criteria assessment of insulating materials with a group decision framework incorporating outranking preference model and characteristic class profiles. *Group Decision and Negotiation*, 27 (1):33–59, Nov. 2017. doi: 10.1007/s10726-017-9549-3. URL `https://doi.org/10.1007/s10726-017-9549-3`, DOI: `10.1007/s10726-017-9549-3`.

Number of citations[1]:

- according to Web of Science: 19
- according to Google Scholar: 29

[P2] L. Rocchi, M. Kadziński, M. Menconi, D. Grohmann, G. Miebs, L. Paolotti, and A. Boggia. Sustainability evaluation of retrofitting solutions for rural buildings through life cycle approach and multi-criteria analysis. *Energy and Buildings*, 173:281–290, Aug. 2018. doi: 10.1016/j.enbuild.2018.05.032. URL `https://doi.org/10.1016/j.enbuild.2018.05.032`, DOI: `10.1016/j.enbuild.2018.05.032`.

Number of citations[1]:

- according to Web of Science: 40
- according to Google Scholar: 56

[P3] K. Govindan, M. Kadziński, R. Ehling, and G. Miebs. Selection of a sustainable third-party reverse logistics provider based on the robustness analysis of an outranking graph kernel conducted with ELECTRE i and SMAA. *Omega*, 85:1–15, June 2019. doi: 10.1016/j.omega.2018.05.007. URL `https://doi.org/10.1016/j.omega.2018.05.007`, DOI: `10.1016/j.omega.2018.05.007`.

Number of citations[1]:

[1] as on September 25, 2023

- according to Web of Science: 95
- according to Google Scholar: 137

[P4] G. Miebs, M. Mochol-Grzelak, A. Karaszewski, and R. A. Bachorz. Efficient strategies of static features incorporation into the recurrent neural network. *Neural Processing Letters*, 51(3):2301–2316, Jan. 2020. doi: 10.1007/s11063-020-10195-x. URL `https://doi.org/10.1007/s11063-020-10195-x`, DOI: `10.1007/s11063-020-10195-x`.

Number of citations[2]:

- according to Web of Science: 7
- according to Google Scholar: 8

[P5] A. Oppio, M. Dell'Ovo, F. Torrieri, G. Miebs, and M. Kadziński. Understanding the drivers of urban development agreements with the rough set approach and robust decision rules. *Land Use Policy*, 96: 104678, July 2020. doi: 10.1016/j.landusepol.2020.104678. URL `https://doi.org/10.1016/j.landusepol.2020.104678`, DOI: `10.1016/j.landusepol.2020.10467`

Number of citations[2]:

- according to Web of Science: 10
- according to Google Scholar: 14

[P6] G. Miebs and M. Kadziński. Heuristic algorithms for aggregation of incomplete rankings in multiple criteria group decision making. *Information Sciences*, 560:107–136, June 2021. doi: 10.1016/j.ins.2021.01.055. URL `https://doi.org/10.1016/j.ins.2021.01.055`, DOI: `10.1016/j.ins.2021.01.055`.

Number of citations[2]:

- according to Web of Science: 9
- according to Google Scholar: 14

[P7] M. Cinelli, M. Kadziński, G. Miebs, M. Gonzalez, and R. Słowiński. Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system. *European Journal of Operational Research*, 302(2):633–651, Oct. 2022. doi: 10.1016/j.ejor.2022.01.011. URL `https://doi.org/10.1016/j.ejor.2022.01.011`, DOI: `10.1016/j.ejor.2022.01.011`. Number of citations[2]:

- according to Web of Science: 31
- according to Google Scholar: 62

[2] as on September 25, 2023

[P8] M. Jurczak, G. Miebs, and R. A. Bachorz. Multi-criteria human resources planning optimisation using genetic algorithms enhanced with MCDA. *Operations Research and Decisions*, 32(4), 2022. doi: 10.37190/ord220404. URL https://doi.org/10.37190/ord220404, DOI: 10.37190/ord220404.

[P9] G. Miebs, M. Wójcik, A. Karaszewski, M. Mochol-Grzelak, P. Wawdysz, and R. A. Bachorz. Predicting a time-dependent quantity using recursive generative query network. *International Journal of Neural Systems*, 32(11), Oct. 2022. doi: 10.1142/s0129065722500563. URL https://doi.org/10.1142/s0129065722500563, DOI: 10.1142/S0129065722500563.

[P10] J. Gehrlein, G. Miebs, M. Brunelli, and M. Kadziński. An active preference learning approach to aid the selection of validators in blockchain environments. *Omega*, 118:102869, July 2023. doi: 10.1016/j.omega.2023.102869. URL https://doi.org/10.1016/j.omega.2023.102869, DOI: 10.1016/j.omega.2023.102869. Number of citations[3]:

- according to Google Scholar: 1

[3] as on September 25, 2023

Contents

Chapter 1

Introduction

The field of Multiple Criteria Decision Analysis (MCDA) tackles problems with alternatives described by at least two criteria reflecting their different, usually contradicting, characteristics. In such a scenario, most alternatives are incomparable, forming the so-called Pareto Front consisting of non-dominated alternatives. To effectively solve this class of problems, additional preference information reflecting the value system of a Decision Maker (DM) is needed. This information, together with alternatives' descriptions, is used to build a mathematical model, which is later exploited to provide a recommendation.

MCDA distinguishes three main types of decision problems: (a) choice, where a subset of the most preferred alternatives have to be selected; (b) sorting or classification, where alternatives have to be assigned to predefined classes; (c) ranking, where alternatives are to be ordered from the best to the worst. In addition, a clustering problem can be considered where alternatives have to be divided into groups based on similarities and patterns. Those types of problems can be observed in many areas like economy [Greco et al., 1998], cybersecurity [Ganin et al., 2017], chemistry [Palacios et al., 2021], or logistics [Govindan et al., 2017].

Preference information acquired from a DM can take different forms. The most popular are weights of each criterion [Németh et al., 2019], pairwise comparisons [Siskos et al., 2016], and assignments for a subset of alternatives to predefined classes [Pawlak, 1982]. The process of obtaining preference information is not error-proof. There are multiple sources of possible data imperfection [Parsons, 1996] like measurement error, human error, or process error. What is more, at this stage, imperfection can have different forms, such as lack of precise information, distorted value, or conflicting preferences. Each type of imperfection requires a different form of tackling the problem.

Apart from preference information, other types of input are usually needed. Some algorithms require specifying their metaparameters. Also, the choice

of the MCDA algorithm is an input. Both input types can be a source of imperfection, affecting the whole decision-making process. As such, they also require special treatment to lower the chances of introducing imperfection or to minimize its impact.

This dissertation is motivated by the observed unsatisfied needs of real-world problems. Most well-known MCDA methods are suitable only for theoretical problems but cannot deal with imperfect data, which is unavoidable in most real applications [Parsons, 1996]. There are different sources of data imperfection like measurement error, imprecise data available (e.g., in the form of ranges), or imperfect preference information coming from a DMs. Each type of imperfection requires different treatment. What is more, sometimes, it is possible to eliminate the source of imperfections.

The dissertation presents the theoretical backgrounds of developed algorithms as well as their applications to real-world use cases. The studies aimed to provide new methods suitable for application to real complex decision-making problems and, therefore, increase the usability and popularity of MCDA.

The remainder of this doctoral thesis is organized in the following way. Chapter 2 discusses the MCDA basic concepts together with methods and their applications. Chapter 3 describes proposed algorithms robust to learning data imperfections. Chapter 4 summarizes the thesis.

Chapter 2

Decision support

This chapter describes MCDA, explains basic theory, introduces the most popular algorithms that served as the base for this dissertation, and whose modifications are presented in Chapter 3.

2.1 Decision problems

Decision problems in MCDA involve a set of n alternatives A described on m criteria. Alternatives reflect objects or actions considered in a given decision-making problem. Criteria are used to evaluate their performance. The evaluation can be deterministic when a precise scalar value is provided or uncertain with a probability distribution.

Criteria can be of a gain type (the greater the performance, the better) or of a cost type (the lesser the performance, the better). Formally, they can be represented as functions g_j defined on the set of alternatives A, with $g_j(a)$ denoting the performance of alternative $a \in A$ on criterion $j, 1 \leq j \leq m$. On a cost-type criterion function, g is non-increasing and respectively non-decreasing for a gain-type criterion. Different scales can be used to express performance on criteria such as ordinal, interval, or ratio.

One or more DMs take part in a decision problem. Capturing the preference of a DM and involving them in working out the recommendation is one of the most important parts of the MCDA. This phase gets even more complex when there is more than one DM, and they have conflicting preferences.

2.2 Preference information

To work out a recommendation in a way proposed by most MCDA methods, additional preference information reflecting the value system of DMs is required. This information is later used to build a mathematical model con-

sistent with a DM's preferences. Such a model can be exploited to produce a recommendation.

The structure of the preference information varies between methods. One of the most popular ways to define preferences is to associate a weight with each criterion. This vector of weights can be the only information provided by a DM, like in a TOPSIS [Hwang and Yoon, 1981] method, or can be supplemented with additional information, like in ELECTRE Roy [1968] methods. Weights can be provided as precise numbers or, e.g., intervals. Also, methods like the revised Simos (SRF) procedure [Figueira and Roy, 2002] are used where much simpler information in the form of a ranking of criteria is transformed to derive a weight vector.

Providing examples of decisions is a different approach to acquiring preference information. Depending on the context, this way of providing preference data might be simpler for a DM. Often, those examples are already available as historical data, or DMs had to solve the decision problem manually without any support. Thus, they are capable of providing such information. For sorting/classification problems, such examples would be assignment of reference alternatives to predefined classes. Another form of preference information of a similar type would be defining relations between reference alternatives. It can take the form of pairwise comparisons or defining a ranking.

The concept of thresholds is used by two prominent families of MCDA methods, namely ELECTRE Roy [1968] and PROMETHEE Brans and Mareschal [2005]. Three types of thresholds are most common. Indifference threshold denotes maximal negligible performance difference. If alternatives' performances differ by no more than this value, then they are treated as indifferent. A preference threshold is used to specify a minimal difference, justifying a strict preference. If one alternative is better by no less than this threshold, then it is rigidly preferred. Those two thresholds are called intracriteria ones because they affect evaluation within a single criterion. The last of three main thresholds, called a veto threshold, is an inter-criteria threshold since it affects an overall evaluation of an alternative. The veto threshold denotes a minimal difference invalidating preference. If one alternative is worse than the other by at least the veto threshold, then it cannot be preferred no matter how good is its performance on the remaining criteria.

No matter the structure of the preference information, the aim is the same, i.e., acquiring data about the value system of a DM to build a model that will aid the decision-making process respecting the user's preferences.

2.3 Methods and applications

Through the years, hundreds of MCDA algorithms have been developed. However, many of them can be assigned to one of a few groups of methods.

Most methods are just extensions of previously existing ones and share with them the same concept.

ELECTRE

ELECTRE method introduced in Roy [1968] was a subject of many modifications and extensions. Currently, variants from this family of algorithms can be used to solve all three main types of decision problems: choice, sorting, and ranking. This method is based on an outranking relation, denoted by S, where aSb means that alternative a is no worse than b. From outranking, three different relations can be inferred [Roy, 1991]. When aSb and bSA both alternatives are indifferent I, if $\neg aSb$ and $\neg bSa$ then they are incomparable ?, if aSb and $\neg bSa$ then a is preferred P, and b is preferred when $\neg aSb$ and bSa.

To calculate the outranking relation apart from the evaluation of alternatives $a \in A$ on criteria $g \in G$, additional preference information in the form of weights w_i as well as values of thresholds associated with each criterion are required. Typically, ELECTRE methods use three thresholds: indifference q_i, preference p_i, and veto v_i.

As the first step, the marginal concordance function $c_j(a, b)$ is calculated on each criterion. This value denotes how data supports a outranking b. The value is equal to one if a is not worse than b by more than the indifference threshold ($g_j(a) - g_j(b) \geq -q_j$) and equal to zero when a is worse than b by more than the preference threshold ($g_j(a) - g_j(b) < -p_j$). Between indifference and preference thresholds value of $c_j(a, b)$ changes in a linear manner ($\frac{g_j(a)-g_j(b)+p_j}{p_j-q_j}$). Then the comprehensive concordance index $C(a, b) = \frac{\sum_{j=1}^{m} w_j \cdot c_j(a,b)}{\sum_{j=1}^{m} w_j}$ is calculated as a weighted average of all marginal concordance functions. This value denotes the strength of a outranking b on all criteria.

In opposition to the marginal concordance function $c_j(a, b)$, a marginal discordance function $d_j(a, b)$, which presents the strength of negation of a outranking b is calculated. The value is equal to one if a is worse than b by more than the veto threshold ($g_j(a) - g_j(b) < -v_j$) and equal to zero when a is not worse than b by more than the preference threshold ($g_j(a) - g_j(b) \geq -p_j$). Between indifference and preference thresholds value of $c_j(a, b)$ changes in a linear manner ($\frac{g_j(b)-g_j(a)-p_j}{v_j-p_j}$). This value is later used to lower the comprehensive concordance index, e.g., by multiplication.

Due to a high number of developed variants [Figueira et al., 2016, Cinelli et al., 2022], a proper ELECTRE method can be found for various types of decision problems with their constraints. ELECTRE methods have been widely applied in domains like finances [Emamat et al., 2022], transport and logistics [Jurczak et al., 2022, Govindan et al., 2019], or sustainability

[Kadziński et al., 2017, Rocchi et al., 2018].

UTA

UTA method [Jacquet-Lagreze and Siskos, 1982] is based on a Multiple Attribute Value Theory (MAVT) [Keeney and Raiffa, 1993]. The main idea of this approach is to define a marginal value function u_j for each criterion mapping performance on a given criterion to usability [Siskos et al., 2016]. Then for each alternative $a \in A$, a comprehensive value $U(a)$ is calculated as a sum of marginal values $U(a) = \sum_{j=1}^{n} u_j(a)$. Such a function can be exploited to compare alternatives since an alternative with a higher value is preferred to an alternative with a lower value of U $U(a) \geq U(b) \quad \Leftrightarrow \quad a \succsim b \quad \forall a, b \in A$, where $a_i \succsim a_k$ means that a is at least as good or weakly preferred to b.

In UTA methods, the marginal value functions are usually piecewise linear interpolations of characteristic points. The function is non-decreasing for gain type criteria and non-increasing for cost type criteria. Typically, in UTA methods, this knowledge is provided in the form of pairwise alternative comparison. DM provides pairs of alternatives with information on which one is preferred or if they are indifferent. This information is reflected in the utility functions, which are obtained using the linear programming technique. Different goal functions might be used in this task [Wójcik et al., 2023]. Often, the aim is to maximize the difference of scores between pairs provided by the DM - if the DM said a is preferred to b, then we are searching for a model where $U(a) - U(b)$ is maximized. The model has to meet constraints based on the preference information. If the DM defined two alternatives as indifferent, then their scores have to be equal $aIb \Rightarrow U(a) = U(b)$. Furthermore, the utility of a preferred alternative has to be higher than the dominated one $aPb \Rightarrow U(a) \geq U(b)$.

UTA method can be applied to different domains and decision problems. It suits problems DMs had to solve manually [Gehrlein et al., 2023] before since it is natural for them to perform pairwise comparisons. The method has been successfully applied to recommend validators in a blockchain environment [Gehrlein et al., 2023], select stocks based on investors' preferences [touni et al., 2019], evaluate risk in maritime transport [Stavrou et al., 2017], or optimize urban planning [Chhipi-Shrestha et al., 2018].

DRSA

Dominance-based Rough Set Approach (DRSA) [Greco et al., 2001] is an extension of the rough set approach [Pawlak, 1982]. It was designed for a classification task with predefined ordered classes. Alternatives are described on criteria of either cost or gain type. Thus, the novel approach can incorporate dominance relation. The main idea behind this algorithm is to not assign a dominated alternative to a more preferred class.

Preference information provided by a DM takes the form of reference alternatives' assignments to predefined decision classes. Those assignments are firstly preprocessed to detent inconsistencies where a worse alternative is assigned to a higher class. Then, a clean dataset is used to produce a mathematical model, which will be later exploited to classify the remaining alternatives. The model usually takes the form of decision rules. However, different algorithms can be used as well [Minz and Jain, 2003].

The main advantage of decision rules is their simplicity, allowing for high interpretability while still achieving results of high quality. They are stored in an if-then form where the if part consists of a conjunction of simple constraints on each criterion independently, and then the part represents an assignment either to one particular class or to a union of classes.

This algorithm has been widely applied to many real-world problems. It was used to detect financial fraud [Błaszczyński et al., 2021], evaluate credit risk [Silva et al., 2021], selection of speed limits [Augeri et al., 2015], or urban development planning [Oppio et al., 2020].

2.4 Imperfect learning data

Data imperfections can take various forms and be implied for different reasons. Nevertheless, they negatively impact the decision-making process. In what follows, imperfectness regarded in this thesis is described.

Preference information tends to be incomplete, especially when it is provided as examples of decisions. This knowledge is ambiguous by its very nature. No matter if the information is provided as an assignment of reference alternatives to the predefined classes or a pairwise comparison of alternatives, in most cases, an infinite number of models reflecting reference examples can be created. These models can lead to different outcomes, and picking just one of them might lead to results poorly reflecting the properties of all remaining ones.

Group decision-making involves more than one DM, which results in more than one point of view. Preferences of DMs are often conflicting to some extent, which makes them imperfect. To work out a group compromise recommendation, performing an aggregation on the input or the output is necessary. What is more, DMs do not have to be equal. The hierarchy of the DMs has to be included in the aggregation process.

Providing weights associated with each criterion is a popular way to obtain preference information used by various algorithms. DM is supposed to either provide a precise number, which is a cognitively hard task, or select weights from the predefined set, which is imprecise. Some procedures allow DM to provide a ranking of criteria instead of weights. This task is less demanding cognitively. However, the provided information is ambiguous.

Chapter 3

Results

3.1 Incomplete preference information

One form of acquiring preference information from DMs is to ask them to provide examples of decisions. They can be expressed depending on the type of problem, for example, by assigning alternatives to classes or defining relations between reference alternatives. Such information is inherently incomplete since it concerns a small subset of all alternatives. Moreover, there are usually multiple mathematical models leading to the same evaluation of alternatives assessed by the user, while varying in the evaluation of the remaining alternatives.

One of the results of the dissertation is the development of a robust procedure for the induction of decision rules based on sample assignments of alternatives to classes presented in publication **P5**. In the proposed approach, rules are induced independently from each reduct. Then, the relevance of individual rules is weighted by the quality of the classification of a given reduct. This approach is characterized by greater robustness of the solution compared to the classical approach, where only one set of equivalent rules is produced. The result of the developed method is the acceptability index of assignments to each class. Moreover, the algorithm also works correctly in the case of incomplete preference information regarding the direction of preferences for individual criteria. The decision maker does not have to specify the type of criteria since the methods can work with non-monotonic criteria.

Models are built based on data provided by the user. Therefore, the more sample decisions given by the user, the higher the precision of the created model, and thus, the developed recommendation will be of higher quality. Unfortunately, the size of the given decisions is closely related to the user's time and effort. As the quantity of data increases, the DM gets more fatigued during the process, which might lead to less focus and, as a result, providing examples inconsistent with their real preferences. This observa-

tion was one of the motivations for developing another procedure described in publication **P10** – an active learning algorithm for the UTA algorithm, where the preference information is provided as the pairwise comparisons of alternatives made by the decision maker.

The proposed algorithm algorithm generates questions in the form of a pair of alternatives (a_i, a_k) to be compared by the user. This question is the result of an optimization that aims to maximize the information gain defined as an improvement of the model's quality f obtained after answering this question. The answered questions form a set of constraints $S = \{(a_i, a_k) \in A \times A \,|\, a_i \succ a_k\}$ that a model has to meet. All models satisfying these conditions form a set $\mathcal{U}(S)$. Each model leads to a ranking of alternatives. Similarity of two rankings $R_{U'}, R_{U''}$ can be scored using the Spearman rank correlation coefficient [Kokoska and Zwillinger, 2000] $\rho\left(R_{U'}, R_{U''}, A\right) = 1 - \dfrac{6 \cdot \sum_{a_i \in A}[R_{U'}(a_i, A) - R_{U''}(a_i, A)]^2}{m(m^2-1)}$. To measure the quality of the current set of constraints, the minimal value of the Spearman rank correlation coefficient between two rankings in the allowed set is used $f_\rho\left(\mathcal{U}(S)\right) = \min_{U',U'' \in \mathcal{U}(S)} \rho(R_{U'}, R_{U''}, A)$. The question after which this metric is the highest is selected.

$$
\begin{aligned}
H\left(\mathcal{U}(S), A\right) &= \underset{\{a_i,a_k\} \in Q(\mathcal{U}(S))}{\arg\max} f_\rho\left(a_i, a_k, \mathcal{U}(S)\right) \\
&= \underset{\{a_i,a_k\} \in Q(\mathcal{U}(S))}{\arg\max} \min\left\{f_\rho(\mathcal{U}(S \cup (a_i, a_k)), f_\rho(\mathcal{U}(S \cup (a_k, a_i)))\right\}.
\end{aligned}
$$

This procedure requires fewer answers from the DM to produce a high-quality model compared to the classical approach, in which the DM creates the questions and answers them. What is more, a variant of this method in which questions are based not on real alternatives but on artificial objects generated just for this purpose has also been developed. It allows us to find the optimal question, maximizing the information gain. Thus, an even lower number of questions is needed.

Both variants were tested on a set of more than 100 real DMs who were experts in the domain of the problem (in this case, the choice of cryptocurrency validators), however, with no knowledge related to MCDA algorithms. Despite this fact, they could effectively use the developed tool, demonstrating the low entry-level required to use the methods and obtain high-quality results properly. The relevance of these observations is enhanced by the fact that in the experiment conducted, users had money at their disposal, which was invested for a certain time according to the recommendations developed by the novel algorithm. Thus, it motivated rational behavior and profit maximization. For both variants, the algorithm's recommendations outperformed manual choices done by the experts.

Table 3.1: Definition of distances $\delta(R'_{ij}, R''_{ij})$ between different pairs of relations Roy and Slowinski [1993].

R'_{ij}/R''_{ij}	$a_iP^+a_j$	$a_iP^-a_j$	a_iIa_j	$a_i?a_j$
$a_iP^+a_j$	0	4	2	3
$a_iP^-a_j$	4	0	2	3
a_iIa_j	2	2	0	2
$a_i?a_j$	3	3	2	0

3.2 Group decision-making dealing with inconsistency

In many real cases, there is more than one decision-maker interested in solving a decision problem. Moreover, they may represent various points of view, making their preferences contradicting. This leads to inconsistent preference information, which must be processed accordingly, and the final recommendation should be a compromise solution.

Within the field of MCDA, many methods have been developed. Some of them have become very popular and have many extensions which have been applied to solve different problems. These methods are well-tested and trusted by many users. Unfortunately, they are mostly dedicated to decision-making problems in which there is only one decision maker; hence, they cannot be used directly to solve group decision problems.

The goal of the solution proposed in publication **P6** was to create a family of procedures in which methods suitable for problems with a single decision maker can be used for problems with multiple decision makers. This is a more generic approach than creating another method dedicated to one specific type of problem. It allows well-tested and popular methods to be used in a broader context. In addition, decision makers do not need to familiarize themselves with the characteristics of the new method, as they can continue to use the one they have been using so far.

The developed solution focuses on the ranking problems. We considered both complete and partial rankings. The latter allows for the incomparability of alternatives. In this approach, any MCDA algorithm returning a ranking is applied independently for each DM, and the aggregation phase takes place on the resulting rankings without interfering with the ranking process itself. The problem was modeled as an optimization task, which aims to find the ranking closest to the rankings obtained for all DMs. To measure the distance between two rankings R' and R'' a metric introduced in Roy and Slowinski [1993] was used. It is decomposed into relations R'_{ij} and R''_{ij}, respectively, for all pairs of alternatives $a_i, a_j \in A$, such that $i < j$. Then the distance can be computed as follows $\sum_{i,j\ i<j} \delta(R'_{ij}, R''_{ij})$.. Distances proposed by Roy and Slowinski [1993] are presented in Table 3.1

Two objective functions were considered. A utilitarian approach, in which we look for a solution that minimizes the average distance from all rankings [Govindan et al., 2017]

$$\sum_{R\in\mathcal{R}} \frac{1}{|\mathcal{R}|} \sum_{i,j\ i<j} \delta(R^U_{ij}, R_{ij}).$$

Furthermore, weights can be introduced denoted as w_R for each ranking $R \in \mathcal{R}$, we account for the following function

$$\sum_{R\in\mathcal{R}} \frac{1}{|\mathcal{R}|} w_R \sum_{i,j\ i<j} \delta(R^U_{ij}, R_{ij}).$$

. The other function is an egalitarian approach, in which the maximum distance is minimized

$$max_{R\in\mathcal{R}}\{ \sum_{i,j\ i<j} \delta(R^E_{ij}, R_{ij})\}.$$

Appropriately adapted metaheuristic algorithms, such as simulated annealing or genetic approaches, were used to solve this optimization task, as well as methods developed specifically for this problem. These algorithms were tested on multiple sets of rankings with specific characteristics generated specifically for testing how these methods perform under different circumstances. Data from a real decision-making problem were also used to verify these methods' quality. The results obtained were of high quality, as well as efficient in terms of computation time. Therefore, it is also possible to apply this procedure not only to group decision-making problems, where the total number of rankings does not exceed a few dozen but also to rankings derived from the Stochastic Multi-criteria Acceptability Analysis (SMAA), where their number can reach thousands.

Besides two main metrics, namely utilitarian and egalitarian distances, time of computations was also used to evaluate algorithms. There is a clear trade-off between the time and the quality of obtained results. Methods returning better results regarding distance between the compromise ranking and input ones tend to require more time than algorithms returning worse solutions. A combination of local search and genetic approaches dominated the remaining algorithms when taking the quality of the results but was also the slowest method. On the other hand, simple heuristics was the fastest while producing the poorest solutions.

Publications **P1** tackles the problem of sorting in group decision-making. The proposed approach is based on applying a suitable method separately to each DM and then aggregating results. To solve the sorting problem SMAA with ELECTRE TRI-rC [Kadziński et al., 2015] algorithm was used. This procedure returns acceptability indexes for each class. Firstly, the class

range stochastic acceptability index $CRSAI^k(a, [h_L, h_R])$ is defined as the proportion of weights compatible with the DM_k's preferences leading to the assignment of the alternative a to the range of classes $[h_L, h_R]$ was calculated. Then, it is used to compute the cumulative class stochastic acceptability index $CuCSAI^k(a, h)$ defined as a fraction of weights assigning alternative a to class h. Finally, after repeating this procedure separately for each DM, the indexes are aggregated to compute a cumulative group class stochastic acceptability index $CuCSAI(a, h) = \frac{\sum_{k=1}^{K} CuCSAI^k(a,h)}{K}$

3.3 Lack of preference information

Lack of preference information is a specific type of imperfect learning data. This situation was considered in publication **P8**, tackling multiple criteria optimization of human resource scheduling in public transport in the city of Poznan. The goal functions reflect economic criteria like the percentage of shifts to which an employee is assigned or equality of deficiencies and social like home depot indicator or shifts distribution. What is more, multiple constraints, including those defined in the labor Code or the drivers' Hours of Work Act, must be met. For this purpose, the NSGA-II Deb et al. [2002] genetic multicriteria optimization algorithm was adapted to find a Pareto front containing a wide cross-section of solutions. It was necessary to include solutions satisfying different sets of DMs' preferences since they were unavailable at this stage. A variant of the ELECTRE method tailored to this type of problem and the form of preferential information transfer that suited the decision-maker was then applied to this front. This method accounts for decision thresholds, namely indifference, preference, and veto threshold, allowing modeling data impreciseness.

3.4 Relevant method selection

Over the years, hundreds of MCDA methods have been developed and documented, along with their extensions and variants. Each algorithm is dedicated to a specific class of decision-making problems and requires input data of a specific type and format. Due to the large number of methods and strictly defined conditions for their use, the very problem of selecting a decision support method can be considered a multiple criteria selection problem, in which it is the user who needs support. The imperfect nature of the data in this problem is due to the limited ability of the user to define the requirements.

To solve this problem, the MCDA-MSS system, presented in publication **P7**, was developed and implemented to support the selection of an appropriate method based on the description of the problem and the user's expectations. More than 200 MCDA algorithms have been classified in this system, and each method has been described on more than 150 attributes.

In most cases, the user will not be able to precisely specify his requirements by defining all of them. Moreover, the number of methods cannot cover all possible combinations of values on attributes. For this reason, it is possible that no method fully meets the decision-maker's preferences. The developed system addresses both described situations.

In case no method meets all the criteria given by the user, the key constraints that must be met and optional ones that can be ignored if needed. This procedure continues until there is at least one method that meets all the key constraints. The system guides the user to avoid situations where overly detailed constraints are provided that no method meets. An additional benefit of this mechanism is the possibility to identify areas where there are no suitable methods, making it possible to direct research into the development of new algorithms accordingly.

When the description of the requirements is incomplete, the system can return all classified methods that meet the given constraints. At the same time, a question minimizing the search space is recommended to reduce the number of matching algorithms as much as possible. Each question and attribute is described in detail and in an exhaustive way, making it easy for users to provide their preferences without having extensive specialized knowledge of the various methods and their parameters. The system also has other user interaction mechanisms to facilitate its usage. The number of corresponding methods is shown in real-time, and the user is also aware of how the number will change when further restrictions are given.

This system helps reduce the number of situations in which the wrong method is selected or misused. It can help increase the popularity of the entire field of MCDA since a DM without a huge knowledge of the subject can choose the right method for a given problem and apply it accordingly.

3.5 Robust weight calculation methods

An extension of the Simos-Roy-Figueira (SRF) procedure for determining criteria weights was proposed in publications **P2** and **P3**. In the original approach, preference information in the form of criteria ranking and a precisely stated ratio of the importance of the most important to the least important criterion is arbitrarily transformed into a precise vector of criteria weights according to an established procedure. This result is only one among potentially infinitely many vectors of weights consistent with the DM's preferences. The imperfectness here is the form of the preference information, which does not define the weights precisely. In addition, the need to provide an exact value for the ratio of the criteria weights may result in the DMs providing a value that is not an accurate reflection of their preferences, as this is a cognitively intensive task.

In the developed approach, the output is not one particular vector of weights but a set of all weights vectors satisfying constraints based on the DM's preferences. Moreover, the decision-maker does not have to precisely define the ratio of importance between the first and last criteria. The preferential information is used to determine a space of weights consistent with it. This space is then sampled using Monte Carlo simulation. Each vector of weights obtained within this simulation is later used in a classical MCDA algorithm requiring criteria weights as an input. Results of multiple runs for different weights are then aggregated to determine acceptability indices, which are then translated into a final robust solution.

Preference information is provided as a ranking of criteria. The procedure of working the ranking out involves using cards with criteria names and blank cards, which might be used to differentiate the intensity of preference between different groups. This procedure leads to successive groups of criteria L_s and L_{s+1} with e_s blank cards inserted between them. Each criterion g_j belongs to exactly one group. Criteria denoted by the DM as indifferent belong to the same group. The user must also provide the ratio of importance Z between the most significant group L_v and the least significant group L_1. Finally, the minimal λ_* and maximal λ^* credibility threshold λ must be provided. The whole preference information is transformed into a set of constraints E defining an allowed space of weights and credibility thresholds $(w, \lambda)^{DM}$ compatible with the preferences of the DM.

$$(w,\lambda)^{DM} = \begin{cases} w_i > w_j & g_i \in L_t, g_j \in L_s, t > s \\ w_i = w_j & g_i, g_j \in L_s \\ w_{j+1} - w_j > w_{p+1} - w_p & e_j > e_p \\ w_i = Z * w_j & g_i \in L_v, g_j \in L_1 \\ w_j > 0 & j = 1, ..., m \\ \sum_{j=1}^{m} w_j = 1 & \\ \lambda_* \leq \lambda \leq \lambda^* & \end{cases}$$

The approach was adapted to two types of problems. The first problem involved a selection problem using the ELECTRE I method. In this variant, acceptability indexes were based on the comparison of pairs of alternatives and the relationships between them. These were used to build an outranking graph, from which the graph kernel was extracted. Both the graph and its kernel obtained in this way are characterized by significantly higher robustness compared to the approach based on a single vector of weights. The motivation for developing this solution was the problem of selecting a logistics service provider.

The second problem was the assignment of alternatives to predefined ordered classes. Acceptability indices describe the assignment to classes for different alternatives. Thus, all models consistent with the DMs' preferences

were considered not just an arbitrarily selected one, potentially leading to a solution being an outlier. The adapted procedure was used with the ELECTRE TRI-rC method. In this task, 13 insulating materials described on six criteria had to be assigned to one of three classes denoting its sustainability level based on the preferences of 38 rural buildings' owners playing the role of the DMs.

3.6 Neural networks dealing with imperfect data

Current decision support processes are multidisciplinary and benefit from development in many areas. Tools designed to aid decision processes can be based on machine learning algorithms. These algorithms are data-driven, so they rely almost exclusively on training data. The quality of the data directly translates into the quality of the model and its results. This is especially evident in modern neural networks, which, in order to unlock their full potential, need large volumes of training data of good quality. Otherwise, the models might learn the imperfection present in the data and treat it as an actual signal.

In publications **P4** and **P9**, two recurrent neural architectures that have greater robustness than classical approaches were proposed. In both cases, it was motivated by the real decision-making problem to be solved. The first was to predict electricity consumption for enterprises of different types. Because of the wide variety of their energy usage, the standard approach would be to train an independent model for each company. In this approach, when imperfect learning data are used for one entity, the model would not have a chance to work correctly. For this purpose, a special architecture of a recurrent neural network was proposed, in which static metadata are processed simultaneously along with time series. In the considered problem, data like size and type of enterprise or its location were used as the static input, while the energy consumption time series were used as a classical dynamic input. With this approach, it was possible to expose the model simultaneously to multiple enterprises, as the additional static features allowed the model to distinguish between them. In addition, it was possible to interpolate using these values, so even if the data was wrong for one company, the model could still make the correct prediction based on the energy consumption curves for neighboring values. It is worth noting that even though this approach was developed for a specific problem, it can be generalized and applied to problems in another domain with similar characteristics.

In classical Recurrent Neural Network (RNN) models [Rumelhart et al., 1986], the current input x_t is used together with the hidden state of the previous time step h_{t-1} to generate the current hidden state h_t.

$$h_t = \sigma_h(W_h x_t + U_h h_{t-1} + b_h),$$

where W_h, U_h and b_h denote trainable parameters of the neural network and function σ_h stands for the activation function. One of the most popular variants of RNN called Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] introduces a more complex model of a single neuron.

$$
\begin{aligned}
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i), \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o), \\
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f), \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c), \\
h_t &= o_t \circ \sigma_h(c_t),
\end{aligned}
$$

where the activation function σ_g is a sigmoid function, while σ_c and σ_h are hyperbolic tangent functions. Vectors c and h are the outputs called cell and hidden states, respectively. When processing a given timestamp, output obtained for the previous one is used as well in the calculation. It cannot be acquired when the very first timestamp is introduced. In the default implementation, an initial state with zero-value vectors is used in this situation.

$$
\begin{aligned}
c_{\text{-1}} &= (0, 0, \ldots, 0), \\
h_{\text{-1}} &= (0, 0, \ldots, 0).
\end{aligned}
$$

In the new approach, two additional Feed-forward Neural Networks (FNNs) are introduced (FNN_c, FNN_h) to derive the initial state from static input data s. Both FNNs take static data as an input and return initial states as an output.

$$
\begin{aligned}
c_{\text{-1}} &= \mathrm{FNN}_c(\mathbf{s}), \\
h_{\text{-1}} &= \mathrm{FNN}_h(\mathbf{s}).
\end{aligned}
$$

It allows firstly to properly transform static data since the transition is part of the overall model training, and secondly to assert the proper shape of the data since the size of the output if easily configurable by the number of units in the last layer.

This architecture was used as a component of another model described in publication **P9**, which aimed to be part of a leak detection and localization system in pipelines transporting liquid fuels. Most methods for this problem are sensitive to various types of data imperfections. In practice, issues with sensors located along the pipeline or with communication are possible. A neural network architecture was developed, in which a set of observations with their description is given as input. In this case, it was the location of the sensor on the pipeline and the time of measurement. Based on this, the model builds a picture of the current situation on the pipeline in the so-called latent variable. In the second phase, this image can be used, and based on a

query, also in the form of location and time, the model generates the corresponding time series. Due to the fact that any number of observations can be given as input, the model is immune to problems with one or even several sensors. It is sufficient to use several combinations of sensors independently for this purpose. Also, this model can be successfully applied to problems from another domain, like predicting traffic congestion.

Chapter 4

Summary

The field of MCDA aims to support DM in solving decision problems. Alternatives described on more than one criterion, together with DM's preferences, are used to build a model, which will later be exploited to produce a recommendation. Real-world decision problems are characterized by a higher uncertainty than theoretical ones. Provided data are often imprecise, which might affect both alternatives description as well as the preferences. In such situations, algorithms must be aware of these possibilities and be robust to data imperfection.

In the thesis, different extensions and procedures addressing various data imperfections were proposed. Notably, the development of novel algorithms was motivated by real-world decision problems and data issues. The new methods' quality was tested on real problems, and they were proven to work as intended.

One of the tackled problems involves group decision-making. When more than one DM is responsible for making a decision, a method has to deal with conflicting preferences, which are additionally often imprecise. For a ranking problem, a generalized framework was proposed within which any MCDA method producing a ranking of alternatives can be applied separately for each DM. Then, results are aggregated to produce a compromise ranking based on the preferences of all DMs. A similar technique was used for a problem with imprecise information in the form of a criteria ranking provided by multiple DMs. Such information was used to derive a space of acceptable weights in a SMAA. Then, robust indices obtained for each DM were aggregated into a group recommendation.

The next topic covered in the thesis is elaborating robust methods for weight calculation. Instead of arbitrarily selecting one vector of weights, the whole space of acceptable weights was used to produce a robust recommendation based on the SMAA. Such results address the imperfectness related to the indetermination of an assumed preference model. This approach was

applied to the choice problem as well as the sorting problem.

Furthermore, providing example decisions is one of the easiest ways to obtain preference information from a DM. However, the information in this form is incomplete. In the thesis, efficient methods of acquiring preference information using active learning to generate pairs of alternatives to be compared by the DM were proposed. The question results from an optimization process aiming to maximize the information gain and, therefore, the quality of a model. Additionally, robust techniques for handling data expressed as class assignments were proposed.

Selecting a proper method for a given decision problem is a complex issue. Choosing the appropriate algorithm might lead to incorrect operation and, therefore, wrong recommendations. In the thesis, a system allowing for efficient searching for MCDA algorithms meeting given requirements was proposed. Based on the problem descriptions, a suitable method is proposed.

A neural network is a mathematical model that can be used in decision-making. This approach relies on a vast amount of high-quality data. Two neural architectures were proposed for more robust time series transformation using static data. The novel approach minimizes the impact of imperfect data by including additional static descriptions in the learning phase.

Nowadays, more and more complex problems are processed in computer systems. The high availability of data allows for the wide use of artificial intelligence and computer-based decision-solving tools. Unfortunately, classical algorithms can rarely be used in practice due to their limitations. Often, they are suitable for small theoretical problems with well-defined structures and high-quality data. Due to these constraints, the use of MCDA methods in practice is limited. Novel techniques presented in this thesis were successfully applied to real problems with all their flaws. Thanks to them, the field of MCDA is open to new directions, such as big data, and will increase its usability in practice.

Bibliography

M. G. Augeri, P. Cozzo, and S. Greco. Dominance-based rough set approach: An application case study for setting speed limits for vehicles in speed controlled zones. *Knowledge-Based Systems*, 89:288–300, Nov. 2015. doi: 10.1016/j.knosys.2015.07.010. URL https://doi.org/10.1016/j.knosys.2015.07.010.

J. Błaszczyński, A. T. de Almeida Filho, A. Matuszyk, M. Szeląg, and R. Słowiński. Auto loan fraud detection using dominance-based rough set approach versus machine learning methods. *Expert Systems with Applications*, 163:113740, Jan. 2021. doi: 10.1016/j.eswa.2020.113740. URL https://doi.org/10.1016/j.eswa.2020.113740.

J.-P. Brans and B. Mareschal. Promethee methods. In *International Series in Operations Research & Management Science*, pages 163–186. Springer-Verlag, 2005. doi: 10.1007/0-387-23081-5˙5. URL https://doi.org/10.1007/0-387-23081-5_5.

G. Chhipi-Shrestha, M. Kaur, K. Hewage, and R. Sadiq. Optimizing residential density based on water–energy–carbon nexus using UTilités additives (UTA) method. *Clean Technologies and Environmental Policy*, 20(4):855–870, Feb. 2018. doi: 10.1007/s10098-018-1506-6. URL https://doi.org/10.1007/s10098-018-1506-6.

M. Cinelli, M. Kadziński, G. Miebs, M. Gonzalez, and R. Słowiński. Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system. *European Journal of Operational Research*, 302(2):633–651, Oct. 2022. doi: 10.1016/j.ejor.2022.01.011. URL https://doi.org/10.1016/j.ejor.2022.01.011.

K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002. doi: https://doi.org/10.1109/4235.996017.

M. S. M. M. Emamat, C. M. de Miranda Mota, M. R. Mehregan, M. R. S. Moghadam, and P. Nemery. Using ELECTRE-TRI and FlowSort meth-

ods in a stock portfolio selection context. *Financial Innovation*, 8(1), Feb. 2022. doi: 10.1186/s40854-021-00318-1. URL https://doi.org/10.1186/s40854-021-00318-1.

J. Figueira and B. Roy. Determining the weights of criteria in the ELECTRE type methods with a revised simos' procedure. *European Journal of Operational Research*, 139(2):317–326, June 2002. doi: 10.1016/s0377-2217(01)00370-8. URL https://doi.org/10.1016/s0377-2217(01)00370-8.

J. R. Figueira, V. Mousseau, and B. Roy. ELECTRE methods. In *Multiple Criteria Decision Analysis*, pages 155–185. Springer New York, 2016. doi: 10.1007/978-1-4939-3094-4˙5. URL https://doi.org/10.1007/978-1-4939-3094-4_5.

A. A. Ganin, P. Quach, M. Panwar, Z. A. Collier, J. M. Keisler, D. Marchese, and I. Linkov. Multicriteria decision framework for cybersecurity risk assessment and management. *Risk Analysis*, 40(1):183–199, Sept. 2017. doi: 10.1111/risa.12891. URL https://doi.org/10.1111/risa.12891.

J. Gehrlein, G. Miebs, M. Brunelli, and M. Kadziński. An active preference learning approach to aid the selection of validators in blockchain environments. *Omega*, 118:102869, July 2023. doi: 10.1016/j.omega.2023.102869. URL https://doi.org/10.1016/j.omega.2023.102869.

K. Govindan, M. Kadziński, and R. Sivakumar. Application of a novel PROMETHEE-based method for construction of a group compromise ranking to prioritization of green suppliers in food supply chain. *Omega*, 71:129 – 145, 2017.

K. Govindan, M. Kadziński, R. Ehling, and G. Miebs. Selection of a sustainable third-party reverse logistics provider based on the robustness analysis of an outranking graph kernel conducted with ELECTRE i and SMAA. *Omega*, 85:1–15, June 2019. doi: 10.1016/j.omega.2018.05.007. URL https://doi.org/10.1016/j.omega.2018.05.007.

S. Greco, B. Matarazzo, and R. Slowinski. A new rough set approach to evaluation of bankruptcy risk. In *Operational Tools in the Management of Financial Risks*, pages 121–136. Springer US, 1998. doi: 10.1007/978-1-4615-5495-0˙8. URL https://doi.org/10.1007/978-1-4615-5495-0_8.

S. Greco, B. Matarazzo, and R. Slowinski. Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research*, 129(1):1–47, Feb. 2001. doi: 10.1016/s0377-2217(00)00167-3. URL https://doi.org/10.1016/s0377-2217(00)00167-3.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, Nov. 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

C.-L. Hwang and K. Yoon. *Multiple Attribute Decision Making*. Springer Berlin Heidelberg, 1981. doi: 10.1007/978-3-642-48318-9. URL https://doi.org/10.1007/978-3-642-48318-9.

E. Jacquet-Lagreze and J. Siskos. Assessing a set of additive utility functions for multicriteria decision-making, the UTA method. *European Journal of Operational Research*, 10(2):151–164, 1982.

M. Jurczak, G. Miebs, and R. A. Bachorz. Multi-criteria human resources planning optimisation using genetic algorithms enhanced with MCDA. *Operations Research and Decisions*, 32(4), 2022. doi: 10.37190/ord220404. URL https://doi.org/10.37190/ord220404.

M. Kadziński, T. Tervonen, and J. R. Figueira. Robust multi-criteria sorting with the outranking preference model and characteristic profiles. *Omega*, 55:126–140, Sept. 2015. doi: 10.1016/j.omega.2014.06.004. URL https://doi.org/10.1016/j.omega.2014.06.004.

M. Kadziński, L. Rocchi, G. Miebs, D. Grohmann, M. E. Menconi, and L. Paolotti. Multiple criteria assessment of insulating materials with a group decision framework incorporating outranking preference model and characteristic class profiles. *Group Decision and Negotiation*, 27(1):33–59, Nov. 2017. doi: 10.1007/s10726-017-9549-3. URL https://doi.org/10.1007/s10726-017-9549-3.

R. L. Keeney and H. Raiffa. *Decisions with Multiple Objectives*. Cambridge University Press, July 1993. doi: 10.1017/cbo9781139174084. URL https://doi.org/10.1017/cbo9781139174084.

S. Kokoska and D. Zwillinger. *CRC Standard Probability and Statistics Tables and Formulae, Student Edition*. CRC Press, 2000.

G. Miebs and M. Kadziński. Heuristic algorithms for aggregation of incomplete rankings in multiple criteria group decision making. *Information Sciences*, 560:107–136, June 2021. doi: 10.1016/j.ins.2021.01.055. URL https://doi.org/10.1016/j.ins.2021.01.055.

G. Miebs, M. Mochol-Grzelak, A. Karaszewski, and R. A. Bachorz. Efficient strategies of static features incorporation into the recurrent neural network. *Neural Processing Letters*, 51(3):2301–2316, Jan. 2020. doi: 10.1007/s11063-020-10195-x. URL https://doi.org/10.1007/s11063-020-10195-x.

G. Miebs, M. Wójcik, A. Karaszewski, M. Mochol-Grzelak, P. Wawdysz, and R. A. Bachorz. Predicting a time-dependent quantity using recursive generative query network. *International Journal of Neural Systems*, 32 (11), Oct. 2022. doi: 10.1142/s0129065722500563. URL https://doi.org/10.1142/s0129065722500563.

S. Minz and R. Jain. Rough set based decision tree model for classification. In *Data Warehousing and Knowledge Discovery*, pages 172–181. Springer Berlin Heidelberg, 2003. doi: 10.1007/978-3-540-45228-7˙18. URL https://doi.org/10.1007/978-3-540-45228-7_18.

B. Németh, A. Molnár, S. Bozóki, K. Wijaya, A. Inotai, J. D. Campbell, and Z. Kaló. Comparison of weighting methods used in multicriteria decision analysis frameworks in healthcare with focus on low- and middle-income countries. *Journal of Comparative Effectiveness Research*, 8(4):195–204, Apr. 2019. doi: 10.2217/cer-2018-0102. URL https://doi.org/10.2217/cer-2018-0102.

A. Oppio, M. Dell'Ovo, F. Torrieri, G. Miebs, and M. Kadziński. Understanding the drivers of urban development agreements with the rough set approach and robust decision rules. *Land Use Policy*, 96:104678, July 2020. doi: 10.1016/j.landusepol.2020.104678. URL https://doi.org/10.1016/j.landusepol.2020.104678.

O. Y. S. Palacios, P. C. N. Rincón, M. Camargo, J.-P. Corriou, and C. Fonteix. Chemical product design integrating mcda: Performance prediction and human preferences modelling. *The Canadian Journal of Chemical Engineering*, 99(S1), Jan. 2021. doi: 10.1002/cjce.23956. URL https://doi.org/10.1002/cjce.23956.

S. Parsons. Current approaches to handling imperfect information in data and knowledge bases. *IEEE Transactions on Knowledge and Data Engineering*, 8(3):353–372, June 1996. doi: 10.1109/69.506705. URL https://doi.org/10.1109/69.506705.

Z. Pawlak. Rough sets. *International Journal of Computer & Information Sciences*, 11(5):341–356, Oct. 1982. doi: 10.1007/bf01001956. URL https://doi.org/10.1007/bf01001956.

L. Rocchi, M. Kadziński, M. Menconi, D. Grohmann, G. Miebs, L. Paolotti, and A. Boggia. Sustainability evaluation of retrofitting solutions for rural buildings through life cycle approach and multi-criteria analysis. *Energy and Buildings*, 173:281–290, Aug. 2018. doi: 10.1016/j.enbuild.2018.05.032. URL https://doi.org/10.1016/j.enbuild.2018.05.032.

B. Roy. Classement et choix en présence de points de vue multiples. *Revue française d'informatique et de recherche opérationnelle*, 2(8):57–75,

1968. doi: 10.1051/ro/196802v100571. URL https://doi.org/10.1051/ro/196802v100571.

B. Roy. The outranking approach and the foundations of electre methods. *Theory and Decision*, 31(1):49–73, July 1991. doi: 10.1007/bf00134132. URL https://doi.org/10.1007/bf00134132.

B. Roy and R. Slowinski. Criterion of distance between technical programming and socio-economic priority. *RAIRO - Operations Research*, 27(1): 45–60, 1993. doi: 10.1051/ro/1993270100451. URL https://doi.org/10.1051/ro/1993270100451.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct. 1986. doi: 10.1038/323533a0. URL https://doi.org/10.1038/323533a0.

J. C. S. Silva, D. F. de Lima Silva, L. Ferreira, and A. T. de Almeida-Filho. A dominance-based rough set approach applied to evaluate the credit risk of sovereign bonds. *4OR*, 20(1):139–164, Jan. 2021. doi: 10.1007/s10288-020-00471-w. URL https://doi.org/10.1007/s10288-020-00471-w.

Y. Siskos, E. Grigoroudis, and N. F. Matsatsinis. UTA methods. In *Multiple Criteria Decision Analysis*, pages 315–362. Springer New York, 2016. doi: 10.1007/978-1-4939-3094-4˙9. URL https://doi.org/10.1007/978-1-4939-3094-4_9.

D. I. Stavrou, E. Y. Siskos, N. P. Ventikos, and J. E. Psarras. Robust evaluation of risks in ship-to-ship transfer operations: Application of the STOCHASTIC UTA multicriteria decision support method. In *International Series in Operations Research & Management Science*, pages 175–218. Springer International Publishing, Oct. 2017. doi: 10.1007/978-3-319-62338-2˙8. URL https://doi.org/10.1007/978-3-319-62338-2_8.

Z. touni, A. Makui, and E. Mohammadi. A mcdm-based approach using utastrar method to discover behavioral aspects in stock selection problem. *International Journal of Industiral Engineering & Producion Research*, 30 (1), Mar. 2019. doi: 10.22068/ijiepr.30.1.93. URL https://doi.org/10.22068/ijiepr.30.1.93.

M. Wójcik, M. Kadziński, and K. Ciomek. Selection of a representative sorting model in a preference disaggregation setting: A review of existing procedures, new proposals, and experimental comparison. *Knowledge-Based Systems*, 278:110871, Oct. 2023. doi: 10.1016/j.knosys.2023.110871. URL https://doi.org/10.1016/j.knosys.2023.110871.

Publication reprints

Publication [P1]

Kadziński, M., Rocchi, L., Miebs, G., Grohmann, D., Menconi, M. E., and Paolotti, L. (2017). Multiple Criteria Assessment of Insulating Materials with a Group Decision Framework Incorporating Outranking Preference Model and Characteristic Class Profiles. In Group Decision and Negotiation (Vol. 27, Issue 1, pp. 33–59). Springer Science and Business Media LLC., DOI: 10.1007/s10726-017-9549-3.

Number of citations[1]:

- according to Web of Science: 19
- according to Google Scholar: 29

Contribution of co-authors (excluding the author and the supervisor of this dissertation):

- Lucia Rocchi:
  - Co-authorship of the concept of sorting algorithms and of the considered case study (Section 3)
  - Co-authorship of the text of the publication
- David Grohmann:
  - Consultation of the context of case study and interpretation of its results
  - Corrections to the first draft of the manuscript
- Maria Elena Menconi
  - Consultation of the context of case study and interpretation of its results
  - Corrections to the first draft of the manuscript
- Luisa Paolotti
  - Consultation of the context of case study and interpretation of its results
  - Corrections to the first draft of the manuscript

[1] as on September 25, 2023

Group Decis Negot
https://doi.org/10.1007/s10726-017-9549-3

Multiple Criteria Assessment of Insulating Materials with a Group Decision Framework Incorporating Outranking Preference Model and Characteristic Class Profiles

Miłosz Kadziński[1] · Lucia Rocchi[2] · Grzegorz Miebs[1] · David Grohmann[2] · Maria Elena Menconi[2] · Luisa Paolotti[2]

Abstract We present a group decision making framework for evaluating sustainability of the insulating materials. We tested thirteen materials on a model that was applied to retrofit a traditional rural building through roof's insulation. To evaluate the materials from the socio-economic and environmental viewpoints, we combined life cycle costing and assessment with an adaptive comfort evaluation. In this way, the performances of each coating material were measured in terms of an incurred reduction of costs and consumption of resources, maintenance of the cultural and historic significance of buildings, and a guaranteed indoor thermal comfort. The comprehensive assessment of the materials involved their assignment to one of the three preference-ordered sustainability classes. For this purpose, we used a multiple criteria decision analysis approach that accounted for preferences of a few tens of rural buildings' owners. The proposed methodological framework incorporated an outranking-based preference model to compare the insulating materials with the characteristic class profiles while using the weights derived from the revised Simos procedure. The initial sorting recommendation for each material was validated against the outcomes of robustness analysis that combined the preferences of individual stakeholders either at the output or at the input level. The analysis revealed that the most favorable materials in terms of their overall sustainability were glass wool, hemp fibres, kenaf fibres, polystyrene foam, polyurethane, and rock wool.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10726-017-9549-3) contains supplementary material, which is available to authorized users.

Miłosz Kadziński
milosz.kadzinski@cs.put.poznan.pl

[1] Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland

[2] Department of Agricultural, Food and Environmental Sciences, University of Perugia, Borgo XX Giungno, 74, 06121 Perugia, Italy

Published online: 18 November 2017

Keywords Multiple criteria decision analysis · Group decision · Characteristic profiles · ELECTRE TRI-rC · Insulating materials · Sustainability

1 Introduction

This paper presents a group decision framework for evaluating sustainability of the insulating materials to retrofit traditional rural buildings. The importance of this research derives from the previous studies on both retrofitting solutions tailored to traditional rural buildings as well as judging an overall desirability of coating materials (see, e.g., Krarti 2015; Fabbri et al. 2012; Ma et al. 2012; Yung and Chan 2012; Martínez-Molina et al. 2016). These studies prove that energy efficiency and thermal comfort are crucial for the maintenance of historic buildings.

The context of the study is that of a typical farmhouse in central Italy. The incorporated building model derives from the analysis of over 800 farmhouses surveyed by the census of the scattered rural buildings of the municipality of Perugia (Umbria region). The high landscape values of traditional buildings and the legislation about their preservation prevent external alterations (Mazzarella 2015). Therefore, the most viable solutions are to intervene on the roof of these structures, increasing their thermal inertia with coating materials (Verbeeck and Hens 2005; Kumar and Suman 2013; Taylor et al. 2000).

We comprehensively evaluate the materials for the roof insulation by considering economic, social, and environmental viewpoints. For this purpose, we incorporate a life cycle costing (LCC) approach, a life cycle assessment (LCA), and a dynamic thermal simulation for the evaluation of energy savings and thermal comfort. As such, we aim at identifying the materials that guarantee the indoor thermal comfort, at the same time reducing the consumption of resources in their entire life cycle as well as maintaining cultural and historic significance of the buildings. In this perspective, we differentiate from the vast majority of previous studies concerning coating materials which incorporate a mono-disciplinary approach (Copiello 2017).

To provide an overall sustainability assessment of coating materials, we incorporate Multiple Criteria Decision Analysis (MCDA). MCDA offers a diversity of approaches designed for providing the decision makers (DMs) with a recommendation concerning a set of alternatives evaluated in terms of multiple conflicting points of view. Few applications of MCDA methods for the evaluation of building materials, which are reported in the literature (Ginevicius et al. 2008) deal mainly with the environmental sustainability of materials (Papadopoulos and Giama 2007; Khoshnava et al. 2016). Some combinations of LCA and MCDA were considered by Santos et al. (2017) and Piombo et al. (2016). Applications which included both LCC and LCA for the definition of criteria to be used in MCDA are still rare (Piombo et al. 2016). Decision analysis methods used in the above-mentioned studies involved different variants of AHP (Motuziene et al. 2016; Khoshnava et al. 2016), PROMETHEE II (Kumar et al. 2017), Weighted Sum, TOPSIS (Čuláková et al. 2013), VIKOR, and COPRAS (Ginevicius et al. 2008).

From the viewpoint of MCDA, our study differs from the aforementioned ones in terms of the following major aspects:

- We formulate the considered problem in terms of multiple criteria sorting, thus aiming at assigning the materials to a set of pre-defined and ordered sustainability classes (categories) rather than at ordering them from the best to the worst;
- We assess the insulating materials while taking into account preferences of multiple DMs (owners of rural houses), thus incorporating group decision making tools into the evaluation framework;
- The adopted assignment procedure builds upon outranking-based comparison of the insulating materials with the characteristic profiles composed of the per-class most representative performances on all criteria (Kadziński et al. 2015b);
- The research results are validated against the outcomes of robustness analysis that takes into account all sets of weights compatible with either the ranking of criteria provided by each DM within the revised Simos (SRF) procedure (Figueira and Roy 2002) or a group compromise ranking of criteria that is constructed with an original procedure proposed in this paper.

The remainder of the paper is organized in the following way. In the next section, we review the existing group decision making methods for multiple criteria sorting. Section 3 describes a three-stage decision aiding method that has been used to evaluate the insulating materials while taking into account preferences of a group of stakeholders. Section 4 exhibits comprehensive results of multiple criteria assessment of the insulating materials. The last section concludes.

2 Review of Multiple Criteria Sorting Group Decision Methods

The objective of the case study presented in this paper is to give an easily interpretable comprehensive assessment of the insulating materials' sustainability. This is achieved by assigning them to a set of pre-defined and ordered decision classes based on their performances on multiple criteria (Kadziński et al. 2015b). While computing the sorting recommendation, we account for the preferences of a group of experts and stakeholders. This requires implementation of a group decision making framework.

As real-world situations often involve multiple stakeholders, some methods have been proposed to support groups in making collective sorting decisions (Daher and Almeida 2010). These approaches can be distinguished at different levels. In particular, they differ in terms of a preference model employed to represent preferences of the DMs. Furthermore, an underlying classification rule may involve analysis of a single preference model instance or all sets of parameters compatible with the DMs' preference information. Moreover, sorting methods can be divided with respect to the level on which individual viewpoints are aggregated (Dias and Climaco 2000). Finally, some approaches account for the importance degrees of the involved DMs, while other methods assume that all DMs play the same role in the committee.

Among multiple criteria sorting group decision methods, outranking-based approaches are prevailing. Most decision support systems in this stream incorporate Electre TRI-B (Yu 1992; Roy 1996). For example, Dias and Climaco (2000) proposed an approach that admits each DM to specify imprecise constraints on the parameters of an outranking model, then exploits a set of compatible parameters using robust assignment rule, and finally aggregates individual perspectives in a disjunctive or conjunctive

manner (thus, not accounting for the DMs' powers). The former accepts an assignment if it is justified by at least one DM, whereas the latter confirms some classification only if it is consistent with the preferences of all DMs. In this way, a group may agree on some result even if its members do not share the same model parameters. This idea was extended by Damart et al. (2007) to an interactive preference disaggregation approach that accepts assignment examples provided by different DMs. The method incorporates robustness analysis by deriving for each DM the possible class assignments (confirmed by at least one compatible preference model instance) and guides the group on sorting exemplary alternatives by exhibiting the levels of consensus between the DMs. Analogously, Shen et al. (2016) developed an adaptive approach under intuitionistic fuzzy environment that allows to reach a classification with an acceptable individual and group consensus levels. Moreover, de Morais Bezerra et al. (2017) enriched Electre TRI-B with the tools for visualizing the comparison of individual results and procedures for guiding the changes of model parameters for deriving a better consensus.

Furthermore, Jabeur and Martel (2007) proposed a framework, which derives a collective sorting decision at the output level from the individual non-robust classifications by additionally accounting for the relative importance of group members. Then, Morais et al. (2014) used a stochastic variant of Electre TRI-B, called SMAA-TRI, to consider uncertainty in criteria weights and to derive for each DM the shares of the relevant parameter vectors that assign a given alternative to a certain category. An overview of thus obtained individual results leads to a collective recommendation. Conversely, Cailloux et al. (2012) employed assignment examples provided by multiple DMs for reaching an agreement at the input level. In particular, they proposed some linear programming models for deriving a joint set of boundary class profiles and veto thresholds.

As far as outranking-based sorting approaches incorporating a model typical for PROMETHEE are concerned, Nemery (2008) extended the FlowSort method to group decision making. His proposal derives an assignment for each alternative from its relative comparison (strength and weakness) against the boundary or central class profiles specified by each individual DM. A similar idea was implemented by Lolli et al. (2015) in FlowSort-GDSS. The underlying procedure derives class assignments by comparing comprehensive (global) net flows of alternatives and reference profiles. The proposed sorting rules distinguish between scenarios in which analysis of the individual assignments leads to either univocal or non-unanimous recommendation. Although the viewpoints of different DMs are aggregated at the output level, the method defines some consistency conditions on the preference information (in particular, reference profiles) provided by the individual DMs.

The majority of existing value-based approaches derive a sorting recommendation incorporating robustness analysis and not differentiating between the roles played by the DMs. In particular, the UTADIS^{GMS}-GROUP method (Greco et al. 2012) accounts for the assignment examples provided by each DM and derives collective results that concern two levels of certainty. The first level refers to the necessary and possible consequences of individual preference information, which is typical for Robust Ordinal Regression (ROR) (Greco et al. 2010; Kadziński et al. 2015b). The other level is related to the necessity or possibility of a support that a particular assignment is given in the

set of DMs. This method was further adapted by Liu et al. (2015) to account for the uncertain evaluations represented with the evidential reasoning approach, to provide some measures on the agreement between the DMs, and to derive a collective univocal assignment.

Conversely, Kadziński et al. (2013) aimed at a joint representation of assignment examples provided by all DMs by a set of additive value functions and investigating the necessary and possible consequences of applying the latter on the set of alternatives. When there is no value function compatible with preferences of all DMs, some linear programming techniques can be used to remove a minimal subset of inconsistent assignment examples. A similar approach was proposed by Cai et al. (2012), though additionally accounting for the DMs' priorities. The latter ones intervene in the selection of a representative value function and in resolving inconsistency in the provided assignment examples. These priorities are updated with the progressive preference elicitation process to reflect the preciseness, quantity and consistency of the example decisions supplied by each DM.

Finally, when it comes to using "if …then …" decision rules for representing preferences of the DMs, one proposed various extensions of the Dominance-based Rough Set Approach (DRSA) (Greco et al. 2001). These accept preference information in form of individual assignment examples. First, Greco et al. (2006) introduced some concepts (e.g., multi-union and mega-union) related to dominance with respect to minimal profiles of evaluations provided by different DMs. Then, Chen et al. (2012) proposed to aggregate the recommendations suggested by individual linguistic decision rules into an overall assignment be means of a Dempster–Shafer Theory. The crucial concepts incorporated in the DRSA sorting method proposed by Sun and Ma (2015) are a dominance relation on the set of multiple sorting decisions (each provided by an individual DM) and a multi-agent conflict analysis framework. Furthermore, Chakhar and Saad (2012) and Chakhar et al. (2016) illustrated how to combine individual approximations of class unions and derive collective decision rules that permit classification of all alternatives in a way consistent with the judgments of all DMs. These approaches measure the contribution of each expert to the collective assignment in terms of the individual quality of classification. Finally, Kadziński et al. (2016) adapted the principle of ROR to a group decision framework with DRSA, thus considering all sets of rules compatible with the individual assignment examples and combining their indications only at the output level.

In this paper, we propose an outranking-based group decision approach that incorporates Electre TRI-rC. Thus, it derives the assignments by comparing alternatives with the characteristic class profiles rather than with the boundary profiles as in Electre TRI-B. The basic procedure we use takes into account a single preference model instance (incorporating criteria weights derived from the SRF procedure) for each DM and aggregates the individual viewpoints at the output level. While still aggregating the preferences at the output level, we extend the basic framework to offer results of robustness analysis with multiple sets of parameters compatible with the DMs' value systems. Additionally, we propose a new algorithm for constructing a group compromise ranking of criteria, hence offering aggregation of the individual viewpoints also at the input level. At all stages, we assume that the involved stakeholders have the same importance degrees. Moreover, instead of providing precise assignments,

our framework offers acceptability indices indicating the support that is given to the assignment of each alternative to various classes by different DMs and/or preference model instances compatible with their preferences.

3 Multiple Criteria Decision Analysis Method for the Assessment of Insulating Materials

This section describes a three-stage multiple criteria decision analysis method that has been used to evaluate the insulating materials while taking into account preferences of a group of stakeholders. Firstly, we discuss the Electre TRI-rC method (Kadziński et al. 2015b) that has been employed to assign the materials to a set of pre-defined and ordered classes. It incorporates the SRF procedure to compute the criteria weights (Figueira and Roy 2002). The method has been extended to a group decision setting to derive for each material some group class acceptability indices, which indicate the proportion of stakeholders that accept an assignment of the material to a given class. Secondly, we have adapted Stochastic Multi-criteria Acceptability Analysis (SMAA; Lahdelma and Salminen 2001; Tervonen and Figueira 2008; Tervonen et al. 2007) to the context of Electre TRI-rC and SRF procedure. It has been used to conduct robustness analysis (Roy 2010) for the results obtained in the first part, i.e., to validate their certainty while avoiding the arbitrary choice of criteria weights, which is conducted by the SRF procedure. Thirdly, we have proposed an algorithm for constructing a group compromise ranking of criteria based on the orders provided by the individual DMs. This ranking of criteria has been used as an input for SMAA to offer yet another view on the stability of computed results.

Let us use the following notation (Kadziński et al. 2015a):

- $A = \{a_1, a_2, \ldots, a_n\}$ is a set of alternatives (insulating materials);
- $G = \{g_1, g_2, \ldots, g_m\}$ is a family of evaluation criteria that represent relevant points of view on the quality of assessed alternatives;
- $g_j(a)$ is the performance of alternative a with respect to criterion g_j, $j = 1, \ldots, m$ (when presenting the method, without loss of generality, we assume that all criteria are of gain type, i.e., the greater the performance, the better);
- $C_1, C_2, \ldots, C_p$ are the preference ordered classes to which alternatives should be assigned; we assume that C_h is preferred to C_{h-1} for $h = 2, \ldots, p$.

3.1 Assessment of Insulating Materials Within a Group Decision Framework Incorporating Electre TRI-rC and the SRF Procedure

In this section, we present the Electre TRI-rC method (Kadziński et al. 2015b) that is used to assign the materials to a set of pre-defined and ordered classes. The method derives for each material a possibly imprecise assignment by constructing and exploiting an outranking relation S (Figueira et al. 2013). This relation quantifies an outcome of the comparison between the materials and a set of characteristic class profiles (Rezaei et al. 2017). In what follows, we discuss the main steps of the incorporated approach.

Step 1 For each class C_h, provide the most typical (representative) performances on all criteria g_j, $j = 1, \ldots, m$, thus specifying the characteristic profiles b_h, $h = 1, \ldots, p$ (Almeida Dias et al. 2010). Defining such profiles was found intuitive and manageable by the involved experts, which was the main reason for incorporating Electre TRI-rC in the study. The set of all characteristic profiles is denoted by B.

Steps 2–7 are conducted separately for each Decision Maker ($DM_k, k = 1, \ldots, K$) in $\partial^{\mathrm{K}} = \{DM_1, DM_2, \ldots, DM_K\}$.

Step 2 Determine the weight w_j^k of each criterion g_j, $j = 1, \ldots, m$, using the SRF procedure (Figueira and Roy 2002). This method expects DM_k to:

- Assign some importance rank $l^k(j)$ to each criterion g_j; this is attained by ordering the cards with criteria names from the least to the most important (the greater $l^k(j)$, the greater w_j^k; some criteria can be assigned the same rank, thus being judged indifferent);
- Quantify a difference between importance coefficients of the successive groups of criteria judged as indifferent, L_s^k and L_{s+1}^k, by inserting e_s^k white (empty) cards between them (the greater e_s^k, the greater the difference between the weights assigned to the criteria contained in L_{s+1}^k and L_s^k);
- Specify ratio Z^k between the importances of the most and the least significant criteria denoted by $L_{v(k)}^k$ and L_1^k.

These inputs are used to derive the criteria weights as follows (Figueira and Roy 2002; Corrente et al. 2016):

$$w_j^k = 1 + \frac{\left(Z^k - 1\right)\left[l^k(j) - 1 + \sum_{s=1}^{l(j)-1} e_s^k\right]}{v(k) - 1 + \sum_{s=1}^{v-1} e_s^k}.$$

Steps 3–6 are conducted for each pair consisting of alternative a and profile b_h.

Step 3 For each criterion g_j compute a marginal concordance index $c_j^k(a, b_h)$ defined as follows:

$$c_j^k(a, b_h) = \begin{cases} 1 & \text{if } g_j(a) - g_j(b_h) \geq 0, \\ 0 & \text{if } g_j(a) - g_j(b_h) < 0. \end{cases}$$

The index quantifies a degree to which a is at least as good as b_h on g_j. Let us remark that in our study the experts defined the performances of characteristic profiles on all criteria by selecting them from the performances of the considered materials. This facilitated the preference elicitation process when dealing with a set of criteria with heterogeneous performance scales. In this perspective, when comparing the alternatives with the characteristic class profiles, we decided to exploit only the ordinal character of criteria and not use the discrimination (indifference and preference) thresholds, which can be, in general, employed in Electre. That is, in our application, the outranking of alternative a over profile b_h on g_j means that $g_j(a)$ is at least as good as the most typical (representative) performance for class C_h on g_j of some considered material.

Step 4 Compute a comprehensive concordance index $\sigma^k(a, b_h)$ defined in the following way:

$$\sigma^k(a, b_h) = \frac{\sum_{j=1}^{m} w_j^k c_j^k(a, b_h)}{\sum_{j=1}^{m} w_j^k}.$$

The index quantifies a joint strength of a subset of criteria supporting the hypothesis about a outranking b_h ($aS^k b_h$). Note that in our study, no criterion was judged strong enough to be attributed a power to veto against the outranking relation. Thus, no discordance effect has been considered.

Step 5 Specify the cutting level λ^k (also called majority threshold), and compare $\sigma^k(a, b_h)$ with λ^k to verify the truth of a crisp outranking relation $aS^k b_h$ in the following way:

$$\sigma^k(a, b_h) \geq \lambda^k \Rightarrow aS^k b_h.$$

The truth of relation $b_h S^k a$ can be verified analogously.

Step 6 Use information on the truth or falsity of $aS^k b_h$ and $b_h S^k a$ to check the validity of:

- a being preferred to b_h ($aS^k b_h \wedge not\left(b_h S^k a\right) \Rightarrow a \succ_k b_h$);
- b_h being preferred to a ($b_h S^k a \wedge not\left(aS^k b_h\right) \Rightarrow b_h \succ_k a$);
- a being indifferent with b_h ($aS^k b_h \wedge b_h S^k a \Rightarrow a \sim_k b_h$);
- a being incomparable with b_h ($not\left(aS^k b_h\right) \wedge not\left(b_h S^k a\right) \Rightarrow a?_k b_h$).

Step 7 For alternative a determine its desired class interval $C^k(a) = \left[C_L^k(a), C_R^k(a)\right]$ by applying the assignment rules of ELECTRE TRI-rC (Kadziński et al. 2015b). To compute the worst class $C_L^k(a)$, compare a successively to b_h, for $h = p-1, \ldots, 1$, seeking the first (i.e., the best) characteristic profile b_h such that:

$$a \succ_k b_h \wedge \sigma^k(a, b_{h+1}) > \sigma^k(b_h, a),$$

and select $C_L^k(a) = C_{h+1}$. When no such a profile is found, $C_L^k(a) = C_1$.

To compute the best class $C_R^k(a)$, compare a successively to b_h, for $h = 2, \ldots, p$, seeking the first (i.e., the worst) characteristic profile b_h such that:

$$b_h \succ_k a \wedge \sigma^k(b_{h-1}, a) > \sigma^k(a, b_h),$$

and select $C_R^k(a) = C_{h-1}$. In case no such a profile is found, $C_R^k(a) = C_p$.

Step 8 Combine the individual class assignments for all DMs into *group class acceptability indices* $E^{\partial}(a, h)$ (Damart et al. 2007; Kadziński et al. 2016). These are defined as the proportion of DMs (stakeholders) that accept an assignment of alternative a to class C_h, i.e.:

$$E^{\partial^{\mathrm{K}}}(a, h) = \frac{\sum_{k=1}^{K} E^k(a, h)}{K},$$

where for $k = 1, \ldots, K$:

$$E^k(a, h) = \begin{cases} 1 & if\ C_h \in C^k(a), \\ 0 & if\ C_h \notin C^k(a). \end{cases}$$

This measure indicates a cumulative support given to the assignment of a to C_h by all group members.

3.2 Stochastic Multi-criteria Acceptability Analysis with Electre TRI-rC

The SRF procedure derives the precise weight values from the ranking of criteria, intensities of preference, and ratio between the most and the least important criteria provided by DM_k applying some arbitrary rule (Figueira and Roy 2002). However, there exist multiple weight vectors compatible with such incomplete preference information. Recently, many researchers have raised the robustness concern in view of the SRF procedure to quantify the impact of uncertainty in the selection of an arbitrary weight vector on the stability of computed recommendation. In particular, Siskos and Tsotsolas (2015) proposed a set of robustness rules for the SRF procedure to obtain tangible and adequately supported results. Then, Govindan et al. (2017) suggested to exploit the whole set of compatible weight vectors to construct the necessary and possible results being confirmed by, respectively, all or at least one compatible vector. Further, Corrente et al. (2017) adapted the stochastic analysis of recommendation with the SRF procedure to the context of Electre III. We follow the latter research direction and integrate Stochastic Multi-criteria Acceptability Analysis (Lahdelma and Salminen 2001; Tervonen et al. 2007) to handle possibly imprecise weight values compatible with the ranking of criteria and to derive robust recommendation with Electre TRI-rC.

SMAA applies the Monte Carlo simulation to provide each DM with the acceptability indices which measure the variety of different preferences (in particular, weight vectors) that confirm the validity of particular elements of the recommendation. In our case, the space $w^k(SRF)$ of weight vectors compatible with preferences of DM_k is defined by the following constraint set $E^k(SRF)$:

$$\left.\begin{array}{c} [O1]\ w_i^k > w_j^k,\ \text{for all } g_i \in L_t^k,\ g_j \in L_s^k \ \text{and } t > s, \\ [O2]\ w_i^k = w_j^k,\ \text{for all } g_i,\ g_j \in L_s^k, \\ [O3]\ w_i^k = Z^k w_j^k,\ \text{for all } g_i \in L_{v(k)}^k,\ g_j \in L_1^k, \\ [O4]\ w_{j+1}^k - w_j^k > w_{p+1}^k - w_p^k,\ \text{if } e_j^k > e_p^k, \\ [O5]\ \sum_{j=1}^m w_j^k = 1, \end{array}\right\} E^k(SRF)$$

where the interpretation of different constraints is as follows:

- $[O1]$ ensures that criteria ranked better by DM_k will be assigned greater weight;
- $[O2]$ guarantees that criteria deemed indifferent by DM_k will be assigned equal weights;
- $[O3]$ sets the ratio Z between weights of the most and the least significant criteria;

- $[O4]$ respects the intensities of preference for different pairs of criteria that have been quantified with the number of inserted empty cards;
- $[O5]$ normalizes the weights.

These constraints also ensure that all weights are positive. For each DM_k, each weight vector $w \in w^k(SRF)$ and each alternative $a \in A$, we compute the resulting class assignment $C^k_w(a) = \left[C^k_{w,L}(a), C^k_{w,R}(a)\right]$ with Electre TRI-rC.

We define the *class range stochastic acceptability index* $CRSAI^k(a, [L, R])$ (Kadziński et al. 2013) on a range of classes $\left[C^k_L(a), \ldots, C^k_R(a)\right]$ with $L \leq R$ as the proportion of compatible weights $w \in w^k(SRF)$ that assign alternative a precisely to the range of classes $\left[C^k_L(a), \ldots, C^k_R(a)\right]$. Formally, the index is computed as follows:

$$CRSAI^k(a, [h_L, h_R]) = \int_{w \in w^k(SRF)} m(w, a, [h_L, h_R])\, dw,$$

where $m(w, a, [h_L, h_R])$ is the class range membership function:

$$m(w, a, [h_L, h_R]) = \begin{cases} 1, & if\ C^k_{w,L}(a) = C_{h_L} \text{ and } C^k_{w,R}(a) = C_{h_R}, \\ 0, & \text{otherwise.} \end{cases}$$

Further, we compute the proportion of $w \in w^k(SRF)$ for which C_h is within $\left[C^k_{w,L}(a), C^k_{w,R}(a)\right]$, i.e., the proportion of weights that either precisely or imprecisely assign a to C_h (Kadziński and Tervonen 2013; Kadziński et al. 2014). Let us define such a *cumulative class stochastic acceptability index* $CuCSAI^k(a, h)$ as:

$$CuCSAI^k(a, h) = \sum\nolimits_{[h_L, h_R]: h \in [h_L, h_R]} CRSAI^k(a, [h_L, h_R]).$$

We estimate $CRSAI$s with acceptable error bounds by sampling the space $w^k(SRF)$ with the Hit-And-Run (HAR) algorithm (Tervonen et al. 2013). Overall, $CRSAI^k(a, [h_L, h_R])$ and $CuCSAI^k(a, h)$ can be interpreted as a support given by DM_k to the assignment of a to, respectively, $\left[C_{h_L}, C_{h_R}\right]$ or C_h.

To measure a cumulative support given to the assignment of a to C_h by all DMs in ∂^{K}, we consider a *cumulative group class stochastic acceptability index* $CuCSAI^{\partial^{\mathrm{K}}}(a, h)$, defined as follows (Kadziński et al. 2016, 2018):

$$CuCSAI^{\partial^{\mathrm{K}}}(a, h) = \frac{\sum_{k=1}^{K} CuCSAI^k(a, h)}{K}.$$

3.3 Selection of a Group Compromise Ranking of Criteria

In this section, we introduce a procedure for deriving a *compromise complete ranking of criteria* based on the rankings provided individually by each DM_k within the SRF procedure. The procedure builds on the algorithm that was introduced by Govindan et al. (2017) for constructing a utilitarian ranking of alternatives. Hence, we adopt an

Table 1 Definition of distances $\delta\left(R_{k'}^{jl}, R_{k''}^{jl}\right)$ between different pairwise relations

$R_{k'}^{jl} \backslash R_{k''}^{jl}$	$g_j \succ_{k''} g_l \left(\succ_{k''}^{jl}\right)$	$g_j \prec_{k''} g_l \quad \left(\prec_{k''}^{jl}\right)$	$g_j \sim_{k''} g_l \left(\sim_{k''}^{jl}\right)$
$g_j \succ_{k'} g_l \left(\succ_{k'}^{jl}\right)$	0	2	1
$g_j \prec_{k'} g_l \left(\prec_{k'}^{jl}\right)$	2	0	1
$g_j \sim_{k''} g_l \left(\sim_{k'}^{jl}\right)$	1	1	0

idea of minimizing a sum of of distances between the compromise ranking and all individual rankings.

When considering a complete ranking of criteria for DM_k, for each pair (g_j, g_l) one of the three relations holds: g_j is preferred to $g_l \quad (g_j \succ_k g_l)$, or g_j is indifferent with $g_l \quad (g_j \sim_k g_l)$, or g_l is preferred to $g_j \quad (g_j \prec_k g_l)$. Let $R_{k'}^{jl}$ and $R_{k''}^{jl}$ denote the relations holding between g_j and g_l in the rankings provided by, respectively, $DM_{k'}$ and $DM_{k''}$ (e.g., $R_{k'}^{jl}$ is $\succ_{k'}^{jl}$ or $\sim_{k'}^{jl}$ or $\prec_{k'}^{jl}$). The distances $\delta(R_{k'}^{jl}, R_{k''}^{jl})$ between $R_{k'}^{jl}$ and $R_{k''}^{jl}$ are provided in Table 1 (for a detailed justification of these values, see Roy and Słowiński 1993). A distance between two rankings of criteria provided by $DM_{k'}$ and $DM_{k''}$ involving all ordered pairs of criteria (g_j, g_l) is defined as follows:

$$\sum_{j,l:j<l} \delta\left(R_{k'}^{jl}, R_{k''}^{jl}\right).$$

In what follows, we present a Binary Linear Program (BLP) for constructing a compromise ranking of criteria for group ∂^{K} involving K DMs. Following Govindan et al. (2017), for each pair of criteria (g_j, g_l), we introduce two binary variables p_∂^{jl} and i_∂^{jl} (see constraint $[R1]$ in E^∂ (SFR)) with the following interpretation:

- p_∂^{jl} represents a weak preference of g_j over g_l in the compromise ranking (i.e., in case $p_\partial^{jl} = 1$, then $g_j \succ_\partial g_l$ or $g_j \sim_\partial g_l$); note that p_∂^{jl} and p_∂^{lj} can be used to instantiate one of the three relations $\succ_\partial^{jl}$, $\sim_\partial^{jl}$, or $\prec_\partial^{jl}$ for g_j and g_l; that is, if $p_\partial^{jl} = 1$ and $p_\partial^{lj} = 0$, then $g_j \succ_\partial g_l$; if $p_\partial^{jl} = 0$ and $p_\partial^{lj} = 1$, then $g_j \prec_\partial g_l$; if $p_\partial^{jl} = 1$ and $p_\partial^{lj} = 1$, then $g_j \sim_\partial g_l$;
- i_∂^{jl} represents an indifference $\sim_\partial$ between g_j and g_l (i.e., in case $p_\partial^{jl} = 1$ and $p_\partial^{lj} = 1$, then $i_\partial^{jl} = 1$ and $g_j \sim_\partial g_l$; see [R3]).

Since we impose completeness and transitivity on a weak preference relation, we require that $p_\partial^{jl} = 1$ or $p_\partial^{lj} = 1$ (see $[R2]$) and that $p_\partial^{jr} = 1$ and $p_\partial^{rl} = 1$ imply $p_\partial^{jl} = 1$ (see $[R4]$). When constructing a utilitarian complete ranking of criteria, we aim at minimizing a comprehensive distance between relations ($\succ_\partial$, $\prec_\partial$, or $\sim_\partial$) instantiated for all pairs of criteria in the compromise ranking and relations observed for these pairs in the individual DMs' rankings (for DM_k, the relation between g_j and g_l $(j < l)$ is denoted by R_k^{jl}):

$$\min \sum_{j,l:j<l} \sum_{k=1}^{K} \left[p_{\partial}^{jl} \delta \left(R_k^{jl}, \succ_{\partial}^{jl} \right) + p_{\partial}^{lj} \delta \left(R_k^{jl}, \prec_{\partial}^{jl} \right) + i_{\partial}^{jl} \left[\delta \left(R_k^{jl}, \sim_{\partial}^{jl} \right) - \delta \left(R_k^{jl}, \succ_{\partial}^{jl} \right) - \delta \left(R_k^{jl}, \prec_{\partial}^{jl} \right) \right] \right]$$

$$\left.\begin{array}{l} [RI] \text{ for all } j, l = 1, 2, \ldots, m : j \neq l \\ [R1]\ p_{\partial}^{jl}, i_{\partial}^{jl} \in \{0, 1\}, \\ [R2]\ p_{\partial}^{jl} + p_{\partial}^{lj} \geq 1, \\ [R3]\ i_{\partial}^{jl} = p_{\partial}^{jl} + p_{\partial}^{lj} - 1, \\ [RII] \text{ for all } j, l, r = 1, 2, \ldots, m : j \neq l \neq r \\ [R4]\ p_{\partial}^{jl} \geq p_{\partial}^{jr} + p_{\partial}^{rl} - 1.5. \end{array}\right\} E^{\partial}\ (SFR)$$

If $g_j \succ_{\partial} g_l\ (p_{\partial}^{jl} = 1, \quad p_{\partial}^{lj} = 0, \text{and}\, i_{\partial}^{jl} = 0)$, $g_j \prec_{\partial} g_l\ (p_{\partial}^{jl} = 0, \quad p_{\partial}^{lj} = 1, i_{\partial}^{jl} = 0)$, or $g_j \sim_{\partial} g_l\ (p_{\partial}^{jl} = 1, \quad p_{\partial}^{lj} = 1, i_{\partial}^{jl} = 1)$ has been instantiated in the compromise ranking, it contributes with, respectively, $\sum_{k=1}^{K} \delta \left(R_k^{jl}, \succ_{\partial}^{jl} \right)$, $\sum_{k=1}^{K} \delta \left(R_k^{jl}, \prec_{\partial}^{jl} \right)$ or $\sum_{k=1}^{K} \delta \left(R_k^{jl}, \sim_{\partial}^{jl} \right)$ to a value of the objective function (for a detailed explanation, see Govindan et al. 2017).

Once a group compromise ranking of criteria is constructed, we conduct robustness analysis with SMAA in the same way as described in the previous section for an individual DM. This leads us to deriving *cumulative group compromise class stochastic acceptability indices* $CuCCSAI^{\partial K}\ (a, h)$.

3.4 Decision Aiding with the Proposed Approach

Multiple criteria sorting decisions can be aided with the proposed group decision making framework through the process illustrated in Fig. 1. It starts with specifying the sets of alternatives, criteria, and ordered classes as well as the alternatives' evaluations (performances) on the criteria.

Then, the preference information is elicited from the involved experts and/or stakeholders. Each stakeholder is required to provide a cutting level as well as a ranking of criteria that incorporates the intensities of preference and the ratio between the importance coefficients of the most and the least significant criteria, as required by the SRF procedure. Moreover, the experts are expected to define a characteristic profile for each class. In our study, the profiles were agreed by multiple experts, but, in general, the methodological framework admits that each stakeholder provides his/her individual set of profiles.

Further, the method derives three types of results. These indicate a support that is given to the assignment of considered alternatives to different classes via the application of Electre TRI-rC for different sets of weights and cutting levels compatible with the preferences of the involved experts. In two cases, the preferences of the individual stakeholders are aggregated only at the output level. Depending on whether these individual preferences are processed using the SRF procedure or the Monte Carlo simulation, the method computes, respectively, group class acceptability indices or

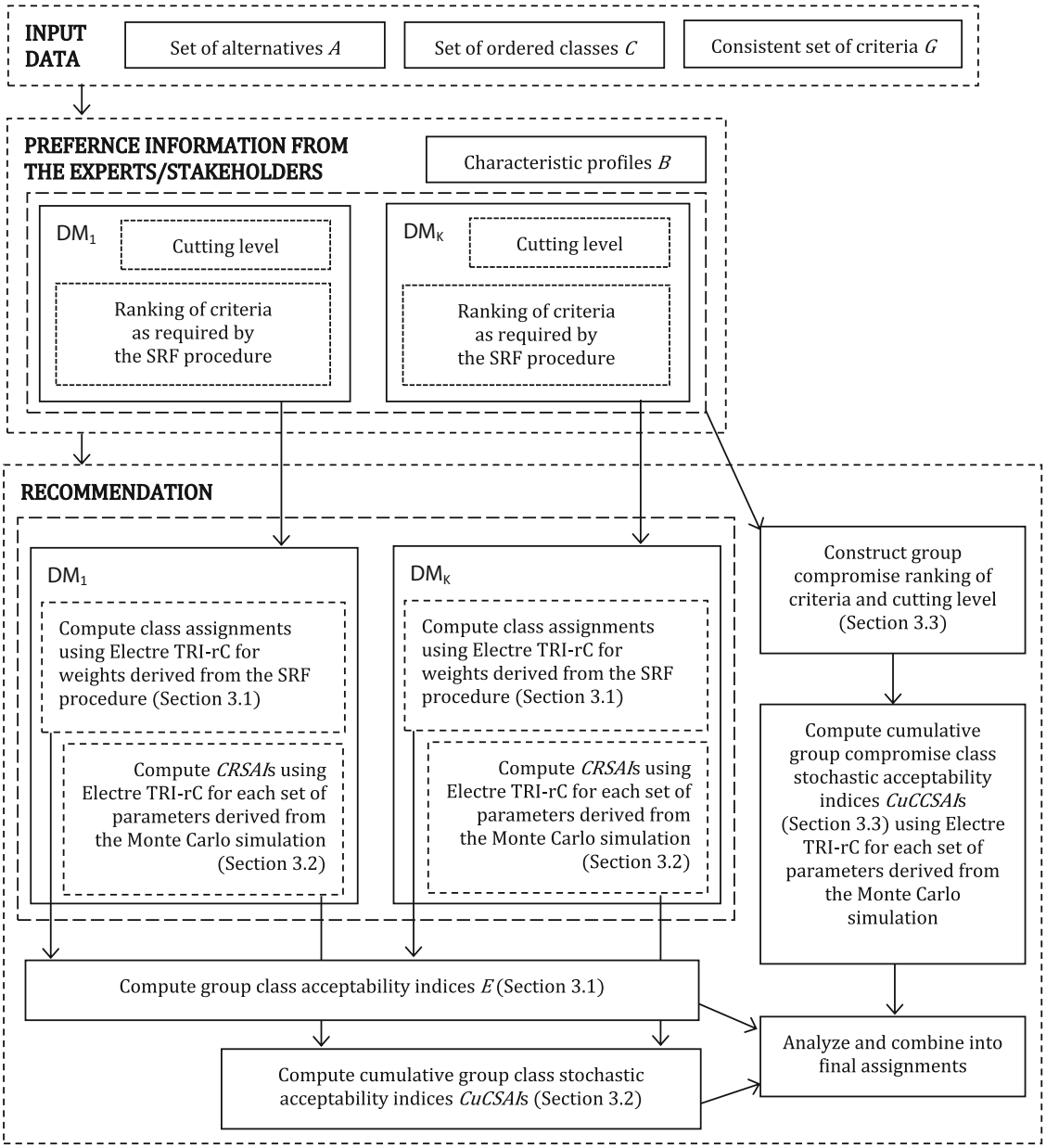


Fig. 1 Decision aiding process with the proposed group decision methodological framework

cumulative group class stochastic acceptability indices. In the third case, the preferences are aggregated at the input level by constructing a group compromise ranking of criteria. Then, the method applies SMAA to derive cumulative group compromise class stochastic acceptability indices.

Finally, these three types of outcomes should be analyzed and combined into the recommended assignments. This is straightforward in case the support given to the assignment of alternatives to decision classes by different results is similar. In case of ambiguous indications by different procedures, the inconsistency needs to be raised by a decision analyst.

Obviously, it is not required to use all three types of procedures and respective results for each study. This may be useful when offering different viewpoints on the robustness of sorting recommendation is desired. Otherwise, one can employ just a

single procedure for processing the experts' preferences depending on whether they should be aggregated at the input or output level and whether the robustness analysis should be incorporated into a particular study.

4 Results of Multiple Criteria Assessment of Insulating Materials with the Outranking Preference Model and Characteristic Class Profiles

The study aims at evaluating overall sustainability of coating materials used in buildings retrofitting. We consider 13 materials listed in Table 2 (they are denoted by $A = \{a_1, a_2, \ldots, a_{13}\}$). All materials having a thickness of 15cm were placed internally on the roof of a model building typical for central Italy, and evaluated from the socio-economic and environmental viewpoints. The six relevant criteria which have been used to assess the materials are: *hour of discomfort* (g_1; DH),CO_2 *avoidance* (g_2); *Net Present Value* (g_3; NPV), *human health* (g_4); *ecosystem quality* (g_5), and *consumed resources* (g_6). In what follows, we explain their meaning.

Discomfort degree Hour (g_1; the less, the better) evaluates a thermal performance of a building on an annual basis (CEN 2007) in accordance with the EN 15251 standard. Thus defined, it serves as a measure of comfort. The performance on g_1 is quantified as an overall time during which the temperature falls outside the second comfort category that was considered in the study (Carlucci and Pagliano 2012), and then weighing it by how much the limit has been exceeded. For this purpose, we have used the following equation:

$$g_1(a) = \sum_{i=1}^{8760} \frac{10}{60} |CC_2 - OT_i|$$

where CC_2 is the lower or upper limit of the assumed comfort category, OT_i is the operative temperature at hour *ì*, and the multiplier $\frac{10}{60}$ refers to an employed time step of 10 minutes.

CO_2 *avoidance* (g_2; the more, the better) measures the energy saved during the building life by using a particular insulating material when compared to the case of no insulation in the following way:

$$g_2(a) = \frac{ES * 277.78 * 406.31}{10^6}$$

where ES is the estimated Energy Saved in GJ at time t with a time horizon of 25 years, 277.78 is a conversion factor to GJ in kWH, while 406.31 is the conversion factor for Italy from kWH to kg of CO_2 per year (EIA, 2015). Therefore, the CO_2 avoided refers only to the use phase, which is not considered in the LCA study.

Net present value (g_3; the more, the better) is the difference between the present values of cash outflows and inflows. On one hand, the outflows involve Primary Energy Input (PEI) cost, installation cost I at time $t = 0$, and the dismissing cost EL_T after the lifespan T of the investment (25 years). On the other hand, the inflows refer to the Cost of Energy Saved ES_t in different time periods t. Overall, we have computed

NPV as follows:

$$g_3(a) = -PEI - I + \sum_{t=0}^{T} \frac{ES_t}{(1+i)^t} - \frac{EL_T}{(1+i)^T}$$

where i is the discount rate. For a detailed justification of this measure, see Menconi and Grohmann (2014). Thus defined, NPV can be seen as an outcome of Life Cycle Costing, which is an economic methodology for assessing the profitability of using different alternatives by taking into account the costs they incur at different stages of a life cycle (e.g., construction, operations, and maintenance).

For the assessment of environmental impacts, we have used the Eco-indicator 99 method (Goedkoop and Spriensma 2001) implemented in the *SimaPro* software (Product Ecology Consultants 1990). The method aggregates the results of Life Cycle Assessment into a set of parameters that can be interpreted as damage categories. In general, LCA is useful for identifying the environmental implications of a given alternative through the quantification of consumed resources (e.g., energy, raw materials, water) and related emissions (e.g., emissions into the air, water and soil, waste and co-products) (Paolotti et al. 2017). We used the following three environmental Eco-indicators expressed on a dedicated point scale:

- *Human health* (g_4; the less, the better) which is derived from the analysis of the following normalized impact categories: carcinogens, respiratory organics and inorganics, climate change, radiation, and ozone layer;
- *Ecosystem quality* (g_5; the less, the better) which is made up by the following three normalized impact categories: ecotoxicity, acidification/eutrophication, and land use;
- *Resources* (g_6; the less, the better) which aggregates two normalized impact categories: minerals and fossil fuels.

The LCA focused on the production phase, starting from the production of a raw material to the obtaining of its complete version. We omitted the use and disposal phases, hence implementing an LCA "from cradle to gate" (Paolotti et al. 2016). All the impacts were calculated considering a functional unit of 1 m^3 of insulating material.

The performances of 13 insulating materials with respect to 6 criteria are provided in Table 2. For all materials but hemp fibres, Ecoinvent Database (Ecoinvent 2010) was used as a source of foreground and background data related to both production and assembly processes as well as to the transport, electricity and fuel consumption. Instead, for the hemp processes the underlying data was derived from Zampori et al. (2013).

The objective of the case study is to give an easily interpretable comprehensive assessment of the materials' sustainability. This is achieved by assigning them to a set of three pre-defined and ordered classes: C_1 (low sustainability), C_2 (medium sustainability), and C_3 (high sustainability).

The study involved elicitation of preferences from the two groups of stakeholders. On one hand, a characteristic profile b_h for each class C_h, $h = 1, 2, 3$, has been collectively specified by the experts from the university-based engineering team specialized

Table 2 Performances of 13 insulating materials with respect to 6 criteria

Insulating material	a	g_1	g_2	g_3	g_4	g_5	g_6
Performance unit	–	Hours	kg of CO_2		Points	Points	Points
Autoclave aerated complete	a_1	4889.339	158.63	283.41	0.009703	0.000636	0.015876
Corkslab	a_2	3974.451	178.49	282.01	0.022122	0.018376	0.040660
Expanded perlite	a_3	3893.646	179.11	326.26	0.006451	0.000759	0.043280
Fibreboard hard	a_4	3657.799	185.29	243.45	0.039111	0.014516	0.136345
Glass wool	a_5	3681.898	187.35	316.92	0.010608	0.001307	0.033364
Gypsum fibreboard	a_6	7051.231	103.24	135.88	0.047131	0.003916	0.070469
Hemp fibres	a_7	3921.449	182.59	334.10	0.002336	0.003079	0.008207
Kenaf fibres	a_8	3685.510	186.82	341.79	0.004760	0.015137	0.003079
Mineralized wood	a_9	4392.808	167.63	245.45	0.042932	0.004548	0.083149
Plywood	a_{10}	7636.502	87.58	71.26	0.095717	0.201332	0.126167
Polystyrene foam	a_{11}	3750.482	187.13	322.02	0.002801	0.000217	0.016521
Polyurethane	a_{12}	3357.309	194.18	330.35	0.013225	0.000564	0.043280
Rock wool	a_{13}	3659.441	188.45	346.14	0.019183	0.000825	0.009846

Table 3 Performances of the characteristic profiles for three classes

Profile	g_1	g_2	g_3	g_4	g_5	g_6
b_1	7051.231	158.63	135.88	0.042932	0.015137	0.083149
b_2	4392.808	182.59	283.41	0.013225	0.003079	0.043280
b_3	3659.441	187.35	330.35	0.004760	0.000636	0.009846

in the materials and retrofitting of rural buildings. On the other hand, the preferences on the importance of individual criteria have been elicited individually from multiple stakeholders who were owners of rural buildings interested in a renovation of their houses for improving the energetic performance. Thus, they can be perceived as potential consumers of the insulating materials.

When it comes to the characteristic profiles, the experts decided to define them by indicating one of the performances observed in the set of materials. The consensus between the experts on the most typical performance levels for each class has been reached during an interactive focus group. These levels are summarized in Table 3.

4.1 Results of Multiple Criteria Assessment of the Insulating Materials Within a Group Decision Framework Incorporating Electre TRI-rC and the SRF Procedure

The weights representing the importance of individual criteria have been elicited from the rural buildings' owner. In what follows, we call them stakeholders. Overall, we approached 63 owners by explaining them the characteristics of different

materials, the interpretation of all criteria and their relation to different phases of the materials' life cycle. Among them, 38 stakeholders (let us denote them by $\partial^{\mathrm{K}} = \{DM_1, DM_2, \ldots, DM_{38}\}$) claimed to understand the meaning and role of different criteria, and expressing their willingness to provide preferences on the criteria importance.

In Table 4, we present the incomplete preference information required by the SRF procedure, which was provided by three selected stakeholders. We also report the computed weights w_j^k and cutting level λ^k. All stakeholders agreed that λ^k should be equal to the sum of weights of the three most important criteria. The complete data for all group members is provided in the supplementary material available as an e-Appendix (the same remark applies to the results discussed in the following sections).

The results of a comprehensive comparison between 13 materials and 3 characteristic profiles are quantified with the comprehensive concordance indices. In Table 5, we present such indices for four exemplary materials for DM_1. Table 5 exhibits also the justification of delivered assignment for the exemplary materials. For instance, a precise assignment of a_6 to C_1 can be explained with b_2 being preferred to a_6 and there existing sufficiently strong support in favor of b_1 outranking a_6 ($\sigma^1(a_6, b_2) = 0.000 < \sigma^1(b_1, a_6) = 0.524$).

In Table 6, we report the assignments obtained for all materials for different DMs. In particular, for DM_1 there are 6 materials assigned to the best class ($a_5, a_7, a_8, a_{11}, a_{12}, a_{13}$), 3 materials whose quality is evaluated as medium (a_1, a_2, a_3), and 4 materials judged as bad (a_4, a_6, a_9, a_{10}). The assignments for DM_5 are the same except for a_4 being imprecisely assigned to $[C_1, C_2]$.

The spaces of consensus and disagreement with respect to the assignments obtained for all DMs are quantified with the group class acceptability indices $E^{\partial}(a, h)$ (see Table 7). For example, for a_1 none stakeholder confirmed its assignment to the worst class C_1, 36 out of 38 stakeholders supported its assignment to the medium class C_2, and 3 stakeholders suggested the assignment of a_1 to the best class C_3. These numbers have been translated to the following group acceptability indices: $E^{\partial}(a_1, 1) = \frac{0}{38} = 0$, $E^{\partial}(a_1, 2) = \frac{36}{38} = 0.95$, and $E^{\partial}(a_1, 3) = \frac{3}{38} = 0.08$. On the contrary, for a_2 all stakeholders agreed with respect to its assignment to C_2 ($E^{\partial}(a_2, 2) = \frac{38}{38} = 1.0$), while the results obtained for 6 of them additionally indicated hesitation in terms of its assignment to C_1 ($E^{\partial}(a_2, 1) = \frac{6}{38} = 0.16$).

The analysis of $E^{\partial}(a, h)$ leads to indicating the assignments which are necessary (in case $E^{\partial}(a, h) = 1$), possible (if $E^{\partial}(a, h) > 0$), and impossible (if $E^{\partial}(a, h) = 0$) in terms of the support they are provided in the group of stakeholders. Additionally, these results clearly indicate the most and the least probable assignments. In particular, for each material we are able to indicate the class with the greatest support among all stakeholders. It is C_1 for a_6, a_9 and a_{10}, C_2 for a_1, a_2, a_3 and a_4, or C_3 for a_5, a_7, a_8, a_{11}, a_{12}, and a_{13}. The support which is given to the assignment of the materials to other classes is significantly smaller. For clarity of presentation, in all tables exhibiting stochastic acceptability indices (Tables 7, 8, 9 and 11), the text in bold indicates the class with the greatest support for a given material.

Table 4 The order of cards with criteria names (ranks $l^k(j)$), white cards e_s^k, and ratio Z^k provided by the three selected DMs in the SRF procedure, the weights w_j^k derived from the SRF procedure, and the cutting level λ^k

$DM_1(Z^1 = 10, \lambda^1 = 0.714)$				$DM_2(Z^2 = 5, \lambda^2 = 0.696)$				$\cdots$	$DM_{38}(Z^{38} = 5, \lambda^{38} = 0.682)$			
g_j	$l^1(j)$	e_s^1	w_j^1	g_j	$l^2(j)$	e_s^2	w_j^2	$\cdots$	g_j	$l^{38}(j)$	e_s^{38}	w_j^{38}
g_1	1		0.024	g_3	1		0.049	$\cdots$	$g_1, \ g_3$	1		0.045
		1		g_1	2		0.088				2	
g_3	2		0.085			1		$\cdots$	$g_2, \ g_4, \ g_5, \ g_6$	2		0.227
		2		g_4	3		0.167	$\cdots$				
g_2	3		0.177	g_2	4		0.206	$\cdots$				
		1		$g_5, \ g_6$	5		0.245	$\cdots$				
$g_4, \ g_5, \ g_6$	4		0.238					$\cdots$				

Table 5 Credibility indices and class assignments obtained with ELECTRE TRI-rC for four exemplary materials for DM_1 (cutting level $\lambda^1 = 0.714$)

	b_1	b_2	b_3	$\left[C_L^1(a), C_R^1(a)\right]$		b_1	b_2	b_3	$\left[C_L^1(a), C_R^1(a)\right]$
a_1	$\succ$	$\succ$	$\prec$	$[C_2, C_2]$	a_6	?	$\prec$	$\prec$	$[C_1, C_1]$
$\sigma^1(a_1, b_h)$	1.000	0.799	0.238		$\sigma^1(a_6, b_h)$	0.585	0.000	0.000	
$\sigma^1(b_h, a_1)$	0.177	0.286	1.000		$\sigma^1(b_h, a_6)$	0.524	1.000	1.000	
a_{11}	$\succ$	$\succ$	?	$[C_3, C_3]$	a_{12}	$\succ$	$\succ$	$\prec$	$[C_3, C_3]$
$\sigma^1(a_{11}, b_h)$	1.000	1.000	0.476		$\sigma^1(a_{12}, b_h)$	1.000	1.000	0.524	
$\sigma^1(b_h, a_{11})$	0.000	0.000	0.524		$\sigma^1(b_h, a_{12})$	0.000	0.476	0.738	

Table 6 Class assignments obtained with Electre TRI-rC for all materials and different stakeholders

a	DM_1	DM_2	DM_3	DM_4	DM_5	DM_6	DM_7	DM_8	DM_9	DM_{10}	$\cdots$	DM_{38}
a_1	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$\cdots$	$[C_2,C_2]$
a_2	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$\cdots$	$[C_2,C_2]$
a_3	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$[C_2,C_2]$	$\cdots$	$[C_2,C_2]$
a_4	$[C_1,C_1]$	$[C_2,C_2]$	$[C_1,C_1]$	$[C_2,C_2]$	$[C_1,C_2]$	$[C_1,C_1]$	$[C_2,C_3]$	$[C_2,C_2]$	$[C_2,C_3]$	$[C_2,C_2]$	$\cdots$	$[C_2,C_2]$
a_5	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$\cdots$	$[C_3,C_3]$
a_6	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$\cdots$	$[C_1,C_1]$
a_7	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_2,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_2,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$\cdots$	$[C_3,C_3]$
a_8	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$\cdots$	$[C_3,C_3]$
a_9	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_2,C_2]$	$[C_1,C_1]$	$[C_2,C_2]$	$[C_1,C_1]$	$\cdots$	$[C_1,C_1]$
a_{10}	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$[C_1,C_1]$	$\cdots$	$[C_1,C_1]$
a_{11}	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$\cdots$	$[C_3,C_3]$
a_{12}	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$\cdots$	$[C_3,C_3]$
a_{13}	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_2,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$[C_3,C_3]$	$\cdots$	$[C_3,C_3]$

Table 7 Group class acceptability indices $E^{\partial}(a, h)$

h \ a	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}
1	0.00	0.16	0.00	0.34	0.00	**1.00**	0.00	0.00	**0.76**	**1.00**	0.00	0.00	0.00
2	**0.95**	**1.00**	**1.00**	**0.76**	0.00	0.00	0.21	0.16	0.24	0.00	0.00	0.00	0.08
3	0.08	0.00	0.00	0.24	**1.00**	0.00	**0.97**	**1.00**	0.00	0.00	**1.00**	**1.00**	**1.00**

4.2 Results of Stochastic Multi-criteria Acceptability Analysis with Electre TRI-rC

To validate the recommendation for insulating materials against the arbitrary choice of weights conducted with the SRF procedure, we applied SMAA. For each stakeholder, we considered a sample of 10000 uniformly distributed weight vectors compatible with the ranking of criteria (s)he provided within the SRF procedure.

Table 8 Class range stochastic acceptability indices $CRSAI^1\,(a, [L, R])$ and cumulative class stochastic acceptability indices $CuCSAI^1\,(a, h)$ for all materials for DM_1

	CRSAIs						*CuCSAIs*		
a	$[C_1,C_1]$	$[C_1,C_2]$	$[C_2,C_2]$	$[C_1,C_3]$	$[C_2,C_3]$	$[C_3,C_3]$	C_1	C_2	C_3
a_1	0.000	0.000	0.825	0.000	0.000	0.175	0.000	**0.825**	0.175
a_2	0.000	0.175	0.825	0.000	0.000	0.000	0.175	**1.000**	0.000
a_3	0.000	0.000	1.000	0.000	0.000	0.000	0.000	**1.000**	0.000
a_4	0.717	0.000	0.283	0.000	0.000	0.000	**0.717**	0.283	0.000
a_5	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	**1.000**
a_6	1.000	0.000	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000
a_7	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	**1.000**
a_8	0.000	0.000	0.000	0.000	0.175	0.825	0.000	0.175	**1.000**
a_9	1.000	0.000	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000
a_{10}	1.000	0.000	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000
a_{11}	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	**1.000**
a_{12}	0.000	0.000	0.000	0.000	0.175	0.825	0.000	0.175	**1.000**
a_{13}	0.000	0.000	0.000	0.000	0.175	0.825	0.000	0.175	**1.000**

The analysis of class range stochastic acceptability indices $CRSAI^k\,(a, [L, R])$ and cumulative class stochastic acceptability indices $CuCSAI^k\,(a, h)$ indicates the potential variability of the recommendation that can be obtained for each DM for different compatible weight vectors. For illustrative purpose, in Table 8 we provide these indices for DM_1. For some materials, all compatible weight vectors confirm the same assignment. These parts of the recommendation can be deemed as robust (e.g., $CRSAI^1\,(a_3, [2, 2]) = 1$ or $CRSAI^1\,(a_9, [1, 1]) = 1$). The same conclusion can be derived from the analysis of the indices which are equal to zero, thus excluding the possibility of the respective assignment. Further, for some other materials the acceptability indices express hesitation with respect to the recommended class though often offering greater support to a particular assignment. For example, although both C_2 and C_3 are possible for a_1, the probability of the previous (C_2) is significantly greater than of the latter (C_3). Finally, the recommendation obtained for various compatible weight vectors can be different, but their intersection can be non-empty. Then, a robust recommendation is confirmed with $CuCSAI^1\,(a, h)$=1. It is the case for, e.g., a_{13} which is assigned imprecisely to $[C_2,C_3]$ or precisely to C_3, thus always confirming C_3 as the possible assignment.

When it comes to a group decision perspective, the cumulative group class stochastic acceptability indices $CuCSAI^{\partial^{\mathrm{K}}}\,(a, h)$ are presented in Table 9. Their values are very similar to the group class acceptability indices $E^{\partial}\,(a, h)$ reported in the previous section. The main differences concern a slightly increased support given to the minority class for some alternatives (see, e.g., a_1 to C_3, or a_2 to C_1, a_8, and a_{12} to C_2).

Overall, the prevailing assignments for all materials are the same as in Sect. 4.1. In this regard, let us emphasize that $CuCSAI^{\partial^{\mathrm{K}}}\,(a, h) = 1$ (see, e.g., a_{10} to C_1, a_3 to C_2, or a_5 to C_3) confirms an agreement with respect to assignment of a to C_h for all weight

Table 9 Cumulative group class stochastic acceptability indices $CuCSAI^{\partial^{\mathrm{K}}}(a, h)$ for all materials

$h\backslash a$	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}
1	0.018	0.226	0.000	0.338	0.000	**1.000**	0.000	0.000	**0.794**	**1.000**	0.000	0.000	0.000
2	**0.897**	**0.995**	**1.000**	**0.760**	0.000	0.000	0.188	0.222	0.206	0.000	0.000	0.063	0.107
3	0.131	0.000	0.000	0.213	**1.000**	0.000	**0.986**	**0.998**	0.000	0.000	**1.000**	**1.000**	**0.999**

Table 10 The numbers of stakeholders indicating a preference or indifference (in the round brackets) for all pairs of criteria

g_j	g_1	g_2	g_3	g_4	g_5	g_6
g_1	–	14 (2)	22 (8)	15 (2)	11 (2)	10 (2)
g_2	22 (2)	–	25 (2)	17 (9)	4 (16)	3 (16)
g_3	8 (8)	11 (2)	–	12 (1)	9 (2)	9 (1)
g_4	21 (2)	12 (9)	25 (1)	–	6 (10)	7 (10)
g_5	24 (2)	18 (16)	27 (2)	22 (10)	–	1 (31)
g_6	26 (2)	19 (16)	28 (1)	21 (10)	6 (31)	–

vectors compatible with preferences of all stakeholders. Thus, such a recommendation needs to be treated with certainty. Conversely, $CuCSAI^{\partial^{\mathrm{K}}}(a, h) = 0$ (e.g., a_2 to C_3, a_3 to C_1, or a_9 to C_3) indicates the no classification model of any stakeholder confirmed the respective assignment. This makes it excluded from the potential recommendation.

4.3 Results of Stochastic Multi-criteria Acceptability Analysis for a Group Compromise Ranking of Criteria

The results presented in the previous sections were derived by aggregating the outcomes obtained individually for each stakeholder. In this section, we offer another perspective on the stability of results by searching for a compromise between different stakeholders already at the stage of provided preferences. In Table 10, we report the numbers of DMs indicating preference or indifference for all pairs of criteria in the ranking they provided for the purpose of applying the SRF procedure. For example, 14 out of 38 stakeholders preferred g_1 to g_2, 22 stakeholders opted for an inverse preference, and only 2 stakeholders judged this pair indifferent. Conversely, when comparing g_5 to g_6, 31 experts opted for an indifference, and only one claimed that g_5 was more important than g_6.

The information from the DMs' individual rankings has been used as an input for the algorithm constructing a compromise utilitarian ranking of criteria, i.e., the one which is on average the closest to 38 individual rankings. In this way, the following group compromise order of criteria has been constructed:

$$g_5 \sim_\partial g_6 \succ_\partial g_2 \sim_\partial g_4 \succ_\partial g_1 \succ_\partial g_3.$$

Thus, the greatest importance has been attributed to *ecosystem quality* (g_5) and *resources* (g_6), while the least important criteria are *NPV* (g_3) and *hour of discomfort* (g_1). The relation instantiated for different pairs of criteria is consistent with the opin-

Table 11 Cumulative group compromise class stochastic acceptability indices $CuCCSAI^{\partial^{\mathrm{K}}}(a, h)$ for all materials

$h\backslash a$	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}
1	0.000	0.419	0.000	**0.533**	0.000	**1.000**	0.000	0.000	**1.000**	**1.000**	0.000	0.000	0.000
2	**0.581**	**1.000**	**1.000**	**0.467**	0.000	0.000	0.000	0.419	0.000	0.000	0.000	0.000	0.000
3	**0.419**	0.000	0.000	0.000	**1.000**	0.000	**1.000**	**1.000**	0.000	0.000	**1.000**	**1.000**	**1.000**

ion expressed by the significant number of stakeholders. For example, 24 stakeholders ranked g_5 and g_6 as the two most important criteria, while 19 of them ranked this pair tied for the first place. Furthermore, 21 stakeholders judged g_3 as the least important criterion.

Obviously, one needs to bear in mind that the compromise ranking of criteria minimizes the sum of distances between relations observed for all pairs of criteria in all individual rankings. In this perspective, it may not be considered representative by all individuals (see, e.g., DM_7, DM_9, DM_{12}, DM_{17}, DM_{19}, DM_{20}, or DM_{36}) whose preferences are represented in the compromise ranking to a marginal degree (i.e., an overall distance between their ranking and the compromise one is substantial).

Such a compromise ranking of criteria has been used to simulate DMs' joint preferences within SMAA. Consistently with the previous sections, the cutting level λ was assumed to be equal to the sum of weights of the three most significant criteria. The results of robustness analysis are materialized with the cumulative group compromise class stochastic acceptability indices $CuCCSAI^{\partial^{\mathrm{K}}}(a, h)$ (see Table 11).

For most materials, the variability of results is lesser than in case of deriving the recommendation by aggregating the individual viewpoints. Indeed, for 11 out of 13 materials there is some class which is recommended with certainty (then, $CuCCSAI^{\partial^{\mathrm{K}}}(a, h) = 1$). Also, for all materials but a_4 the class assignments with the greatest support have not changed with respect to those reported in the previous sections. The main differences concern a lesser support for the assignment of a_1, a_4 and a_9 to C_2 in favor of judging the quality of a_1 as high (C_3) and the quality of a_4 or a_9 as low (C_1). Finally, although the assignments of a_2 and a_8 to, respectively, C_2 and C_3 are robust, the acceptability for their assignment to some worse classes (C_1 and C_2, respectively) has increased to 0.419.

4.4 Summary

In view of the results derived from an application of a three-stage multiple criteria decision aiding method to our study (see Tables 7, 9, and 11), we recommended the following assignments for the insulating materials:

- *Low* (C_1): gypsum fibreboard (a_6), mineralized wood (a_9) and plywood (a_{10});
- *Low* (C_1) or medium (C_2): fibreboard hard (a_4);
- *Medium* (C_2): autoclave aerated complete (a_1), corkslab (a_2), and expanded perlite (a_3);

Table 12 Subsets of criteria on which the materials attain at least as good performances as these of the characteristic profiles b_1, b_2, and b_3 of three decision classes

Insulating material	a	b_1	b_2	b_3
Autoclave aerated	a_1	$g_1, g_2, g_3, g_4, g_5, g_6$	g_3, g_4, g_5, g_6	g_5
Corkslab	a_2	g_1, g_2, g_3, g_4, g_6	g_1, g_3, g_6	
Expanded perlite	a_3	$g_1, g_2, g_3, g_4, g_5, g_6$	g_1, g_3, g_4, g_5, g_6	
Fibreboard hard	a_4	g_1, g_2, g_3, g_4, g_5	g_1, g_2, g_3	g_1
Glass wool	a_5	$g_1, g_2, g_3, g_4, g_5, g_6$	$g_1, g_2, g_3, g_4, g_5, g_6$	g_2
Gypsum fibre board	a_6	g_1, g_3, g_6		
Hemp fibres	a_7	$g_1, g_2, g_3, g_4, g_5, g_6$	$g_1, g_2, g_3, g_4, g_5, g_6$	g_3, g_4, g_6
Kenaf fibres	a_8	$g_1, g_2, g_3, g_4, g_5, g_6$	g_3, g_4, g_6	g_3, g_4, g_6
Mineralized wood	a_9	$g_1, g_2, g_3, g_4, g_5, g_6$	g_1, g_3	
Plywood	a_{10}			
Polystyrene foam	a_{11}	$g_1, g_2, g_3, g_4, g_5, g_6$	$g_1, g_2, g_3, g_4, g_5, g_6$	g_4, g_5
Polyurethane	a_{12}	$g_1, g_2, g_3, g_4, g_5, g_6$	$g_1, g_2, g_3, g_4, g_5, g_6$	g_1, g_2, g_3, g_5
Rock wool	a_{13}	$g_1, g_2, g_3, g_4, g_5, g_6$	$g_1, g_2, g_3, g_4, g_5, g_6$	g_1, g_2, g_3, g_6

- *High* (C_3): glass wool (a_5), hemp fibres (a_7), kenaf fibres (a_8), polystyrene foam (a_{11}), polyurethane (a_{12}), and rock wool (a_{13}).

The probability of other assignments was often non-negligible though significantly lower than for the above indicated classes. Nevertheless, the results obtained from the stochastic analysis allowed to nullify the risk of a false declaration that some material was assigned to a class which was not confirmed by any compatible set of weights for any expert.

For each insulating material, the recommended decision can be justified by comparing its performances on different criteria with those of the characteristic class profiles. In Table 12, we indicate the subsets of criteria on which the materials outrank (i.e., are at least as good as) the characteristic profiles. In this regard, let us explicitly explain the most likely assignments suggested for some materials:

- a_{10} is worse than b_1 on all criteria, thus being assigned to the worst class C_1; in the same spirit, a_6 is worse than b_1 on g_2, g_4, and g_5 (thus, on 3 out of 4 considered environmental criteria), and not better than b_2 on any criterion, which makes C_1 its most desired class;
- a_3 is better than b_1 and worse than b_3 on all criteria, which makes its performance vector typical for C_2;
- a_{12} and a_{13} are at least as good as b_2 on all criteria and better than b_3 on four criteria (g_1, g_2, g_3, g_5 or g_1, g_2, g_3, g_6, respectively (note that both scenarios include two accounted socio-economic criteria, g_1 and g_3)), which makes their assignment to C_3 the most justified.

5 Conclusions

We considered a multiple criteria problem of sustainability assessment of insulating materials. We combined Life Cycle Costing, Life Cycle Assessment, and adaptive comfort evaluation to derive performances of these materials on six socio-economic and environmental criteria. The comprehensive assessment of the materials involved their assignment to three preference-ordered sustainability classes. The classification was performed with a group decision counterpart of the Electre TRI-rC method that compares alternatives with the characteristic class profiles defined by the experts.

To derive a recommendation that would reflect viewpoints of a wide spectrum of potential customers, we accounted for the preference information of a few tens of rural buildings' owners being interested in the roof's insulation. The initial recommendation was derived by computing the proportion of stakeholders who accepted an assignment of a particular material to a given class. These results were subsequently validated against the outcomes of a two-fold robustness analysis realized with the Monte Carlo simulation. The latter exploited the space of all criteria weights compatible with either each stakeholder's preference information provided in the SRF procedure or collective ranking of criteria that was derived with an original algorithm proposed in this paper.

The three-stage analysis revealed that the most sustainable materials were glass wool, hemp fibres, kenaf fibres, polystyrene foam, polyurethane, and rock wool. This was mainly due to their favorable performances quantified with the Net Present Value and Eco-indicators. On the contrary, gypsum fibreboard, mineralized wood and plywood were assessed as the least sustainable materials. This can be justified in terms of their poor performances on thermal comfort, human health, and ecosystem quality. Overall, the proposed method provided greater clarity for decision making and guaranteed credibility in the eyes of the traditional rural houses' owners. Moreover, all research results—concerning both materials' performances on the individual criteria and comprehensive sorting recommendation—were well perceived by the experts on insulating materials in Italy.

The proposed framework can be applied to other decision contexts than that of a typical farmhouse in central Italy. This would require, however, accounting for a comfort model as well as warm and cold periods suitable to a particular geographical context, specification of a relevant lifespan for the investment, and adapting life cycle assessment to the reality of a particular study.

From the methodological viewpoint, we envisage the following future developments. Firstly, we plan to extend the SRF procedure to a group decision context so that it tolerates intensities of preference for different pairs of criteria and accepts information on different roles (weights) of the decision makers. Secondly, we aim at extending the proposed group decision framework to methods dealing with choice and ranking problems. This would require elaboration of the algorithms for deriving a compromise recommendation that would appropriately combine results of robustness analysis computed individually for each stakeholder.

Acknowledgements The work of Miłosz Kadziński and Grzegorz Miebs was supported by the Polish Ministry of Science and Higher Education under the *Iuventus Plus* program in 2016–2019 Grant Number IP2015 029674 - 0296/IP2/2016/74.

References

Almeida Dias J, Figueira JR, Roy B (2010) Electre Tri-C: a multiple criteria sorting method based on characteristic reference actions. Eur J Oper Res 204(3):565–580

Cai F-L, Liao X, Wang K-L (2012) An interactive sorting approach based on the assignment examples of multiple decision makers with different priorities. Ann Oper Res 197(1):87–108

Cailloux O, Meyer P, Mousseau V (2012) Eliciting ELECTRE TRI category limits for a group of decision makers. Eur J Oper Res 223(1):133–140

Carlucci S, Pagliano L (2012) A review of indices for the long-term evaluation of the general thermal comfort conditions in buildings. Energy Build 53:194–205

CEN (2007) Indoor environmental input parameters for design and assessment of energy performance of buildings: addressing indoor air quality, thermal environment lighting, and acoustics. EN15251. Comite Europeen de Normalisation, Brussels

Chakhar S, Saad I (2012) Dominance-based rough set approach for groups in multicriteria classification. Decis Support Syst 54(1):372–380

Chakhar S, Ishizaka A, Labib A, Saad I (2016) Dominance-based rough set approach for group decisions. Eur J Oper Res 251(1):206–224

Chen Y, Kilgour D, Hipel K (2012) A decision rule aggregation approach to multiple criteria-multiple participant sorting. Group Decis Negot 21(5):727–745

Copiello S (2017) Building energy efficiency: a research branch made of paradoxes. Renew Sustain Energy Rev 69:1064–1076

Corrente S, Figueira JR, Greco S, Roy B, Słowiński R (2017) A robust ranking method extending ELECTRE III to hierarchy of interacting criteria, imprecise weights and stochastic analysis. Omega 73:1–17

Corrente S, Greco S, Słowiński R (2016) Multiple criteria hierarchy process for ELECTRE TRI. Eur J Oper Res 252(1):191–203

Čuláková M, Vilčeková S, Katunská J, Burdová EK (2013) Multi-criteria analysis of material selection in order to reduce environmental impact. Chem Eng Trans 35:379–384

Damart S, Dias LC, Mousseau V (2007) Supporting groups in sorting decisions: methodology and use of a multi-criteria aggregation/disaggregation DSS. Decis Support Syst 43(4):1464–1475

Daher SFD, Almeida AT (2010) Recent patents using group decision support systems: a short review. Recent Pat Comput Sci 3(2):81–90

de Morais Bezerra F, Melo P, Costa JP (2017) Reaching consensus with VICA-ELECTRE TRI: a case study. Group Decis Negot 26(6):1145–1171

Dias LC, Climaco JN (2000) Electre Tri for groups with imprecise information on parameter values. Group Decis Negot 9(5):355–377

Ecoinvent. (2010). Ecoinvent Data V2.2

Fabbri K, Zuppiroli M, Ambrogio K (2012) Heritage buildings and energy performance: mapping with GIS tools. Energy Build 48:137–145

Figueira JR, Greco S, Roy B, Słowiński R (2013) An overview of ELECTRE methods and their recent extensions. J Multi-Criteria Decis Anal 20:61–85

Figueira JR, Roy B (2002) Determining the weights of criteria in the ELECTRE type methods with a revised Simos' procedure. Eur J Oper Res 139(2):317–326

Ginevicius R, Podvezko V, Raslanas S (2008) Evaluating the alternative solutions of wall insulation by multicriteria methods. J Civ Eng Manag 14(4):217–226

Goedkoop, M, Spriensma, R (2001) The Eco-indicator 99–a damage oriented method for life cycle impact assessment. Methodology report third edition. (Product Ecology Consultants)

Govindan K, Kadziński M, Sivakumar R (2017) Application of a novel PROMETHEE-based method for construction of a group compromise ranking to prioritization of green suppliers in food supply chain. Omega 71:129–145

Greco S, Kadziński M, Mousseau V, Słowiński R (2012) Robust ordinal regression for multiple criteria group decision: UTA^{GMS}-GROUP and $UTADIS^{GMS}$-GROUP. Decis Support Syst 52(3):549–561

Greco S, Matarazzo B, Słowiński R (2001) Rough sets theory for multi criteria decision analysis. Eur J Oper Res 129(1):1–47

Greco S, Matarazzo B, Słowiński R (2006) Dominance-based rough set approach to decision involving multiple decision makers. In: Greco S, Hata Y, Hirano S, Inuiguchi M, Miyamoto S, Nguyen H, Słowiński R (eds) Proceedings of the 5th international conference rough sets and current trends in computing (RSCTC2006), Kobe, Japan, November 6–8, vol 4259. Lecture notes in artificial intelligence. Springer-Verlag, Berlin, pp 306–317

Greco S, Mousseau V, Słowiński R (2010) Multiple criteria sorting with a set of additive value functions. Eur J Oper Res 207(3):1455–1470

Jabeur K, Martel JM (2007) An ordinal sorting method for group decision-making. Eur J Oper Res 180(3):1272–1289

Kadziński M, Cinelli M, Ciomek K, Coles SR, Nadagouda MN, Varma RS, Kirwan K (2018) Stochastic ordinal regression for multiple criteria sorting problems. Eur J Oper Res 264(2):472–490

Kadziński M, Ciomek K, Słowiński R (2015a) Modeling assignment-based pairwise comparisons within integrated framework for value-driven multiple criteria sorting. Eur J Oper Res 241(3):830–841

Kadziński M, Greco S, Słowiński R (2013) Selection of a representative value function for robust ordinal regression in group decision making. Group Decis Negot 22(3):429–462

Kadziński M, Greco S, Słowiński R (2014) Robust ordinal regression for dominance-based rough set approach to multiple criteria sorting. Inf Sci 283:211–228

Kadziński M, Słowiński R, Greco S (2016) Robustness analysis for decision under uncertainty with rule-based preference model. Inf Sci 328:321–339

Kadziński M, Tervonen T (2013) Stochastic ordinal regression for multiple criteria sorting problems. Decis Support Syst 55(1):55–66

Kadziński M, Tervonen T, Figueira JF (2015b) Robust multi-criteria sorting with the outranking preference model and characteristic profiles. Omega 55:126–140

Khoshnava SM, Rostami R, Valipour A, Ismail M, Rahmat AR (2016) Rank of the green building material criteria based on the three pillars of sustainability using the hybrid multi criteria decision making method. J Cleaner Prod. https://doi.org/10.1016/j.jclepro.2016.10.066

Krarti M (2015) Evaluation of large scale building energy efficiency retrofit program in Kuwait. Renew Sustain Energy Rev 50:1069–1080

Kumar A, Suman BM (2013) Experimental evaluation of insulation materials for walls and roofs and their impact on indoor thermal comfort under composite climate. Build Environ 59:635–643

Kumar V, Hewage K, Haider H, Sadiq R (2017) Sustainability evaluation framework for building cooling systems: a comparative study of snow storage and conventional chiller systems. Clean Technol Environ Policy 19(1):137–155

Lahdelma R, Salminen P (2001) SMAA-2: stochastic multicriteria acceptability analysis for group decision making. Oper Res 49(3):444–454

Liu J, Liao X, Yang J-B (2015) A group decision-making approach based on evidential reasoning for multiple criteria sorting problem with uncertainty. Eur J Oper Res 246(3):858–873

Lolli F, Ishizaka A, Gamberini R, Rimini B, Messori M (2015) FlowSort-GDSS—A novel group multi-criteria decision support system for sorting problems with application to FMEA. Expert Syst Appl 42(17–18):6342–6349

Ma Z, Cooper P, Daly D, Ledo L (2012) Existing building retrofits: methodology and state-of-the-art. Energy Build 55:889–902

Martínez-Molina A, Tort-Ausina I, Cho S, Vivancos JL (2016) Energy efficiency and thermal comfort in historic buildings: a review. Renew Sustain Energy Rev 61:70–85

Mazzarella L (2015) Energy retrofit of historic and existing buildings. The legislative and regulatory point of view. Energy Build 95:23–31

Menconi ME, Grohmann D (2014) Model integrated of life-cycle costing and dynamic thermal simulation (MILD) to evaluate roof insulation materials for existing livestock buildings. Energy Build 81:48–58

Morais DC, de Almeida AT, Figueira JR (2014) A sorting model for group decision making: a case study of water losses in Brazil. Group Decis Negot 23(5):937–960

Motuziene V, Rogoza A, Lapinskiene V, Viletiene T (2016) Construction solutions for energy efficient single-family house based on its life cycle multi-criteria analisys: a case study. J Clean Prod 112(1):532–541

Nemery, P (2008) Extensions of the FlowSort sorting method for group decision-making. In: 19th International conference on multiple criteria decision making (MCDM 2008), Auckland

Paolotti L, Boggia A, Castellini C, Rocchi L, Rosati A (2016) Combining livestock and tree crops to improve sustainability in agriculture: a case study using the life cycle assessment (LCA) approach. J Clean Prod 131:351–363

Paolotti L, Martino G, Marchini A, Boggia A (2017) Economic and environmental assessment of agro-energy wood biomass supply chains. Biomass Bioenerg 97:172–185

Papadopoulos AM, Giama E (2007) Environmental performance evaluation of thermal insulation materials and its impact on the building. Build Environ 42(5):2178–2187

Piombo O, Allacker K, Rivela B, Neila J (2016) Sustainability assessment of energy saving measures: a multi-criteria approach for residential buildings retrofitting—a case study of the Spanish housing stock. Energy Build 116:384–394

Product Ecology Consultants. (1990) SimaPro LCA Software. Plotterweg

Rezaei J, Kadziński M, Vana C, Tavasszy L (2017) Embedding carbon impact assessment in multi-criteria supplier segmentation using ELECTRE TRI-rC. Ann Oper Res. https://doi.org/10.1007/s10479-017-2454-y

Roy B (1996) Multicriteria methodology for decision analysis. Kluwer, Dordrecht

Roy B (2010) Robustness in operation research and decision aiding: a multi-faceted issue. Eur J Oper Res 200(3):629–639

Roy B, Słowiński R (1993) Criterion of distance between technical programming and socio-economic priority. RAIRO Oper Res Recherche Operationnelle 27(1):45–60

Santos J, Flintsch G, Ferreira A (2017) Environmental and economic assessment of pavement construction and management practices for enhancing pavement sustainability. Resour Conserv Recycl 116:15–31

Shen F, Xu J, Xu Z (2016) An outranking sorting method for multi-criteria group decision making using intuitionistic fuzzy sets. Inf Sci 334–335:338–353

Siskos E, Tsotsolas N (2015) Elicitation of criteria importance weights through the Simos method: a robustness concern. Eur J Oper Res 246(2):543–553

Sun B, Ma W (2015) Rough approximation of a preference relation by multi-decision dominance for a multi-agent conflict analysis problem. Inf Sci 315:39–53

Taylor PB, Mathews EH, Kleingeld M, Taljaard GW (2000) The effect of ceiling insulation on indoor comfort. Build Environ 35(4):339–346

Tervonen T, Figueira J (2008) A survey on stochastic multicriteria acceptability analysis methods. J Multi-Criteria Decis Anal 15(1–2):1–14

Tervonen T, Lahdelma R, Almeida Dias J, Figueira J, Salminen P (2007) SMAA-TRI. In: Linkov I, Kiker GA, Wenning RJ (eds) Environmental security in harbors and coastal areas. NATO security through science series (series C: environmental security). Springer, Dordrecht

Tervonen T, van Valkenhoef G, Basturk N, Postmus D (2013) Hit-and-Run enables efficient weight generation for simulation-based multiple criteria decision analysis. Eur J Oper Res 224(3):552–559

Verbeeck G, Hens H (2005) Energy savings in retrofitted dwellings: economically viable. Energy Build 37(7):747–754

Yu W (1992) ELECTRE TRI. Aspects méthodologiques et guide d'utilisation. Document du LAMSADE, vol 74. Université Paris-Dauphine, Paris

Yung EHK, Chan EHW (2012) Implementation challenges to the adaptive reuse of heritage buildings: towards the goals of sustainable, low carbon cities. Habitat Int 36(3):352–361

Zampori L, Dotelli G, Vernelli V (2013) Life cycle assessment of hemp cultivation and use of hemp-based thermal insulator materials in buildings. Environ Sci Technol 47(13):7413–7420

Publication [P2]

Rocchi, L., Kadziński, M., Menconi, M. E., Grohmann, D., Miebs, G., Paolotti, L., and Boggia, A. (2018). Sustainability evaluation of retrofitting solutions for rural buildings through life cycle approach and multi-criteria analysis. In Energy and Buildings (Vol. 173, pp. 281–290)., DOI: 10.1016/j.enbuild.2018.05.032.

Number of citations[2]:

- according to Web of Science: 40
- according to Google Scholar: 56

Contribution of co-authors (excluding the author and the supervisor of this dissertation):

- Lucia Rocchi:
  - Co-authorship of the concept of to the considered case study
  - Data collection and analysis of results
  - Co-authorship of the text of the publication
- Maria Elena Menconi
  - Consultation of the context of case study and interpretation of its results
  - Editing of the manuscript
- David Grohmann:
  - Consultation of the context of case study and interpretation of its results
  - Editing of the manuscript
- Luisa Paolotti
  - Consultation of the context of case study and interpretation of its results
  - Editing of the manuscript
- Antonio Boggia
  - Consultation of the context of case study and interpretation of its results
  - Editing of the manuscript

[2] as on September 25, 2023

Energy & Buildings 173 (2018) 281–290

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild

Sustainability evaluation of retrofitting solutions for rural buildings through life cycle approach and multi-criteria analysis

L. Rocchi[a,*], M. Kadziński[b], M.E. Menconi[a], D. Grohmann[a], G. Miebs[b], L. Paolotti[a], A. Boggia[a]

[a] Department of Agricultural, Food and Environmental Sciences, University of Perugia, Borgo XX Giugno, 74, Perugia 06121, Italy
[b] Institute of Computing Science, Poznan University of Technology, Piotrowo 2, Poznan 60-965, Poland

ARTICLE INFO

Article history:
Received 16 November 2017
Revised 5 May 2018
Accepted 19 May 2018
Available online 31 May 2018

Keywords:
Multi-criteria decision making
Decision analysis
Sustainability analysis
Energy efficiency
Life cycle approach
Multi-criteria sorting

ABSTRACT

In an international context that is faced with several environmental issues, especially in relation to energy efficiency, the finding of sustainable solutions for retrofitting existing buildings is a challenging issue. Comprehensive evaluation of the applied materials is required to understand their true impact on the environment as well as their economic suitability. Therefore, a life cycle-based approach is required. This work presents a hybrid method, including energy and comfort optimization, Life Cycle Assessment (LCA) and Life Cycle Costing (LCC) analysis, applied for retrofitting a traditional rural building and then combined together in a multi-criteria approach. The case study considers the assessment of the sustainability of several solutions for roof insulation according to seven criteria derived from the hybrid method developed: energy saving, non-renewable energy, comfort performance, global warming, ozone layer depletion, respiratory inorganics and Net Present Value. The multi-criteria approach used to combine these criteria is the Electre TRI-rC method. The results obtained with this method show the division of the materials into three categories with overall bad, medium or good performance. The stability of results is verified against a two-fold sensitivity analysis.

1. Introduction

Modern society's concern about energy and environmental issues has grown in recent decades. Global warming, ozone layer depletion and the uncontrolled accumulation of waste are international trending topics [1]. In this scenario, the building sector plays a crucial role in both energy and natural resource consumption, as well as in the release of emissions [1–3]. The massive adoption of energy saving measures in the building sector could contribute significantly to the reduction of greenhouse gases (GHG) [2,4]. In the last ten years, the European Union has adopted various directives[1] for obligating Member States to move towards the energy efficient, eco-designed and environmentally sustainable buildings, considering both new and existing structures.

Within this framework, one of the most challenging issues is the improvement of the energy and environmental performance of historic and traditional buildings, which are usually characterized by bad performance [5]. At the same time, these buildings, including rural structures, are an integral part of the European cultural heritage, and thus they must be preserved [6]. Traditional farmhouses are representative of the identity of a community, and many rural landscapes have high conservation values thanks to their vernacular buildings [7].

Although the potentiality of improving energy efficiency in traditional buildings is enormous, their high landscape values, protected by law, hamper external alterations [5]. Therefore, the most viable solutions for harmonizing the needs of proper preservation with those of energy saving consist in acting on building enclosures, increasing their thermal inertia by means of insulating materials added internally, and optimizing the operation of mechanical installations such as heat recovery ventilation [8,9]. When the goal of the retrofitting solution is to minimize energy consumption, it is important to insulate the building envelope, and the most influential component for comfort is roof insulation [8,10–12].

The aim of this paper is to evaluate different envelope solutions for insulating traditional rural buildings, using a hybrid multi-criteria method, which combines energy and thermal comfort optimization with the environmental and economic life cycle analysis. The case study concerns a farmhouse in central Italy. The analysis considered the impact of materials production and their application to a typical rural building, as well as the overall sustainability of different options. According to Cinelli et al. [13,14],

* Corresponding author.
E-mail address: lucia.rocchi@unipg.it (L. Rocchi).

[1] Directives 2009/125/EC; 2010/31/EU; 2012/27/EU.

https://doi.org/10.1016/j.enbuild.2018.05.032
0378-7788/

the assessment of criteria is a key issue in sustainability analysis. Therefore, the approach presented here accounts for different viewpoints, which are important throughout the entire life cycle of the building and materials [4]. As was proved by Agarski et al. [15], Life Cycle Approach and Multi-Criteria Decision Analysis (MCDA) are full of analogies and complement each other. Despite this, combined applications of LCA and MCDA are still rare [see, e.g., 16–19], and a fully integrated approach, which also incorporates economic evaluation and energy optimization, has been applied only in Motiuziene et al. [20].

The paper is structured as follows: Section 2 presents a hybrid approach used in the sustainability analysis; Section 3 deals with the case study concerning evaluation of the retrofitting materials used in a typical Italian farmhouse; Section 4 discusses the results of the study; Section 5 presents some conclusions.

2. Methods

This work is based on a multi-methodological approach, described by the following three main steps:

- Stage 1: development of a dynamic energy simulation for an existing building and assessment of the performance of different insulation materials added to the building model (Section 2.1);
- Stage 2: implementation of a Life Cycle Costing analysis and a Life Cycle Assessment, for evaluating the economic and environmental impacts of the different insulating materials, within the context of traditional buildings in central Italy (Section 2.2);
- Stage 3: implementation of a decision-making model, for a comprehensive evaluation of insulation solutions based on the results of previous stages, as well as of the overall sustainability of the materials (Section 2.3).

2.1. Dynamic thermal simulation and comfort models in mixed-mode buildings

The first step toward making a building sustainable is to minimize its energy consumption. A dynamic thermal simulation is thus required to analyze its behavior, using an integrated approach. We used *EnergyPlus* 8.5 as the simulation engine for building performance. The Model Building has one thermal zone per floor and the simulation period is one year, during which the building performance is evaluated every 10 min (52,560 times). The indoor environment is determined by different energy sources that evolve in different ways and speeds. Three main sources must be considered: the outdoor climate, the occupants causing an unpredictable energy intake and using electrical equipment, and the auxiliary systems that provide heating, cooling or ventilation of the indoor environment [21,12].

In order to design the building model, information is needed on the characteristics of the envelope (geometry, shape, materials), while the elements necessary for determining the indoor environment are a “typical year” weather file for the location, the schedule of the presence and the activities of the building’s occupants and the technical specifications of the equipment. The model developed is a mixed-mode building: the model uses heating, ventilation and air conditioning (HVAC) only during the worst periods, namely those in which the temperature conditions are farthest from the optimal comfort.

Varying the thermal inertia of construction elements (roof, wall, etc.) is a method for improving the performance of a building [21,22]. Thermal inertia is a measure of the responsiveness of a material to variations in temperature, and it represents the material’s ability to accumulate heat and then dispose of it. Thermal inertia (I) depends on the thermal conductivity (k), density (ρ), and specific heat (c) of the material, as given by (1):

$$I = (k \cdot \rho \cdot c)^{1/2} \tag{1}$$

In the mixed-mode building, people’s actual thermal sensations and the acceptance of thermal conditions vary when the building shifts from an air conditioning mode to a natural ventilation mode, or vice versa [23]. Since some authors recommended that the occupants’ thermal responses to mixed-mode buildings should be evaluated separately [23,24], the model developed uses the adaptive comfort approach only during the natural ventilation period [25–28] and Fanger’s Predicted Mean Vote/Predicted Percentage Dissatisfied model during the HVAC period [29]. For a more detailed description of these comfort models, see Orosa and Oliviera [22], Peter et al. [30], and Carlucci and Pagliano [31].

The scientific literature and some well-accepted standards offer different long-term thermal discomfort indices [32,33]. In this work, we adopted the Degree Hours (DH) criterion [33] to evaluate comfort on an annual basis according to the EN 15251 standard [32]. DH accounts for the total time during which the temperature falls outside a specific comfort category, which is then weighed by how much the limit has been exceeded. The category used in this study is equivalent to 3 °C below and above the comfort temperature in the adaptive model, and corresponds to the temperature related to a predicted mean vote of +/- 0.5 for the Fanger model. DH was calculated according to (2):

$$\mathrm{DH} = \sum_{i=1}^{8760} 10/60 \cdot \left| \mathrm{CC}_2 - \mathrm{OT}\grave{\imath} \right|, \tag{2}$$

where CC_2 is the lower or upper limit of the second comfort category and OT is the operative temperature at hour “ì”. Eq. (2) differs from the model proposed in EN1525 (which was also reported by Carlucci and Pagliano [31]) by introducing the multiplier 10/60. This multiplier was incorporated because the applied simulation engine processes data with a time step of 10 min.

2.2. Life cycle costing and life cycle assessment

Although their names are very similar, Life Cycle Costing (LCC) and Life Cycle Assessment (LCA) provide answers to very different questions [34]. LCC is an economic method for assessing projects, which takes into consideration all costs that occur during their lifetime [35]. LCA is a scientific methodology useful for identifying the environmental implications of a product (or process, or service), considering all the stages in its life cycle, through the quantification of resources (energy, raw materials, water) and of emissions into the environment (emissions into the air, water and soil, waste and co-products) associated with the system being assessed [36]. The product is therefore analyzed along its entire lifespan, from the extraction of the raw material to its final disposal [37]. As the two methods focus on the entire life cycle and are complementary, combined applications of LCA and LCC have increased in the last 20 years [35,38–40].

The functional unit used for both LCC and LCA in the study was 1 m^3 of the insulating material or, for some of the materials analyzed, the corresponding amount in kg, calculated according to their density. The quantity of material needed for insulating the roof was then calculated within the case study.

For the economic analysis, we applied the Net Present Value (NPV), which is one of the LCC methods most widely applied in both the scientific literature and everyday business analysis [35]. NPV corresponds to the present monetary value of a long-term project, discounting back future net cash flows minus the initial investment. The most challenging aspect regarding the proper use of this method is the definition of both the time horizon and the discount rate to be considered [34]. Most applications of the NPV

to buildings use time spans ranging from 20 to 70 years, although it is not rare to find longer periods [35, 41]. Following [42], we considered a time horizon of 25 years, since a longer period would make no significant difference in the final assessment [21]. As regards the discount rate, it usually ranges from 2% to 8%, sometimes also including the inflation rate [41]. In this paper, we applied a discount rate of 3%, following the guidelines of the Italian Federation for the Rational Use of Energy.[2] Moreover, we used a real cost approach, thus not accounting for the inflation rate, as suggested by Islam et al. [41] and Gluch and Baumann [43]. Calculation of the NPV in this work is based on [21] according to the following formula:

$$NPV = -PEI - \mathrm{I} + \sum_{(t=0)}^{T} \frac{ES_t}{(1+i)^t} - \frac{EL_T}{(1+i)^T} \quad (3)$$

where *PEI* is the Primary Energy Input cost, *I* is the installation cost at time *t*, ES_t is the cash flow at time *t*, and EL_T is the dismissing cost. The lifespan of the investment is *T*, while *i* is the discount rate.

As concerns the environmental analysis, we performed the LCA according to ISO 14040 [44] and ISO 14044 [45], which define the guidelines to be followed, setting the following four basic phases: goals and scope definition; inventory analysis; impact assessment; interpretation. Several authors have applied LCA in the specific context of building materials, in order to understand their environmental performance [1,46] or to compare traditional materials with ecological materials [47].

Goal and scope definition – The goal of the LCA study was to evaluate the environmental impacts of different materials usable as envelope solutions for insulating traditional rural buildings. The materials analyzed were those subsequently described in the case study (12 materials in total, see Table 3), and had both synthetic and natural origins, with widespread utilization in the study area. The functional unit used within the study was, as mentioned, $1 m^3$ of the insulating material or, for some of the materials analyzed, the corresponding amount in kg, calculated according to their density. The choice of such a functional unit was due to the major simplicity in calculating it, and the higher comparability among the different materials, and with available literature.

In relation to the system boundaries, we considered all the main phases of the life cycle, starting from the production of a raw material to the obtaining of the complete insulating material. To focus the analysis on the production phase, an LCA "from cradle to gate" can be performed [48], omitting the use and disposal phases. For what concerns the topic of multifunctional processes, no allocation processes were needed.

Life Cycle Inventory – For the majority of the materials, foreground data related to the main production and assembly processes, as well as background data for transport, electricity and fuel consumption were taken directly from the Ecoinvent Database [49]. Data about processes of production, for the materials chosen within the Database, presented standard European characteristics, considering, for these types of envelope materials, the most recent available technologies. The choice of using these items for the LCA study was due to the higher comparability with available literature, and also to the fact that in some cases local data about the assembly processes were not complete or precise. For the hemp processes instead, which were no present within the Database, the data were derived from Zampori et al. [50]. In this case, the main inputs considered were the quantities of water and energy needed to assembly the panel, starting from the hemp woody core. Moreover, all the operations needed to obtain the hemp woody core were also considered (i.e. all the necessary agricultural practices like sowing, soil tillage, fertilizing, baling, exc.). In addition, the transport phase was taken into account for each material, estimating a distance between 50 and 200 km within the study area, from the production area of each material to the building.

Table 1
Impact categories used for LCA (method: IMPACT 2002 +).

IMPACT CATEGORY	UNIT
Global Warming	kg CO_2 eq
Ozone Layer Depletion	kg CFC-11 eq
Respiratory Inorganics	kg PM2.5 eq
Non-renewable energy	MJ primary

Impact Assessment – For the assessment of environmental impacts, the IMPACT 2002 + method [51] was applied. The method proposes a feasible implementation of a combined midpoint/damage approach, linking all types of life cycle inventory results (elementary flows and other interventions) via several midpoint categories to several damage categories. The impact categories used in this work are given in Table 1. For each category, higher values indicate worse environmental performance [52]. Simapro software[3] (version 8.2.0) was used to perform the calculations.

Interpretation – The results of the environmental evaluation were briefly reported in the subsequent sections (described in paragraph 4.1 and Table 4). As the LCA study was only a part of the wider model implemented, and LCA results were included into the multi-criteria analysis together with other types of results, discussion and interpretation are directly embraced into the final discussion of the multi-criteria analysis results.

2.3. Multi-criteria analysis

A comprehensive evaluation of the different options for roof insulation needs to account for the energy and thermal optimization as well as the life cycle costing and assessment. To perform such an evaluation, we employed a multi-criteria ordinal classification (sorting) approach, called Electre TRI-rC [53]. The method consists of two main phases. In the first phase, it constructs an outranking relation by comparing the alternatives with the characteristic class profiles provided by the experts. In the second stage, this relation is exploited to identify for each alternative its desired assignment to some pre-defined decision classes.

Within the study, we consider a set $A = \{a_1, a_2, \ldots, a_{12}\}$ of twelve alternatives representing different insulating materials. Their quality is assessed from different viewpoints formalized as a set of seven criteria, $G = \{g_1, g_2, \ldots, g_7\}$. The performance (evaluation) of material a on criterion g_j is denoted by $g_j(a)$. The study aims at assigning the materials to three preference-ordered classes C_1, C_2, C_3, such that C_3 is considered better than C_2, which, in turn, is preferred over C_1.

2.3.1. Preference information

In the variant of Electre TRI-rC that was used, the experts involved in the study were asked to provide two types of preference information. On one hand, for each class C_h, $h = 1, 2, 3$, they had to specify a characteristic (central) profile b_h. This profile consists of the performances judged to be the most typical for a particular category.

On the other hand, they supplied information on the relative importance of criteria, which was required by the SRF (Simon-Roy-Figueira) procedure [54]. In line with the assumptions of this procedure, the experts were given a set of cards with the criteria

[2] Available at https://www.fire-italia.org/lcca/.

[3] Product Ecology Consultants. 1990. SimaPro LCA Software. Plotterweg.

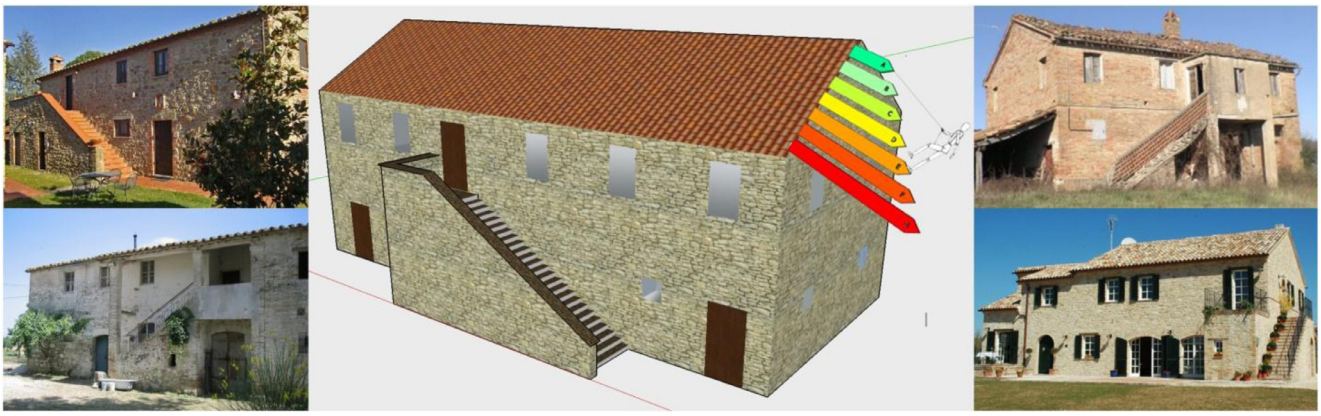


Fig. 1. Typical farmhouses of central Italy.

names as well as some blank (white) cards. First, they were expected to order the cards corresponding to various criteria from the least to the most important. In this way, each criterion g_j was assigned rank $l(j)$, such that the greater the rank, the more significant g_j is. The blank cards were used to differentiate the intensity of preference between groups of criteria judged indifferently (i.e. assigned the same rank). A greater number e_s of blank cards among successive groups of criteria implied a more significant difference between their weights. Lastly, ratio Z between the weights of the most and the least important criteria needs to be defined. These inputs, i.e. the ranks $l(j)$, empty cards e_s, and ratio Z, were used to derive normalized weights w_j, $j = 1, \ldots, 7$, according to Corrente et al. [55]

2.3.2. Construction of an outranking relation

In the first stage of Electre TRI-rC, the alternatives are compared against the characteristic class profiles to construct an outranking relation S. This process is conducted in three steps. First, for each criterion g_j, the marginal concordance index $c_j(a, b_h)$ quantifies an outcome of a comparison between alternative a and profile b_h. Specifically, if a is at least as good as b_h on g_j, then $c_j(a, b_h) = 1$, and, otherwise, i.e. if a is worse than b_h on g_j, then $c_j(a, b_h) = 0$.

Second, the marginal concordance indices $c_j(a, b_h)$ are aggregated into a comprehensive concordance index using the normalized weights w_j, $j = 1, \ldots, 7$, derived from the SRF procedure. The aggregation is performed as follows: $\sigma(a, b_h) = \sum_{j=1}^{7} w_j \cdot c_j(a, b_h)$. Thus defined, $\sigma(a, b_h)$ represents a valued outranking relation, quantifying a strength of the subset of criteria which confirm that a outranks b_h.

Third, a crisp outranking relation S is constructed by comparing $\sigma(a, b_h)$ with a pre-defined concordance threshold λ, which indicates a minimal value of $\sigma(a, b_h)$ that would validate relation aSb_h. Thus, $\sigma(a, b_h) \geq \lambda \Rightarrow aSb_h$. The truth of b_hSa is verified analogously.

Knowing whether aSb_h and b_hSa, we can determine whether a and b_h are indifferent ($aSb_h \wedge b_hSa \Rightarrow a \sim b_h$), or incomparable ($not(aSb_h) \wedge not(b_hSa) \Rightarrow aRb_h$), or if a is preferred to b_h ($aSb_h \wedge not(b_hSa) \Rightarrow a \succ b_h$), or b_h is preferred to a ($not(aSb_h) \wedge b_hSa \Rightarrow b_h \succ a$).

2.3.3. Assignment procedures

In the other stage of Electre TRI-rC, the outranking relation and comprehensive concordance indices are exploited to derive for each alternative a its desired class assignment interval $C(a) = [C_L(a), C_R(a)]$ [53]. On one hand, the worst class $C_L(a)$ for alternative a is set to C_h in case b_{h-1} is the best characteristic profile, such that a is preferred to b_{h-1} and the evidence in favor of a being at least as good as b_h is greater than the support given to the hypothesis about b_{h-1} outranking a:

$$a \succ b_{h-1} \wedge \sigma(a, b_h) > \sigma(b_{h-1}, a). \quad (4)$$

On the other hand, the best class $C_R(a)$ for alternative a is set to C_h in case b_{h+1} is the worst characteristic profile, such that b_{h+1} is preferred to a and the support given to the assertion b_hSa is greater than the evidence in favor of aSb_{h+1}:

$$b_{h+1} \succ a \wedge \sigma(b_h, a) > \sigma(a, b_{h+1}). \quad (5)$$

The recommended assignment $C(a)$ is precise if $C_L(a) = C_R(a)$, and otherwise, if $C_L(a) < C_R(a)$, it is imprecise [56].

2.3.4. Justification for applying Electre TRI-rC and the SRF procedure

The use of Electre TRI-rC and the SRF procedure in a case study was motivated by the following four characteristics. First, Electre TRI-rC is suitable for dealing with ordinal classification, which agrees with the formulation of the problem considered. Indeed, the study aimed at classifying the insulating materials into a set of pre-defined and ordered classes, which prevents the use of multi-criteria ranking methods such as PROMETHEE, TOPSIS, or AHP.

Second, Electre Tri-rC derives the assignments from the comparison of alternatives with some characteristic class profiles. In this way, the results depend on some pre-defined norms rather than on the relative comparison between the alternatives. Furthermore, the experts involved in the study found it intuitive to provide the characteristic profiles.

Third, the incompleteness of the preference information on the relative importance of criteria by the SRF procedure was also found to be appealing. Moreover, the interpretation of weights derived from this procedure is consistent with the one adopted in the Electre methods. Indeed, both approaches assume that the weights are interpreted as voting powers assigned to different criteria. Finally, Electre TRI-rC does not require any prior transformation or normalization of the performances of the materials, which makes it applicable with heterogeneous scales of the seven criteria considered in the study.

3. Case study

We applied the proposed methodological approach to the typical Italian farmhouse. Farmhouses are scattered across the country: their forms, materials and original functions are closely linked to the socio-geographic context in which they occur. Instead of studying the performance of a real building, the characteristics of 860 buildings of this typology were statistically analyzed (the totality of farmhouses in a central Italian Municipality) and a representative model of the results was developed (Fig. 1). The 860 studied

Table 2
Case study building: characteristics of the envelope.

Total building area (m^2): 288	Ground floor volume (m^3) 532	First floor volume (m^3) 641.6		Window-wall ratio (%): 5.98	Window opening area (m^2): 20.16
ROOF					
Material	**Thickness (m)**	**Conductivity (W/m-K)**	**Density (kg/m^3)**	**Specific heat (J/kg-K)**	**Transmittance (W/m2-K)**
Clay tiles	0.015	0.72	1800	840	2.5
Airgap	**Thermal resistance (m2-K/W):** 0.2				
Clay tiles	0.03	0.72	1800	840	
LOAD BEARING WALLS - FIRST FLOOR					
Sandstone	0.12	2.3	2600	1000	2.08
Mortar	0.03	0.72	1800	840	
Sandstone	0.12	2.3	2600	1000	
Plaster	0.02	0.8	1600	1000	
FLOOR					
Clay tiles	0.03	0.72	1800	840	2.89
Sand and gravel	0.15	2.3	2600	1000	
Clay tiles	0.03	0.72	1800	840	
LOAD BEARING WALLS – GROUND FLOOR					
Sandstone	0.16	2.3	2600	1000	1.81
Mortar	0.19	0.8	1600	1000	
Sandstone	0.16	2.3	2600	1000	
Plaster	0.02	0.8	1600	1000	
GROUND FLOOR					
Ceramic tiles	0.015	1.3	2300	1000	5.76
Concrete	0.3	1.35	2000	1000	
LOAD BEARING WALLS – INTERNAL WALLS					
Plaster	0.02	0.8	1600	1000	1.98
Sandstone	0.12	2.3	2600	1000	
Mortar	0.16	0.8	1600	1000	
Sandstone	0.12	2.3	2600	1000	
Plaster	0.02	0.8	1600	1000	
DIVIDING WALLS - INTERNAL WALLS					
Plaster	0.02	0.8	1600	1000	2.45
Brick	0.15	0.72	1800	840	
Plaster	0.02	0.8	1600	1000	
DOORS					
Timber	0.05	0.13	500	1600	1.87
WINDOWS					
Clear 3mm	0.003	0.9			
Air 8mm	0.0079				
Clear 3mm	0.003	0.9			

buildings are the farmhouse surveyed in the census of the rural buildings scattered throughout the municipality of Perugia (Region of Umbria) .[4]

The ground floor traditionally fulfilled multiple purposes: barn, warehouse, cellar and tool shed. The building is generally oriented with the stairway on the long side facing south. The ground floor windows are square shaped and smaller in size, because the ground floor was originally used as a shelter for animals. The characteristics of the envelope are described in Table 2. The geometry and the materials of the model respect the original characterization, with the exception of the windows, as all retrofit solutions already performed resulted in the replacement of the typical single glazing with double glazed windows.

The analysis was focused only on the roof insulation, since empirical evidence shows that it is the most influential component for energy efficiency [8,10], particularly in both temperate zones climates [57]. In particular, the important role of the roof for this model of farmhouses was demonstrated in a previous paper [12] where the authors have studied all the components of the building envelope. The insulating material was placed internally, because the laws on the preservation of Italian traditional farmhouses restrict external alterations [5]. Table 3 gives the 12 insulating materials considered. The list includes both inorganic (e.g., glass and rock wool), organic (e.g., cellulose, polystyrene foam), and unconventional (kenaf and hemp fibers) materials [4].

[4] http://istituzionale.comune.perugia.it/pagine/vincoli (in Italian).

Table 3
Insulation materials compared in the case study.

Insulating material for roof (thickness: 15 cm)	Conductivity (W/mK)	Density (Kg/m^3)	Specific heat (j/kgK)
a_1 - Hard fiberboard	0.04	130	2100
a_2 - Mineralized wood	0.075	400	2100
a_3 - Polystyrene foam slab	0.036	10	800
a_4 - Cork slab	0.052	150	1900
a_5 - Rock wool	0.034	50	840
a_6 - Glass wool	0.036	75	1030
a_7 - Kenaf fibers	0.037	50	2050
a_8 - Hemp fibers	0.044	50	1700
a_9 - Expanded perlite	0.045	95	837
a_{10} - Polyurethane	0.025	30	1500
a_{11} - Expanded vermiculite	0.065	95	1000
a_{12} - Cellulose	0.039	40	1400

4. Results of the multi-criteria assessment of the insulating materials

Twelve insulating materials were applied to the roof internally, and their performance was evaluated as discussed in the first and second phases of our approach. The computed results are shown in Table 4 and given in Sections 4.1 and 4.2.

Table 4
Performance of 12 materials with respect to 7 criteria (g1 - global warming; g2 - ozone layer depletion; g_3 - respiratory inorganics; g_4 - non-renewable energy; g_5 - Net Present Value; g_6 - comfort performance; g_7 - energy saving).

Material	*Code*	g_1	g_2	g_3	g_4	g_5	g_6	g_7
		Kg CO2 eq	kg CFC-11eq	Kg PM2.5 eq	MJ primary	€	HrDis	GJ
Hard fiberboard	a_1	12,096.012	0.0012312	5.5728	246,240.24	2951.50	3657.79	65.67
Mineralized wood	a_2	15,746.557	0.0006328	5.6160	139,537.39	2670.14	4392.80	59.41
Polystyrene foam slab	a_3	797.040	0.0000283	0.4168	22,680.02	2980.71	3750.48	66.32
Cork slab	a_4	3650.403	0.0003348	4.0392	81,216.08	2843.18	3974.48	63.26
Rock wool	a_5	1531.441	0.0000697	2.3328	23,328.02	3001.83	3659.44	66.79
Glass wool	a_6	2332.802	0.0003974	1.5465	73,008.07	2984.31	3681.89	66.40
Kenaf panel	a_7	855.360	0.0000861	0.9050	12,074.41	2975.77	3685.51	66.21
Hemp panel	a_8	630.720	0.0000743	0.3456	10,821.61	2908.35	3921.44	64.71
Expanded perlite	a_9	1995.841	0.0004881	0.9374	33,264.03	2853.07	3893.64	60.76
Polyurethane	a_{10}	2354.402	0.0000012	2.7432	65,880.06	3093.07	3357.30	68.82
Expanded vermiculite	a_{11}	788.400	0.0000956	1.2787	12,182.41	2730.82	4478.60	63.48
Cellulose	a_{12}	311.040	0.0000347	0.4039	6177.60	2954.19	3800.44	65.73

Table 5
Performances of the characteristic profiles for three classes.

Profile	g_1	g_2	g_3	g_4	g_5	g_6	g_7
b_1	12,096.012	0.0003348	4.0392	139,537.39	2730.82	4392.80	60.76
b_2	2332.802	0.0000956	1.2787	33,264.03	2853.07	3800.44	63.48
b_3	630.720	0.0000283	0.4039	10,821.61	2980.71	3659.44	66.21

4.1. Life cycle assessment and life cycle cost

The results of the environmental evaluation are reported in Table 4 (see columns: g_1 - global warming; g_2 - ozone layer depletion; g_3 - respiratory inorganics; g_4 - non-renewable energy). Since they represent impact categories, they must be minimized. Results show that cellulose (a_{12}) is always among the top ranked materials for all impact categories. Hemp panels (a_8) also perform relatively well. Among the non-natural materials, polystyrene (a_3) achieves the best results. Fiberboard (a_1) and mineralized wood (a_2) are ranked among the bottom three materials for all categories.

The evaluation of the insulating materials according to the LCC analysis is reported in column g_5 - Net Present Value (the lower the NPV, the better). The base year for the analysis and costs evaluation is 2015. The most favorable materials from an economic perspective are mineralized wood (a_2) and expanded vermiculite (a_{11}), whereas the highest cost is associated with polyurethane (a_{10}) and rock wool (a_5).

4.2. Dynamic thermal and energy simulation results

Results of energy and thermal comfort performances are presented in Table 4 (see columns g_6 and g_7). The comfort performance indicator (g_6) represents the hours of discomfort, thus it needs to be minimized. Criterion g_7 represents the energy saving during one year (to be maximized) resulting from the use of each insulating material compared to no insulation. According to both indicators, the use of polyurethane (a_{10}) is the most advantageous solution, whereas mineralized wood (a_2) and expanded vermiculite (a_{11}) are deemed the worst options.

4.3. Multi-criteria decision analysis

The performances of the twelve materials in terms of the seven evaluation criteria (see Table 4) were used for the application of the multi-criteria analysis. The objective of a comprehensive assessment of the materials' sustainability was to assign them to a set of three pre-defined and ordered classes: C_1 (low/bad quality), C_2 (medium quality), and C_3 (high/good quality).

4.3.1. Elicitation of preference information

Each class C_h, $h = 1, 2, 3$, is defined by a characteristic profile b_h. Three experts participating in a focus group meeting decided to define the performances of these profiles by indicating one of the performances observed in the set of materials. These characteristic class performances are summarized in Table 5.

Fig. 2 shows a representation of the three characteristic profiles (bad (b_1), medium (b_2), and good (b_3)) and three selected materials (hard fiberboard (a_1), cork slab (a_4), and rock wool (a_5)). Note that the figure accounts for different criteria preference directions (only g_5 and g_7 are of gain type). This is reflected by the least and the most preferred performances for the particular criteria denoted in Fig. 2 (i.e. the most preferred performances are greater than the respective least preferred performances only for g_5 and g_7).

When taking into account all seven criteria, the materials attain performances that are typical for different classes (see Fig. 2). For example, the performances of hard fiberboard (a_1) on g_1-g_4 are not better than the performances of characteristic profile b_1 for the bad class (C_1), but its performances for g_5-g_7 are more favorable than the performances of b_2. Furthermore, for all criteria but g_3, rock wool (a_5) is more preferred than the characteristic profile b_2 for the medium class (C_2), but for g_5-g_7 it is already at least as good as b_3. In view of such diverse results for the comparisons of alternatives and characteristic class profiles for different criteria, in order to obtain a decisive comprehensive recommendation in terms of the assignment of materials to pre-defined classes, we used Electre TRI-rC.

The preferences regarding the importance of individual criteria were elicited from the DM being an owner of the rural building. For this purpose, we used the SRF procedure. In Table 6, we present the ranking of criteria, the number of blank cards inserted, and ratio $\boldsymbol{Z}$. We also give the derived weights $\boldsymbol{w_j}$ and concordance threshold $\boldsymbol{\lambda}$. As agreed by the stakeholders involved, the latter was assumed to be equal to the sum of the weights of the three most significant criteria ($\boldsymbol{\lambda} = 0.7415$).

4.3.2. Class assignments derived from the application of Electre TRI-rC

The derived weights and concordance threshold were used to comprehensively compare the materials against the set of characteristic profiles. The results of this comparison are quantified with

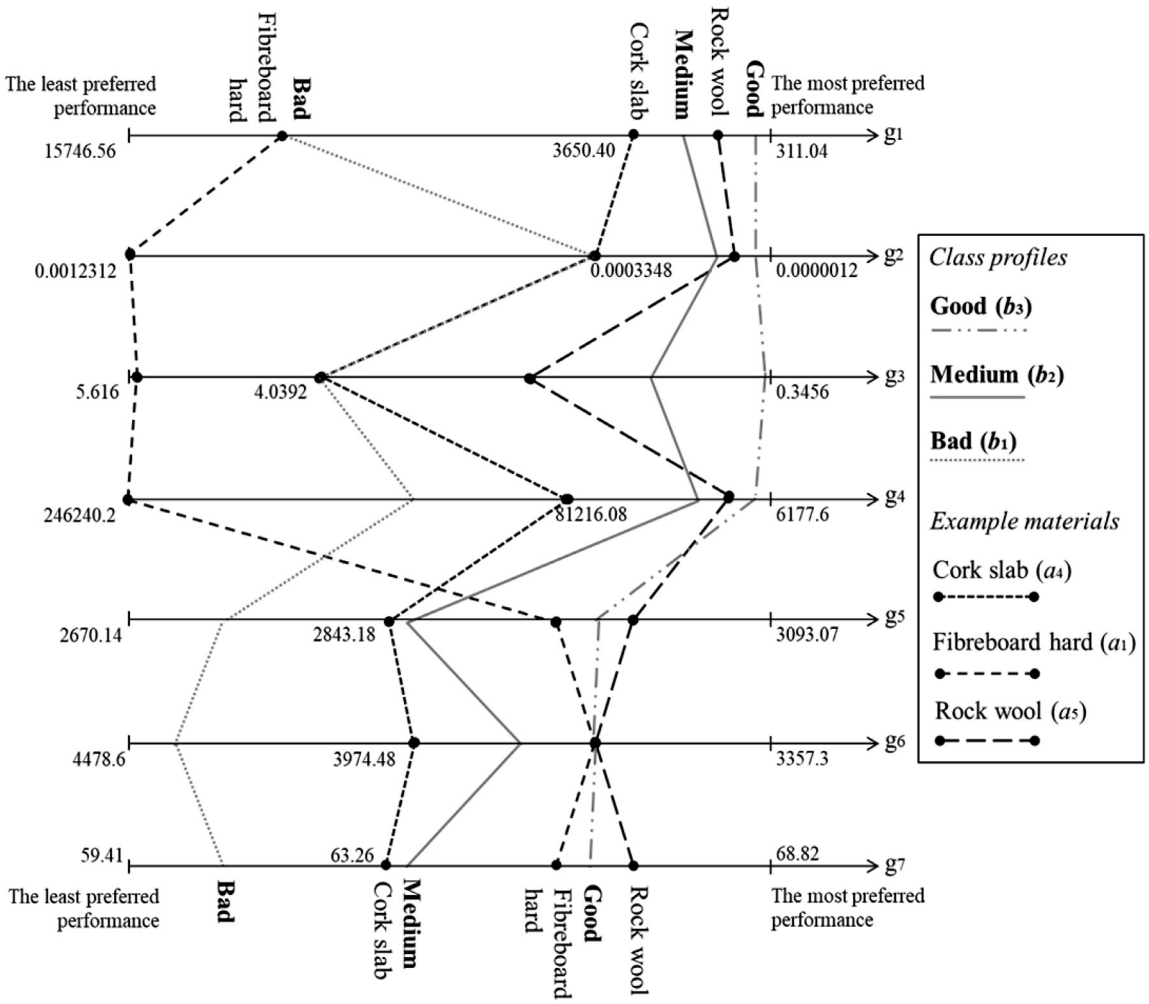


Fig. 2. Representation of the characteristic profiles and three selected materials.

Table 6
The order of cards with criteria names and blank cards provided by the expert in the SRF procedure ($l(j)$ (rank; the higher, the better), e_s (the number of inserted blank cards); Z (ratio between the weights of the most and the least important criteria); the weights w_j derived from the SRF procedure and the concordance threshold λ.

$Z = 8$, $\lambda = 0.7415$			
g_j	$l(j)$	e_s	w_j
g_1, g_6	1		0.0408
		0	
g_5, g_7	2		0.0884
		1	
g_4	3		0.1837
		0	
g_2	4		0.2313
		1	
g_3	5		0.3265

the comprehensive concordance degrees and a crisp outranking relation. In Table 7, we present these outcomes for a pair of exemplary materials.

Table 7 also shows the justification of the assignment given to these exemplary materials. On one hand, an assignment of a_1 to C_1 can be explained in terms of a_1 being judged comprehensively worse than b_1 ($b_1 \succ a_1$). On the other hand, an assignment of a_5 to C_2 can be justified with a_5 being preferred to b_1 ($a_5 \succ b_1$) and worse than b_3 ($b_3 \succ a_5$), and there being strong evidence that a_5 outranks b_2 and vice versa ($\sigma(a_5, b_2) = 0.6735 > \sigma(b_1, a_5) = 0.0$ and $\sigma(b_2, a_5) = 0.3265 > \sigma(a_5, b_3) = 0.2177$).

For eleven out of twelve insulating materials, the assignments suggested by ELECTRE TRI-rC are precise. This leads us to recommend the following classification for these materials:

- low quality (C_1): hard fiberboard (a_1), mineralized wood (a_2), cork slab (a_4), and glass wool (a_6);
- medium quality (C_2): rock wool (a_5), expanded perlite (a_9), and expanded vermiculite (a_{11});
- high quality (C_3): polystyrene foam slab (a_3), kenaf panel (a_7), hemp panel (a_8), and cellulose (a_{12}).

Rigid polyurethane foam (a_{10}) is assigned to [C_2, C_3], which excludes its assignment to the lowest class (C_1). These assignments are summarized in Table 8.

4.3.3. Sensitivity analysis

A two-fold sensitivity analysis was performed in order to verify the stability of the recommendation given. First of all, to validate the above assignments against the arbitrary choice of weights conducted with the SRF procedure, we sampled 1000 uniformly distributed weight vectors compatible with the order of criteria, intensities of preference, and ratio $Z = 8$ provided by the expert. For all materials except a_6 and a_9, all parameter sets considered (weights and concordance thresholds) unanimously confirmed the recommendation obtained for the weights derived with the SRF

Table 7
Comprehensive concordance indices and class assignments obtained with ELECTRE TRI-rC for a pair of exemplary materials (concordance threshold $\lambda = 0.7415$).

	b_1	b_2	b_3	$[C_L(a), C_R(a)]$		b_1	b_2	b_3	$[C_L(a), C_R(a)]$
a_1	$\prec$	$\prec$	$\prec$		a_5	$\succ$	$_R$	$\prec$	
$\sigma(a_1, b_h)$	0.2585	0.2177	0.0408	$[C_1, C_1]$	$\sigma(a_5, b_h)$	1.0	0.6735	0.2177	$[C_2, C_2]$
$\sigma(b_h, a_1)$	0.7415	0.7823	0.9592		$\sigma(b_h, a_5)$	0.0	0.3265	0.8231	

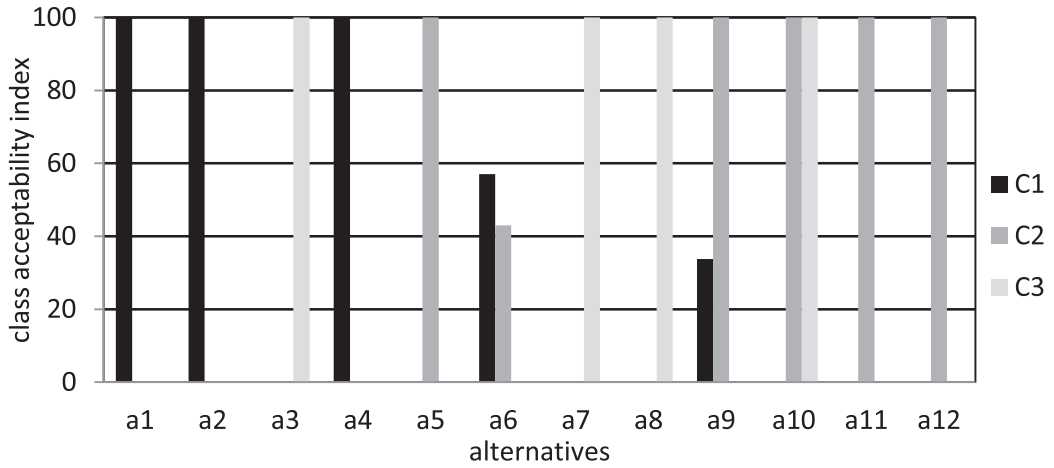


Fig. 3. Class acceptability indices (in %) for weight vectors consistent with the order of criteria provided by the expert in the SRF procedure and $Z = 8$.

Table 8
Class assignments obtained with ELECTRE TRI-rC for twelve insulating materials.

Low quality (C_1)	Medium quality (C_2)	High quality (C_3)
hard fiberboard (a_1)	rock wool (a_5)	polystyrene foam slab (a_3)
mineralized wood (a_2)	expanded perlite (a_9)	kenaf panel (a_7)
cork slab (a_4)	expanded vermiculite (a_{11})	hemp panel (a_8)
glass wool (a_6)		cellulose (a_{12})
polyurethane foam (a_{10})		

procedure (see Table 9). As concerns a_6, the recommendation provided by 57% of the sampled weights was C_1, while the remaining 43% models suggested C_2. For a_9, the majority of models (66.2%) opted for a precise assignment to C_2, while the rest of models recommended an imprecise assignment to $[C_1, C_2]$. As a result, a_9 was assigned either precisely or imprecisely to C_2 by all the sampled weight vectors.

The shares of weights presented in Table 9 can be interpreted as class acceptability indices, which indicate the support that is given to the assignment of each material to a particular class by various feasible parameters of an assumed preference model. The graphic presentation of these indices (see Fig. 3) clearly confirms the existence of a robust assignment to a particular class or a class interval for all materials except a_6, being confirmed by 100% uniformly distributed weight vectors consistent with the order of criteria provided by the expert in the SRF procedure and $Z = 8$.

Second, we investigated the recommendation obtained with the SRF procedure by considering 1000 different values of ratio Z uniformly distributed within the range [6,10]. The shares of weights thus derived for which material a is either precisely or imprecisely assigned to each class C_h are provided in Table 10 and Fig. 4. For the vast majority of materials, the class acceptability indices obtained are the same as for weight vectors consistent with the order of criteria provided by the expert in the SRF procedure and $Z = 8$ (see Table 9). The only exceptions in this regard concern a_6 (an increased support for the bad class (C_1) and a decreased support for the medium class (C_2)) and a_9 (a nullified support for the bad class (C_1)).

Overall, the assignments initially derived (a_1, a_2, a_4, and a_6 to C_1; a_5, a_9, and a_{11} to C_2; a_3, a_7, a_8 and a_{12} to C_3; a_{10} to C_2 and C_3) were confirmed to be robust, being supported either by all or by the vast majority of the sampled weight vectors. Moreover, the two-fold sensitivity analysis indicated that for a_6 and a_9, the probability of being assigned respectively to C_2 and C_1 was non-zero, though significantly lower when compared to the support given to the aforementioned recommendations.

5. Discussion

The presence of a complex system of technical, economic and environmental criteria required a multi-criteria approach for performing an integrated evaluation. The partial evaluation obtained from the energy and thermal comfort optimization and life cycle approach did not produce univocal results. By contrast, the application of ElectreTRI-rC and the underlying sensitivity analysis indicated that polyurethane, polystyrene foam, hemp and kenaf fibers were the most preferred materials.

The profiles of these materials are very diverse, which is useful for showing the paradox related to the topic of building energy efficiency. According to Copiello [58], adopting energy-intensive materials in order to achieve substantial energy savings in operation sometimes means using materials with a great amount of embodied energy. For example, solutions such as kenaf and hemp fibers are not the best according to operational energy saving, but are good as concerns the embodied energy.

The proposed methodology can be applied to any building model and to different parts of the envelope in order to identify the most favorable insulating materials for thermal comfort with good LCA and LCC performance. In comparison with other works in the building sector, the suggested approach has the advantage of incorporating different aspects, including the competitive characteristics. Several authors accounted for the environmental performance, applying LCA. For example, Frenette et al. [59] derived an environmental index for evaluating assemblies for exterior walls in residential buildings. The reviews of Islam et al. [41] and Zuo et al. [35] confirm an extensive use of the combined LCA and LCC

Table 9
Share of weights (in %) obtained for uniformly distributed weight vectors consistent with the order of criteria provided by the expert in the SRF procedure and $Z = 8$, for which material a is assigned precisely or imprecisely to class C_h.

$h \setminus a$	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}
1	**100.0**	**100.0**	0.0	**100.0**	0.0	**57.0**	0.0	0.0	33.8	0.0	0.0	0.0
2	0.0	0.0	0.0	0.00	**100.0**	43.0	0.0	0.0	**100.0**	**100.0**	**100.0**	**100.0**
3	0.0	0.0	**100.0**	0.00	0.0	0.0	**100.0**	**100.0**	0.0	**100.0**	0.0	0.0

Table 10
Share of weights (in %) obtained with the SRF procedure for Z uniformly distributed within the range [6,10], for which material a is assigned precisely or imprecisely to class C_h.

$h \setminus a$	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}
1	**100.0**	**100.0**	0.0	**100.0**	0.0	**75.0**	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.00	**100.0**	25.0	0.0	0.0	**100.0**	**100.0**	**100.0**	**100.0**
3	0.0	0.0	**100.0**	0.00	0.0	0.0	**100.0**	**100.0**	0.0	**100.0**	0.0	0.0

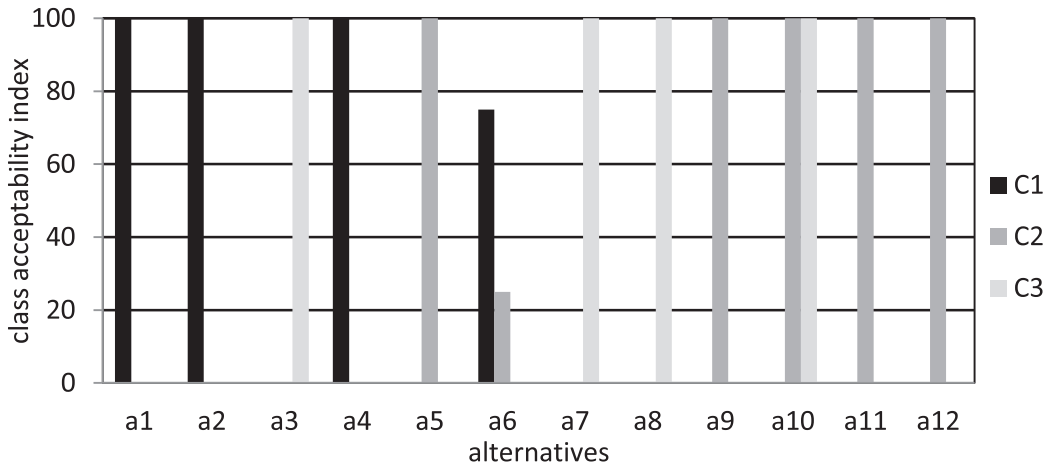


Fig. 4. Class acceptability indices (in %) for weights obtained with the SRF procedure for Z uniformly distributed within the range [6,10].

approaches in the building sector. Some combinations of LCA and MCDA were also considered by De Felice et al. [60]; Vilcekova et al. [61] and Santos et al. [62]. However, in comparison with the aforementioned works, our paper presents a more holistic approach, allowing one to consider environmental impact, economic feasibility, and comfort performance. A similar method was suggested by Motuziene et al. [20], although using a limited number of scenarios and giving no consideration to comfort.

6. Conclusions

The paper presents a hybrid approach, combining different methods, for a comprehensive assessment of the insulating materials applied to the roof of a traditional rural building in central Italy. The results offer some interesting solutions regarding the optimal materials to be used in the insulation of typical rural buildings, which are still poorly investigated. In particular the results show the presence of both natural (hemp and kenaf fibers) and not natural materials (polyurethane and polystyrene foam) as top solution, according to the DM involved (Farmhouse owner).

The results of our study could be further validated by eliciting preferences from more experts. This would require the extending of the proposed approach to a group decision-making framework Kadziński et al. [63]. Furthermore, the method can be extended to tolerate uncertainty in the performance of the insulating materials. Lastly, a comparison with other MCDA methods can be established to demonstrate further benefits of using ELECTRE TRI-rC in comparison with AHP, TOPSIS, or PROMETHEE, which so far have been the most popular MCDA methods in applications concerning buildings.

Acknowledgment

Miłosz Kadziński and Grzegorz Miebs acknowledge support from the Iuventus Plus program (0296/IP2/2016/74).

References

[1] L.F. Cabeza, L. Rincon, V. Vilarino, G. Perez, A. Castell, Life cycle assessment (LCA) and life cycle energy analysis (LCEA) of buildings and the building sector: a review, Renewable Sustainable Energy Rev. 29 (2014) 394–416.
[2] J. Li, M. Colombier, Managing carbon emissions in China through building energy efficiency, J. Environ. Manag. 90 (8) (2009) 2436–2447.
[3] J. Liang, Y. Qiu, B.L. Ruddell, M. Dalrymple, S. Earl, A. Castelazo, T. James, Do energy retrofits work? Evidence from commercial and residential buildings in Phoenix, J. Environ. Manag. (2017) in press, doi:10.1016/j.jeem.2017.09.001.
[4] S. Schiavoni, F. D'Alessandro, F. Bianchi, F. Asdrubali, Insulation materials for the building sector: a review and comparative analysis, Renewable Sustainable Energy Rev. 62 (2016) 988–1011.
[5] L. Mazzarella, Energy retrofit of historic and existing buildings. The legislative and regulatory point of view, Energy Build. 95 (2015) 23–31.
[6] M. Cellura, G. Ciulla, S. Longo, F. Guarino, Historical buildings in protected areas in Italy: a re-design study of a rural building, in: Proceedings of BSA 2017, 2017.
[7] J. Fischer, T. Hartel, T. Kuemmerle, Conservation policy in traditional farming landscapes, Conserv. Lett. 5 (3) (2012) 167–175.
[8] G. Verbeck, H. Hens, Energy savings in retrofitted dwellings: economically viable, Energy Build. 37 (2005) 747–754.
[9] E. Dascalaki, M. Santamouris, On the potential of retrofitting scenarios for offices, Build. Environ. 37 (2002) 557–567.
[10] P.B. Taylor, E.H. Mathews, M. Kleingeld, G.W. Taljaard, The effect of ceiling insulation on indoor comfort, Build. Environ. 35 (2000) 339–346.
[11] A. Kumar, B.M. Suman, Experimental evaluation of insulation materials for walls and roofs and their impact on indoor thermal comfort under composite climate, Build. Environ. 59 (2013) 635–643.
[12] M.E. Menconi, M. Chiappini, J.L.M. Hensen, D. Grohmann, Thermal comfort optimisation of vernacular rural buildings: passive solutions to retrofit a typical farmhouse in central Italy, J. Agric. Eng. 48 (3) (2017) 127–136.

[13] M. Cinelli, S.R Coles, K. Kirwan, Analysis of the potentials of multicriteria decision analysis methods to conduct sustainability assessment, Ecol. Indic. 46 (2014) 138–148.
[14] M. Cinelli, S.R. Coles, O. Sadik, B. Karn, K. Kirwan, A framework of criteria for sustainability assessment of nanoproducts, J. Cleaner Prod. 126 (2016) 277–287.
[15] B. Agarski, I. Budak, D. Vukelic, J. Hodolic, Fuzzy multi-criteria-based impact a category weighting in life cycle assessment, J. Cleaner Prod. 112 (2016) 3256–3266.
[16] A.R. Domingues, P. Marques, R. Garcia, F. Freire, L.C. Dias, Applying multi-criteria decision analysis to the life-cycle assessment of vehicles, J. Cleaner Prod. 107 (2015) 749–759.
[17] S.M. Khoshnava, R. Rostami, A. Valipour, M. Ismail, A. Rahmat, Rank of green building material criteria based on the three pillars of sustainability using the hybrid multi criteria decision making method, J. Cleaner Prod. (2018) 1–18.
[18] J. Rezaei, T. Nispeling, J. Sarkis, L. Tavasszy, A supplier selection life cycle approach integrating traditional and environmental criteria using the best worst method, J. Cleaner Prod. 135 (2016) 577–588.
[19] V. Kumar, K. Hewage, H. Haider, R. Sadiq, Sustainability evaluation framework for building cooling systems: a comparative study of snow storage and conventional chiller systems, Clean Technol. Environ. Policy 19 (2017) 137–155.
[20] V. Motuzienė, A. Rogoza, V. Lapinskiene, T. Vilutiene, Construction solutions for energy efficient single-family house based on its life cycle multi- criteria analysis: a case study, J. Cleaner Prod. 112 (2016) 532–541.
[21] M.E. Menconi, D. Grohmann, Model integrated of life cycle costing and dynamic thermal simulation (MILD) to evaluate roof insulation materials for existing livestock buildings, Energy Build. 81 (2014) 48–58.
[22] J.A. Orosa, A.C. Oliveira, A new thermal comfort approach comparing adaptive and PMV models, Renewable Energy 36 (2011) 951–956.
[23] M. Luo, B. Cao, J. Damiens, B. Lin, Y. Zhu, Evaluating thermal comfort in mixed–mode buildings: a field study in a subtropical climate, Build. Environ. 88 (2015) 46–54.
[24] M.P. Deuble, R.J. de Dear, Mixed-mode buildings: a double standard in occupants' comfort expectations, Build. Environ. 54 (2012) 53–60.
[25] J.F. Nicol, M.A. Humphreys, Thermal comfort as part of a self-regulating system, Build. Res. Pract. 6 (3) (1973) 191–197.
[26] J.F. Nicol, M.A. Humphreys, Adaptive thermal comfort and sustainable thermal standards for buildings, Energy Build. 34 (2002) 563–572.
[27] J.F. Nicol, M.A. Humphreys, Derivation of the adaptive equations for thermal comfort in free-running buildings in European standard EN15251, Build. Environ. 45 (2010) 11–17.
[28] R.J. de Dear, T. Akimoto, E.A. Arens, G. Brager, C. Candido, K.W.D. Cheong, N. Nishihara, S.C. Sekhar, S. Tanabe, J. Toftum, H. Zhang, Y. Zhu, Review Article - Progress in thermal comfort research over the last twenty years, Indoor Air 23 (2013) 442–461.
[29] P. Fanger, Thermal comfort: Analysis and Application in Environmental Engineering, Danish Technology Press, Copenhagen, 1970.
[30] L.F.R. Peeters, R. de Dear, J.L.M. Hensen, W. D'Haeseleer, Thermal comfort in residential buildings: comfort values and scales for building energy simulation, Appl. Energy 86 (5) (2009) 772–780.
[31] S. Carlucci, L. Pagliano, A review of indices for the long-term evaluation of the general thermal comfort conditions in buildings, Energy Build. 53 (2012) 194–205.
[32] CEN. 2007. Indoor environmental input parameters for design and assessment of energy performance of buildings: addressing indoor air quality, thermal environment, lighting and acoustics. EN15251.
[33] ISO- International Organization For Standardization, ISO 7730. Analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria, Ergon. Thermal Environ. (2005).
[34] G.A. Norris, Integrating life cycle cost analysis and LCA, International J. Lie Cycle Anal. 2 (6) (2001) 118–120.
[35] J. Zuo, S. Pullen, R. Rameezdeen, H. Bennetts, Y. Wang, G. Mao, Z. Zho, H. Du, H. Duan, Green building evaluation from a life cycle perspective in Australia: a critical review, Renewable Sustainable Energy Rev. 70 (2017) 358–368.
[36] L. Paolotti, G. Martino, A. Marchini, A. Boggia, Economic and environmental evaluation of trasporting importaed pellet: a case study, Biomass Energy 83 (2015) 340–353.
[37] L. Paolotti, G. Martino, A. Marchini, A. Boggia, Economic and environmental assessment of agro-energy wood biomass supply chains, Biomass Bioenergy 97 (2017) 172–185.
[38] M. Carlsson Reich, Economic assessment of municipal waste management systems- case studies using a combination of life cycle assessment (LCA) and life cycle costing (LCC), J. Cleaner Prod. 13 (2005) 253–263.
[39] R. Heijungs, E. Settanni, J. Guinee, Towards a computational structure for life cycle sustainability analysis: unifying LCA and LCC, Int. J. Life Cycle Assess. 18 (2013) 1722–1733.
[40] M. Ristimaki, A. Saynajoki, J. Heinonen, S. Junnila, Combining life cycle costing and life cycle assessment for an analysis of a new residential district energy system design, Energy 63 (2013) 168–179.
[41] H. Islam, M. Jollands, S. Setunge, Life cycle assessment and life cycle cost implication of residential buildings - A review, Renewable Sustainable Energy Rev. 42 (2015) 129–140.
[42] RU Halwatura, MTR Jayasinghe, Influence of insulated roof slabs on air conditioned spaces in tropical climatic conditions—a life cycle cost approach, Energy Build. 41 (6) (2009) 678–686.
[43] P. Gluch, H. Baumann, The life cycle costing (LCC) approach: a conceptual discussion of its usefulness for environmental decision-making, Build. Environ. 39 (2004) 571–580.
[44] ISO - International Organization For Standardization a. 2006. ISO 14040. Environmental management - Life cycle assessment: Principles and framework.
[45] ISO - International Organization For Standardization b. 2006. ISO 14044. Environmental management - Life cycle assessment: Requirements and Guidelines.
[46] D. Khasreen, P. Banfill, G. Menzies, Life cycle assessment and the environmental impact of buildings: a review, Sustainability 1 (3) (2009) 674–701.
[47] I.Z. Bribian, A.V. Capilla, A.A. Uson, Life cycle assessment of building materials: comparative analysis of energy and environmental impacts and evaluation of eco-efficiency improvement potential, Build Environ. 46 (2011) 1133–1140.
[48] L. Paolotti, A. Boggia, C. Castellini, L. Rocchi, A. Rosati, Combining livestock and tree crops to improve sustainability in agriculture: a case study using the Life Cycle Assessment (LCA) approach, J. Cleaner Prod. 131 (2016) 351–363.
[49] Ecoinvent, Ecoinvent Data V2.2 (2010).
[50] L. Zampori, G. Dotelli, V. Vernelli, Life cycle assessment of hemp cultivation and use of hemp-based thermal insulator materials in buildings, Environ. Sci. Technol. 47 (2013) 7413–7420.
[51] O. Jolliet, M. Margni, R. Charles, S. Humbert, J. Payet, G. Rebitzer, R. Rosenbaum, IMPACT 2002+: a new life cycle impact assessment methodology, Int. J. Life Cycle Assess. 8 (2003) 324.
[52] C. Castellini, A. Boggia, L. Paolotti, G.J. Thoma, K. Dae-soo, Environmental impacts and life cycle analysis of organic meat production and processing, in: S.C. Ricke, E.J. Van Loo, M.G. Johnson, C.A. O'Bryan (Eds.), Organic Meat Production and Processing, Wiley-Blackwell, 2012, pp. 113–136.
[53] M. Kadziński, T. Tervonen, J.R. Figueira, Robust multi-criteria sorting with the outranking preference model and characteristic profiles, Omega 55 (2015) 126–140.
[54] J.R. Figueira, B. Roy, Determining the weights of criteria in the ELECTRE type methods with a revised Simos' procedure, Eur. J. Oper. Res. 139 (2) (2002) 317–326.
[55] S. Corrente, S. Greco, R. Słowiński, Multiple criteria hierarchy process for ELECTRE TRI, Eur. J. Oper. Res. 252 (1) (2016) 191–203.
[56] J. Rezaei, M. Kadziński, C. Vana, L. Tavasszy, Embedding carbon impact assessment in multi-criteria supplier segmentation using ELECTRE TRI-rC, Ann. Oper. Res. (2017) in press, doi:10.1007/s10479-017-2454-y.
[57] J. Lucero- Alvarez, N.A. Rodriguez-Munoz, R.M. Martin-Dominguez, The effects of roof and wall insulation on the energy costs of low income housing in Mexico, Sustainability 8 (2016) 590.
[58] S. Copiello, Building energy efficiency: a research branch made of paradoxes, Renewable Sustainable Energy Rev. 69 (2017) 1064–1076.
[59] C.C. Frenette, B. Beauregard, R. Salenikovich, D. A., Derome, Using life cycle assessment to derive an environmental index for light- frame wood wall assemblies, Build. Environ. 45 (2010) 2111–2122.
[60] F. De Felice, F. Campagioni, A. Petrillo, Economic and environmental evaluation via an integrated method based on LCA and MCDA, *Procedia- Social and Behav. Sci.* 99 (2013) 1–10.
[61] S. Vilcekova, E. Burdova, M. Culakova, R. Nagy, Evaluation of material compositions of internal bearing walls from embodied energy and greenhouse gas emissions, J. Multidiscip. Eng. Sci. Technol. 2 (12) (2015) 3542–3545.
[62] J. Santos, G. Flintsch, A. Ferreira, Environmental and economic assessment of pavement construction and management practices for enhancing pavement sustainability, Resour. Conserv. Recycl. 116 (2017) 15–31.
[63] M. Kadziński, L. Rocchi, G. Miebs, D. Grohmann, M.E Menconi, L. Paolotti, Multiple criteria assessment of insulating materials with a group decision framework incorporating outranking preference model and characteristic class profiles, Group Decis. Negot. 27 (2018) 33–59.

Publication [P3]

Govindan, K., Kadziński, M., Ehling, R., and Miebs, G. (2019). Selection of a sustainable third-party reverse logistics provider based on the robustness analysis of an outranking graph kernel conducted with ELECTRE I and SMAA. In Omega (Vol. 85, pp. 1–15). Elsevier BV., DOI: 10.1016/j.omega.2018.05.007.

Number of citations[3]:

- according to Web of Science: 95
- according to Google Scholar: 137

Contribution of co-authors (excluding the author and the supervisor of this dissertation):

- Kannan Govindan:
  - Co-authorship of the idea underlying the paper (Section 1)
  - Data and preference collection for the case study and interpretation of its results
  - Co-authorship of the text of the publication
- Ronja Ehling:
  - Data and preference collection for the case study and interpretation of its results
  - Corrections to the first draft of the manuscript

[3] as on September 25, 2023

Omega 85 (2019) 1–15

Contents lists available at ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

Selection of a sustainable third-party reverse logistics provider based on the robustness analysis of an outranking graph kernel conducted with ELECTRE I and SMAA☆

Kannan Govindan[a,*], Miłosz Kadziński[b], Ronja Ehling[a], Grzegorz Miebs[b]

[a] Center for Sustainable Supply Chain Engineering, Department of Technology and Innovation, University of Southern Denmark, Campusvej 55, Odense, DK-5230, Denmark

[b] Institute of Computing Science, Faculty of Computing, Poznan University of Technology, Poznań, Piotrowo 2, 60-965, Poland

ARTICLE INFO

Article history:
Received 20 April 2017
Accepted 14 May 2018
Available online 19 June 2018

Keywords:
Third-party logistics provider
Sustainability
Multiple criteria decision analysis
Robustness analysis
Outranking method
Reinforced preference

ABSTRACT

Pressure from legislation and customers has motivated companies to consider reverse logistics (RL) in their operations. Since it is a complex procedure that requires an adequate system, the recent trend consists in outsourcing RL to third-party reverse logistics providers (3PRLPs). This paper provides the background of sustainable triple bottom line theory with focus on economic, environmental, and social aspects under 3PRL concerns. The relevant sustainability criteria are used in a case study conducted in cooperation with an Indian automotive remanufacturing company. To select the most preferred service provider, we use a hybrid method combining a variant of ELECTRE I accounting for the effect of reinforced preference, the revised Simos procedure, and Stochastic Multi-criteria Acceptability Analysis. The incorporated approach exploits all parameters of an outranking model compatible with the incomplete preference information of the Decision Maker. In particular, it derives the newly defined kernel acceptability and membership indices that can be interpreted as a support given to the selection of either a particular subset of alternatives or a single option. The proposed ELECTRE-based method enriches the spectrum of multiple criteria decision analysis approaches that can be used to effectively approach the problem of the 3PRLP selection. As indicated by the extensive review presented in the paper, this application field was so far dominated by Analytic Hierarchy Process and TOPSIS, whose weaknesses can be overcome by applying the outranking methods.

1. Introduction

Reverse logistics (RL) is defined as "the process of planning, implementing, and controlling the efficient, cost effective flow of raw materials, in-process inventory, finished goods and related information from the point of consumption to the point of origin for the purpose of recapturing value or proper disposal" [96]. Research attention to RL has increased in the past years for several reasons. With the recent focus on sustainability, organizations are mandated to take back end-of-life (EOL) products as part of their environmental service demands [18]. Moreover, the growing popularity of online shopping implies that more and more products need to be returned to their points of origin. Indeed, the return rate of all online purchased items lies between 8 and 12%. The costs of handling them in the reverse supply chain can exceed the costs that were necessary in the forward logistics processes [110].

As noted by Rogers and Tibben-Lembke [97], if the focus of forward logistics is the movement of a material from the point of origin toward the point of consumption, then the focus of reverse logistics should be the movement of a material from the point of consumption toward the point of origin. Thus, the requirements for forward logistics and reverse logistics are clearly different. In this perspective, Decision Makers (DMs) must rethink their strategies when RL issues are adopted, because the methods associated with forward logistics are no longer applicable. Concepts such as RL capacity and RL practices may sound similar to logistics capacity and logistics practices, but the reverse logistics terms have different constraints. For example, logistics capacity generally considers the logistics infrastructure and focuses on collection, delivery, information, and cash flow. Reverse logistics capacity includes more operations such as partial remanufacturing, recycling, and disposal. In addition, the customers' willingness and active engagement play a

☆ This manuscript was processed by AE C. Chen.

* Corresponding author.
E-mail addresses: kgov@iti.sdu.dk (K. Govindan), milosz.kadzinski@cs.put.poznan.pl (M. Kadziński), re@iti.sdu.dk (R. Ehling), grzegorz.miebs@protonmail.ch (G. Miebs).

https://doi.org/10.1016/j.omega.2018.05.007
0305-0483/

vital role in the difference between forward and reverse logistics capacity.

The company can handle its reverse logistics activities in three different ways [94]. Firstly, it can manage the service in-house. Secondly, it could own logistics subsidiaries through setting up or buying a logistics firm [15]. The last option is to outsource the function and buy the service [94]. Indeed, outsourcing the logistics operations is nowadays one of the foremost management strategies [40]. Nevertheless, many companies struggle with the implementation of RL, because they do not have any adequate systems in place to handle the reverse supply chain in-house [53]. Since the implementation of RL is an irreversible decision, manufacturers need to recognize its gravity. It is risky due to involving financial and operational aspects that have a long-term effect on the company [93]. Furthermore, the management of returns is complicated by uncertainties in timing, volume, and condition that may be difficult to predict [110].

If the company chooses to outsource its operations, it has to choose a reliable third-party reverse logistics provider (3PRLP). The provider needs to suit the type of reverse logistics network through an adequate information system, transportation, and material handling equipment, as well as warehousing facilities [42]. Moreover, evaluation of the 3PRLPs is by nature a multiple criteria problem. Formulating a set of relevant attributes for the context of outsourcing the logistics operations is not straightforward. Furthermore, dealing with their multiplicity and conflicting character requires the use of dedicated Multiple Criteria Decision Analysis (MCDA) methods. In this perspective, the contribution of this paper to the literature on the 3PRLP evaluation and selection with MCDA is four-fold.

Firstly, we provide the background of sustainable triple bottom line theory [29,66] with focus on the economic, environmental, and social aspects under 3PRLP concerns. Our in-depth literature review indicates that although all three sustainability dimensions should be considered in the evaluation of 3PRLPs, companies often place more emphasis on some of them (in particular, solely on the economic criteria). Drawing from the literature, we propose a classification that divides the criteria relevant for 3PRLP into the three sustainable concerns. The classification involves fourteen major evaluation categories (e.g., costs and RL capacity in the economic dimension, RL practices and green level within the environmental concerns, and micro- and macro-social impacts among the social aspects), and over seventy elementary viewpoints that can be relevant for a particular problem. Indeed, the importance of defining the appropriate criteria in the context of 3PRLP evaluation has been raised in many studies (see, e.g., [46]). Our classification can be interpreted in terms of critical success factors (CSFs) describing the key areas in business that can ensure competitive performance for a company [95] and can be useful in planning implementations such as outsourcing of the logistics operations. Above all, we detail the significance of sustainability in company's reverse logistics and, in general, supply chain.

Secondly, we provide a comprehensive summary of the journal papers on third-party reverse logistics. We note that their main focus is on prioritizing the importance of relevant criteria, designing reverse logistics networks, and selecting the most preferred provider. In fact, the majority of studies deal with a selection of a suitable RL provider [40,42,79]. In this context, we list the main application fields and note that the reported case studies were conducted in countries with the world's largest economies. Finally, we indicate that when dealing with multiple criteria evaluation of 3PRLPs, the most prevailing methods include different variants of Analytic Hierarchy Process (AHP) [104] and TOPSIS [51], or, in case an efficiency of different solutions had to be examined, Data Envelopment Analysis (DEA) [16]. Conversely, little research has focused on employing outranking methods in the context of sustainable 3PRLP selection although their characteristics have proven useful in other application fields [38].

In this perspective, the third contribution of the paper consists in proposing an ELECTRE-based approach [32] that can be used to support companies to evaluate the third-party logistics providers. When collecting inter-criteria preference information, the method admits incompleteness in the provided statements. On one hand, the preferences of the DM on the importance of particular criteria are collected using the revised Simos procedure [34] (note that it is also called the Simos–Roy–Figueira (SRF) procedure). Precisely, (s)he is expected to rank the criteria from the least to the most important while additionally differentiating the intensity of preference between different pairs of attributes. On the other hand, we tolerate an imprecise range of admissible values for the credibility threshold. As far as intra-criterion preference information is concerned, apart from the comparison thresholds traditionally considered in ELECTRE [102], we account for the effect of reinforced preference [103]. So far, it has not been used in the context of any real-world case study. The effect justifies an additional bonus for the alternative which is very strongly preferred over another alternative.

Since there exist multiple parameter values of an outranking model compatible with thus provided preferences, we tolerate all of them in the exploitation phase and use the results of robustness analysis for the recommendation of the most preferred option [100]. For this purpose, we couple ELECTRE I [98,99] that allows to indicate the most prevailing subset of alternatives with Stochastic Multi-criteria Acceptability Analysis [70,119]. The latter computes some acceptability indices by considering recommendation derived for different feasible sets of parameters.

In ELECTRE I, the most preferred subset of alternatives corresponds to a kernel of an outranking graph. Consequently, our robustness analysis is focused on the stability of a kernel obtained with different compatible instances of the preference model. In particular, we define the kernel acceptability and membership indices that can be interpreted as a support given to the selection of, respectively, a particular subset of alternatives or a single option. Nevertheless, we additionally address the robustness concerns by computing the respective results for pairwise relations and one against all comparisons. A comprehensive analysis of such results allows us to indicate the most preferred provider as well as some optional choices.

The fourth contribution of this paper consists in reporting the results of a case study that was conducted in cooperation with a company in an Indian automotive remanufacturing industry. In this way, we explore the status of 3PRL in the specific context of India, where it is still undeveloped. The study concerned evaluation of five service providers in terms of several economic, environmental, and social criteria. The justification for applying the proposed hybrid approach to support the case company in selecting the most preferred 3PRLP(s) derives from the favourable characteristics of the incorporated elementary approaches. These include:

- An operational simplicity and acceptance of the incomplete preference information on the relative importance of criteria (with the possibility of differentiating the intensity of preference for various pairs of criteria) by the revised Simos procedure [24,34,112];
- accounting for the imperfect knowledge of data and the arbitrariness when building the criteria by ELECTRE [32];
- deriving the recommendation from comparing the alternatives pairwise (without scoring the available options) and indicating the most preferred subset of alternatives by ELECTRE I [99] which is consistent with the type of a problem considered by the company;

- tolerating the incompleteness of preference information and verifying the stability of the choice recommendation by SMAA [119].

The paper is structured in the following way. Section 2 provides relevant literature resources on reverse logistics and third-party reverse logistics. The need for outsourcing logistics operations and the relevant criteria for the evaluation of 3PRLPs are discussed in Section 3. A new hybrid MCDA approach combining Stochastic Multi-criteria Acceptability Analysis with ELECTRE I is presented in Section 4. Section 5 presents a case study where the proposed method is applied to a real-world problem considered by an Indian auto parts remanufacturing company. The concluding section summarizes the managerial implications and draws some avenues for future research.

2. Review of literature on third-party reverse logistics

In this section, we present a literature review concerning third-party reverse logistics. We emphasize that scholarly attention for reverse logistics and 3PRLP has increased in the past years due to the demand of manufacturing companies.

Indeed, more and more firms consider the reverse supply chain as a management strategy that can lead to a competitive advantage [89]. Pokharel and Mutha [88] indicated that research on reverse logistics had been growing since the 1960s. In their review, Agrawal et al. [1] covered more than 240 papers on this topic published within the time frame 1986–2015, whereas Govindan et al. [44] reviewed over 150 works on RL that were published only between 2007 and 2013. The interest in reverse logistics is constantly increasing for legal, environmental, social, and economic reasons. In particular, Steven et al. [113] demonstrated its positive impact on the economic and ecological performance of a company. In the same spirit, Ramirez and Morales [91] analysed how RL could affect costs and organization performance of the companies. [122] dealt with the impact of reverse logistics on supply chain management performance in order to analyse its relations with the order and inventory variance amplification. A focus on the green operations in RL was presented by Kumar and Kumar [68].

Research on reverse logistics has been further developed in connection with the outsourcing of logistics operations to third-party reverse logistics providers. Selecting the appropriate reverse logistics provider is challenging since many criteria influence this decision. These criteria can have environmental, industrial, or operational characteristics [43]. Recent research has been done in the areas of supplier selection and outsourcing decisions, and different solution methodologies for the problem have been presented.

In Table 1, we summarize the journal papers on third party reverse logistics. Their scope varies from prioritizing criteria relevant for 3PRLP through applying various MCDA methods for supplier selection to designing reverse logistics networks. The reported application domains include electronics (e.g., mobile phones or computers) industry, plastic, paper, and carpet recycling, tire, battery, and medical device manufacturing, steel enterprise, and petrol retailing. The majority of studies on selecting the provider for outsourcing forward logistics operations were performed in USA, Brazil, China, Turkey, UK, India, and Iran. The prevailing techniques used to evaluate the service providers include different variants of AHP [104], TOPSIS [51], and Data Envelopment Analysis [16].

The review on 3PRLP proves that little research has included sustainability criteria. Since issues of corporate social responsibility (CSR) and sustainability were introduced to RL, a shift towards sustainable performance models has occurred [85]. Nevertheless, studies that accounted for all three pillars of sustainability (environmental, economic, and social) are rare. For example, Wang and Zhu [125] included the environmental area in their study when they combined the status of domestic third-party reverse logistics and the environmental requirements of a low-carbon economy. Razzaque and Sheng [94] accounted for the social aspect in their work and emphasized the importance of human factors. Social issues were also acknowledged by Boyson et al. [12], who pointed out the importance of human resource policy and labor conditions for their employees. However, according to Kafa et al. [62], there is the need for more research on 3PRLP that includes the triple bottom line of environmental, economic, and social goals. This paper contributes to closing this research gap by providing the background of sustainable triple bottom line theory under 3PRLP concerns. The usefulness of such an approach is demonstrated with a case study in the specific context of India, where the status of third-party logistics is still undeveloped.

3. Framework development

In this section, we discuss the reasons for outsourcing the logistics operations instead of maintaining them in-house. Since this paper emphasizes sustainability, we draw a strong connection with the third-party reverse logistics. We also discuss the critical success factors for the selection of 3PRLP. These had been used as criteria to evaluate the service providers within the case study reported in Section 5.

3.1. Outsourcing of logistics operations

Outsourcing has been defined as “an agreement in which one company contracts out a part of their existing internal activity to another company” [77, p. 68]. There is an ongoing trend in many industries to outsource certain activities, including supply chain management and logistics operations. The primary reasons for this involve globalization of business, rapid growth in global marketplaces [72], cost savings, enhancing revenue potential, and operational benefits.

Since reverse logistics performance generally takes up less than 5% of a company’s performance, it is natural for many companies to delegate this task to third-party reverse logistics providers [12]. According to Bradley [13], company’s logistics costs may be reduced by using a 3PRLP, since the experienced outsourced providers are often more efficient than the company itself. Moreover, production costs can be lowered by specialization effects and the proper utilization of core competencies. Companies can also take advantage of economies of scale by converting RL functions into an activity where profit is created [36]. Outsourcing can have further advantages, such as higher logistics performance, higher quality, optimized asset use, increased flexibility, and reduction in strategic and operational risk management [71]. In choosing the right provider, the DM’s selection can be strategic and can result in the achievement of new technologies, knowledge, and new markets [78].

Furthermore, the providers specialize in managing the reverse flow of the returned products as well as providing services such as remanufacturing, refurbishing, and recycling [96]. Due to the specialization, they can motivate companies to reach environmental sustainability goals, since they offer useful solutions for sustainable supply chains. Moreover, the activity of greening the supply chain may be connected with a better reputation for the company [123]. Consequently, RL providers are considered a key operational element for the development of a sustainable supply chain.

Nevertheless, implementing reverse logistics also carries some risks. These include loss of control over the logistics operations [126], difficulties in implementing environmentally practices due

Table 1
Summary of journal papers on third party reverse logistics published after 2000 (criteria: E = economic, V = environmental, S = social).

No.	Source	Type of study/ Application field / Country	Contribution	Tools	Issues addressed		
					E	V	S
1	[90]	Case study / Electronics industry	Proposed an integrated model based on Fuzzy AHP for evaluation and prioritization of selection criteria and Fuzzy TOPSIS for the selection and development of reverse logistics partners	MCDM – Fuzzy AHP and Fuzzy TOPSIS	X	X	
2	[118]	Case study / Composite pipe manufacturer / USA	Used the Analytic Network Process (ANP) and proposed an analytical framework to systematically model the complex nature of interactions among the selection factors of 3PRLPs	MCDM – ANP	X	X	
3	[81]	Case Study	Proposed a multiple objective additive network DEA model to evaluate and select the most preferred 3PRLPs	Network data envelopment analysis and MONLP	X		
4	[46]	Review / Brazil	Identified the main criteria, the systematic methods that can be used in order to select the most appropriate 3PRLP, and proposed a framework based on a multiple criteria decision aid approach to select 3PRLP		X	X	
5	[109]	Case Study / Plastic recycling / India	Evaluated the most efficient Reverse Logistics Contractor (RLC) through a proposed model, a hybrid method using AHP and the Fuzzy TOPSIS	MCDM – AHP and Fuzzy TOPSIS	X	X	
6	[115]	Numerical example / Copier remanufacturing and paper recycling / Turkey	Presented two hybrid simulation-analytical modelling approaches for the RL network design of the 3PRLP	Stochastic simulation model, Generic method	X		
7	[52]	Case study / Mobile phone company / India	Developed decision support system to assist the top management of the company in selection and evaluation of 3PRLPs by a hybrid approach using AHP and TOPSIS methods	MCDM – TOPIS and AHP	X	X	
8	[7]	Numerical example	Proposed an innovative approach based on a free disposable hull (FDH) to select the most preferred 3PRLP	DEA – FDH	X		
9	[130]	Case study / Supply chain & logistics company / Iran	Investigated the feedback and relationships among attributes, and identified the most important attributes in the evaluation of 3PRLP using ANP	ANP, Intuitionistic fuzzy set (IFS) and Grey Relation Analysis (GRA)	X	X	
10	[75]	Case Study / ELV company / Iran	Defined suitable assumptions given the situation of ELV's management in Iran to model the problem as a 3PRL network	Integer linear programming	X		
11	[41]	Case Study / Tire industry / India	Illustrated the interactions between the attributes for the 3PRLP development using Interpretive Structural Modelling (ISM)	MCDM – ISM	X	X	
12	[128]	Numerical example / Steel enterprise / China	Analysed the closed-form analytic expressions for both united optimization strategies in a centralized closed-loop system and the Stackelberg strategies in a decentralized system	Stochastic modelling	X		
13	[74]		Proposed a model for the logistics business of a large third party service provider; incorporated both forward and reverse product flows for the company, including price, transportation mode, and outsourcing cost		X		
14	[92]	Case study / Computer company / India	Proposed a model to efficiently assist the DMs in determining the most appropriate third-party reverse logistics provider using a combination of AHP and TOPSIS methods	MCDM – AHP and TOPSIS	X	X	
15	[9]	Case studies / Medical device remanufacturing and carpet recycling	Presented a multiple criteria approach for the RL network design	AHP	X		
16	[106]	Numerical example	Proposed a model for dealing with selecting 3PRLPs in the presence of both dual-role factors and imprecise data	DEA	X	X	
17	[5]	Numerical example	Proposed a new chance-constrained data envelopment analysis (CCDEA) approach to support the DMs in determining the most preferred 3PRLPs in the presence of both dual-role factors and stochastic data	DEA	X		
18	[6]	Numerical example	Proposed a new model (output-oriented super slack-based measure (SBM) model in the presence of stochastic data along with non-linear program was derived and further converted to quadratic program) for 3PRLP selection	Output-oriented super SBM model	X		
19	[40]	Case study / Battery company / India	Proposed a structured model for the selection of a 3PRLP under fuzzy environment for the battery industry, which establishes the relative weights for attributes and sub-attributes	Fuzzy extent analysis	X	X	
20	[105]		Proposed a model for selecting 3PRLP in the presence of multiple dual-role factors	DEA	X		
21	[108]	Numerical example	Introduced a heuristic based approach for solving the Vehicle Routing Problem (VRP) of 3PRLs which can be described as the problem of designing optimal routes from one depot to a number of customers subject to constraints	Tabu search, Clarke/Wright algorithm	X		
22	[64]	Case study / Battery recycling industry / India	Developed a multi-criteria group decision making (MCGDM) model in a fuzzy environment to guide the selection process of the most preferred 3PRLP through analysis of the interactions between criteria	MCDM – ISM and TOPSIS	X		

(*continued on next page*)

Table 1 (*continued*)

No.	Source	Type of study/ Application field / Country	Contribution	Tools	Issues addressed		
					E	V	S
23	[107]	Numerical example	Introduced a methodology (imprecise DEA) to select the most efficient 3PRLP in the conditions for both ordinal and cardinal data	IDEA	X		
24	[50]	Case study / China	Proposed a benchmarking decision making model to assist in the entry of 3PRLPs		X		
25	[63]	Case Study / Battery industry / India	Proposed a structured model for evaluating and selecting the most preferred 3PRLP under a fuzzy environment for the battery industry	MCDM – AHP and Fuzzy AHP	X	X	
26	[65]	Case Study / Tire manufacturing / India	Proposed a structured, MCDM model for evaluating and selecting the most preferred 3PRLP using Fuzzy TOPSIS	MCDM – Fuzzy TOPSIS	X	X	
27	[129]		Proposed a grey comprehension model to evaluate the 3PRLPs	Grey-AHP and Grey Relational Theory	X		
28	[28]	Numerical example	Supported the DMs in determining the most preferred 3PRLP using a two-phase model based on artificial neural networks and fuzzy logic in a holistic manner	Artificial neural networks, MCDM – Delphi and Fuzzy AHP	X		
29	[80]	Numerical example	Proposed a mixed-integer programming model and a genetic algorithm that can solve the reverse logistics problem involving the location and allocation of repair facilities for 3PLs	Genetic algorithm	X		
30	[54]	Numerical example	Selected the logistics service provider	ANP	X		
31	[55]	China	Analysed the interactions among major barriers that hinder or prevent the application of reverse logistics in the Chinese industries	AHP, TOPSIS, Grey Relative Analysis	X		
32	[79]	Numerical example	Evaluated 3PRLPs accounting for the factors such as end-of-life product organization within a decision framework	MCDM – ANP	X		
33	[67]	Review and field study / Third party transportation company / USA	Examined the issues and processes that an organization (reverse logistics provider) has to address to engage in the reverse logistics business		X	X	
34	[11]	Case study – cross case analysis / Petrol retailers / UK	Identified the factors which influence outsourcing decisions and the supply chain implications of outsourcing strategies		X		

to lack of capabilities and resources [35], and the need for maintaining a complex relationship between the company and the provider [124]. The latter needs to account for the trust between the involved parties as well as the exchange of information about materials, waste management, shared profits, and savings [4].

3.2. Critical success factors

The critical success factors (CSF) define key elements that are required to ensure the success of any business and allow company to achieve its goals [95]. If the areas identified as CSFs receive careful attention from the top management, they have the potential to create a competitive advantage for the company. Talib and Hamid [116] summarized the CSFs in various supply chain management fields such as the reverse supply chain, the outsourcing of the logistics operations, and the third-party logistics evaluation. However, they neglected sustainability aspects to a large extent. Instead, Kafa et al. [62] claimed that reverse logistics outsourcing should be environmentally, economically, and socially applied. This is consistent with the findings of Almeida [2] and Govindan et al. [43] who claimed that when selecting the RL provider, the outsourcing contract price is no longer the sole relevant criterion.

In Table 2, we propose a classification that divides the criteria relevant for 3PRLP into the three sustainable concerns. The environmental, economic, and social aspects include the criteria that fall under each category as well as the examples for these viewpoints seen as sub-criteria. The latter ones were derived from a literature review. The criteria listed in Table 2 lay the groundwork for the selection of the most preferred 3PRLP within the case study, but they can be also adopted for dealing with other problems in the same application area.

4. Multiple criteria decision analysis method for the assessment of the third-party reverse logistics providers

This section describes a multiple criteria decision analysis method that has been used to evaluate the third party logistics providers within the case study, to advance in solving the problem, and to select the most preferred alternative. The incorporated approach employs an outranking model to represent preferences of the DM, and investigates the impact of using different parameter values compatible with the DM's value system on the choice recommendation. In this regard, it combines the ELECTRE I method [99] used to select the most preferred subset of providers for a particular set of parameters of an outranking model, the revised Simos procedure [34] to derive the ranking of criteria, and Stochastic Multi-criteria Acceptability Analysis to conduct robustness analysis while avoiding the arbitrary selection of parameters [120].

The review presented in Section 2 proved that the most prevailing MCDA methods used in the context of 3PRLP evaluation were AHP [104], TOPSIS [51], and Data Envelopment Analysis [16]. The popularity of AHP is mainly due to its intuitiveness, the natural appeal of a semantic scale it employs for expressing relative importance, availability of the user-friendly software, and a hierarchical decomposition of the multiple criteria problems being efficient from both operational and computational viewpoints [10]. Furthermore, TOPSIS is appreciated by the practitioners for a sound logic that represents the rationale of human choice involving the comparisons of each alternative with both the ideal and anti-ideal options, as well as for a simple computation process that can be programmed even in a spreadsheet [111]. Finally, the main advantages of DEA are that it is a non-parametric approach not requiring any functional forms, it can simultaneously handle heterogeneous inputs and outputs, and it provides means for identifying the sources of inefficiency that can be analysed and quantified [20].

Table 2
Sustainability criteria for the evaluation of third-party logistics providers.

Aspect	Criteria	Sub-criteria	References
Economic	Costs	RL costs, cost of relationship, cost of service, cost reduction	[3,62,73,79]
	Quality	Quality of product / service / people, product performance, quality improvement	[37,62]
	RL Capacity	Financial capacity, specialized infrastructures, skilled professionals, capability under uncertainty, RSC performance	[48,62,83,114]
	Technology	Technology capacity, warehouse management, transportation management, inventory management, information technology, demand forecasting, investment in IT	[3,12,37,48,83]
	Relationship	Effective communication, mutual commitment, flexibility, fairness, trust, channel relationship, service quality	[48,117]
	Financial performance	Assets, reputation, primary customer loyalty, understanding business needs, market share, profitability	[3,12,73,82,117]
	Management of risk	Setting standard and monitoring, external and internal communication, supply chain integration, government policy, complaint management, shipping and tracking, order management	[17,19,48,56,83,94]
Environmental	RL practices	Collection, sorting, treatment, redistribution, take back policy, packing, storage, delivery	[17,62,94]
	Organizational role	Reclaim, recycle, remanufacturing, reuse, disposal, treatment	[25,62]
	Green level	Environmental management, pollution, resource consumption	[62,86]
	Low-carbon	Oil consumption, cleaning materials and clean energy use, carbon emissions, average volume of air emission pollutants	[49,86]
	Environmental management system	ISO 14000, environmental policies, environmental objectives, checking of environmental activities	[8,66,69]
Social	Micro-social impact	Employee satisfaction, customer satisfaction, stakeholder satisfaction, overall working relations	[12,62]
	Macro-social impact	Health and safety, local community, human factors	[62,94]

However, each of these approaches has also some major weaknesses. In particular, the completion of the model used by AHP is time consuming because the number of required pair-wise comparisons may become troublesome. Moreover, it applies an arbitrary transformation of the linguistic terms to real numbers [14]. Furthermore, TOPSIS requires an arbitrary normalization of the original evaluations on different criteria scales into a common scale, which affects the ranking of alternatives [131]. Besides, score- and distance-based MCDA methods admit compensation between different criteria. When it comes to Data Envelopment Analysis, it is often criticized for its poor discriminative power (for a discussion on the weaknesses of DEA, see [58]).

ELECTRE methods may be used to overcome the aforementioned problems [32]. In fact, they tolerate qualitative nature of some criteria and heterogeneous criteria scales. Also, they apply non-compensatory aggregation of multiple criteria, additionally offering the possibility to model the effects of strong advantage (i.e., the effect of reinforced preference [103]) or critical weakness (i.e., the veto effect) in the comparison of a pair of alternatives. Moreover, although ELECTRE admits incomparability, it offers a wide spectrum of tools to discriminate between different alternatives. Finally, the proposed method explicitly deals with the incompleteness of the DM's preference information by considering all compatible preference models, which enhances her/his trust in the provided recommendation. Let us use the following notation:

- $A = \{a_1, a_2, \ldots, a_n\}$ is a set of alternatives (third-party logistics suppliers);
- $G = \{g_1, g_2, \ldots, g_m\}$ is a family of evaluation criteria that represent relevant points of view on the quality of assessed alternatives; $J = \{1, 2, \ldots, m\}$;
- $g_j(a)$ is the performance of alternative $a \in A$ with respect to criterion g_j, $j = 1, \ldots, m$ (when presenting the method, without loss of generality, we assume that all criteria are of gain type (i.e., the greater the performance, the better)).

4.1. Preference elicitation

In this section, we discuss the meaning of parameters related to the formulation of an outranking preference model that need to be elicited from the DM. Their use allows to first make the alternatives more comparable and then to identify the most preferred subset of options.

Step P1: Elicit the ranking of criteria g_j, $j = 1, \ldots, m$ using the SRF procedure [34], and determine the constraints on the admissible criteria weights w_j. The SRF procedure assumes the DM would rank the cards with criteria names from the least important to the most important, while admitting that some criteria are deemed indifferent. Then, the DM is asked to quantify the intensity of preferences between successive groups of criteria L_s and L_{s+1} through e_s blank cards inserted between these groups. As a result, each criterion g_j is assigned some importance rank l_j so that the greater the rank, the better. Finally, the DM needs to specify ratio Z between the importances of the most and the least significant criteria denoted by L_v and L_1, respectively [24].

Step P2: Specify the indifference q_j and preference p_j thresholds for each criterion g_j, $j = 1, \ldots, m$. These technical parameters reflect the discriminating character of g_j in face of imperfect knowledge [27]. They indicate, respectively, the maximal performance difference that is negligible on g_j and the minimal performance difference that induces a strict preference of one alternative over another on g_j [102]. Setting $p_j > q_j \geq 0$ allows considering a gradual transition from full concordance (in case $g_j(a) - g_j(b) \geq -q_j$) to no concordance (in case $g_j(a) - g_j(b) \leq -p_j$).

Step P3: Specify the reinforced preference threshold rp_j and the reinforcement factor ω_j for a subset of criteria on which a very strong preference of one alternative over another ($g_j(a) - g_j(b) \geq rp_j > p_j$) should justify an additional bonus ($\omega_j > 1$) when compared with the case where the preference is not that strong [103].

Step P4: Specify the veto threshold v_j for a subset of criteria g_j, $j = 1, \ldots, m$, that should be attributed a sufficient power to impose a strong opposition to outranking of a over b. This threshold represents the minimal performance difference $g_j(b) - g_j(a)$ that would justify not considering a at least as good as b even if it is more advantageous on all remaining criteria.

Step P5: Elicit the minimal λ_* and maximal λ^* value of the credibility threshold λ, i.e., the admissible range of values for the credibility of an outranking relation that would justify the truth of a crisp outranking relation.

4.2. Construction and exploitation of an outranking relation for selecting the most preferred subset of alternatives

In this section, we present a variant of ELECTRE I that has been used in the study to select the most preferred reverse logistics service provider. The method constructs an outranking relation S and represents it with an outranking graph G_S. The latter is exploited to identify the graph kernel that corresponds to the most advantageous subset of options. When recalling the main steps of the method, we assume that a set of precise parameters values ($w_j, q_j, p_j, rp_j, \omega_j, v_j$ for $j = 1, \ldots, m$, and λ) related to the formulation of an outranking model is given.

Step O1: For each criterion g_j, $j = 1, \ldots, m$, define the marginal concordance function $c_j(a, b)$ indicating for each pair of alternatives a and b a degree to which a outranks b on g_j. We use $c_j(a, b)$ defined as follows:

$$c_j(a,b) = \begin{cases} \omega_j, & \text{if } g_j(a) - g_j(b) \geq rp_j, \\ 1, & \text{if } rp_j > g_j(a) - g_j(b) \geq -q_j, \\ 0, & \text{if } g_j(a) - g_j(b) < -p_j, \\ \dfrac{g_j(a) - g_j(b) + p_j}{p_j - q_j}, & \text{otherwise.} \end{cases}$$

Thus, if a is very strongly preferred to b on g_j, then $c_j(a, b) = \omega_j > 1$; if a is at least as good as b on g_j, but its superiority is not that strong, then $c_j(a, b) = 1$; if b is weakly preferred to a, then $c_j(a, b) \in (0, 1)$, whereas in case b is strictly preferred to a, there is no concordance with aSb.

Step O2: Compute the comprehensive concordance index $C(a, b)$ indicating the strength of the coalition of criteria supporting the hypothesis about a being at least as good as b:

$$C(a,b) = \frac{\sum_{j=1}^{m} w_j \cdot c_j(a,b)}{\sum_{j \in F^{RP}} w_j \cdot \omega_j + \sum_{j \in J \setminus F^{RP}} w_j} \in [0,1],$$

where $F^{RP}(a, b) = \{j \in J : g_j(a) - g_j(b) \geq rp_j\}$ is a subset of criteria for which a reinforced preference of a over b occurs [103]. In this way, the contribution of a subset of criteria $F^{RP}(a, b)$ in the support of aSb is greater than their contribution in case the effect of reinforced preference was not considered. Clearly, the greater $C(a, b)$, the greater the concordance with the hypothesis that a outranks b.

Step O3: For each criterion g_j for $j = 1, \ldots, m$, define the marginal discordance function $d_j(a, b)$ indicating for each pair of alternatives a and b a degree to which g_j opposes to outranking aSb. We use $d_j(a, b)$ defined as follows:

$$d_j(a,b) = \begin{cases} 1, & \text{if } g_j(b) - g_j(a) \geq v_j, \\ 0, & \text{if } g_j(b) - g_j(a) \leq p_j, \\ \dfrac{g_j(b) - g_j(a) - p_j}{v_j - p_j}, & \text{otherwise.} \end{cases}$$

Thus, if a is critically worse than b on g_j, then $d_j(a, b) = 1$; if a is not worse than b by more than p_j, there is no discordance, whereas in case b is strictly preferred to a on g_j, but the performance difference is not that critical, then $d_j(a, b) \in (0, 1)$. Also, when v_j was not specified for g_j, $d_j(a, b) = 0$ for all pairs of alternatives.

Step O4: Combine the comprehensive concordance index and the marginal discordance indices into a credibility $\sigma(a, b)$ of an outranking relation aSb in the following way [84]:

$$\sigma(a,b) = C(a,b) \cdot [1 - \max_{j \in J} d_j(a,b)].$$

Thus, in case there is no discordance on any criterion, then $\sigma(a, b) = C(a, b)$, whereas in case $d_j(a, b) > 0$ for some $j = 1, \ldots, m$, then the credibility is decreased with respect to $C(a, b)$. In particular, if some criterion strongly opposes to aSb ($d_j(a, b) = 1$), then $\sigma(a, b) = 0$.

Step O5: For each pair of alternatives a and b, compare $\sigma(a, b)$ with the credibility threshold λ to verify the truth of a crisp outranking relation aSb in the following way:

$$\sigma(a,b) \geq \lambda \Rightarrow aSb.$$

If $\sigma(a, b) < \lambda$, a does not outrank b ($aS^C b$). Knowing if aSb and bSa hold, one can verify the truth of preference $\succ$, indifference $\sim$, and incomparability R relations in the following way:

$$aSb \wedge bS^C a \Rightarrow a \succ b,$$
$$aSb \wedge bSa \Rightarrow a \sim b,$$
$$aS^C b \wedge bS^C a \Rightarrow aRb.$$

Note that while constructing an outranking relation S, we have replaced the original concordance and discordance tests considered in ELECTRE I by their more recent implementations. This allowed us to better represent the preferences of the DM involved in the case study. Firstly, when conducting the concordance test, we considered the effect of reinforced preference, because at the stage of problem structuring we identified some criteria on which a very strong preference of one alternative over another should justify some bonus with respect to the case where the preference is not that strong. Secondly, we employed a fuzzified marginal discordance so that to account for the criteria weakly opposing to an outranking relation ($0 < d_j(a, b) < 1$). The binary marginal discordance with $d_j(a, b)$ defined solely with respect to the veto threshold v_j was found too restrictive for the considered study. It would simply occur too rarely to imply that the arguments against the outranking have an actual impact on the obtained results. Such an impact was found appealing by the involved experts. Thirdly, when computing the comprehensive discordance, we accounted for the greatest observed marginal discordance rather than all marginal discordances. Let us note that our study involved 14 criteria, but only 5 of them were attributed a power to veto against the outranking. In this perspective, a maximal marginal discordance was found as an adequate measure quantifying a power of such a limited subset of criteria potentially opposing to an outranking. Note that under such a setting, in case there were multiple criteria weakly opposing to an outranking, their impact on the outranking credibility would be reduced as only the strongest opposition against the outranking would be taken into account. Nonetheless, when there was at least one criterion strongly opposing to an outranking, the credibility would be nullified. Fourthly, the implemented outranking credibility clearly distinguishes the ways concordance and discordance are implemented as the criteria that negatively intervene in the credibility are not restricted to those for which the discordance is greater than the comprehensive concordance as, e.g., in the ELECTRE TRI method [32].

Step O6: Represent the outranking relation S with an outranking graph G_S whose nodes correspond to the alternatives and arcs reflect the truth of S. Select the most preferred subset of alternatives in A by identifying the kernel K of graph G_S [98]. K is defined as a subset of alternatives (nodes) which are incomparable in terms of S, and the alternatives not contained in K are outranked by at least one alternative belonging to K [101,127]. If G_S is acyclic, there exists a unique kernel. Otherwise, the cycle(s) need to be reduced before K is identified. For this purpose, we aggregate all elementary nodes in a cycle into a singleton (an artificial node) inheriting all in- and out-arcs from its component nodes (for an example, see Fig. 1(a)). Consequently, all alternatives which form a cycle are considered indifferent.

Note that the interpretation of an outranking graph kernel as the most preferred subset of alternatives derives from its definition. On one hand, a condition of an internal stability implies that the alternatives contained in the kernel do not outrank each other. This means that they are incomparable in terms of S, and hence there are no sufficiently strong reasons to judge one alternative from the kernel as more advantageous than another. On the other

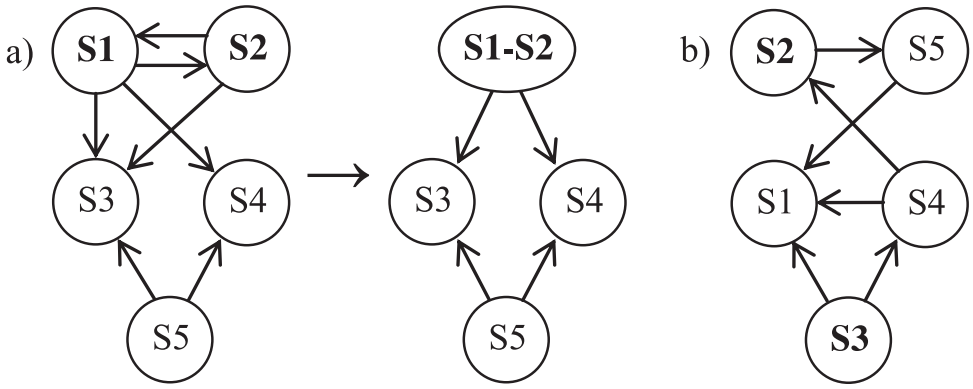


Fig. 1. Example outranking graphs: a) elimination of a cycle involving $S1$ and $S2$, b) identification of the kernel $\{S2, S3\}$.

hand, a condition of an external stability implies that the alternatives not contained in the kernel are outranked by at least one alternative from the kernel. In this perspective, the alternatives contained in the kernel jointly offer sufficiently strong arguments to neglect the remaining alternatives. However, as noted by Figueira et al. [33], in practical decision aiding, the computation of an outranking graph kernel can be treated as a pre-processing step. Then, the kernel is not treated as the most preferred subset of alternatives, but rather as a limited subset of options, from which the best compromise solution could be subsequently selected. In our study, we refer to both above interpretations. That is, when discussing the results obtained for a particular set of parameters, we treat the kernel as the most preferred subset of alternatives. However, since we consider multiple sets of parameters compatible with the DM's preferences, the analysis of all respective graph kernels is used to identify the most preferred alternative.

In what follows, we present the steps of an algorithm for identifying a kernel in an acyclic graph:

I. Construct a table with three columns (denoted as status, key and value) and n rows, where n is the number of nodes (alternatives). For each row, the status is initially empty, the key corresponds to the identifier of some node and the respective value is composed of all predecessors of this node.
II. Repeat until the status of each row is marked with either Y or N:
   a) Mark with Y all rows with the empty values (i.e., nodes with no predecessors that would be in the kernel; note that each value, which contains solely the crossed out identifiers, is also treated as empty).
   b) Mark with N all rows whose values contain at least one key of a row marked with Y (i.e., nodes which have some predecessor in the kernel).
   c) Cross out the keys of rows marked with N from all values of the rows which have not been yet marked with Y or N.
III. The graph kernel contains all keys of rows marked with Y.

In Table 3, we report the elementary steps of the above algorithm applied to the graph illustrated in Fig. 1(b). The kernel is composed of nodes $S2$ and $S3$.

4.3. Robustness analysis with stochastic multi-criteria acceptability analysis applied to ELECTRE I

In this section, we discuss the results that can be derived from the robustness analysis incorporating all parameter values related to the formulation of an outranking model that are compatible with the preference information provided by the DM. In this regard, let us remind that the plurality of different sets of weights compatible with the preference of the DM expressed within the SRF procedure can be handled in different ways. In particular, [112] proposed different robust rules for selection of the precise weights consistent with the DMs' rankings of criteria. Moreover, [39] computed the variety of results that can be obtained for the whole set of compatible weight vectors by means of Linear Programming techniques. Finally, [22] exploited this set using Stochastic Multi-criteria Acceptability Analysis [70,120], while additionally extending the SRF procedure to handling a hierarchical structure of criteria as well as an imprecision in the number of blank cards inserted between successive subsets of criteria and in the ratio between the most and the least important criteria. We follow the last approach by incorporating the Monte Carlo simulation [121] to estimate the values of acceptability indices measuring the variety of different preferences that confirm a particular choice recommendation.

Step R1: Define the space $(w, \lambda)^{DM}$ of weights and credibility thresholds compatible with preferences of the DM by considering the following constraint set $E(w, \lambda)$:

$$E(w,\lambda)\left\{\begin{array}{l}
[C1]\ w_i > w_j,\ \text{for all } g_i \in L_t, g_j \in L_s \text{ and } t > s,\\
[C2]\ w_i = w_j,\ \text{for all } g_i, g_j \in L_s,\\
[C3]\ w_{j+1} - w_j > w_{p+1} - w_p,\ \text{if } e_j > e_p,\\
[C4]\ w_i = Z \cdot w_j,\ \text{for all } g_i \in L_v, g_j \in L_1,\\
[C5]\ \sum_{j=1}^{m} w_j = 1,\\
[C6]\ w_j > 0,\ j = 1, \ldots, m,\\
[C7]\ \lambda_* \le \lambda \le \lambda^*,
\end{array}\right.$$

where constraints [C1], [C2], [C3] and [C4] reproduce the ranking of criteria provided by the DM while accounting for the preference intensities (see [C3]) as well as the ratio between the weights of the most and the least significant criteria (see [C4]), [C5] normalizes the sum of weights, [C6] guarantees that all weights are positive, and [C7] sets the bounds for the credibility threshold.

Note that constraint [C3] interprets the number of blank cards inserted between successive subsets of criteria in terms of preference intensities. For example, a difference between the weights of all pairs of criteria separated by two blank cards should be greater than a difference between the weights of all pairs of criteria separated by a single blank card. This interpretation is consistent with both an intuitive understanding of the SRF procedure by the experts involved in our case study and a discussion provided by [112]. However, it implies that values assigned to blank cards inserted between various groups of criteria may differ. Note that even though [22] postulated that these values should be equal, they also admitted that the differences between weights of criteria separated by the same number of blank cards could be different.

Let us remark that assigning exactly the same value to each blank card by means of equalities rather than inequalities would imply that – in case there is no imprecision in the number of blank

Table 3
Elementary steps of the algorithm for identifying a kernel in an example acyclic graph.

Step I			Step IIa			Step IIb		
Status	Key	Value	Status	Key	Value	Status	Key	Value
	S1	S3, S4, S5		S1	S3, S4, S5	N	S1	**S3**, S4, S5
	S2	S4		S2	S4		S2	S4
	S3	–	Y	S3	–	Y	S3	–
	S4	S3		S4	S3	N	S4	**S3**
	S5	S2		S5	S2		S5	S2
	Step IIc			Step IIa			Step IIb	
Status	Key	Value	Status	Key	Value	Status	Key	Value
N	S1	S3, S4, S5	N	S1	S3, S4, S5	N	S1	**S3**, S4, S5
	S2	~~S4~~	Y	S2	~~S4~~	Y	S2	~~S4~~
Y	S3	–	Y	S3	–	Y	S3	–
N	S4	S3	N	S4	S3	N	S4	S3
	S5	S2		S5	S2	N	S5	**S2**

cards between successive subsets of criteria and/or in the ratio between the most and the least important criteria – there exists just a single compatible weight vector. This, in turn, would prevent the need for conducting robustness analysis.

Step R2: For each $(w, \lambda) \in (w, \lambda)^{DM}$, construct an outranking relation $S^{(w,\lambda)}$ and exploit it to derive a respective graph kernel $K^{(w,\lambda)}$.

Step R3: Compute a set of stochastic acceptability indices by exploiting the consequences of applying all compatible sets of parameters of an outranking model $(w, \lambda)^{DM}$ on the set of alternatives A:

- *Acceptability indices derived from the pairwise comparisons*
  Let us define a *relation acceptability index RelAI*(a, b, Rel) for $Rel \in \{S, S^C, \succ, \prec, \sim, R\}$ as the share of compatible weights and credibility threshold $(w, \lambda) \in (w, \lambda)^{DM}$ for which *Rel* holds for the comparison of a and b. Formally, the index is computed as an integral over the space $(w, \lambda)^{DM}$ of uniformly distributed parameters of an outranking model:

$$RelAI(a, b, Rel) = \int_{(w,\lambda)\in(w,\lambda)^{DM}} m((w, \lambda), a, b, Rel)\, d(w, \lambda),$$

  where $m((w, \lambda), a, b, Rel)$ is the confirmation function of a specific relation $Rel \in \{S, S^C, \succ, \prec, \sim, R\}$:

$$m((w, \lambda), a, b, Rel) = \begin{cases} 1, & \text{if } aRel^{(w,\lambda)}b, \\ 0, & \text{otherwise.} \end{cases}$$

  The definition of *RelAI*(a, b, Rel) can be adapted to the case of S, S^C, $\succ$, $\prec$, $\sim$, or R. Let us call the respective indices as *outranking* (*OAI*), *non-outranking* (*NOAI*), *preference* (*PAI*), *inverse preference* (*IPAI*), *indifference* (*IAI*), and *incomparability* (*IRAI*) *acceptability indices*. For example:

$$OAI(a, b) = \int_{(w,\lambda)\in(w,\lambda)^{DM}} m((w, \lambda), a, b, S)\, d(w, \lambda),$$

  where $m((w, \lambda), a, b, S) = 1$, if $aS^{(w,\lambda)}b$, and 0, otherwise. Further:

$$IRAI(a, b) = \int_{(w,\lambda)\in(w,\lambda)^{DM}} m((w, \lambda), a, b, R)\, d(w, \lambda),$$

  where $m((w, \lambda), a, b, R) = 1$, if $aS^{C,(w,\lambda)}b$ and $bS^{C,(w,\lambda)}a$, and 0, otherwise. Note that for each $(a, b) \in A \times A$: $OAI(a, b) + NOAI(a, b) = 1$ and $PAI(a, b) + IPAI(a, b) + IAI(a, b) + IRAI(a, b) = 1$.
- *Acceptability indices derived from the comparison of one alternative against all remaining ones:*
  - *Comprehensive outranking index COI*(a) is the share of $(w, \lambda)^{DM}$ for which a outranks all remaining alternatives $b \in A\backslash\{a\}$ jointly, i.e.:

$$COI(a, A, S) = \int_{(w,\lambda)\in(w,\lambda)^{DM}} m((w, \lambda), a, A, S)\, d(w, \lambda),$$

    where $m((w, \lambda), a, A, S) = 1$, if $aS^{(w,\lambda)}b$ for all $b \in A \setminus \{a\}$, and 0, otherwise.
  - *Comprehensive non-outranked index CNOI*(a) is the share of $(w, \lambda)^{DM}$ for which a is not outranked by any other alternative $b \in A\backslash\{a\}$, i.e.:

$$CNOI(a, A, S^C) = \int_{(w,\lambda)\in(w,\lambda)^{DM}} m((w, \lambda), a, A, S^C)\, d(w, \lambda),$$

    where $m((w, \lambda), a, A, S^C) = 1$ if $bS^{C(w,\lambda)}a$ for all $b \in A \setminus \{a\}$, and 0, otherwise.

  Thus defined *COI*(a, A, S) and *CNOI*(a, A, S^C) indicate the joint superiority of a over all remaining alternatives while taking into account, respectively, the positive (S) or negative (S^C) arguments.
- *Acceptability indices related to the stability of a graph kernel*
  The *kernel acceptability index KAI*(A') for each subset of alternatives $A' \subseteq A$ is the share of $(w, \lambda)^{DM}$ that indicate A' as the graph kernel $K^{(w,\lambda)}$, i.e.:

$$KAI(A') = \int_{(w,\lambda)\in(w,\lambda)^{DM}} m((w, \lambda), K, A')\, d(w, \lambda),$$

  where $m((w, \lambda), K, A')$ is the kernel membership function:

$$m((w, \lambda), K, A') = \begin{cases} 1, & \text{if } K^{(w,\lambda)} = A', \\ 0, & \text{otherwise.} \end{cases}$$

  Further, we compute the share of $(w, \lambda) \in (w, \lambda)^{DM}$ for which $a \in A$ is in the graph kernel $K^{w,\lambda}$, i.e., the share of parameters of an outranking model confirming that a is contained in the most preferred subset of options. Let us define such a *kernel membership index KMI*(a) as:

$$KMI(a) = \sum_{a\in A'\subseteq A} KAI(A').$$

  Overall, *KAI*(A') and *KMI*(a) can be interpreted as a support given to the selection of, respectively, A' or a.

5. Case study

Remanufacturing auto parts is one of the most successful businesses in the Indian scenario. We made a web-based review along with the references of Indian industrial (official) data books to identify the focal firms that practice auto parts remanufacturing. Our research revealed that many auto parts remanufacturing units

Table 4
The evaluation of 5 service providers in terms of 14 criteria provided by the DM (↑ and ↓ denote, respectively, the maximizing and minimizing criteria).

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}	g_{13}	g_{14}
	↓	↑	↑	↑	↑	↑	↑	↑	↑	↑	↓	↑	↑	↑
S1	3	8	7	6	6	7	8	7	6	6	4	8	6	8
S2	5	8	6	7	7	6	6	8	7	6	4	8	9	8
S3	2	8	8	8	9	8	7	6	5	5	4	7	7	3
S4	4	7	8	7	6	7	8	6	7	6	3	8	6	6
S5	4	8	8	8	8	7	8	9	8	8	2	9	8	6

Table 5
The order of cards with criteria names and blank cards provided by the DM within the SRF procedure (the higher $l(j)$, the more important criterion g_j).

$l(j)$	1	2	3	4	5	6	7	8	9	10
g_j	g_1, g_4	g_6	g_5	g_{13}, g_{14}	g_3, g_8	g_9	g_{10}, g_{11}	g_{12}	g_2	g_7
e_s		1			1			1		

were active in India, but only eleven companies followed the formal remanufacturing process and, in addition, had foreign customers. The latter increased the chances of third-party reverse logistics. The research proposal was sent via mail to the selected companies along with the preliminary conceptualization of the study. Five companies replied to the inquiry. In this section, we report the results of a case study that was conducted in cooperation with one of these companies.

The case company was run in 1973 in Spain as a manufacturer of replacement auto parts by offering a wide range of products, including steering gears, pumps, air conditioning compressors, and electronic control units. Due to the influence of globalization development, the firm extended its branches in various countries including France, Germany, and India. Within the study, we dealt with a subsidiary started in Chennai (India) in 2013. This plant was limited to the production of steering gears and pumps. Nevertheless, they were aware of the complexity of third-party reverse logistics processes in their parent company. Due to the versatile cultural developments in the Indian context, they were less aware about the selection of third-party logistics service providers. However, we confirmed the company's interest in concentrating on their logistics systems, mainly because their customers were reputed ones in the automotive field, including companies such as Tata, Suzuki, Hyundai, Honda, Toyota, and Ford.

The study focused on three different dimensions of sustainability in the third-party reverse logistics service provider selection, which would be likely to improve the effectiveness of the company's logistics processes. Five potential third-party reverse logistics providers $SUP = \{S1, S2, S3, S4, S5\}$ were identified. These were rated in terms of the following criteria: g_1 (costs, min), g_2 (quality, max), g_3 (RL capacity, max), g_4 (technology, max), g_5 (relationship, max), g_6 (financial performance, max), g_7 (management of risk, max), g_8 (RL practices, max), g_9 (organizational role, max), g_{10} (green level, max), g_{11} (low-carbon, min), g_{12} (environmental management, max), g_{13} (micro-social, max), and g_{14} (macro-social, max). Note that "max" and "min" denote, respectively, gain- and cost-type criteria. Thus, the family of constructed criteria involved economic, environmental, and social aspects. The performances of 5 service providers in terms of 14 criteria are provided in Table 4.

5.1. Preference information

The results of the elicitation process that was conducted with the revised Simos procedure for the case study are presented in Table 5. The DM divided 14 criteria into 10 groups, thus, judging some pairs of criteria as equally important. For example, g_1 and g_4 were grouped together as the least significant criteria. The DM differentiated the intensity of preference between different groups by introducing 0 or 1 blank card between them. Although the most important criteria g_2 (quality) and g_7 (management of risk) represented economic viewpoints, the environmental aspects were of great importance for the case company as confirmed by the high ranks of g_8, g_9, g_{10}, g_{11}, and g_{12}. The ratio between the weights of the most and the least significant was set to $Z = 7$.

The indifference, preference, reinforced preference, and veto thresholds, and reinforcement factors elicited from the DM are provided in Table 6. When it comes to q_j and p_j, these differ from one criterion to another. They have been elicited so that to relate the difference in performance levels on a given criterion with the desired outranking degree. For example, with $q_j = 0$ and $p_j = 0$, the concordance index $c_j(a, b)$ fully agrees that provider a outranks provider b when the performance of a is at least as good as the performance of b and does not agree with the hypothesis about outranking if the performance of a is worse than that of b. On the other hand, in case $q_j = 0.5$ and $p_j = 1.5$, $c_j(a, b)$ partially agrees with the concordance if the performance of a is one level below the performance of b.

When it comes to the effect of reinforced preference, it was admitted only for the five most important criteria (g_7, g_8, g_{10}, g_{11}, g_{12}) for which the variation of performances between the providers was great enough so that a considerable performance difference could imply a very strong preference of one provider over another. The elicited reinforcement factors ranged between 1.2 and 1.5. In this way, a very large performance difference on these criteria was judged meaningful for considering them as more important in the coalition supporting the outranking relation by increasing their weights.

Analogously, the veto thresholds were not specified for the criteria that were judged the least important by the DM nor for the attributes with low differentiation of performances. Consequently, the discordance effect was considered in the context of five criteria (g_9, g_{10}, g_{11}, g_{12}, g_{14}) and elicited by asking the DM to provide the number of levels that would be sufficient for assessing one supplier critically worse than another supplier irrespective of their performances on all remaining criteria. For all discordance criteria, v_j was significantly greater than p_j. Finally, the credibility threshold was allowed to vary in the range [0.5, 0.9]. In this way, the weighted majority of criteria was always required to validate the truth of a crisp outranking relation ($\lambda_* \geq 0.5$), but the support of all criteria was required in none of the considered scenarios ($\lambda^* < 1$).

5.2. Results

The robustness analysis involved a set of criteria weights and credibility thresholds compatible with the preferences of the DM. The stochastic indices were derived from 10,000 uniformly dis-

Table 6
Indifference, preference, reinforced preference, and veto thresholds along with the reinforcement factors elicited from the DM for 14 criteria ("-" means that the respective parameter was not specified).

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}	g_{13}	g_{14}
q_j	0	0	0.5	1	0.5	0	0.5	1	0.5	1	0.5	1	0	0
p_j	1	1	1.5	2	2.5	1	1.5	2	2	2	1.5	1	1	1
rp_j	–	–	–	–	–	–	2	3	–	3	3	2	–	–
ω	–	–	–	–	–	–	1.5	1.2	–	1.3	1.3	1.5	–	–
v_j	–	–	–	–	–	–	–	–	4	6	3.5	5	–	6

Table 7
Outranking, preference, indifference, and incomparability acceptability indices for all pairs of service providers.

	Outranking acceptability index					*Preference acceptability index*				
a	$S1$	$S2$	$S3$	$S4$	$S5$	$S1$	$S2$	$S3$	$S4$	$S5$
$S1$	1.000	1.000	0.787	0.974	0.000	0.000	0.303	0.787	0.653	0.000
$S2$	0.697	1.000	0.726	0.494	0.000	0.000	0.000	0.726	0.317	0.000
$S3$	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$S4$	0.321	0.177	0.587	1.000	0.124	0.000	0.000	0.587	0.000	0.000
$S5$	0.592	0.546	1.000	1.000	1.000	0.592	0.546	1.000	0.876	0.000
Indifference acceptability index						*Incomparability acceptability index*				
a	$S1$	$S2$	$S3$	$S4$	$S5$	$S1$	$S2$	$S3$	$S4$	$S5$
$S1$	1.000	0.697	0.000	0.321	0.000	0.000	0.000	0.213	0.026	0.408
$S2$	0.697	1.000	0.000	0.177	0.000	0.000	0.000	0.274	0.506	0.454
$S3$	0.000	0.000	1.000	0.000	0.000	0.213	0.274	0.000	0.413	0.000
$S4$	0.321	0.177	0.000	1.000	0.124	0.026	0.506	0.413	0.000	0.000
$S5$	0.000	0.000	0.000	0.124	1.000	0.408	0.454	0.000	0.000	0.000

Table 8
Kernel acceptability indices for different subsets of third-party reverse logistics providers.

A'	{$S5$}	{$S1$, $S5$}	{$S1$, $S2$, $S5$}	{$S1$, $S2$, $S4$, $S5$}	All other subsets
KAI	0.468	0.303	0.124	0.105	0.0

Table 9
Kernel membership, comprehensive outranking and non-outranked indices for all service providers.

a	$S1$	$S2$	$S3$	$S4$	$S5$
KMI	0.532	0.229	0.0	0.124	1.0
COI	0.0	0.0	0.0	0.124	0.546
CNOI	0.303	0.0	0.0	0.0	0.876

tributed values of the admissible parameters of an outranking model. In Table 7, we report the respective relation acceptability indices for all pairs of service providers. Let us discuss the most representative results when referring to a pairwise relation perspective.

For three pairs of service providers ($S1$, $S2$), ($S5$, $S3$), and ($S5$, $S4$), the outranking acceptability index is equal to one. Thus, these relations need to be treated with certainty, and three providers $S2$, $S3$, and $S4$ can be deemed as less advantageous as there is some other alternative, which necessarily outranks them. In this same spirit, $S3$ never outranks any other supplier ($OAI(S3, \cdot) = 0$), while $S1$, $S2$, and $S3$ do not outrank $S5$ for any compatible set of parameters.

When analysing *RelAIs* for different pairs of service providers, one can indicate the relations observed for the vast majority of feasible parameters (e.g., $PAI(S5, S4) = 0.876$ or $IAI(S1, S2) = 0.697$ and these which are extremely unlikely, being confirmed by few feasible weights (e.g., $IRAI(S1, S4) = 0.026$). Such direct pairwise comparisons of the providers offer means to suggest benchmarks for guiding the less advantageous suppliers to improve their performance by attaining the levels derived from the better rated 3PRLPs. For other pairs of suppliers, the observed relation varies depending on the admissible parameters used for the comparison (e.g., $PAI(S2, S4) = 0.317$, $IAI(S2, S4) = 0.177$, and $IRAI(S2, S4) = 0.506$).

For all feasible parameters, we ran the ELECTRE I method and identified a respective graph kernel. Note that a kernel can be interpreted as the most advantageous subset of 3PRLPs. In Table 8, we present the kernel acceptability indices. There are only four subsets of service providers for which this index is greater than zero: {$S5$}, {$S1$, $S5$}, {$S1$, $S2$, $S5$}, and {$S1$, $S2$, $S4$, $S5$}. For the greatest number of considered scenarios (46.8%), $S5$ is the unique supplier contained in the kernel and for other 30% of feasible parameter values it is accompanied only by $S1$. In Fig. 2, we present four outranking graphs with different kernels that were obtained for at least one feasible set of parameters. These graphs support comprehension of the conditions under which each subset of suppliers was selected.

In Table 9, we present the kernel membership indices along with the comprehensive outranking and non-outranked indices for all 3PRLPs. The robust conclusions that can be derived from their analysis are as follows:

- $S5$ was included in all kernels ($KMI(S5) = 1$); in fact, for the majority of considered scenarios $S5$ was not outranked by any other service provider ($CNOI(S5, A, S^C) = 0.876$) and it was outranking all remaining suppliers jointly ($COI(S5, A, S) = 0.546$);
- $S3$ was not contained in any kernel ($KMI(S3) = 0$); what is more, $S3$ was never deemed at least as good as any other service provider ($OAI(S3, \cdot) = 0$);
- $S1$ was found in the majority of kernels ($KMI(S3) = 0.532$); in most cases, this followed a scenario in which $S1$ was incomparable with $S5$; moreover, for over 30% of considered parameter values, $S1$ was not outranked by any other supplier ($CNOI(S1, A, S^C) = 0.303$), while proving its superiority over other alternatives;
- for $S2$ and $S4$ the kernel membership indices are positive though significantly lower than for $S5$ and $S1$; on one hand, $S2$ was contained in the kernel only when it became indiffer-

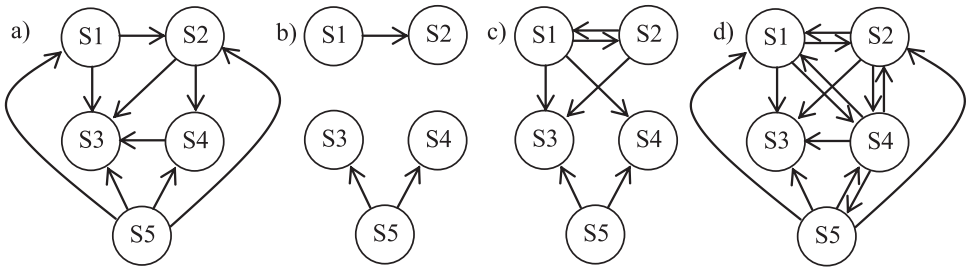


Fig. 2. Outranking graphs with four different kernels: a) {*S*5}, b) {*S*1, *S*5}, c) {*S*1, *S*2, *S*5} (after eliminating a cycle involving *S*1 and *S*2), and d) {*S*1, *S*2, *S*4, *S*5} (after eliminating a cycle involving *S*1, *S*2, *S*4, and *S*5).

ent with *S*1; on the other, for 12.4% of considered settings, *S*4 was assessed to be at least as good as all remaining service providers (being in this way indifferent with *S*5 and, thus, contained in the kernel); in both cases, it was sufficient to increase the credibility threshold to eliminate *S*2 and *S*4 from the respective kernels.

5.3. Recommendation

The robust conclusions led to the following recommendations that were submitted to the case company:

- If a single 3PRLP needed to be selected, *S*5 should be retained to take over the reverse logistics activities;
- in case two suppliers were acceptable (which was considered as the upper limit by the company), *S*1 could be considered jointly with *S*5;
- *S*3 should be eliminated with certainty, whereas *S*2 and *S*4 could be neglected being always outranked by, respectively, *S*1 and *S*5.

The recommendation was followed by the company with selecting *S*5 as the third-party logistics provider, and considering *S*1 as an optional/supplementary choice.

6. Conclusions

A supply chain is only as strong as its weakest part [30]. It is even more true in case products are returned into the supply chain. In order to have a more sustainable supply chain, to fulfil legislation requirements, and to increase customer satisfaction, companies need to investigate reverse logistics. Since most companies do not have the adequate systems or space to meet the requirements for reverse logistics activities, the specialized service providers offer exactly these opportunities. Many firms have learned not to see this option as a threat to their profits, but rather as a marketplace advantage against their competitors based on resources that the 3PRLP can offer. This paper highlighted the increase in attention of this topic by the literature review and discussed the opportunities of outsourcing the logistics activities to 3PRLP.

When companies begin to consider outsourcing their reverse logistics activities, they should be supported by information on a suitable decision making model [43]. This includes details on the viewpoints that are relevant for the evaluation of 3PRLPs and multiple criteria analysis methods for selecting the appropriate service provide while taking into account the preferences of the involved stakeholders.

As far as the criteria relevant for the context of 3PRL are concerned, we provided the background of sustainable triple bottom line theory with focus on economic, environmental, and social aspects. If companies seek to outsource their reverse logistics activities under the consideration of sustainability criteria, there is a potential for gaining a competitive advantage. Based on the literature review, we pointed out that the economic and environmental viewpoints are often included in the sustainable models, but the social aspect is commonly left out and should attract more attention. In this perspective, the classification of criteria proposed in this paper may help the company's management involved in the strategic decisions to implement a framework for evaluating 3PRLPs. In particular, the constructed family of criteria involving three sustainability pillars can be adopted in other case studies. The prioritization of criteria within the study reported in this paper confirmed that there has been a shift towards more acceptance of sustainable criteria in reverse logistics activities. Companies do not only see reverse logistics opportunities to gain a financial advantage, but also include environmental and social aspects in their decision making.

When it comes to the MCDA methods, our extensive literature review showed that different variants of AHP, TOPSIS, and DEA have been so far most widely used in the context of 3PRLPs evaluation and selection. However, these approaches require great cognitive effort from the DMs, need some arbitrary transformation of the performances scales, or offer poor discriminative power. For these reasons, we have proposed a new outranking-based approach that was used to support the 3PRLP selection problem for the Indian manufacturing company.

The introduced method combines a variant of ELECTRE I accounting for the effect of reinforced preference, the revised Simos procedure, and Stochastic Multi-criteria Acceptability Analysis. It indicates the most preferred option by analysing the stability of alternatives' membership in the kernel of an outranking graph obtained for different parameters of an outranking model compatible with the incomplete preference information of the Decision Maker. Let us note that it was the first time both ELECTRE and SMAA were used in the context of 3PRLP, even though the specific areas of past applications of these MCDA methods in logistics and supply chain management already included facility layout and location, supplier selection, inventory decisions, vehicle fleet planning, transportation modes, and supply chain design [38]. Obviously, the approach presented in this paper can be rigorously followed to deal with multiple criteria choice problems also in other application areas such as policy analysis [87], environmental management [61], or energy planning [76].

From the application point of view, the research results in our study could be further validated by collecting data related to the evaluation of 3PRLPs from more respondents. One can also emphasize the role of social criteria such as micro- and macro-social impacts, since they have been comprehensively judged less relevant by the DM than the economic and environmental factors. Moreover, a comparison with other MCDA methods can be established to demonstrate further benefits of using ELECTRE and SMAA in relation to AHP, TOPSIS, and DEA.

From the methodological viewpoint, the proposed approach can be extended to a group decision making framework [47,59]. The

compromise between different DMs may be searched at the input (preference) or at the output (recommendation) levels [26]. Furthermore, the method can be extended to dealing with a hierarchical structure of criteria [23], interactions between criteria [31] as well as to tolerating imprecision in the specification of comparison thresholds [45] or in the judgements provided in the revised Simos procedure [22]. Additionally, the analysis of stochastic acceptability indices can be enriched with selection of a representative robust set of parameters of an outranking model [57] and with consideration of the necessary and possible results [45,61] derived with linear programming. These results indicate the parts of recommendation confirmed by, respectively, all or at least one feasible set of parameters. Finally, if all alternatives need to be ordered from the best to the worst, one can use the net flow rules to compute a unique score based on the relation acceptability indices (see, e.g., [21,22,60]).

Acknowledgements

The authors thank the three anonymous referees for many insightful comments, which helped us to significantly improve the paper. The work of Miłosz Kadziński and Grzegorz Miebs was supported by the Polish Ministry of Science and Higher Education under the Iuventus Plus program in 2016–2019 grant number IP2015 029674 - 0296/IP2/2016/74.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.omega.2018.05.007.

References

[1] Agrawal S, Singh RK, Murtaza Q. A literature review and perspectives in reverse logistics. Resour Conserv Recycl 2015;97:76–92.
[2] Almeida AT. Multicriteria decision model for outsourcing contracts selection based on utility function and ELECTRE method. Comput Oper Res 2007;34(12):3569–74.
[3] Andersson D, Norrman A. Procurement of logistics services – a minutes work or a multi-year project. Eur J Purchasing Supply Manag 2002;8(1):3–14.
[4] Anttonen M, Halme M, Houtbeckers E, Nurkka J. The other side of sustainable innovation: is there a demand for innovative services? J Clean Prod 2013;45:89–103.
[5] Azadi M, Saen RF. A new chance-constrained data envelopment analysis for selecting third-party reverse logistics providers in the existence of dual-role factors. Expert Syst Appl 2011;38(10):12231–6.
[6] Azadi M, Saen RF. Developing an output-oriented super slacks-based measure model with an application to third-party reverse logistics providers. J Multi-Criteria Decis Anal 2011;18(5–6):267–77.
[7] Azadi M, Saen RF. Developing a chance-constrained free disposable hull model for selecting third-party reverse logistics providers. Int J Oper Res Inf Syst 2013;4(4):96–113.
[8] Bai C, Sarkis J. Flexibility in reverse logistics: a framework and evaluation approach. J Clean Prod 2013;47:306–18.
[9] Barker TJ, Zabinsky ZB. A multi-criteria decision making model for reverse logistics using analytical hierarchy process. Omega 2011;39(5):558–73.
[10] Belton V, Stewart T. Multiple criteria decision analysis. An integrated approach. New York: Springer; 2002.
[11] Bolumole YA. The supply chain role of third–party logistics providers. Int J Logist Manag 2001;12(2):87–102.
[12] Boyson S, Corsi T, Dresner M, Rabinovich E. Managing effective third party logistics relationships: what does it take? J Bus Logist 1999;20(1):73–100.
[13] Bradley P.. Contract logistics: it's all about costs. 1994. Purchasing, pp. 56, A3–A14.
[14] Brunelli M. Introduction to the analytic hierarchy process. SpringerBriefs in operations research. Cham: Springer; 2015.
[15] Candler J. You make it, they distribute it. Nations Bus 1994;82(3):46–8.
[16] Charnes A, Cooper W, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res 1978;2(6):429–44.
[17] Cochran J, Ramanujan B. Carrier-mode logistics optimization of inbound supply chains for electronics manufacturing. Int J Prod Econ 2006;103(2):826–40.
[18] Colicchia C, Marchet G, Melacini M, Perotti S. Building environmental sustainability: empirical evidence from logistics service providers. J Clean Prod 2013;59:197–209.
[19] Conlon DE, Murray NM. Customer perceptions of corporate responses to product complaints: the role of explanations. Acad Manag J 1996;39(4):1040–56.
[20] Cook WD, Tone K, Zhu J. Data envelopment analysis: prior to choosing a model. Omega 2014;44:1–4.
[21] Corrente S, Figueira J, Greco S. The SMAA-PROMETHEE method. Eur J Oper Res 2014;239(2):514–22.
[22] Corrente S, Figueira J, Greco S, Słowiński R. A robust ranking method extending ELECTRE III to hierarchy of interacting criteria, imprecise weights and stochastic analysis. Omega 2017;71:1–17.
[23] Corrente S, Greco S, Słowiński R. Multiple criteria hierarchy process with ELECTRE and PROMETHEE. Omega 2013;41(5):820–46.
[24] Corrente S, Greco S, Słowiński R. Multiple criteria hierarchy process for ELECTRE TRI. Eur J Oper Res 2016;252(1):191–203.
[25] Demir I, Orhan M. Reuse of waste bricks in the production line. Build Environ 2003;38:1451–5.
[26] Dias LC, Climaco JN. ELECTRE tri for groups with imprecise information on parameter values. Group Decis Negotiation 2000;9(5):355–77.
[27] Dias LC, Antunes CH, Dantas G, de Castro N, Zamboni L. A multi-criteria approach to sort and rank policies based on delphi qualitative assessments and ELECTRE TRI: the case of smart grids in brazil. Omega 2018;76:100–11.
[28] Efendigil T, Semih O, Elif K. A holistic approach for selecting a third-party reverse logistics provider in the presence of vagueness. Comput Ind Eng 2008;54(2):269–87.
[29] Elkington J. Towards the sustainable corporation: win-win-win business strategies for sustainable development. Calif Manage Rev 1994;36(2):90–100.
[30] Ertay T, Kahveci A, Tabanli RM. An integrated multi-criteria group decision–making approach to efficient supplier selection and clustering using fuzzy preference relations. Int J Computer Integr Manuf 2011;24(12):1152–67.
[31] Figueira J, Greco S, Roy B. ELECTRE Methods with interaction between criteria: an extension of the concordance index. Eur J Oper Res 2009;199(2):478–95.
[32] Figueira J, Greco S, Roy B, Słowiński R. An overview of ELECTRE methods and their recent extensions. J Multi-Criteria Decis Anal 2013;20:61–85.
[33] Figueira J, Mousseau V, Roy B. ELECTRE methods. In: Figueira JR, Greco S, Ehrgott M, editors. Multiple criteria decision analysis: state of the art surveys. New York: Springer; 2016. p. 155–85.
[34] Figueira J, Roy B. Determining the weights of criteria in the ELECTRE type methods with a revised simos' procedure. Eur J Oper Res 2002;139(2):317–26.
[35] Gonzalez-Torre P, Alvarez M, Sarkis J, Adenso-Diaz B. Barriers to the implementation of environmentally oriented reverse logistics: evidence from the automotive industry sector british. J Manag 2010;21(4):889–904.
[36] Govindan K. Fuzzy approach for the selection of third party reverse logistics provider. Asia Pacific J Market Logist 2009;21(3):397–416.
[37] Govindan K, Haq A. Analysis of interactions of criteria and sub-criteria for the selection of supplier in the built-in-order supply chain environment. Int J Prod Res 2007;45(11):3831–52.
[38] Govindan K, Jepsen M. ELECTRE: A comprehensive literature review on methodologies and applications. Eur J Oper Res 2016;250(1):1–29.
[39] Govindan K, Kadziński M, Sivakumar R. Application of a novel PROMETHEE-based method for construction of a group compromise ranking to prioritization of green suppliers in food supply chain. Omega 2017;71:129–45.
[40] Govindan K, Murugesan P. Selection of third-party reverse logistics provider using fuzzy extent analysis. Benchmarking Int J 2011;18(1):149–67.
[41] Govindan K, Palaniappan M, Zhu Q, Kannan D. Analysis of third party reverse logistics provider using interpretive structural modeling. Int J Prod Econ 2012;140(1):204–11.
[42] Govindan K, Pokharel S, Kumar P. A hybrid approach using ISM and fuzzy TOPSIS for the selection of reverse logistics provider. Resour Conserv Recycl 2009;54(1):28–36.
[43] Govindan K, Roohollah K, Ahmad J. A fuzzy multi-criteria approach for measuring sustainability performance of a supplier based on triple bottom line approach. J Clean Prod 2013;47:345–54.
[44] Govindan K, Soleimani H, Kannan D. Reverse logistics and closed loop - supply chain: a comprehensive review to explore the future. Eur J Oper Res 2015;240(3):603–26.
[45] Greco S, Kadziński M, Mousseau V, Słowiński R. ELECTRE^{GKMS}: Robust ordinal regression for outranking methods. Eur J Oper Res 2011;214(1):118–35.
[46] Guarnieri P, Sobreiro VA, Nagano MS, Serrano ALM. The challenge of selecting and evaluating third-party reverse logistics providers in a multicriteria perspective: a brazilian case. J Clean Prod 2015;96:209–19.
[47] Hatami-Marbini A, Tavana M. An extension of the electre i method for group decision-making under a fuzzy environment. Omega 2011;39:373–86.
[48] Hong J, Chin A, Liu B. Logistics outsourcing by manufacturers in china: a survey of the industry. Transp J 2004;43(1):17–25.
[49] Hu Z, Rao C, Zheng Y, Huang D. Optimization decision of supplier selection in green procurement under the mode of low carbon economy. Int J Comput Intell Syst 2015;8(3):407–21.
[50] Hung Lau K, Wang Y. Reverse logistics in the electronic industry of china: a case study. Supply Chain Manag 2009;14(6):447–65.
[51] Hwang CL, Yoon K. Multiple attribute decision making. Methods and applications. New York: Springer; 1981.
[52] Jayant A, Gupta P, Garg SK, Khan M. TOPSIS-AHP Based approach for selection of reverse logistics service provider: a case study of mobile phone industry. Procedia Eng 2014;97:2147–56.
[53] Jayaraman V, Guide VDR, Srivastava R. A closed-loop logistics model for remanufacturing. J Oper Res Soc 1999;50(5):497–508.

[54] Jharkharia S, Shankar R. Selection of logistics service provider: an analytic network process (ANP) approach. Omega 2007;35(3):274–89.
[55] Ji G. Market-motivated value systems, reverse logistics and the evaluation model for the third party reverse logistics providers. Int Logist Trade 2006;4(1):53–92.
[56] Jing AL, Wu Y, Kin KL, Liu K. Optimal ordering policy in a distribution system. Int J Prod Econ 2006;103:527–34.
[57] Kadziński M, Greco S, Słowiński R. Selection of a representative set of parameters for robust ordinal regression outranking methods. Comput Oper Res 2012;39(11):2500–19.
[58] Kadziński M, Labijak A, Napieraj M. Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of polish airports. Omega 2017;67:1–18.
[59] Kadziński M, Rocchi L, Miebs G, Grohmann D, Menconi M, Paolotti L. Multiple criteria assessment of insulating materials with a group decision framework incorporating outranking preference model and characteristic class profiles. Group Decis Negotiation 2018;27(1):33–59.
[60] Kadziński M, Słowiński R, Greco S. Multiple criteria ranking and choice with all compatible minimal cover sets of decision rules. Knowl Based Syst 2015;89:569–83.
[61] Kadziński M, Tervonen T, Figueira J. Robust multi-criteria sorting with the outranking preference model and characteristic profiles. Omega 2015;55:126–40.
[62] Kafa N, Hani Y, Mhamedi AE, B Grabot B, Vallespir S, Gomes A, Bouras, Kiritsis D. A fuzzy multi criteria approach for evaluating sustainability performance of thirdæ party reverse logistics providers. in: advances in production management systems. In: Innovative and knowledge-based production management in a global-local world.. Heidelberg: Springer Berlin; 2014. p. 270–7.
[63] Kannan G. Fuzzy approach for the selection of third party reverse logistics provider. Asia Pacific J Market Logist 2009;21(3):397–416.
[64] Kannan G, Murugesan P, Senthil P, Haq AN. Multicriteria group decision making for the third party reverse logistics service provider in the supply chain model using fuzzy TOPSIS for transportation services. Int J Serv Technol Manag 2009;11(2):162–81.
[65] Kannan G, Pokharel S, Kumar PS. A hybrid approach using ISM and fuzzy TOPSIS for the selection of reverse logistics provider. Resour Conserv Recycl 2009;54(1):28–36.
[66] Kannan D. Role of multiple stakeholders and the critical success factor theory for the sustainable supplier selection process. Int J Prod Econ 2018;195:391–418.
[67] Krumwiede DW, Sheu C. A model for reverse logistics entry by third-party providers. Omega 2002;30(5):325–33.
[68] Kumar NR, Kumar RMS. Closed loop supply chain management and reverse logistics – a literature review. Int J Eng Res Technol 2013;6(4):455–68.
[69] Kuo RJ, Wang YC, Tien FC. Integration of artificial neural network and MADA methods for green supplier selection. J Clean Prod 2010;18(12):1161–70.
[70] Lahdelma R, Salminen P. SMAA-2: Stochastic multicriteria acceptability analysis for group decision making. Oper Res 2001;49(3):444–54.
[71] Lalonde BJ, Maltz AD. Some proposition about outsourcing the logistics function. Int J Logist Manag 1992;3(1):1–11.
[72] Lambert S, Riopel D, Abdul-Kader W. A reverse logistics decisions conceptual framework. Comput Ind Eng 2011;61(3):561–81.
[73] Lao SI, Choy KL, Ho GTS, Tsim YC, Chung NSH. Determination of the success factors in supply chain networks: a Hong Kong-based manufacturer's perspective. Meas Bus Excellence 2011;15(1):34–48.
[74] Lee YJ, Baker T, Jayaraman V. Redesigning an integrated forward reverse logistics system for a third party service provider: an empirical study. Int J Prod Res 2012;50(19):5615–34.
[75] Mahmoudzadeh M, Mansour S, Karimi B. To develop a third-party reverse logistics network for end-of-life vehicles in iran. Resour Conserv Recycl 2013;78:1–14.
[76] Marinakis V, Doukas H, Xidonas P, Zopounidis C. Multicriteria decision support in local energy planning: an evaluation of alternative scenarios for the sustainable energy action plan. Omega 2017;69:1–16.
[77] McCarthy I, Anagnostou A. The impact of outsourcing on the transaction costs and boundaries of manufacturing. Int J Prod Econ 2004;88(1):61–71.
[78] McCarthy I, Silvestre BS, Kietzmann JH. Understanding outsourcing contexts through information asymmetry and capability fit. Prod Plan Contr 2013;24(4–5):277–83.
[79] Meade L, Sarkis J. A conceptual model for selecting and evaluating third-party reverse logistics providers. Supply Chain Manag 2002;7(5):283–95.
[80] Min H, Ko HJ. The dynamic design of a reverse logistics network from the perspectives of third party logistics service providers. Int J Prod Econ 2008;113(1):176–92.
[81] Momeni E, Azadi M, Saen RF. Measuring the efficiency of third party reverse logistics provider in supply chain by multi objective additive network DEA model. Int J Shipping Transp Logist 2015;7(1):21–41.
[82] Monczka RM, Ternt RJ, Callahn TJ. Supply based strategies to maximize supplier performance. Int J Phys Distrib Logist 1994;23(4):42–54.
[83] Mothilal S, Gunasekaran A, Nachiappan SP, Jayaram J. Key success factors and their performance implications in the Indian third-party logistics (3PL) industry. Int J Prod Res 2012;50:2407–22.
[84] Mousseau V, Dias L. Valued outranking relations in ELECTRE providing manageable disaggregation procedures. Eur J Oper Res 2004;156:467–82.
[85] Nikolaou I, Evangelinos K, Allan S. A reverse logistics social responsibility evaluation framework based on the triple bottom line approach. J Clean Prod 2013;56:173–84.
[86] Noci G. Designing green vendor rating systems for the assessment of a supplier's environmental performance. Eur J Purchasing Supply Manag 1997;3(2):103–14.
[87] Pinto FS, Costa AS, Figueira J, Marques RC. The quality of service: an overall performance assessment for water utilities. Omega 2017;69:115–25.
[88] Pokharel S, Mutha A. Perspectives in reverse logistics: a review. Resour Conserv Recycl 2009;53(4):175–82.
[89] Prahinski C, Kocabasoglu C. Empirical research opportunities in reverse supply chains. Omega 2006;34(6):519–32.
[90] Prakash C, Barua MK. An analysis of integrated robust hybrid model for third–party reverse logistics partner selection under fuzzy environment. Resour Conserv Recycl 2016;108:63–81.
[91] Ramirez AM, Morales VJG. Improving organisational performance through reverse logistics. J Oper Res Soc 2014;65(6):954–62.
[92] Ravi V. Selection of third-party reverse logistics providers for end-of-life computers using TOPSIS-AHP based approach. Int J Logist Syst Manag 2011;11(1):24–37.
[93] Ravi V, Shankar R. Analysis of interactions among the barriers of reverse logistics. Technol Forecast Soc Change 2005;72(8):1011–29.
[94] Razzaque MA, Sheng CC. Outsourcing of logistics functions: a literature survey. Int J Phys Distrib Logist Manag 1998;28(2):89–107.
[95] Rockart JF. Chief executives define their own data needs. Harv Bus Rev 1979;57(2):81–93.
[96] Rogers DS, Tibben-Lembke RS. Going backwards: reverse logistics trends and practices. Centre for logistics management, University of Nevada, Reno, Reverse logistics executive council 1998.
[97] Rogers DS, Tibben-Lembke RS. An examination of reverse logistics practices. J Bus Logist 2001;22(2):129–48.
[98] Roy B. Classement et choix en presence de points de vue multiples (la methode ELECTRE). La Revue d'Informatique et de Recherche Operationelle (RAIRO) 1968;8:57–75.
[99] Roy B. The outranking approach and the foundations of ELECTRE methods. In: Bana e Costa C, editor. Readings in multiple criteria decision aid. Berlin: Springer-Verlag; 1990. p. 155–83.
[100] Roy B. Robustness in operational research and decision aiding: a multi–faceted issue. Eur J Oper Res 2010;200(3):629–38.
[101] Roy B, Bouyssou D. Aide multicritere a la decision: methodes et cas. Paris: Economica; 1993.
[102] Roy B, Figueira J, Almeida-Dias J. Discriminating thresholds as a tool to cope with imperfect knowledge in multiple criteria decision aiding: theoretical results and practical issues. Omega 2014;43:9–20.
[103] Roy B, Słowiński R. Handling effects of reinforced preference and counter-veto in credibility of outranking. Eur J Oper Res 2008;188:185–90.
[104] Saaty T. The analytic hierarchy process. New York: McGraw-Hill; 1980.
[105] Saen RF. A new model for selecting third-party reverse logistics providers in the presence of multiple dual-role factors. Int J Adv Manuf Technol 2010;46(1–4):405–10.
[106] Saen RF. A decision model for selection third-party reverse logistics providers in the presence of both dual-role factors and imprecise data. Asia-Pacific J Oper Res 2011;28(2):239–54.
[107] Saen RF. A mathematical model for selecting third-party reverse logistics providers. Int J Procurement Manag 2009;2(2):180–90.
[108] Sasikumar P, Kannan G, Haq AN. A heuristic based approach to vehicle routing model for third party reverse logistics provider. Int J Serv Technol Manag 2009;12(1):106–25.
[109] Senthil S, Srirangacharyulu B, Ramesh A. A robust hybrid multi-criteria decision making methodology for contractor evaluation and selection in third–party reverse logistics. Expert Syst Appl 2014;41(1):50–8.
[110] Serrato MA, Gaytan J, Ryan SM. A Markov decision model to evaluate outsourcing in reverse logistics. Int J Prod Res 2007;45(18–19):4289–315.
[111] Shih HS, Shyur HJ, Lee ES. An extension of TOPSIS for group decision making. Math Comput Model 2007;45(78):801–13.
[112] Siskos E, Tsotsolas N. Elicitation of criteria importance weights through the simos method: a robustness concern. Eur J Oper Res 2015;246(2):543–53.
[113] Steven M, Dyckhoff H, Lackes R, Reese J. Networks in reverse logistics. In: Supply chain management and reverse logistics. Berlin Heidelberg: Springer; 2004. p. 163–80.
[114] Stock GN, Greis NP, Kasarda JD. Logistics, strategy and structure: a conceptual framework. Int J Phys Distrib Logist Manag 1999;29:224–39.
[115] Suyabatmaz AC, Altekin FT, Sahin G. Hybrid simulation-analytical modeling approaches for the reverse logistics network design of a third-party logistics provider. Comput Ind Eng 2014;70:74–89.
[116] Talib MSA, Hamid ABA. Application of critical success factors in supply chain management. Supply Chain Manag 2014;3(1):21–33.
[117] Tate K. The elements of a successful logistics partnership. Int J Phys Distrib Logist Manag 1996;26(3):7–13.
[118] Tavana M, Zareinejad M, Santos-Arteaga FJ, Kaviani MA. A conceptual analytic network model for evaluating and selecting third-party reverse logistics providers. Int J Adv Manuf Technol 2016;86(5–8):1705–21.
[119] Tervonen T, Figueira J. A survey on stochastic multicriteria acceptability analysis methods. J Multi-Criteria Decis Anal 2008;15(1–2):1–14.
[120] Tervonen T, Lahdelma R. Implementing stochastic multicriteria acceptability analysis. Eur J Oper Res 2007;178(2):500–13.

[121] Tervonen T, van Valkenhoef G, Basturk N, Postmus D. Hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis. Eur J Oper Res 2013;224(3):552–9.
[122] Turrisi M, Bruccoleri M, Cannelle S. Impact of reverse logistics on supply chain performance. Int J Phys Distrib Logist Manag 2013;43(7):564–85.
[123] Vachon S, Klassen RD. Green project partnership in the supply chain: the case of the package printing industry. J Clean Prod 2006;14(67):661–71.
[124] Wallenburg C.M.. Kundenbindung in der logistik-eine empirische untersuchung zu ihren einflussfaktoren. 2004. Bern, Stuttgart, Wien.
[125] Wang J, Zhu Y. Research on third-party reverse logistics provider selection based on fuzzy clustering in perspective of low-carbon economy. Commun Inf Sci Manag Eng 2012;2(2):63–6.
[126] Wentworth F. Outsourcing services: the case against. Logist Transp Focus 2003;5(2):57–9.
[127] Vincke P. Robust solutions and methods in decision-aid. J Multi-Criteria Decis Anal 1999;8(3):181–7.
[128] Yan NN, Sun BW. Optimal Stackelberg strategies for closed-loop supply chain with third-party reverse logistics. Asia-Pacific J Oper Res 2012;29(05):125–6.
[129] Yin ZH, Lu Q. The grey comprehensive evaluation model of a third-party reverse logistics provider selection. In: ieee international joint conference on artificial intelligence 2009 (JCAI'09); 2009. p. 497–500.
[130] Zareinejad M, Javanmard H. Evaluation and selection of a third-party reverse logistics provider using ANP and IFG-MCDM methodology. Life Sci J 2013;10(6):350–5.
[131] Zielniewicz P. A ranking method based on the aggregate distance measure function in the value space. Int J Inf Technol Decision Making 2017;16(3):685–710.

Publication [P4]

Miebs, G., Mochol-Grzelak, M., Karaszewski, A., and Bachorz, R. A. (2020). Efficient Strategies of Static Features Incorporation into the Recurrent Neural Network. In Neural Processing Letters (Vol. 51, Issue 3, pp. 2301–2316). Springer Science and Business Media LLC., DOI: 10.1007/s11063-020-10195-x.

Number of citations[4]:

- according to Web of Science: 7
- according to Google Scholar: 8

Contribution of co-authors (excluding the author and the supervisor of this dissertation):

- Małgorzata Mochol-Grzelak:
  - Co-authorship of the idea underlying the paper
  - Consultation of the algorithms incorporating static metadata into a recurrent neural network
  - Corrections to the first draft of the manuscript
- Adam Karaszewski:
  - Consultation of the context of case study and interpretation of its results
  - Corrections to the first draft of the manuscript
- Rafał Bachorz:
  - Co-authorship of the idea underlying the paper
  - Consultation of the algorithms incorporating static metadata into a recurrent neural network
  - Consultation of the context of case study and interpretation of its results
  - Co-authorship of the text of the publication

[4] as on September 25, 2023

Neural Processing Letters
https://doi.org/10.1007/s11063-020-10195-x

Efficient Strategies of Static Features Incorporation into the Recurrent Neural Network

Grzegorz Miebs[1] · Małgorzata Mochol-Grzelak[1] · Adam Karaszewski[1] · Rafał A. Bachorz[1]

Abstract
Recurrent neural networks (RNNs) have evolved to become one of the most powerful tools for making predictions on sequenced data, such as time series, textual data, signals, music etc. In many real-life cases, however, sequenced data are additionally characterized by static features which, due to their non-sequential nature, cannot be transferred directly into RNNs. In this paper, we discuss a method which incorporates static features into RNNs in order to influence and generalize the learning process. Furthermore, we will demonstrate that our approach significantly enhances the performance of RNNs, enabling the networks to learn the sequenced data exhibiting varying characteristics and then distinguish between them through the use of static supplementary information. Finally, we will evaluate our model against real energy consumption measurements of energy time series and verify that high-accuracy demand forecasts for different types of customers can be achieved only by way of incorporation of static features.

Keywords Recurrent neural network · Deep learning · Time series · Energy consumption · Static features

1 Introduction

Time series are ordered sets of values recorded at sequenced time intervals. Their internal structure allows to investigate data cycles and trends, or even predict future values against historical entries. Many machine learning (ML) algorithms were developed to date, aiming to improve the accuracy of forecasts and business decisions in general, such as random forests in Breiman [2], or extreme gradient boosting and very promising RNNs in Hochreiter and Schmidhuber [14]. RNNs are dedicated to series data because of their capability of learning long-term dependencies from sequences. This ability makes them suitable for application

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11063-020-10195-x) contains supplementary material, which is available to authorized users.

Rafał A. Bachorz
rbachorz@psi.pl

[1] Advanced Analytics Team, PSI Poland, Towarowa 35, 61-896 Poznań, Poland

Published online: 31 January 2020

in various real-life tasks, such as handwritting recognition [12], speech recognition [13], machine translation [31], robot control [3–5,21], music composition [7], and many other applications; cf. also [26] for an overview. However, real-life time series are usually related to additional static data, which may enhance the forecast and improve the performance of RNNs, yet such auxiliary features are not easily processed by RNNs due to their non-sequential structure. The approach presented in this paper can be considered as a qualitatively new methodology capable of applying the same predictive model to multiple time series, even with quite different characteristics. Simultaneous incorporation of static and dynamic features is a prerequisite for extracting the relations between them in the training phase. In the prediction phase, however, the provided static features guide the forward-pass calculation towards the desired dynamic component of the data. This naturally leads to potential generative use cases. For instance, the static features set can be estimated for a qualitatively new entity, which was not evident in the training phase. Incorporation of these features into an already trained predictive model will stimulate the neural network to generate dynamic sequences with properties predefined by the static component. We believe that these aspects should be considered the primary contribution of our research.

Reliable predictions of energy consumption, discussed below in more detail, are an example of a real-life use case. Depending on the customer size, customer type (businesses/individuals), working hours etc., the trends over the years can differ significantly. This causes serious issues for machine learning algorithms since intrinsic dependencies between the static and dynamic features are essential to achieving a well-performing and generic predictive model. One of the possible solutions would be to group the customers with similar energy consumption characteristics and to train an independent ML model for each cluster. Despite the overall simplicity and straightforwardness of this approach, it also comes with major weaknesses, such as the need to maintain many ML models and to define new customer groups manually, inter alia. Therefore, we propose a more sophisticated solution, developed based on the engineering of artificial neural network (ANN) involving both the recurrent and fully-connected elements. While the former is dedicated to sequenced data, the latter is used for acquiring the associated static information. With a well-thought-out merger of those two areas, the resulting architecture is capable of simultaneous learning of both the static and dynamic data. The unquestionable advantage of the proposed approach is the ability of making predictions for all previously learned customer types using a single ML model. Furthermore, its generic application, which involves any type of sequenced data (such as time series, textual data, etc.), is also noteworthy. The only obvious issue which needs to be addressed would be an intricate feature engineering process for both the static and dynamic feature spaces.

1.1 Structure of the Paper

The paper is organized as follows. In Sect. 2, the current state of the art is discussed. Section 3 presents the complete picture of the methodology used for incorporating static features into RNNs, in particular the neural network architecture and final prediction method. In Sect. 4, all obtained results are discussed, starting from the simplest possible demonstration of the proposed method and gradually moving towards more complex examples, up to the real-life use case of energy consumption. Section 5 summarizes all results and shows possible extensions of the presented approach.

2 Related Works

The RNN concept was developed more than 30 years ago by Rumelhart et al. [25]. Since then, many extensions and improvements were introduced thereto in order to enhance the accuracy of predictive models applying to sequenced data. The RNN architecture enables the encoding of past information relating to the hidden states of neural networks by combining the current input x_t with the hidden state of the previous time step h_{t-1}, i.e.:

$$h_t = \sigma_h(W_h x_t + U_h h_{t-1} + b_h), \tag{1}$$

where W_h, U_h and b_h denote parameter matrices and the vector, respectively, and function σ_h stands for the nonlinear activation function. Formally, the RNN's prediction output vector can be expressed as:

$$y_t = \sigma_y(W_y h_t + b_y), \tag{2}$$

where W_y and b_y again denote the matrix and vector of the model's parameters with the activation function σ_y. However, as pointed out by Bengio et al. [1], it is difficult to learn long-term dependencies for standard RNNs due to the vanishing or exploding gradient problem. To prevent the occurrence of such problems, a gating mechanism was introduced by Hochreiter and Schmidhuber [14] in the form of long short-term memory (LSTM) units. A single LSTM unit contains a memory part (cell) and gates which regulate the information flow inside the unit: an input gate, an output gate and a forget gate. LSTM components can be expressed by the following equations defined by Gers et al. [10]:

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i), \tag{3}$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o), \tag{4}$$
$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f), \tag{5}$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c), \tag{6}$$
$$h_t = o_t \circ \sigma_h(c_t), \tag{7}$$

where the operator $\circ$ denotes an element-wise product and i, o and f are input, output and forget gates, respectively. The activation function σ_g is a sigmoid function, while σ_c and σ_h are hyperbolic tangent functions. The LSTM unit has two output vectors: (a) c_t which is a cell state vector in time step t and (b) h_t which is a hidden state vector. For time $t = 0$, the initial state at time $t = -1$ needs to be defined. In the default implementation of the LSTM unit, initial states are set as zero-value vectors, i.e.:

$$c_{-1} = (0, 0, \ldots, 0), \tag{8}$$
$$h_{-1} = (0, 0, \ldots, 0). \tag{9}$$

Incorporation of static features into the RNN with the aim of improving the quality of sequenced data classification was considered by a number of authors. The very first attempt was made by Esteban et al. [8] in order to predict clinical events of a specific type, combining the static data of patients (blood type, age, etc.) with the dynamic history of their visits, prescribed medicines and laboratory tests. The main purpose was to predict the probability of occurrence of kidney transplantation endpoints for each patient. A similar approach was presented by Yang et al. [32] also in the context of clinical decisions for hierarchical classification. Both authors concatenated hidden vectors coming from the RNN and the latent representation of static features. We adopted the same methodology for laboratory models, i.e. simple sine functions and a prototype of real-life use, which formed the foundations for

further research. However, we found that this approach might be insufficient in the case of more complex problems. In particular, the use case of real-life energy consumption needed significant methodology extensions, which led us to the development of a more sophisticated, novel architecture, described in more detail in Sect. 3 and compared against other methods in Table 2.

A completely different approach was taken by Leontjeva and Kuzovkin [18], where the authors used a sophisticated feature extraction methodology to preprocess static and dynamic features prior to the training of the final model, instead of modifying the RNN's learning process. Although this idea indeed improves the performance of the specific models evaluated by those authors, we did not find it suitable for the data considered in the present study, i.e. the energy consumption time series.

We wish to emphasize that all aforementioned papers concern the classification of sequenced data and, to the best of our knowledge, no regression problems have been properly addressed to date. Moreover, we found no literature discussing the energy-demand time series in the context of using both static and dynamic features in forecasting.

Data-driven methods used in time series forecasting rely on the previously collected historical data after the consumption or generation of energy. ML approaches allow for the efficient extraction of information from existing data sets and are capable of finding the statistical non-linear relations between relevant predictors without any human supervision. The neural network can be viewed as one possible application of machine learning. The advantages and disadvantages associated with the application of ANN-based methods in various areas are well known. One of them is an issue of overfitting and the method potential for generalization. A representative example would be a study of various hybrid forecasting models of micro-grids by Liu et al. [20], where the accuracy of forecasts varies substantially for objects with different capabilities and load characteristics. Other studies related to the application of ANNs to energy demand time series forecasting with an emphasis on short-term load forecasting (STFL) can be found in the research conducted by Zheng et al. [33]; Khosravani et al. [16]; Kuo and Huang [17]. Apart from ANNs, also other Machine Learning or, more generally, Computational Intelligence approaches were extensively used in this domain. In particular, support vector machines in Kaytez et al. [15]; Selakov et al. [27]; Niu and Dai [22]; Liang et al. [19], random forests in Dudek [6], fuzzy logistic methods in Suganthi et al. [28], and Bayesian neural networks in Niu et al. [23] are examples of the successful application of these methodologies in the area of the energy-demand prediction. It should be noted that the majority of available research papers are focused on short-term energy load forecasting, while the problem of medium- and long-term load forecasting seems to be generally neglected and is mentioned only in several studies, e.g. in Ringwood et al. [24]; Feilat and Bouzguenda [9]; González-Romera et al. [11]. It is also worth mentioning that the presented method is independent of the forecasting horizon. The technical details associated with the prediction phase are discussed in paragraph 3.3 and the forecast errors obtained against time are shown in paragraph 4.3.

3 Methods

3.1 Incorporation of Static Features into RNNs

It was found that the incorporation of static information into RNNs requires an in-depth investigation, and any simplistic attempt such as duplication of static features in each

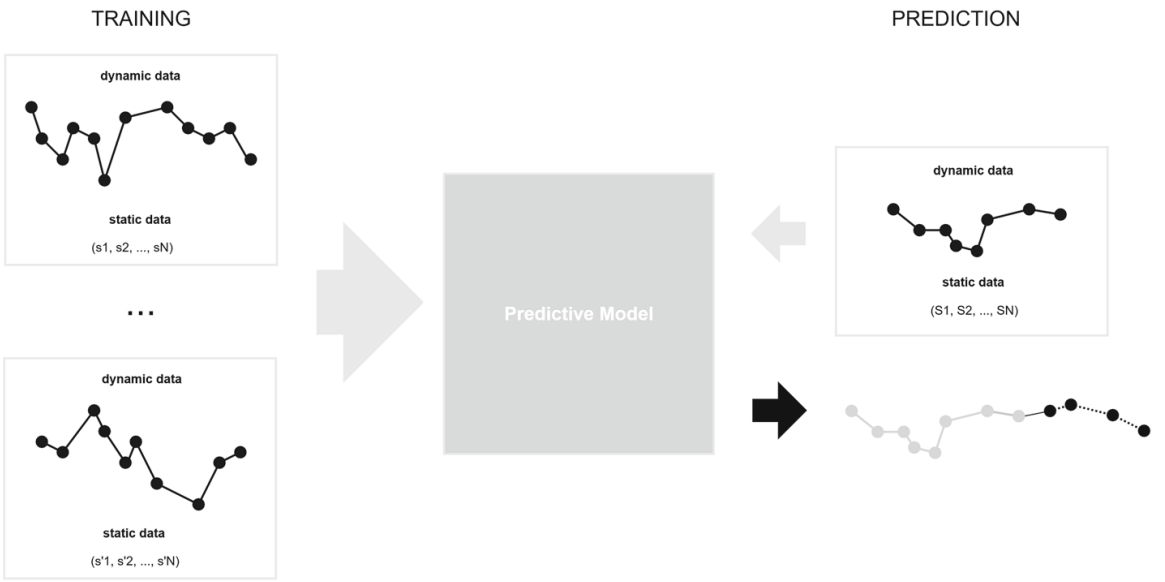


Fig. 1 Diagrammatic representation of the applied methodology. In both phases, i.e. the training phase and the prediction phase, both the dynamic and the static features are introduced into the predictive model. In the prediction phase (right-hand side of the picture), the initial fragment of the time series is supplemented with the expected set of static features, which allows to generate a dynamic feature forecast. As a result, the forecasting process can be influenced in such a way that the obtained dynamic features correspond to their static equivalents

sequence step is not an optimal choice for the RNN architecture. Such an approach leads to a significant increase in the number of RNN parameters and a long feeding time, i.e. the time needed for the RNN to transform data, since duplicated static features are evaluated each time for every sequence. As a consequence, the training of such a network is both memory- and time-consuming. In order to overcome this obstacle, we propose an alternative approach, where the key idea is to inject additional data into the RNN and make its training process dependent on these supplementary features. Nevertheless, static data should be first preprocessed in stage one, similarly to any other application of the ML methodology. The latent representation of static features is then obtained by using the conventional fully-connected layer, and subsequently the hidden state of the RNN is initialized by this quantity.

The general workflow proposed under the present study is presented in Fig. 1. We have distinguished two key use cases, namely the training case and the prediction case. For each of those processes, data is provided in the form of both dynamic and static features. The design of the static component should take into account the properties of the dynamic part, which will be useful during the prediction phase, such as statistical descriptors, aggregates, etc. The training process should be carried out on a representative sample of data, which is required to obtain the predictive model of the desired predictive strength. In the prediction phase, data must be provided in exactly the same form as during the training. The dynamic component, understood as a seed of dynamic features, is accompanied by a complementary static part which plays a role of a supplementary source of information, directing the prediction process towards the expected forecast. The description presenting technical details of the prediction phase of the energy demand forecast can be found in Sect. 3.3.

3.2 Proposed RNN Architecture

Our research focuses on three approaches: (1) concatenation of the resulting RNN hidden state with the latent representation of static data coming from feed-forward neural network (FNN), (2) initialization of the hidden and cell states of the LSTM layer with the latent representation of static features, and (3) initial injection of the hidden and cell states coming from the latent representation of static data into the first recurrent layer, and additional concatenation of the RNN hidden state and the latent representation, as described in point (1). For the corresponding graphs, see Fig. 2. It is worth noting that the architecture presented in Fig. 2c is a combination of the approaches shown in the form of diagrams presented in Fig. 2a and b, i.e. it involves both the initialization of LSTM states (hidden and cell states), as in diagram 2b, as well as a further exposure of the RNN model on the static features, as in Fig. 2a. In standard LSTM implementations, both the cell and hidden states are initiated as zero-valued vectors at time $t = -1$, see Eqs. 8 and 9. However, under novel approaches presented in Fig. 2b and c, these initial states carry the information extracted from static features, i.e.:

$$c_{-1} = \mathrm{FNN}_c(\mathbf{s}), \tag{10}$$

$$h_{-1} = \mathrm{FNN}_h(\mathbf{s}), \tag{11}$$

where $\mathrm{FNN}_{c,h}$ denotes the latent representation coming from FNN layers and $\mathbf{s} = (s_1, s_2, \ldots, s_N)$ is a vector of N static features. The main rationale behind this approach is to incorporate static features into the RNN through relevant initialization of the neuron states which are significantly involved in the overall learning process workflow. One possible option is to initiate the neuron state vectors, e.g. the hidden states and the cell states in the case of LSTM units, by using the quantity carrying the information about static features. The natural choice is to transform the static features vector with the fully-connected sub-architecture (see Fig. 2). Such an approach is naturally embedded in the entire neural network. The weights introduced by the fully-connected component are optimized together with all remaining degrees of freedom; therefore, the associated loss function component is minimized. Moreover, this approach is a very convenient technical solution for the management of the dimensionality of tensor quantities involved in the learning process. To a certain extent, this procedure is inspired by the research into image caption generation, where features extracted from an image are injected into LSTM layers; cf. [30] or [29] for more details. The choice of a specific approach is dictated by the nature of a problem under consideration—namely, the more complex the case, the more sophisticated the neural network architecture. This will become more apparent in the examples given in Sect. 4. Our experience shows that static feature information can be lost during the learning process. This can be considered a side effect of the data transformation process within the RNN layers. In particular, in more complicated cases requiring deeper architectures, i.e. involving many recurrent layers, it might be necessary to introduce static information more than once, since the subsequent introduction points provide the neural network with refreshed information on those static features.

3.3 Prediction Phase

The nature of the prediction problem discussed here differs in comparison with other issues related to the incorporation of static features. We consider a regression problem instead of classification, with a strong emphasis on the energy-demand time series. Moreover, under the

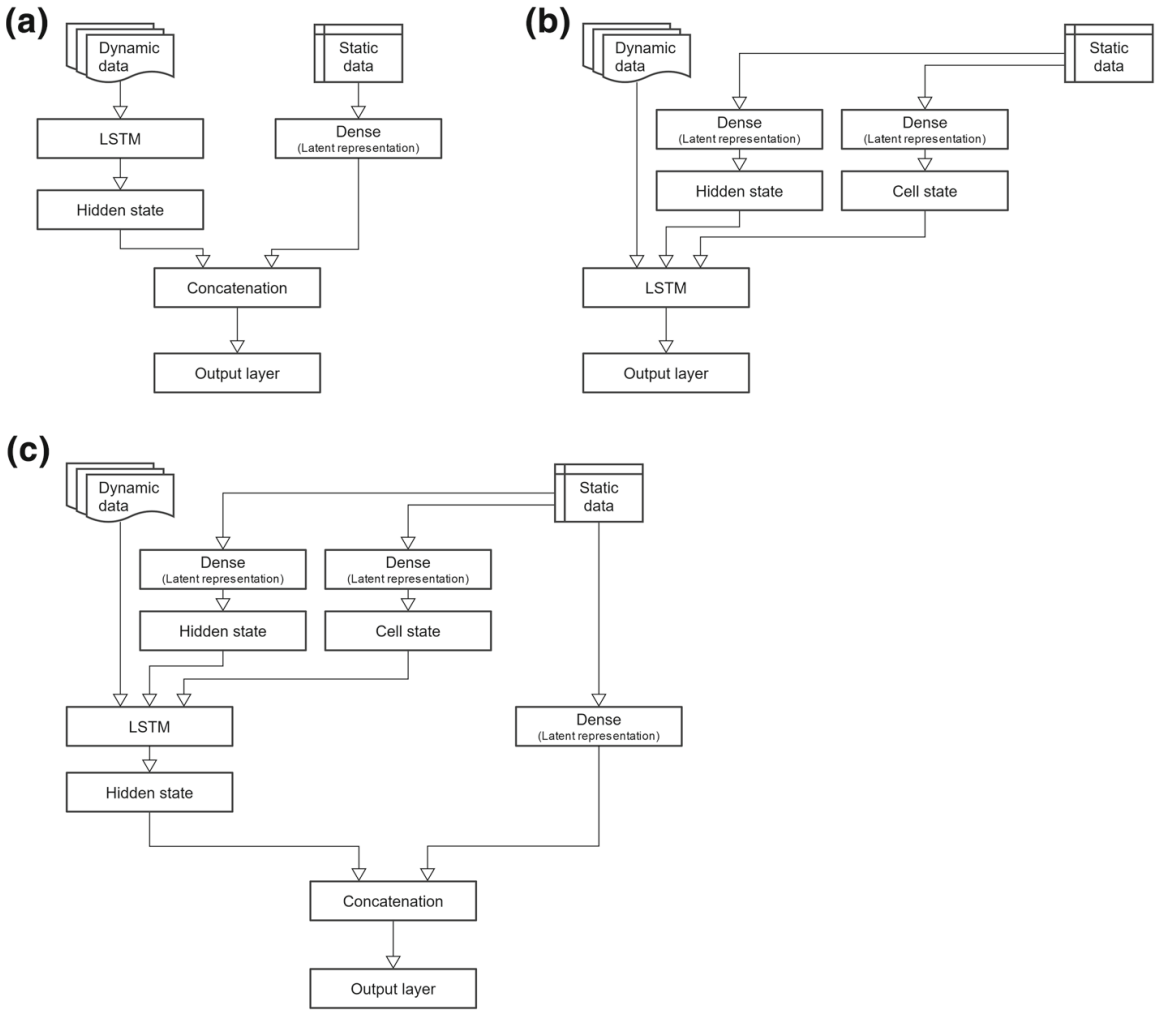


Fig. 2 Schematic diagrams of neural network architectures implemented for the model and the real-life use cases. **a** Diagram of neural network which incorporates static features by concatenation of the RNN hidden state and the latent representation of static data. This architecture was used for simple model cases. **b** Diagram of neural network which incorporates static features as an initial state of the hidden and cell states of the LSTM layer treated as the latent representation. **c** Architecture of NN applied to real-life use case. Static data are injected twice in this case: (1) at the beginning of training as the initial hidden and cell states of the LSTM layer to influence the learning process of RNN, (2) after RNN evaluation as shown in diagram **a**

proposed approach, prediction is not based on a standard sequence-to-sequence (seq2seq) process, in which the final prediction is a sequence whose length is strictly determined and predefined in the RNN architecture design phase. Instead, under the presented approach only several upcoming sequence elements are forecasted and they gradually become predictors for subsequent predictions, as presented in Table 1. Such an approach does not assume any forecast horizon arbitrarily, and the methodology itself can be applied to short-, medium- and long-term forecasting. Moreover, it reflects a standard business use case of the energy consumption forecasting scenario, where historical data are available for a given period of time. As anticipated, explicit utilization of historical time series is therefore severely limited and after a relatively short period of time only previous forecasts remain as predictors.

Table 1 Prediction rules applied in the experiments under study

Prediction step	Predictors	Targets
1	$x_0, x_1, \ldots, x_{k-1}$	$y_0, \ldots, y_{n-1}$
2	$x_n, x_{n+1}, \ldots, x_{k-1}, y_0, \ldots, y_{n-1}$	$y_n, \ldots, y_{2n-1}$
3	$x_{2n}, x_{2n+1}, \ldots, x_{k-1}, y_0, \ldots, y_{2n-1}$	$y_{2n}, \ldots, y_{3n-1}$
…	…	…
m	$y_{(m-1)n-k}, \ldots, y_{(m-1)n-1}$	$y_{(m-1)n}, \ldots, y_{mn-1}$

Parameter k defines the length of the initial sequence, which is an input sequence to the RNN, and n denotes the length of the prediction. In each prediction step, we use a rolling window on the predictors which are gradually substituted by previous-step predictions from the forecasting process

4 Experiments

4.1 Preliminary Studies—A Simple Sine Function

In order to verify the effectiveness and substantiate the underlying concept of the method, we launched experiments for the 'laboratory' case. Let us consider two simple sine functions with different amplitudes, namely 1 and 3. The data were generated for the grid equal to 0.1 radians, and the resulting training data set contained almost 20,000 samples with 50 timesteps each. The recurrent layer has been created based on LSTM units and the resulting neural network architecture has been trained on both sine functions. After network optimization, the predictive model quality was evaluated in order to determine its predictive capabilities. Here, in the prediction phase the first 50-element long window of data was initialized as the predictor's vector. The prediction generated the next value in the sequence. This value was subsequently appended to the vector of predictors, and the prediction process was repeated on the shifted predictor sequence. The static feature was taken as the amplitude value, so that for each sine function there was only one number associated with the sequence.

We tested two models: (1) one based exclusively on dynamic data, (2) the other based both on static features and sequenced data. In case (1), the model was capable of reproducing the high-amplitude sine function, i.e. with the amplitude set at 3 (see Supplementary Figure S2), but it failed in the case of amplitude set at 1 (see Supplementary Figure S1). This outcome could be expected intuitively, since larger values impact the learning process to a greater extent, and therefore the resulting RNN-based predictive model is fitted better to a high-amplitude data. This predictive model has been stimulated to generate the function with an unseen amplitude of 2 (see Supplementary Figure S3), but it was not capable of creating such a qualitatively new function. In case (2), we used a model with a very simple architecture, where incorporation of static features was achieved only through concatenation of the LSTM hidden state with those static features, similarly to the architecture presented on the left-hand side of Fig. 2. The only difference is that the static feature was input directly to the concatenation layer, without any transformation into the latent representation. The resulting predictive model reproduced both sine functions with high accuracy (see Supplementary Figure S1 and Supplementary Figure S2), and performed better in the generative scenario (see Supplementary Figure S3).

This example clearly demonstrates that the incorporation of static features into the RNN-based architecture provides an efficient tool for learning the statistical relations between the static and dynamic components of the analyzed data.

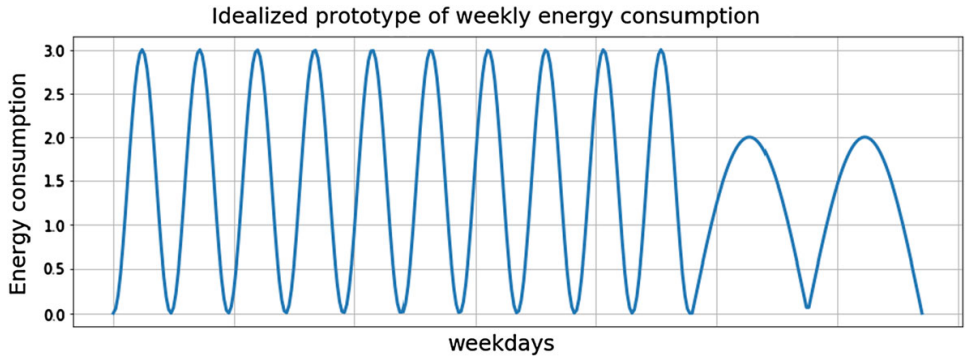


Fig. 3 Weekly energy consumption approximated by trigonometric functions. The higher-frequency part models energy consumption on workdays with two distinct peaks (morning and evening) per each day, while two smaller peaks relate to two weekend days and indicate the maximum consumption of energy during the day

4.2 Idealized Model of a Real-Life Energy Consumption Use Case

The next step, which will take us closer to the real-life use case, consists in the development of an idealized prototype of energy consumption. The prototype should account for the seasonality of energy consumption, which can be broken down into two main components: the weekly component (energy consumption is different on workdays and during weekends) and the yearly component (with seasonal differences throughout the year). For the sake of simplicity, we decided to leave out the yearly component and focus exclusively on weekly dependencies. We have adopted the following assumptions: (1) on working days, daily consumption is modeled by $A_1 \times sin^2(2x)$, which results in two distinct peaks per day, with values ranging between 0 and A_1, which reflects the morning and afternoon/evening energy consumption intensity, (2) the weekend consumption is approximated by $A_2 \times |sin(x)|$, with the maximum value at A_2, which reflects a more uniform energy consumption pattern during the day. Static features contain information on both amplitudes of the sine function, i.e. $[A_1, A_2]$ as presented in Fig. 3.

In order to reproduce such a function, proper setting of the input sequence length was required—namely, up to 6 days of the week to include at least one weekend day in the input sequence, so as to capture information about the weekend consumption. As a consequence, more computational resources were needed for the training process to account for the resulting increase in the number of NN parameters. We have found, however, that adding the binary feature "isWeekend" to the static data indicating weekend days (0 denoted a workday and 1 stood for a weekend day) significantly reduced the input sequence length required to ensure satisfactory accuracy of the predictive model. Thus, it was sufficient to take 1 day of energy consumption as a predictor to maintain the previous performance of the predictive model. Similarly to the simple sine function example, we simulated energy consumption of two different customers by changing the amplitudes of modeling functions, i.e. two sets of static data: $[A_1, A_2, isWeekend]$ and $[A_1', A_2', isWeekend]$ were taken for customer 1 and customer 2, respectively, with the corresponding energy consumption data. We analyzed the architectures shown in Fig. 2a, which were based both on the GRU and LSTM layers, and we concluded that for this particular task the LSTM showed better performance. Latent representation of the static features, which was concatenated with the LSTM output, was calculated by two fully-connected layers with 16 and 8 neurons, respectively. The RNN branch consists of three LSTM layers with 32, 16 and 8 neurons, respectively. Concatenation of the static and dynamic outputs was transformed by three fully-connected layers including 16, 8

and 6 neurons. We have also observed that adding a small stochastic component enhanced the training process in terms of its reproducibility. Moreover, we verified that turning the RNN architecture into the seq2seq mode significantly improved the accuracy of the models. However, the proper output sequence length had to be carefully selected, i.e. it cannot be too short. This effect can be expected intuitively if we consider the curvature of the output sequence. It is clear that the minimum number of points which provides sufficient information on the function curvature is 3. Thus, by increasing the length of the output sequence to 3, we will see a gradual increase in the accuracy level. Any further increase in length still improves the accuracy, but the gains will become progressively smaller.

Just as in the previous example of a simple sine function, adding static features was the key to distinguishing between the energy consumption profiles for each customer (See Supplementary Figure S4 and Supplementary Figure S5). The RNN which was based only on dynamic data was not able to properly learn the customer-specific energy consumption shape and always aimed at reproducing higher energy consumption trends. It is worth emphasizing that the results obtained for the energy consumption prototype clearly substantiate our previous conclusion that the incorporation of static features enhances the RNN capability and performance, and provides more options for the application of RNN in real-life cases. This simple model example also builds the foundation for the real-life use case which is an extension of the prototype, as discussed in the following paragraph. Moreover, the RNN incorporating static features was capable of predicting a new signal, which was not used during training, and was considerably more efficient in this respect compared with the RNN without additional static information (See Supplementary Figure S6).

4.3 Real-Life Use Case: Energy-Consumption Time Series

For the purposes of our analysis, we used the data of four different customers with significantly different weekly and yearly energy consumption trends. This data reflects real energy consumption and has been acquired directly at respective customer sites. In order to ensure the sufficient level of confidentiality, customer data was subject to anonymization. Following the necessary data preprocessing, the time series were used for the end-to-end analysis presented in this section. To the best of our knowledge, those data sets are not commonly available and have not been used in any previous studies. Historical data from the past 1–4 years were used for each customer for the purposes of model training, with the most recent year used as the test dataset. Data of each customer were scaled accordingly to the customer's average energy consumption, i.e. $\tilde{x}_{ij} = x_{ij}/\bar{X}_i$ where i and j stand for the customer and the observation, respectively. The symbol x_{ij} denotes a single observation and $\bar{X}_i$ denotes the mean value for the customer's entire consumption history covered by the training part of the data. In general, all analyzed signals are similar in terms of their overall shape. The time series are sampled every 15 min, which is a standard energy data aggregation level in many countries. However, the main difference concerns the weekend data, where the nature of energy consumption varied significantly depending on the customer type. Due to the lack of real static data representing each customer, we invented a set of features which characterized each time series, namely: (1) a set of percentiles: 0th, 20th, 40th, 60th, 80th and 100th, which provided information about the distribution and variation of energy consumption, (2) one-hot-encoded vector reflecting the calendar features, (3) the weekly trend used to extract the signature of unusual situations such as holidays, etc., and (4) the daily consumption average for each weekday type, extracted from all historical data per customer. It is worth noting that in point (2) the time descriptors are treated as static rather than dynamic features. The reason is that

the inclusion of those additional $24 + 7 + 12$ sequences of one-hot encoded features would increase the number of RNN variational parameters, which implies a non-negligible increase in computational costs. Moreover, some of the proposed time descriptors tend to vary gradually over time, especially those relating to weeks or months; therefore, we have concluded that it is only feasible to neglect their explicit time dependencies and to include them in a static form. This, however, is an arbitrary choice. With regard to point (3), we used the trend that originated from the test data. The rationale for adopting such a procedure is that it is both fully justified and needed to introduce an element of external expert knowledge, which can be used to model known changes in the nature of energy consumption. In particular, an increase in energy consumption can be expected due to business development or other factors. Such a trend feature provides a convenient tool capable of incorporating a priori expert knowledge into the predictive model. In the case of seasonality modeling, we have investigated an approach where seasonal cycles were provided as a set of harmonics obtained from Fourier analysis. In our case, however, all time series revealed very similar cyclic behavior; therefore, we have concluded that these quantities considered as features are not feasible. We wish to emphasize that the process of feature extraction is highly dependent on the problem under consideration and other sequenced data need to be analyzed independently.

Due to the complexity of real data, we have extended the previously adopted approach by injecting static features at the beginning of the RNN learning process (see Fig. 2c). This novel procedure leverages the advantages of the LSTM states, i.e. the hidden and cell states, by defining their initial form as the latent representation of static features obtained from the classical feed-forward sub-architecture. This means that the LSTM learning process is influenced by supplementary static features and allows the RNN to extract the static-dynamic dependencies included in the data. The hidden state of the last LSTM layer is concatenated with another latent representation of static information in a complete analogy with a more simplified model (see Fig. 2a). The latent representation of static features, injected as the hidden and cell states of the LSTM unit, was taken as two fully-connected layers with 128 and 64 neurons. The latent representation of static data which was concatenated with the LSTM output was a single fully-connected layer with 32 neurons. The RNN part consists of 3 LSTM layers with 64, 32 and 32 neurons, respectively. The concatenation of the RNN with static data was then transformed into 3 fully-connected layers with 32, 16 and 6 neurons, where the last layer determines the final length of the predicted sequence.

In the prediction phase, all static data must be provided, i.e. both the trend and the average consumption per weekday as well as the percentiles derived from the entire history of training data.

We have observed that an accurate selection of static features is the key to the forecast quality; therefore, they must be chosen carefully and intentionally so as to reflect all relevant quality metrics of the sequenced data. This observation also supports the main findings of our research - namely, that static features have a significant impact on the learning process and can substantially improve the quality of the forecast. Fig. 4 presents the forecast results for the data sets described above, and clearly shows that all qualitative features of energy consumption profiles have been properly recovered. Moreover, the forecast quality is also satisfactory in quantitative terms. Depending on the customer type and repeatability pattern over time, in two cases (customer 1 and 3, see Fig. 4a, c) the predictive model was very effective in learning the signal shape, whereas in the remaining two cases (customer 2 and 4, see Fig. 4b, d) some discrepancies still occurred. Higher forecast errors might be related to changes in the nature of energy consumption over the forecasting period. In particular, for customer 2 we have observed a major increase in energy consumption, which can be associated e.g. with the company growth. This change was so significant that even taking

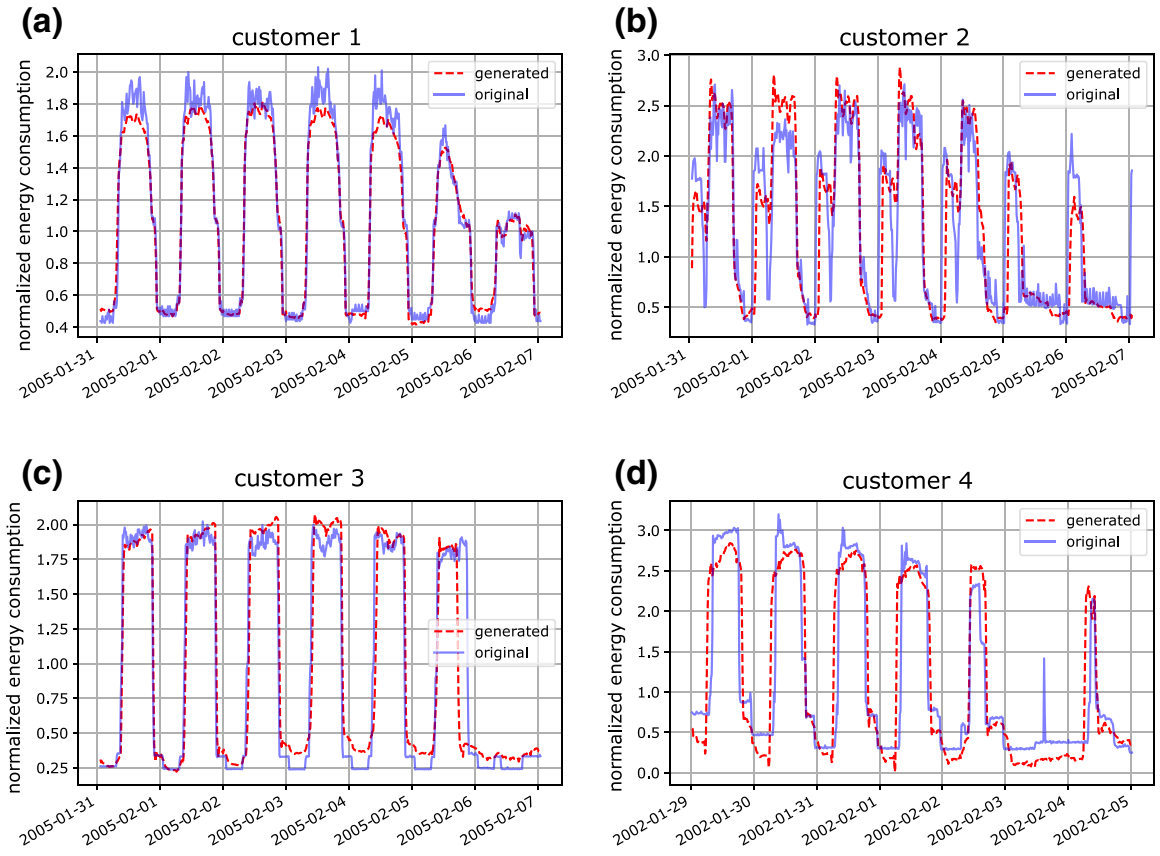


Fig. 4 Results obtained for the architecture involving two injections of static features, i.e. at the beginning of the training and as the concatenation of the latent representation and the hidden state resulting from static and dynamic data, respectively (see Fig. 2c for details). As discussed in the main body of the text, weaker performance for customer 2 and 4 was due to data specifics, i.e. for customer 2 there was a significant change in energy consumption during the year under prediction, which prevented an accurate prediction under the ML model, while for customer 4 historical data was available from only 1 year, so the number of training examples was limited compared to other customers. In this case, we also used the information regarding future trends, which was reasonable in this particular business context

current trends as predictors did not solve the problem altogether. The predictive model of customer 4 has been obtained from a 1-year period of historical data, which led to a limited number of training examples and, subsequently, lower predictive strength. On the contrary, energy consumption in the first two examples (customers 1 and 3) was repeatable and therefore all predictors at the prediction phase are sufficient to obtain an accurate prediction. In the case of customer 1, the obtained forecast accuracy can be directly compared to previously generated results. For instance, Zheng et al. [33] obtained the accuracy of 8.47% in terms of MAPE for 196-h short-term forecasting with the use of pure LSTM-based approach. It is worth noting, however, that the existing approaches focusing on single time series could outperform our methodology in terms of the forecast accuracy.

We also verified that taking into consideration the trend data for the specific year under prediction significantly improves the results, which is especially evident for customer 3 and customer 4, see Fig. 4. Moreover, the proposed neural network architecture allowed to obtain higher accuracy of the model in comparison with the architecture used previously; see Table 2 for more details.

The most remarkable outcome, however, is the fact that all the signals were correctly recognized. In particular, the signal shape during weekends was well reproduced for each customer, even though it constituted the main difference between various energy consumption

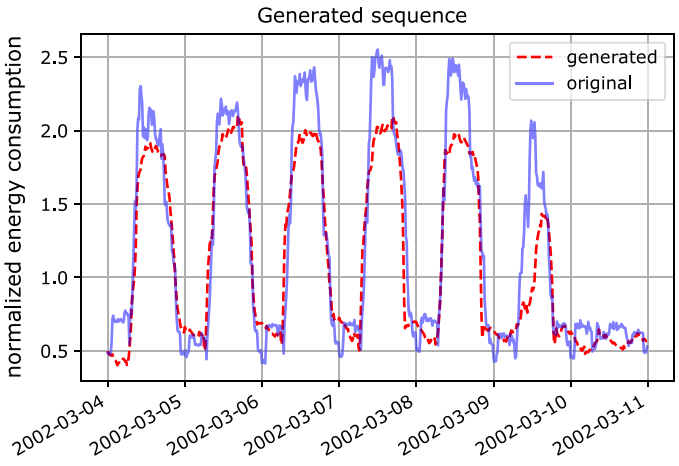


Fig. 5 Energy consumption generated from static features. Despite a noticeable discrepancy, the model was able to reproduce the characteristics of energy consumption for new data

profiles. We also demonstrated that the use of a standard RNN with static data treated as the signal (constant sequences) generated much worse results (see Table 2 for details). Two cases were examined: (1) all static data were included as the signal, (2) only calendar features and trend information were attached to the sequence. In both cases, the predictive model was not capable of reproducing energy consumption with the quality comparable to other approaches discussed earlier. Additionally, the training of the network required much more memory than in the case of other approaches.

Another capability ensured by the strategies proposed herein is the option to follow a generative approach, i.e. to use the model with a view to generating energy consumption forecasts for a given set of static features. Such a generative mode allows to obtain energy demand predictions even if no historical data for a given customer is available. The previously trained predictive model is equipped with all necessary static features as predictors, together with a seed of dynamic data, and therefore it delivers the predicted time series. For demonstration purposes, we generated new data, which was not seen in the training phase, by applying a set of transformations to old data sets, e.g. through scaling by the factor, additive components, etc. We subsequently extracted static features from the signal and prepared input sequence to obtain the desired forecast from the model. The results are presented in Fig. 5. In the generative mode, the predictive model accurately recovered the main features of the target time series. Furthermore, it is worth mentioning that the amount of data to which the model has been exposed during training was rather limited, i.e. it gained the knowledge about the consumption profiles for only 4 customers. Nevertheless, it was possible to generate qualitatively new data of an acceptable quality level. In this case, the set of static features was an instrument which guided the generation process towards the expected prediction.

It is also worth noting that the prediction error does not increase with time, which means that the proposed methodology can be used independently of the forecast time horizon without compromising accuracy. An example of time evolution of the Mean Absolute Percentage Error (MAPE) is presented in Fig. 6 for customer 3. A clearly noticeable peak in April corresponds to the unmanaged behavior associated with Easter holidays. Similar behavior is also shown in Table 2 as an increase in MAPE for customer 2. Excluding the problematic period from the forecasts results in a significant decline in the error figure (values in parentheses).

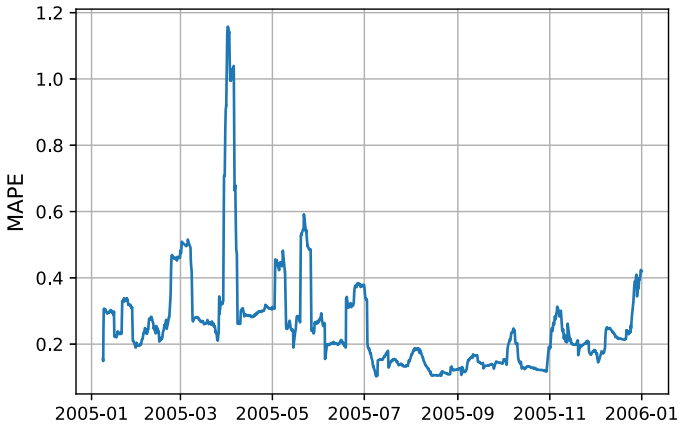


Fig. 6 MAPE of the energy consumption forecast for customer 3. The high April peak corresponds to an unexpected event, which is difficult to reproduce in the ML model. The overall trend, however, seems to be stable and time-independent, which indicates that the presented approach will be applicable also in long-term forecasts

Table 2 The mean absolute percentage error (MAPE) for each customer

id	(1)	(2)	(3)	(4)	(5)
1	0.0797	0.0868	0.0960	0.1590	0.2850
2	0.4296 (0.2279)	0.4307 (0.2509)	0.4190 (0.2356)	0.4090 (0.3260)	0.4870 (0.4920)
3	0.1727	0.1760	0.1746	0.2370	0.3080
4	0.3221	0.3543	0.3452	0.3710	0.5010

We evaluated the following 5 architectures: (1) the merged model, i.e. the architecture from Fig. 2c which leverages the advantages of both approaches, (2) concatenation of the hidden state from the last LSTM with the latent representation originating from the fully-connected layers (presented in Fig. 2b), (3) injection of static features as the initial hidden and cell states of the first LSTM layer without any further concatenation of the layer's output (presented in Fig. 2c), (4) the RNN with all static features included in the form of a sequenced signal, and (5) the RNN with calendar and trend features in the form of a sequenced signal. The proposed merged model achieved the best performance in comparison with other approaches. It is worth noting that for customer 2 we have observed irregular and unexpected data behavior, which cannot be learned properly by the NN. Removing this problematic period from the forecast resulted in much better performance, as shown in parentheses

5 Conclusions

In the present paper, we have discussed the methodology for incorporating static features into RNN architectures. We have demonstrated that the proper handling of static features significantly enhances the training process, and the resulting predictive models exhibit much higher predictive strength. The analyzed strategies allow to expose the same predictive model to multiple time series, and therefore the model maintenance cost can be significantly reduced. We have also shown the evolution of the presented concept, starting from the simplest possible example of sine functions where we applied the architectures already discussed in the literature. We then moved on towards a more sophisticated exemplary use case of a basic predictive model for energy consumption. Finally, we presented a real-life example of energy consumption case, where static features are incorporated twice: (1) as the initial hidden and cell states of the first LSTM layer, and (2) as the latent representation concatenated with the resulting hidden state of the last LSTM layer. Such a meticulous treatment of static features ensured the efficient handling of statistical relationships between the static and dynamic part of the time series under consideration. We have also demonstrated that the proposed methodologies can be successfully applied under generative schemes. As an illustration, we have stimulated the predictive model to generate qualitatively new time series. It is worth emphasizing that the incorporation of accurately selected static features provided efficient

capability to direct the generative process. The methodology discussed in the present study will be extended to generative approaches based on the variational autoencoder concept. The decoder latent representation space will be carefully explored in order to find the areas responsible for the generation of dynamic features with expected properties.

References

1. Bengio Y, Frasconi P, Simard P (1993) The problem of learning long-term dependencies in recurrent networks. In: IEEE international conference on neural networks, IEEE. https://doi.org/10.1109/ICNN.1993.298725, https://ieeexplore.ieee.org/document/298725
2. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
3. Chen D, Li S, Lin FJ (2019) New super-twisting zeroing neural-dynamics model for tracking control of parallel robots: a finite-time and robust solution. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2019.2930662
4. Chen D, Li S, Lin FJ, Wu Q (2019b) Rejecting chaotic disturbances using a super-exponential-zeroing neurodynamic approach for synchronization of chaotic sensor systems. IEEE Trans Cybern 19(1):74. https://doi.org/10.3390/s19010074
5. Chen D, Li S, Wu Q, Luo X (2019) New disturbance rejection constraint for redundant robot manipulators: an optimization perspective. IEEE Trans Ind Inf. https://doi.org/10.1109/TII.2019.2930685
6. Dudek G (2015) Short-term load forecasting using random forests, vol 323. Springer, Berlin, pp 821–828
7. Eck D, Schmidhuber J (2002) Learning the long-term structure of the blues, vol 2415. Springer, Berlin, pp 284–289
8. Esteban C, Staeck O, Yang Y, Tresp V (2016) Predicting clinical events by combining static and dynamic information using recurrent neural networks. arXiv: 1602.02685 [cs], arXiv:1602.02685
9. Feilat EA, Bouzguenda M (2011) Medium-term load forecasting using neural network approach. In: 2011 IEEE PES conference on innovative smart grid technologies—Middle East, IEEE, pp 1–5. https://doi.org/10.1109/ISGT-MidEast.2011.6220810, http://ieeexplore.ieee.org/document/6220810/
10. Gers FA, Schmidhuber J, Cummins F (2000) Learning to forget: continual prediction with LSTM. Neural Comput 12(10):2451–2471. https://doi.org/10.1162/089976600300015015
11. González-Romera E, Jaramillo-Morán M, Carmona-Fernández D (2008) Monthly electric energy demand forecasting with neural networks and fourier series. Energy Convers Manag 49(11):3135–3142. https://doi.org/10.1016/j.enconman.2008.06.004
12. Graves A, Fernández S, Liwicki M, Bunke H, Schmidhuber J (2007) Unconstrained online handwriting recognition with recurrent neural networks. In: NIPS'07 Proceedings of the 20th international conference on neural information processing systems, Curran Associates Inc., https://dl.acm.org/citation.cfm?id=2981562.2981635
13. Graves A, Mohamed A, Hinton GE (2013) Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing, pp 6645–6649
14. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
15. Kaytez F, Taplamacioglu MC, Cam E, Hardalac F (2015) Forecasting electricity consumption: a comparison of regression analysis, neural networks and least squares support vector machines. Int J Electr Power Energy Syst 67:431–438. https://doi.org/10.1016/j.ijepes.2014.12.036
16. Khosravani H, Castilla M, Berenguel M, Ruano A, Ferreira P (2016) A comparison of energy consumption prediction models based on neural networks of a bioclimatic building. Energies 9(1):57. https://doi.org/10.3390/en9010057
17. Kuo PH, Huang CJ (2018) A high precision artificial neural networks model for short-term energy load forecasting. Energies 11(1):213. https://doi.org/10.3390/en11010213
18. Leontjeva A, Kuzovkin I (2016) Combining static and dynamic features for multivariate sequence classification. In: 2016 IEEE international conference on data science and advanced analytics (DSAA) pp. 21–30. https://doi.org/10.1109/DSAA.2016.10, arXiv: 1712.08160
19. Liang Y, Niu D, Ye M, Hong WC (2016) Short-term load forecasting based on wavelet transform and least squares support vector machine optimized by improved cuckoo search. Energies 9(10):827. https://doi.org/10.3390/en9100827
20. Liu N, Tang Q, Zhang J, Fan W, Liu J (2014) A hybrid forecasting model with parameter optimization for short-term load forecasting of micro-grids. Appl Energy 129:336–345. https://doi.org/10.1016/j.apenergy.2014.05.023

21. Mayer H, Gome F, Wierstra D, Nagy I, Knoll A, Schmidhuber J (2006) A system for robotic heart surgery that learns to tie knots using recurrent neural networks. In: 2006 IEEE/RSJ international conference on intelligent robots and systems. https://doi.org/10.1109/IROS.2006.282190
22. Niu D, Dai S (2017) A short-term load forecasting model with a modified particle swarm optimization algorithm and least squares support vector machine based on the denoising method of empirical mode decomposition and grey relational analysis. Energies 10(3):408. https://doi.org/10.3390/en10030408
23. Niu D, Shi H, Wu DD (2012) Short-term load forecasting using bayesian neural networks learned by hybrid Monte Carlo algorithm. Appl Soft Comput 12(6):1822–1827. https://doi.org/10.1016/j.asoc.2011.07.001
24. Ringwood JV, Bofelli D, Murray FT (2001) Forecasting electricity demand on short, medium and long time scales using neural networks. J Intell Rob Syst 31(1):129–147. https://doi.org/10.1023/A:1012046824237
25. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536. https://doi.org/10.1038/323533a0
26. Salehinejad H, Sankar S, Barfett J, Colak E, Valaee S (2018) Recent advances in recurrent neural networks. arXiv:1801.01078
27. Selakov A, Cvijetinović D, Milović L, Mellon S, Bekut D (2014) Hybrid pso-svm method for short-term load forecasting during periods with significant temperature variations in city of burbank. Appl Soft Comput 16:80–88. https://doi.org/10.1016/j.asoc.2013.12.001
28. Suganthi L, Iniyan S, Samuel AA (2015) Applications of fuzzy logic in renewable energy systems—a review. Renew Sustain Energy Rev 48:585–607. https://doi.org/10.1016/j.rser.2015.04.037
29. Tanti M, Gatt A, Camilleri KP (2018) Where to put the image in an image caption generator. Nat Lang Eng 24(3):467–489. https://doi.org/10.1017/S1351324918000098
30. Vinyals O, Toshev A, Bengio S, Erhan D (2014) Show and tell: a neural image caption generator. arXiv:1411.4555
31. Wu Y, Schuster M, Chen Z, Le QW, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado GS, Hughes M, Dean J (2016) Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144
32. Yang Y, Fasching PA, Tresp V (2017) Predictive modeling of therapy decisions in metastatic breast cancer with recurrent neural network encoder and multinomial hierarchical regression decoder. In: 2017 IEEE international conference on healthcare informatics (ICHI), IEEE, pp 46–55. https://doi.org/10.1109/ICHI.2017.51, http://ieeexplore.ieee.org/document/8031131/
33. Zheng H, Yuan J, Chen L (2017) Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation. Energies 10(8):1168. https://doi.org/10.3390/en10081168

Publication [P5]

Oppio, A., Dell'Ovo, M., Torrieri, F., Miebs, G., and Kadziński, M. (2020). Understanding the drivers of Urban Development Agreements with the rough set approach and robust decision rules. In Land Use Policy (Vol. 96, p. 104678). Elsevier BV., DOI: `10.1016/j.landusepol.2020.104678`.

Number of citations[5]:

- according to Web of Science: 10
- according to Google Scholar: 14

Contribution of co-authors (excluding the author and the supervisor of this dissertation):

- Alessandra Oppio:
  - Co-authorship of the idea underlying the paper
  - Data and preference collection; interpretation of results
  - Co-authorship of the text of the publication
- Marta Dell'Ovo:
  - Co-authorship of the idea underlying the paper
  - Data and preference collection; interpretation of results
  - Co-authorship of the text of the publication
- Francesca Torrieri:
  - Data and preference collection
  - Corrections to the first draft of the manuscript

[5] as on September 25, 2023

Land Use Policy 96 (2020) 104678

Contents lists available at ScienceDirect

Land Use Policy

journal homepage: www.elsevier.com/locate/landusepol

Understanding the drivers of Urban Development Agreements with the rough set approach and robust decision rules

Alessandra Oppio[a,*], Marta Dell'Ovo[b], Francesca Torrieri[c], Grzegorz Miebs[d], Miłosz Kadziński[d]

[a] Department of Architecture and Urban Studies, Politecnico di Milano, Milan, Italy
[b] Department of Architecture, Built Environment and Construction Engineering, Politecnico di Milano, Milan, Italy
[c] Department of Industrial Engineering, University of Naples Federico II, Naples, Italy
[d] Institute of Computing Science, Poznan University of Technology, Poznań, Poland

ARTICLE INFO

Keywords:
Multiple Criteria Decision Analysis
Urban Development Agreements
Negotiation
Dominance-based Rough Set Approach
Decision rules
Robustness analysis

ABSTRACT

The rise of neoliberalism in the context of urban development has encouraged cooperation between public and private parties. This cooperation is structured by contracts, generally called Urban Development Agreements (UDAs). Being part of the urban regeneration strategies, UDAs aim at achieving durable improvements of abandoned areas, namely brownfields, according to the sustainability principles. Thus, within the negotiation between private developers and public administrations, multiple and conflicting instances have to be faced case by case. Despite the uniqueness of each UDA, it is possible to define a set of pertinent characteristics that play a crucial role in determining the fairness and appropriateness of the public-private partnership. Given this context, we propose a novel variant of the Dominance Rough Set Approach (DRSA) for i) exploring the relationship between condition attributes or criteria and effects of urban development processes and for ii) supporting negotiations according to the detection of a set of relevant features. Specifically, DRSA is applied on a sample of UDAs recently concluded in the Lombardy Region (Northern Italy), and then tested on a sample of other UDAs under the negotiation phase. The analysis involves five dimensions represented by attributes and criteria related to urban, institutional, negotiation, development and economic contexts. The inferred decision rules provide useful knowledge for supporting complex decision processes such as the allocation of costs and benefits within UDAs.

1. Introduction

In land-use planning, the dual relationship between urban regulations and property value has been widely recognized (Alterman, 2012). In particular, land value capture refers to the financial instruments through which public authorities can detect increases in the values of properties that are unrelated to the actions of landowners (Micelli, 2011; Oppio et al., 2019). Due to traditionally different administrative and tax implementation schemes, several forms of value capture mechanisms exist. These include: i) impact fees, ii) joint developments, iii) property or land value taxes, iv) land banking, v) tax increment financing, vi) betterment levies, and vii) development agreements (Youngman and Malme, 1994, 2004; Malme and Youngman, 2001; OECD, 2017). Despite the differences among them, an idea underlying all mechanisms is to impose a levy on new developments in order to fully or partially finance new local infrastructures and/or to improve the existing ones (Reimer et al., 2014).

In some cases, standard charges are defined by local public administrations, while in other cases private contributions are negotiated within integrated programs or complex partnership arrangements in addition to, or instead of, fixed tariffs. In some Italian regions, Urban Development Agreements (UDAs) represent a common value-capture mechanism. Within this kind of agreements, the involved developers provide public services and/or financial contributions for obtaining planning/building permissions or rezoning decisions that allow more profitable development than the one defined by the statutory urban plan. Variation from the local plans in terms of uses, density, volumes or ratio of buildings to open spaces can give rise to the planning gain, being at the basis of the land value capture mechanism within the UDAs. Given the contractual nature of UDAs, they allow public authorities and developers to find the solutions to the specific problems according to a case-by-case approach.

Considering the particular case of Italy, UDAs between public administrations and private developers were first introduced by law at the

* Corresponding author.
E-mail addresses: alessandra.oppio@polimi.it (A. Oppio), marta.dellovo@polimi.it (M. Dell'Ovo), frtorrie@unina.it (F. Torrieri), grzegorz.miebs@protonmail.ch (G. Miebs), milosz.kadzinski@cs.put.poznan.pl (M. Kadziński).

https://doi.org/10.1016/j.landusepol.2020.104678
Received 8 July 2019; Received in revised form 30 March 2020; Accepted 6 April 2020

end of the last century (see article 16 of Law no. 179, 1992 "Integrated Intervention Programs"). They aimed at fostering the involvement of private players and resources in the field of urban development in a period of strong growth of the real estate market. At that time, the returns for developers were high and local authorities were willing to use the real estate expansion as a means for supporting urban improvement (Camagni, 1996), in the wake of the local fiscal crisis and the increasing cuts on resources from the national government.

Indeed, in the context of a profound crisis of public, national and local finances, a fair allocation of the value capture from urban transformations between the public and private sectors represents a promising option for dealing with problems that the traditional urban planning tools have failed to solve (Camagni, 2016). Within these negotiation processes and the consequent agreements between private developers and public bodies, the privates' interest of maximizing profits often diverges from public authorities' objectives of improving urban quality and social well-being. In many cases, the negotiating power of the public administrations turned out to be weaker than the capability of private developers to direct urban transformations towards the achievement of their interests (Camagni, 1996; Oppio et al., 2019).

In turn, in the current market scenario, the investment and real estate development operations show weak returns. As a consequence, the new challenge for developers and investors in the context of urban regeneration projects is to shift the focus from generating short-term financial returns to creating positive social and environmental values in a long-term perspective. The cultural and technological advances are changing the traditional business models and blurring the boundaries between the need to demonstrate good "social purpose" and enhancing shareholder value. Thus, the interests of the public and private sectors converge on generating positive social and environmental effects that can generate adequate returns for investors in the long term (PwC, 2018).

Based on such premises, integrating the interests of all the stakeholders in the urban development processes appears to be crucial to build cooperative negotiation agreements, in which the interests of both public and private sectors converge into a holistic view of creating value (Ottomano Palmisano et al., 2016). While in the past the conflicts between public and private actors were solved on a normative basis (e.g., through the expropriation mechanism), nowadays the multi-dimensional and multi-stakeholder dimension of urban regeneration processes drives all the actors involved to defining an innovative model of cooperative negotiations (Raiffa, 1982; Stanghellini, 2019).

This paper focuses on UDAs with the aim of i) defining a set of pertinent characteristics for setting the problem of the public-private partnership in the field of urban development; special attention is drawn on the value generated by proposals that differ from the statutory plan (planning gain); ii) defining a decision support system for fair allocation between the public and the private; iii) pointing out what are the main drivers of UDAs able to influence the value generated by urban developments; and iv) providing operational recommendations for supporting the on-going and future negotiations according to a co-operative approach. To answer these research questions, we apply a novel variant of the Dominance-based Rough Set Approach (DRSA) to two groups of UDAs in the Lombardy Region (Northern Italy), one concluded and the other still in progress. These groups are composed of, respectively, 15 and 10 UDAs. We aim at analyzing their main features and identifying the drivers determining the outcome of UDAs, in terms of the planning gain generated by the urban transformations.

This method was chosen because of its appropriateness for analyzing the consistency of available data and for inferring the decision rules. The latter can capture some recurring patterns and characteristics observed within the groups of UDAs that correspond to decision classes representing various success levels in UDAs' completion. The importance of this stage is emphasized by the diversity of the pertaining factors taken into account and involving the urban, institutional, negotiation, development and economic contexts. Then, the inferred rules are used to predict the outcomes of the ongoing UDAs to support the negotiation process. The proposed approach handles subsets of both attributes and criteria, respectively, without and with preferential directions on their respective performance scales. Moreover, the basic results are validated against the outcomes of robustness analysis, which involves the results obtained with different sets of decision rules corresponding to all reducts, i.e., the minimal subsets of attributes/criteria capable of fully explaining the classification of concluded UDAs.

The paper is organized in the following way. In Section 2, we focus on the previous studies that dealt with the issue of evaluation in the urban development domain. The Dominance-based Rough Set Approach and its novel variant are described in Section 3. Section 4 introduces the UDAs under analysis and discusses the context and results of the case study. The operational recommendations and conclusions follow in Section 5.

2. Previous studies

As a consequence of the devolution of public policy, local authorities and the private sector often cooperate in defining and implementing urban development projects. This is confirmed by numerous studies on the role played by evaluation in supporting urban development under different kinds of Public-Private Partnerships (PPPs). The negotiation process between public authorities and private developers according to which the definition of a PPP agreement is defined needs to be supported by evaluation evidence. With special reference to UDAs, two main evaluation perspectives are generally considered: i) the private developer's perspective focused on the viability of urban developments, ii) the public perspective aimed at capturing a part of the surplus-value generated by new land uses. This section provides an overview of the most relevant studies according to the following three aspects: i)evaluation of public and private benefits from urban development; ii) ex-post evaluation of successful urban development drivers; iii) definition of decision support system for aiding negotiation; some insights are presented in the conclusions.

2.1. Evaluation of public and private benefits from urban development

Research on evaluating public and private benefits from urban development has evolved according to the innovative forms of cooperation between public bodies and private developers. In the last years, such cooperation has led many European cities to enhance complex urban interventions. This kind of PPP enable the two stakeholders to work together for a common interest in products and services while sharing risks, costs, and resources (Green Paper, 2006; Stanghellini and Copiello, 2011; Han et al., 2019). Concerning the instance of evaluating the urban development attractiveness from the private and public viewpoints, Mambelli and Stanghellini (2003) focused on the role played by an evaluation within the negotiation process and discussed its potential by analyzing the results of a financial and multiple criteria analysis applied to a case study. Micelli (2004; 2011) and Camagni and Modigliani (2013) proposed a model for assessing the overall value generated by an urban development based on the assumption that the value of a development project is equal to the monetary residual or surplus available once a site has been developed. Specifically, it was defined as a percentage of the difference between the land's final value resulting from the urban development, including transformation costs and expenses, and its initial value. For its operational effectiveness, this model has been widely used to evaluate the fairness of land value capture in the urban developments (Oppio and Torrieri, 2018; Oppio et al., 2018, 2019; Calabrò and Della Spina, 2012). Furthermore, Copiello (2012) reflected on the balance between private and public instances as the requirement for legitimating PPPs. A general overview of the advantages and disadvantages of different appraisal methodologies for developments in the frame of PPP was given by Tánczos and Kong (2001).

2.2. Ex-post evaluation of successful urban development drivers

An additional contribution to the methodologies for evaluating urban development under PPPs is represented by the research on the drivers of urban transformations. Copiello (2011) applied the Rough Set Analysis (Pawlak 1982; 1991) to investigate into five categories of attributes (urban context, institutional context, PPP procedure, PPP features, and financial and economics performances), which are the drivers for successful urban developments. In turn, Nijkamp et al. (2002) identified the crucial drivers of urban renewal projects concerning complex factors such as the institutional context, the financial viability, and the presence of spatial externalities. To compare nine selected urban development projects involving different types of PPPs in The Netherlands, they applied the Rough Set Analysis for detecting the most important drivers of successes and failures within recent development plans in Dutch cities. In the same spirit, Xiong et al. (2019) conducted a systematic review of case-study literature on PPP to investigate the dynamics of governance issues, including the relative importance, interrelationships, and connections with PPP success and failure. Furthermore, Abastante et al. (2013) used DRSA and Analytic Network Process (ANP) for assessing the sustainability of different urban transformation scenarios, whereas Chen et al. (2009) classified brownfield redevelopments in the American cities according to their level of effectiveness and future needs.

2.3. Definition of Decision Support System for aiding negotiation

Decision Support Systems (DSSs) can be considered as a prominent field of research within negotiation processes. Urban developments are characterized by the coexistence of multiple and conflicting instances, therefore they are considered as complex, ill-structured problems to decide on. Moreover, joint project investments require transparency to legitimate decisions with respect to public preferences and expectations.

In analyzing PPPs in previous studies, several evaluation methods were applied with the purpose to aid the negotiation process or to give robustness to the choice of different scenarios. In particular, Morano and Tajani (2017) introduced the logic of fuzzy systems for evaluating the coherence of the projects submitted by private operators with the importance of the objectives of redevelopment set by the public administration. The proposed algorithm was able to adapt to the specificities of the project under evaluation and to the changes that, over time, could arise in the preference system of the Decision Maker.

Gan et al. (2018) identified the importance of involving different stakeholders in the decision process by developing multiple-attribute group DSS or through an interactive learning framework (Smajgl and Ward, 2015). To provide recommendations on large-scale projects, Liang and Hu (2018) defined an evaluation framework by using the Fuzzy Analytic Hierarchy Process (FAHP) to investigate different PPPs based on features such as partnership, benefit-sharing, or risk-sharing. Moreover, Bai et al. (2017) applied a Fuzzy Comprehensive Evaluation Model (FCEM) to analyze various sustainability risk factors like culture and society, cost and economy, ecology and environmental, project and organization, and politics and laws. Finally, Cilona and Granata (2015) investigated nine criteria capturing the environmental and aesthetic quality of the investment projects along with their attractiveness and accessibility with the support of the Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE).

2.4. Insights from the previous studies

Concerning the evaluation dimensions described in the previous sub-sections, what deserves to be investigated is how to find a balance between collective benefits and financial sustainability. Even though negotiation procedures emphasize the role of evaluation, from the current practices, a gap within the public decision-making process emerges. Given this instance of sound choices, this paper proposes a methodological contribution for appropriately supporting local administrations and private developers when they enter into negotiation. Thus, the decision problem considered in this paper consists in explaining a classification of the already concluded UDAs and deriving a decision model that could be used to classify programmed or on-going UDAs based on their characteristics, as well as to support ill-structured negotiation processes among public and private parties. To tackle these issues, an effort has been made to point out the drivers of successful UDAs by considering their complexity. To this purpose, the Rough Set-Based methodology has been employed. However, the specificity of the considered case study in terms of a small number of alternatives, a large number of criteria as well as characteristics of evaluation attributes, has required the development of a novel variant of DRSA, specifically tailored to the problem under investigation. To provide an effective contribution to local authorities, the set of criteria and attributes have been developed with the support of experts and Lombardy Region officials. The success of the UDAs has been settled according to decision classes, defined by estimating the overall value generated by the UDAs.

3. Methodology

In this section, we introduce a variant of the DRSA that handles both criteria and attributes in the data structuring phase and a multiplicity of compatible sets of rules in the phase of preference learning. Also, the novel approach differentiates the impacts that different sets of rules have on the recommended assignment by making the weights assigned to these sets dependent on their predictive abilities. These developments have been motivated by the following characteristics of the considered decision problem: a presence of attributes without preferentially ordered performance scales in the set of features describing the UDAs, a relatively large set of 11 attributes and criteria considered in the analysis, and a relatively small set of reference alternatives composed of only 15 UDAs. In what follows, we discuss the original variant of DRSA (see Section 3.1), and then point out the modifications proposed in the novel approach (see Section 3.2).

3.1. The Dominance-based Rough Set Approach

Dominance-based Rough Set Approach is a Multiple Criteria Decision Analysis method originally proposed by Greco et al. (2001) for dealing with multiple criteria sorting (ordinal classification) problems (Zopounidis and Doumpos, 2002). The distinctive features of this approach derive from handling assignments of alternatives that may be inconsistent with respect to the dominance principle and using a transparent preference model in the form of "*if … then …*" decision rules (Greco et al., 2016). In what follows, we introduce the notation and discuss the main steps of DRSA.

3.1.1. Notation

DRSA handles knowledge about the decision problem represented in the form of an information table. Its rows correspond to a set of evaluated alternatives $A = \{a_1, ..., a_i, ..., a_n\}$, whereas the columns are divided into the evaluation criteria $G = \{g_1, ..., g_j, ..., g_m\}$ and the decision attribute D (Cinelli et al., 2019). The preference-ordered criteria in G are used to assess the alternatives from different relevant viewpoints (Roy, 1996). The evaluation (performance) of $a_i \in A$ on $g_j \in G$ is denoted by $g_j(a_i)$. For clarity of presentation, without loss of generality, we assume that all criteria in G have an increasing direction of preference (hence, greater performances are preferred).

The decision attribute D is filled for a subset of reference (training) alternatives $A^R \subseteq A$, hence indicating their overall evaluation. Such a comprehensive evaluation takes the form of an assignment of $a_i \in A^R$ to one of the t pre-defined, preference ordered classes $Cl_1, ..., Cl_r, ..., Cl_t$, where Cl_r is more preferred to Cl_{r-1}, $r = 2, ..., t$ (Zopounidis and Doumpos, 2002). The classification for the non-reference (testing)

alternatives in $A^{NR} = A \setminus A^R$ is unknown and needs to be predicted with a decision model derived from the analysis of reference alternatives.

3.1.2. Approximations of class unions

The construction of a decision model starts with the analysis of inconsistencies in the classifications provided for the reference alternatives. In DRSA, such a consistency check is based on the following dominance principle: "*if* a_i *dominates* a_k *, then* a_i *should be classified at least as good as* a_k*"* (Greco et al., 2001). The dominance relation Δ^G for a family of criteria G is defined as follows (Roy, 1996):

$a_i \Delta^G a_k \Leftrightarrow \forall\, g_j \in G$: $g_j(a_i) \geq g_j(a_k)$.

Given the ordinal classification problem, the consistency analysis is conducted for different class unions. We distinguish the upward $Cl_r^{\geq}$ and downward $Cl_r^{\leq}$ unions corresponding to, respectively, at least or at most class Cl_r. It is relevant to consider only the unions which do not contain all reference alternatives, i.e., $Cl_r^{\geq}$ for $r = 2, ..., t$ and $Cl_r^{\leq}$ for $r = 1, ..., t-1$. For each union, the results of a consistency check are quantified with its lower and upper approximations. On the one hand, the lower approximation contains all reference alternatives certainly belonging to the union, which means no argument would question their membership. On the other hand, the upper approximation is composed of the reference alternatives possibly belonging to the union meaning that there is at least one argument that would confirm their membership. Formally, these approximations (denoted by $\underline{P}$ and $\overline{P}$) can be defined for any subset of criteria $P \subseteq G$ as follows (Greco et al., 2016):

$$\underline{P}(Cl_r^{\geq}) = \{a_i \in A^R: \nexists a_k \in Cl_{r-1}^{\leq},\ a_k \Delta^P a_i\} \text{and} \overline{P}(Cl_r^{\geq}) = \{a_i \in A^R: \exists\ a_k \in Cl_r^{\geq},\ a_i \Delta^P a_k\}$$

$$\underline{P}(Cl_r^{\leq}) = \{a_i \in A^R: \nexists a_k \in Cl_{r+1}^{\geq},\ a_i \Delta^P a_k\} \text{and} \overline{P}(Cl_r^{\leq}) = \{a_i \in A^R: \exists\ a_k \in Cl_r^{\leq},\ a_k \Delta^P a_i\}$$

In particular, the lower approximation of $Cl_r^{\geq}$ ($\underline{P}(Cl_r^{\geq})$) is composed of all reference alternatives, which are not dominated in terms of Δ^P by any reference alternative belonging to a class worse than . In this perspective, let us emphasize that Δ^P is reflexive, and thus none alternative from $Cl_{r-1}^{\leq}$ belongs to $\underline{P}(Cl_r^{\geq})$. The upper approximation of $Cl_r^{\geq}$ ($\overline{P}(Cl_r^{\geq})$) is composed of all reference alternatives, which dominate in terms of Δ^P some reference alternative belonging to union $Cl_r^{\geq}$.

3.1.3. Quality of approximation of classification

A key measure that quantifies the consistency of the assignments provided for the reference alternatives is the quality of approximation of classification $\gamma_P(Cl) \in 0{,}1$. It refers to the number of alternatives contained in the boundaries of all upper or lower class unions defined as the differences between the respective upper and lower approximations (i.e., $Bn_P(Cl_r^{\geq}) = \overline{P}(Cl_r^{\geq}) \setminus \underline{P}(Cl_r^{\geq})$ and $Bn_P(Cl_r^{\leq}) = \overline{P}(Cl_r^{\leq}) \setminus \underline{P}(Cl_r^{\leq})$). Such alternatives are troublesome, underlying the inconsistency in the provided classifications. Hence, if they are more numerous, the quality of approximation should be lowered. Formally:

$$\gamma_P(Cl) = \frac{card(A^R - \bigcup_{r=2,...,t} Bn_P(Cl_r^{\geq}))}{card(A^R)} = \frac{card(A^R - \bigcup_{r=1,...,t-1} Bn_P(Cl_r^{\leq}))}{card(A^R)}$$

Overall, measure $\gamma_P(Cl)$ quantifies the quality of knowledge that can be extracted from the assignments provided for the reference alternatives, where P is a subset of criteria $P \subseteq G$ and Cl is the considered classification task (Słowiński et al., 2015).

3.1.4. Reducts

To investigate the role that different subsets of attributes may have in explaining the classification for the reference alternatives, one could refer to the concept of *reduct* (Susmaga and Słowiński, 2015; Kadziński et al., 2020). It is defined as the minimal subset of criteria $P \subseteq G$ that allows achieving the same quality of classification as the whole set of attributes ($\gamma_P(Cl) = \gamma_G(Cl)$). Hence, the criteria contained in the reduct play a significant discriminatory role, whereas the remaining ones are redundant from the viewpoint of maintaining a consistency of the information table. Let us emphasize, however, that there may exist more than one reduct for the same dataset.

3.1.5. Decision rules

Having structured the data, the lower and upper approximations serve as a basis for deriving a set of logical statements in the form of decision rules. Such "*if …, then …*" statements explain the classification provided for the reference alternatives in terms of the minimal subset of conditions referring to their performances on particular attributes (Greco et al., 2016). In what follows, we focus only on certain rules, which are induced from the lower approximations of class unions. These rules are induced from the upward and downward class unions and have the following form:

$$\textbf{if}(g_{j1}(a_i) \geq r_{j1}\ \textbf{and}\ ...\ \textbf{and}\ g_{jq}(a_i) \geq r_{jq}\ \textbf{and}\ ...)\ \textbf{then}\ a_i \in Cl_r^{\geq}$$

$$\textbf{if}(g_{j1}(a_i) \leq r_{j1}\ \textbf{and}\ ...\ \textbf{and}\ g_{jq}(a_i) \leq r_{jq}\ \textbf{and}\ ...)\ \textbf{then}\ a_i \in Cl_r^{\leq}$$

where g_{j1}, $g_{jq} \in G$. Such rules represent robust knowledge in the form of a conjunction of some performances on a subset of criteria that would justify the assignment of an alternative to a particular class union. In particular, the assignment of a_i to $Cl_r^{\geq}$ could be motivated by attaining some lower performance thresholds on a subset of criteria. For the induction of a non-redundant set of minimal decision rules, one can apply different algorithms (for a review, see Błaszczyński et al. (2011)). Among these algorithms, DOMLEM (Greco et al., 2000) is the one that has been originally coupled with DRSA.

3.1.6. Induction of decision rules

When inducing a minimal set of certain decision rules, the main procedure of DOMLEM (see Algorithm 1) is iteratively repeated for all lower approximations of the upward (downward) class unions. To meet the requirement of non-redundancy, the main procedure for rule induction (see Algorithm 2) is repeated starting from the strongest class union (e.g., in case of upward class union, from $\underline{P}(Cl_t^{\geq})$ to $\underline{P}(Cl_2^{\geq})$).

Algorithm 1: The **DOMLEM** algorithm inducing a minimal set of certain decision rules from the upward unions of decision classes

input: L_{upp} – a family of lower approximations of upward unions of decision classes: $\{\underline{P}(Cl_t^{\geq}), \underline{P}(Cl_{t-1}^{\geq}), ..., \underline{P}(Cl_2^{\geq})\}$;
output: $R_{\geq}$ – a minimal set of minimal decision rules covering all alternatives from L_{upp};

```
begin
    R≥ := ∅; (initialize a set of rules)
    for each B ∈ L_upp do (iteratively induce a minimal set of rules for each lower approximation of class union, starting from the strongest one)
    begin
      E := InduceRules(B);
      for each rule E ∈ E do
        if E is minimal rule then R≥ := R≥ ∪ E; (verify the non-redundancy of each rule)
    end
end
```

In the rule induction algorithm, E denotes a complex (i.e., conjunction of elementary conditions e) being a candidate for a condition part of the rule and $[E]$ denotes a set of alternatives matching E. When evaluating the candidates for the best condition part of a rule using function *evaluate*(E), DOMLEM selects complex E with the highest ratio $|[E] \cap W|/|[E]|$ and in case of a tie between various candidates – it selects complex E with the highest ratio $|[E] \cap W|$, where $W \subseteq B$ is a set of alternatives to be covered by the rules which are yet to be inferred and B is the considered approximation. Complex E is accepted as a condition part of the rule if and only if $[E] = \bigcap_{e \in E} [e] \subseteq B$. Since DOMLEM is a greedy algorithm, it needs to verify whether some potentially redundant conditions from the condition part should be eliminated. The rules are induced iteratively until all alternatives in B are covered by at least one rule.

Algorithm 2: The ***InduceRules*** function

input: B – a set of alternatives to be covered with decision rules; $P \subseteq G$ – a set of criteria;
output: **E** – a set of minimal rules covering set B;

begin
  $W := B$; (a set of alternatives from the given approximation)
  E := ∅; (initialization of a set of rules)
  while $W \neq \emptyset$ ***do*** (continue induction of decision rules until all alternatives are covered)
  begin
    $E := \emptyset$; (starting complex)
    $S := W$; (a set of alternatives currently covered by E)
    while ($E \neq \emptyset$) ***or not*** ($[E] \subseteq B$) ***do***
    begin
      best := ∅; (the best candidate for elementary condition)
      for each *criterion* $g_j \in P$ ***do begin***
        Cond := { $g_j(a) \geq r_{g_j}$ such that $\exists a \in S : g_j(a) = r_{g_j}$ }; (for each alternative from S create an elementary condition)
        for each *elem* ∈ *Cond* ***do***
        if *evaluate*({*elem*} ∪ E) *is better than evaluate*({*best*} ∪ E) ***then*** *elem := best;*
      end *(for)*
      $E := E \cup \{best\}$;
      $S := S \cap [best]$;
    end (***while not*** ($[E] \subseteq B$))
    for each *elementary condition* $e \in E$ ***do***
      if $[E/\{e\}] \subseteq B$ ***then*** $E := E/\{e\}$;
    create a rule based on E;
    E := **E** ∪ {E}; (add the induced rule)
    $W := B / \cup_{E \in \mathbf{E}}[E]$; (remove alternatives covered by the rule)
  end (while $W \neq \emptyset$)
end

In a default setting, when searching for the conditions to be included in a decision rule, DOMLEM is allowed to examine all criteria in G. However, its run may be also limited to account only for the criteria from a subset $P \subseteq G$. In particular, when constraining the application of DOMLEM to some reduct, a set of rules reproducing the assignments of all reference alternatives can be induced successfully, because the attributes contained in each reduct are sufficient to discriminate the alternatives contained in the lower approximation of each class union from the remaining ones.

3.1.7. Classification

The induced rules serve as a decision model that can be employed for deriving a classification for the non-reference alternatives in A^{NR}. The classification algorithm proceeds first by identifying the decision rules, which match the performances of a given alternative. It assumes that for each alternative there is at least one rule that matches its performances. In case this condition is not met, one can consider the rules that are "nearest" to the alternative in terms of a valued closeness relation (Słowiński and Stefanowski, 1994). It is based on the assumption that providing the recommendation based on the analysis of reference alternatives that do not significantly differ from the non-reference alternative under analysis can be the best compromise, instead of offering nothing (Cinelli et al., 2019). For example, alternative a_i with the following description: $g_1(a_i) = 2$, $g_2(a_i) = 2$ and $g_3(a_i) = 2$ matches 2 out of 3 conditions of the following rule, hence being relatively close to it: *if* $g_1(a_i) \geq 1$ and $g_2(a_i) \geq 3$ and $g_3(a_i) \geq 0$ *then* $a_i \in Cl_3^{\geq}$.

Since the assignments suggested by the rules matching a given alternative can be different or even conflicting, one needs to apply a classification scheme that would indicate support given to each potential classification. While there exist different classification procedures (see, e.g., Kadziński et al. (2014)), the so-called *new classification algorithm* (Błaszczyński et al., 2007) is the most widely used in real-world applications involving DRSA. The algorithm uses a set of rules R to compute the strength of confidence $Score_R^{net}(Cl_r, a_i)$ given to the assignment of alternative a to class Cl_r, $r = 1, ..., t$. Such a net score is defined as the difference between the positive $Score_R^{+}(Cl_r, a_i)$ and negative $Score_R^{-}(Cl_r, a_i)$ scores. The former is derived from the analysis of rules supporting the assignment of a to Cl_r, whereas the latter accounts for the rules that suggest the assignment to a class other than Cl_r. For a given alternative, the class with the highest net score is recommended.

Overall, an ability to deal with inconsistent preference data, implementation of an intuitive paradigm of learning from examples as well as an incorporation of a transparent preference model in the form of decision rules imply that DRSA has been widely used for supporting the solution of real-world decision problems in various application domains. Some recent case studies that have been successfully approached with DRSA concern a classification of brownfield remediation projects (Han et al., 2018), an evaluation of requalification strategies of farm building (Ottomano Palmisano et al., 2016), a diagnosis of diabetic retinopathy (Saleh et al., 2018), an analysis of customer behavior in the airline market (Liou and Tzeng, 2010), and a hazard assessment of energy accidents (Cinelli et al., 2019).

Nevertheless, the original DRSA has some weaknesses that limit its applicability and reduce the faithfulness of the provided results. In this paper, we aim at addressing the following three research gaps by proposing a novel variant of DRSA that will be used in a case study concerning the comprehensive evaluation of urban development agreements. Firstly, DRSA has been designed for reasoning about data with monotonicity constraints (Greco et al., 2001). Consequently, it is not suitable for analyzing the nominal attributes for which preference direction cannot be specified. Adjusting the approach for dealing with

both criteria and attributes at the same time requires a revision of the dominance relation and an adaptation of the consistency rule used in the phase of data structuring. Secondly, DRSA has been originally coupled with an application of a single minimal set of decision rules (e.g., inferred with the DOMLEM algorithm). As noted by Kadziński et al. (2014), the example assignments specified for a set of reference alternatives represent indirect preference information, and hence they can be potentially explained with different subsets of conditions. This, in turn, leads to the potential existence of multiple sets of rules that would reproduce the desired assignments for reference alternatives (Kadziński et al., 2016). However, when applied to the classification of non-reference alternatives, these sets could suggest different assignments. Accounting for the multiplicity of such sets of rules is particularly important when dealing with relatively small sets of decision examples. The rules capture some patterns in the learning data and any peculiar pattern can lead to some hardly justifiable recommendation for a non-reference alternative. Such a risk is vastly reduced when different compatible sets of rules and hence various patterns observed in data are exploited. Then, the robustness of different assignments can be examined and the class with the greatest support given the multiplicity of different compatible sets of rules can be finally recommended for each non-reference alternative. Thirdly, even if the multiplicity of minimal sets of rules has been already exploited in some previous works (see, e.g., Kadziński et al., 2014, 2015), the latter ones do not differentiate the impacts of different sets of rules when working out the recommendation for the non-reference alternatives. However, it is natural to make the weights assigned to different sets of rules dependent on their predictive abilities. Specifically, greater weights should be associated with these sets of rules, which prove to make less or smaller errors when predicting the classification for alternatives that were not used at the stage of inferring these sets.

3.2. The novel variant of Dominance-based Rough Set Approach

In this paper, we propose the application of a novel variant of the DRSA, whose modifications with respect to the original one are presented in this section.

3.2.1. Notation

Our formulation of the decision problem involves both preference-ordered criteria $G_C \subset G$ and nominal attributes $G_A \subset G$ for which preference directions cannot be specified ($G_C \cup G_A = G$).

3.2.2. Approximations of class unions

Since the dominance relation Δ^G cannot be analyzed alone in the context of nominal attributes, we consider a dominance-indifference relation that will be used to compare the alternatives in the following way:

$a_i \Delta_I^G a_k \Leftrightarrow \forall\, g_j \in G_C \cap G : g_j(a_i) \geq g_j(a_k)$ and $\forall\, g_j \in G_A \cap G : g_j(a_i) = g_j(a_k)$.

Hence, a pair of alternatives is related by Δ_I^G if one of them is at least as good on all criteria in G_C and they are indifferent in terms of all nominal attributes in G_A. Otherwise, if the order of alternatives' performances on at least two criteria is inverse or their performances on at least one attribute are different, such a pair is deemed as incomparable in terms of Δ_I^G. Then, the approximations of class unions are obtained analogously as in the original DRSA by replacing Δ^P with Δ_I^P with $P \subseteq G$ in the following way:

$$\underline{P}(Cl_r^{\geq}) = \{a_i \in A^R: \nexists a_k \in Cl_{r-1}^{\leq}, a_k \Delta_I^P a_i\} \text{and} \overline{P}(Cl_r^{\geq}) = \{a_i \in A^R: \exists\, a_k \in Cl_r^{\geq},\ a_i \Delta_I^P a_k\}$$

$$\underline{P}(Cl_r^{\leq}) = \{a_i \in A^R: \nexists a_k \in Cl_{r+1}^{\geq}, a_i \Delta_I^P a_k\} \text{and} \overline{P}(Cl_r^{\leq}) = \{a_i \in A^R: \exists\, a_k \in Cl_r^{\leq},\ a_k \Delta_I^P a_i\}$$

3.2.3. Decision rules

The form of decision rules that involve conditions referring to the performances on both criteria and attributes is as follows:

$$\textbf{if}(g_{j1}(a_i) \geq r_{j1} \textbf{ and} \ldots \textbf{and} g_{jq}(a_i) = r_{jq} \textbf{ and} \ldots) \textbf{then} a_i \in Cl_r^{\geq}$$

$$\textbf{if}(g_{j1}(a_i) \leq r_{j1} \textbf{ and} \ldots \textbf{and} g_{jq}(a_i) = r_{jq} \textbf{ and} \ldots) \textbf{then} a_i \in Cl_r^{\leq}$$

where $g_{j1} \in G_C$ and $g_{jq} \in G_A$. The rules are induced using a suitably modified variant of DOMLEM (Greco et al., 2000), which constructs the candidates for the elementary conditions (see Algorithm 2) in the following way:

$Cond := \{g_j(a) = r_{g_j} \text{if} g_j \in G_A \text{ or } g_j(a) \geq r_{g_j} \text{if} g_j \in G_C \text{ such that } \exists\, a \in S: g_j(a) = r_{g_j}\}$.

3.2.4. Robustness analysis

In the basic variant of the proposed method, the recommendation for each non-reference alternative is based on the application of a single set of rules arbitrarily selected by a suitably modified DOMLEM coupled with the new classification algorithm (Błaszczyński et al., 2007). However, to verify the robustness of the assignments obtained in this way, we investigate the stability of classification suggested by different sets of rules that can be induced from the set of assignment examples. Specifically, these sets of rules $R \in \mathscr{R}$ are generated by DOMLEM, which is constrained to use the attributes from a single reduct at a time. In this way, each constructed set of rules is guaranteed to refer to a minimal subset of attributes and criteria capable of explaining the assignments of reference alternatives.

To measure the stability of recommendation suggested by all sets of rules in $\mathscr{R}$ for the non-reference alternatives, one normally refers to a Class Acceptability Index (CAI) defined as follows (Kadziński et al., 2016):

$$CAI(a_i, r) = \frac{\sum_{R \in \mathscr{R}} m(R, a_i, r)}{card(\mathscr{R})}$$

where $m(R, a_i, r)$ is the class membership function, which confirms the assignment of alternative a_i to class Cl_r by set of rules R (Kadziński et al., 2018):

$$m(R, a_i, r) = \begin{cases} 1, & \text{if } r = \arg\max_{h=1,\ldots,t} Score_R^{net}(Cl_h, a_i), \\ 0, & \text{otherwise}. \end{cases}$$

The closer $\boldsymbol{CAI}(\boldsymbol{a_i}, \boldsymbol{r})$ is to one, the more robust is the assignment of $\boldsymbol{a_i}$ to $\boldsymbol{Cl_r}$, being confirmed by different sets of rules compatible with the classification examples. Thus defined $\boldsymbol{CAI}$ assumes that equal weights are assigned to all considered sets of rules $\boldsymbol{R} \in \mathcal{R}$. However, it is reasonable to differentiate these weights $\boldsymbol{w_R}$, $\boldsymbol{R} \in \mathcal{R}$, so that greater voting power is assigned to the sets $\boldsymbol{R} \in \mathcal{R}$ with greater predictive ability. To assess the latter one, we apply a leave-one-out cross-validation (Geisser, 1993). Such a choice is motivated by a small cardinality of the learning set in the considered case study. Specifically, when assessing the predictive ability of a set of rules based on a particular reduct, we divide a set of reference alternatives $\boldsymbol{A^R}$ into a training set composed of $\boldsymbol{card}(\boldsymbol{A^R}) - 1$ alternatives and a single validation alternative. For the latter one, the recommendation suggested with a set of rules inferred based on the $\boldsymbol{card}(\boldsymbol{A^R}) - 1$ alternatives is compared with its pre-defined assignment. The same process is applied once for each alternative playing the role of a single-item test set and using all remaining alternatives as a training set. The results are averaged over all such runs. In particular, we consider the mean absolute error (Willmott and Matsuura, 2005) defined as follows:

$$\boldsymbol{MAE}(\boldsymbol{R}) = \frac{\sum_{\boldsymbol{a_i} \in \boldsymbol{A^R}} |\boldsymbol{Cl_i} - \boldsymbol{y_i^R}|}{\boldsymbol{card}(\boldsymbol{A^R})}$$

where Cl_i and y_i^R are indices of classes to which $a_i \in A^R$ is assigned by, respectively, the DM or the classification algorithm based on a set of rules $R \in \mathscr{R}$. Then, weight w_R assigned to $R \in \mathscr{R}$ can be defined as a reciprocal of $MAE(R)$ (i.e., $w_R = 1/MAE(R)$) so that to favor the sets of rules with better predictive abilities. Even though it is possible to use other performance measures than MAE (e.g., classification accuracy or

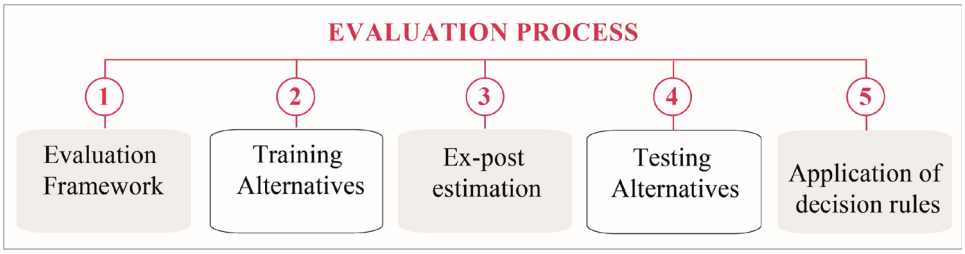


Fig. 1. Main steps of the evaluation process.

mean squared error), the general idea of the proposed weighting scheme remains the same.

Finally, instead of quantifying the results of robustness analysis with *CAI* s, we consider Weighted Class Acceptability Indices (*CAI* s). They are defined for $a_i \in A$ and Cl_r, $r = 1, ..., t$, in the following way:

$$WCAI(a_i, r) = \frac{\sum_{R \in \mathscr{R}} w_R \bullet m(R, a_i, r)}{\sum_{R \in \mathscr{R}} w_R}$$

Overall, $WCAI(a_i, r)$ gets higher in case an assignment of a_i to Cl_r is confirmed by more compatible sets of rules which proved to perform better in the validation experiment.

4. Case study

Firstly introduced in 1986 in the Lombardy Region, UDAs have enhanced urban transformations according to the following main features: i) institutional cooperation between different government levels, ii) subsidiarity, iii) stakeholders' involvement for the definition of strategies and actions, iv) local public-private partnership, v) definition of the agreement between the two parties on the basis of the increase in land value generated by the urban development, vi) public investments' efficiency and effectiveness, vii) functional mix, and viii) achievement of environmental and social objectives (Oppio and Torrieri, 2018). Given these specific characteristics, they show a high degree of complexity with their success being affected by several aspects across long administrative and development processes. When considering the UDAs launched in the last 20 years, only 14 % of them have been concluded, 35 % are under development, and the remaining ones have failed for many reasons involving inadequacy of the basic economic assumptions concerning the real estate market trends over long development phases (almost ten years).

Within this context, DRSA has been applied to point out the main features of UDAs and to find out which characteristics can be considered as the drivers of urban development for the additional value generated for both private developers and public administrations. The research has been developed into the following five main steps (see Fig. 1):

- definition of the evaluation framework;
- evaluation of the performances of a sample of 15 concluded UDAs (called a set of *training alternatives*);
- ex-post estimation of the additional value generated by UDAs;
- application of DRSA to the training alternatives;
- application of the evaluation model and identification of urban development drivers, to support 10 on-going negotiations (called a set of *testing alternatives*) between the private developers and the public administrations.

These steps are described in detail in the following subsections.

4.1. Evaluation framework

According to a constructivist approach, the evaluation framework (see Fig. 2 and Table 1) has been defined by the interaction with a group of urban planning experts and regional delegates, with specific knowledge on this topic (Rita et al., 2018; Brito et al., 2019). Five thematic areas have been identified as essential, and subsequently detailed by a pair of attributes (A2 and B1) and 9 criteria (A1, B2, C1, C2, D1, D2, E1, E2, and E3), respectively, without and with preference directions on the respective performance scales. Criteria and attributes have been selected according to data and information actually available from Lombardy Region according to regional law recommendations.

Specifically, the first thematic area "**Urban Context (A)**" includes *Accessibility* (A1) with respect to the public transport system and private mobility and *Development site* (A2), which considers the sites' previous functions (greenfield, industrial brownfield, railway brownfield, urban brownfield, mixed) and their environmental as well as economic impacts. The "**Institutional context (B)**" accounts for the characteristics of *Developers* (B1) and *Ownership* (B2). In both cases, they can be

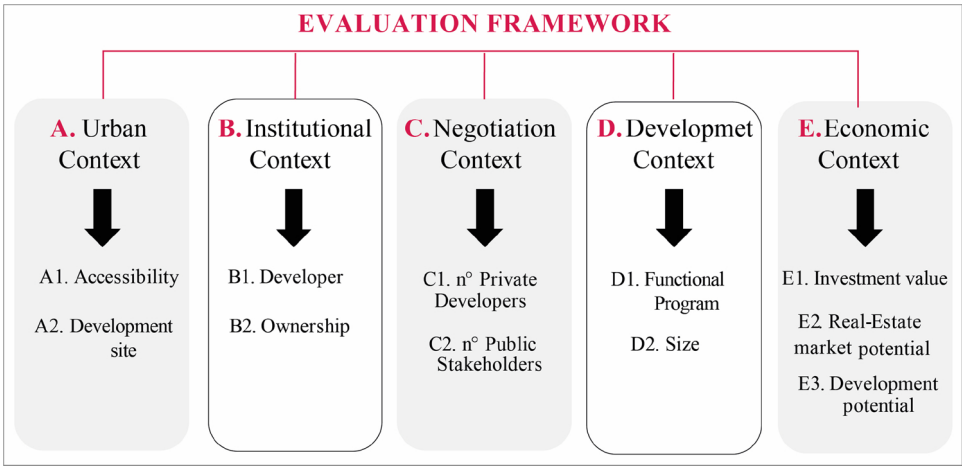


Fig. 2. Evaluation framework used to evaluate the Urban Development Agreements.

Table 1
Structure of the decision problem.

	Attribute / Criterion		Type	Scale / Unit	Definition
A	**Urban Context**				
A1	Accessibility		Gain	Qualitative	The public and private mobility assessed by considering the presence of train stations, metro stations, bus stops, highways, primary and secondary roads according to 10 min isochrones
A2	Development site		–	Nominal	Greenfield, industrial brownfield, railway brownfield, urban brownfield, or mixed
B	**Institutional context**				
B1	Developer		–	Nominal	The nature of the subject involved in the urban development project: public, private or mixed
B2	Ownership	Private	Gain	%	The percentage of public and private bodies owning the site
		Public	Cost	%	
C	**Negotiation context**				
C1	Number of private developers		Gain	n°	The number of private developers
C2	Number of public stakeholders		Cost	n°	The number of public stakeholders
D	**Development context**				
D1	Functional program		Gain	%	The incidence of commercial GFA on the total developed area
D2	Size		Gain	sqm	The GFA of the project, the overall area of the development site and the building ratio index
E	**Economic context**				
E1	Investment value		Cost	€	The total development cost
E2	Real-estate market potential		Gain	Qualitative	Comparison between the average market prices of the municipalities and the market prices of the UDAs areas. The OMI database (Agenzia delle Entrate) has been investigated for the market prices.
E3	Development potential		Gain	cm/sqm	Difference between the building index ratio of the site before and after the development

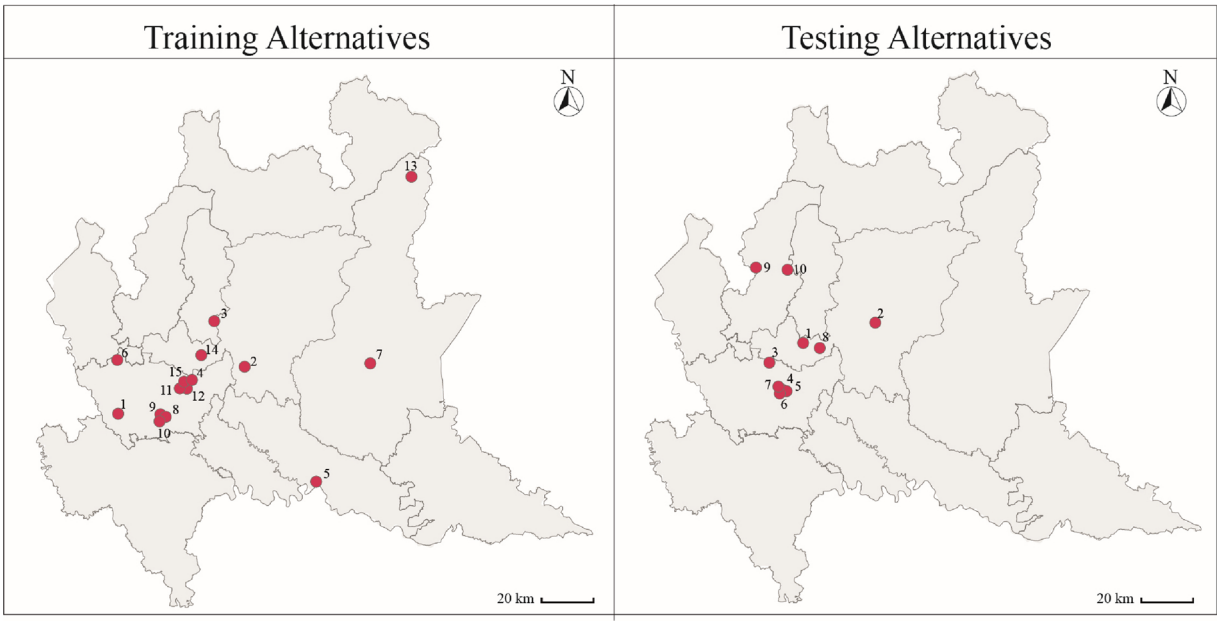


Fig. 3. The location of Urban Development Agreements under evaluation in Lombardy.

private, public or mixed. The "**Negotiation Context (C)**" deals with the complexity of the agreement, given by the *Numbers of involved private developers* (C1) and *Public Stakeholders* (C2). The "**Development Context (D)**" considers the intrinsic features of UDAs such as the *Functional program* (D1), with a special focus on the gross floor area (GFA) of each type of uses (residential, social housing, offices, retail, services, green areas, others) as well as *Size* (D2) measured by the GFA, i.e., the total area of development. Finally, the "**Economic Context (E)**" refers to the *Investment value* (E1) given by the sum of hard and soft costs of urban development, *Real estate market potential* (E2), which measures the attractiveness of the development under investigation with respect to the future market values, and *Development potential* (E3) interpreted as the difference between the current development expectation and the new one defined within the UDA. The UDAs performances in terms of all attributes and criteria were assessed regarding the financial report, which is one of the mandatory documents a private subject has to submit to the municipality and the regional authority together with the development proposal.

The selected criteria have different preference directions (see Table 1). On the one hand, accessibility (A1), the amount of private ownership (B2), the presence of multiple private stakeholders (C1), the percentage of commercial GFA in the total development area (D1), the size of the lot (D2), the positioning of UDAs' area market prices with respect to the average urban values (E2), and increases of building index (E3) are considered as gain criteria (the higher the performance, the better), since they all enhance urban development by reducing the investment risk (B2, C1), attracting users (A1, D1), supporting adequate development in terms of buildable area and potential incomes (D2, E2, E3). On the other hand, the public ownership and the coexistence of multiple public stakeholders (B2, C2) are thought to impact negatively on the administrative and bureaucratic process, thus increasing the investment value due to higher costs and delayed incomes (E1). For these reasons, they are considered as cost criteria with lesser performances being more preferred.

Table 2
Urban Development Agreements (training and testing alternatives) under evaluation: provinces and size.

Concluded (training) UDAs				On-going (testing) UDAs			
UDA	Province	Size (sqm)	Volume (cm)	UDA	Province	Size (sqm)	Volume (cm)
A1	Milan	17.527	40.000	CS1	Milan	19.730	25.290
A2	Bergamo	95.100	53.000	CS2	Bergamo	20.727	11.869
A3	Lecco	9.955	22.000	CS3	Pavia	163.493	106.338
A4	Milan	32.035	54.000	CS4	Milan	266.183	446.286
A5	Cremona	283.440	124.431	CS5	Milan	151.257	238.035
A6	Milan	128.345	180.000	CS6	Milan	1.111.573	1.831.002
A7	Brescia	25.436	37.600	CS7	Milan	230.338	689.079
A8	Milan	97.600	93.262	CS8	Monza Brianza	87.980,31	48.000
A9	Milan	86.204	78.762	CS9	Varese	44.998,58	75.627
A10	Milan	88.539	53.200	CS10	Varese	19.311	26.850
A11	Milan	44.290	57.222				
A12	Milan	138.878	85.815				
A13	Bergamo	11.493	11.176				
A14	Milan	2.090	8.158				
A15	Milan	225.403	329.080				

4.2. Sets of concluded and on-going Urban Development Agreements

In the last years, Lombardy – like many other European regions and cities – has experienced the renewal of abandoned areas and brownfields through the innovative forms of cooperation between the public and private sectors. Based on the analysis of UDAs carried out over the last fifteen years 15 case studies have been selected as a set of training alternatives. They represent successful experiences since they have been completed. A significant part of these case studies are located in the Province of Milan and in 13 different medium-low populated municipalities of Lombardy (see Fig. 3 and Table 2). The private and public functions of these UDAs correspond to, respectively, 88 % and 12 % of the total gross floor area developed. Among the private functions, the greatest share is represented by residential buildings (69 %), retail (17 %), and offices (7%).

The testing sample includes 10 UDAs under agreement. They are located in 5 different municipalities in Lombardy (see Fig. 3 and Table 2). Since the negotiation is still on-going in their context, finding out the drivers of value is essential to address the agreement toward achieving a balanced negotiation between the public and the private.

4.3. Estimation of Urban Development Agreements' value

The early detection of the territorial, institutional, negotiation as well as economic features, which are responsible for the positive gap between potential revenues and costs, being the allocation of this value at stake, can support the agreement between the private developers and the public administrations in favor of fair and balanced choices. To find out the drivers of successful urban development, the planning gain (*UDA Value*) generated by UDAs has been estimated. The estimation of such a value plays a crucial role in negotiation processes, as the agreements between private developers and public administrations establish how to allocate this value between the two parties. On the one hand, private developers consider as legitimate the return given by their proposals and entrepreneurial capabilities. On the other hand, public authorities require a percentage of developers' return, i.e., so-called land value capture, being dependent on cross-synergies and cross-externalities created through investments and decisions of both the public and the private (Camagni, 2008).

Thus, the starting point of the negotiation process is the estimation of UDA Value. According to the notion of planning gain (Bowers, 1992), the *UDA Value* can be estimated as the difference between the values of the area before and after UDA (denoted by, respectively, V_0 and V_f), i.e.: $UDA\ Value = V_f - V_0$. Both V_f and V_0 can be appraised by considering the difference between the potential revenues and the development costs according to both the Residual Value approach and the Discounted Cash Flow Analysis model. While the former is relatively simple in use and widely understood, the latter is appropriate for longer-term development schemes, since it accounts for the distribution of costs and revenues over time. This assumption implies the potential use of a discount rate that reflects the investment's gross risk/return (Oppio et al., 2018).

Table 3 shows an overview of the estimated *UDA Value* (column A), the estimated value captured by public authorities, namely 50 % of the *UDA Value* according to the minimum incidence suggested by the national law (column B), and the land value actually captured by public authorities, calculated as the difference between the planning fees defined by the agreements and the ones due to local planning law and regulations (column C). As it emerges from Table 3, in most cases decisions about the UDA's value allocation rule are made case by case, according to the variations of functional programs defined over long horizon time. This high degree of discretion confirms the difficulty for the public authorities to estimate the adequacy of the value capture,

Table 3
Training sample: overview of UDA Value (A), estimated land value capture (B), real land value capture (C).

UDA	A	B	C
A1	668.673 €	334.337 €	244.853 €
A2	3.625.155 €	1.812.578 €	4.800.056 €
A3	549.695 €	274.848 €	606.006 €
A4	8.290.369 €	4.145.185 €	921.848 €
A5	4.318.798 €	2.159.399 €	7.434.044 €
A6	20.193.173 €	10.096.587 €	8.913.342 €
A7	1.576.193 €	788.097 €	2.745.133 €
A8	1.030.662 €	515.331 €	1.421.428 €
A9	1.734.131 €	867.066 €	2.366.773 €
A10	2.955.971 €	1.477.986 €	2.163.906 €
A11	7.219.064 €	3.609.532 €	2.570.041 €
A12	14.118.188 €	7.059.094 €	32.167 €
A13	769.218 €	384.609 €	125.143 €
A14	1.328.282 €	664.141 €	221.896 €
A15	16.854.746 €	8.427.373 €	5.969.732 €

Table 4
Definition of decision classes in terms of the ratio between the value generated by UDAs and the total incomes.

Decision class	Value/Income
Class 1 (Cl_1)	$\leq$ 10 %
Class 2 (Cl_2)	10 % $<$ x $\leq$ 20 %
Class 3 (Cl_3)	20 % $<$ x $\leq$ 30 %
Class 4 (Cl_4)	x $>$ 30 %

Table 5
Performance matrix of the concluded (training) UDAs.

					Concluded (training) UDAs					
	Condition attributes		G/C	U.M.	A1	A2	A3	A4	A5	A6
A	**Urban Context**									
A1	Accessibility		G	qualitative	high	medium	high	medium	high	high
A2	Development site		/	nominal	Industrial brownfield	Urban brownfield	Industrial brownfield	Industrial brownfield	Industrial brownfield	Industrial brownfield
B	**Institutional context**									
B1	Developer		/	nominal	private	private	private	private	private	private
B2	Ownership	private	G	%	100 %	52 %	100 %	98 %	71 %	98 %
		public	C	%	0%	48 %	0%	2%	29 %	2%
C	**Negotiation context**									
C1	Number of private developers		G	n°	1	1	1	2	2	3
C2	Number of public stakeholders		C	n°	2	2	2	2	3	3
D	**Development context**									
D1	Functional program		G	%	5%	8%	14 %	7%	93 %	24 %
D2	Size		G	sqm	13333,3	17666,7	7333,3	18000,0	41477,0	60000,0
E	**Economic context**									
E1	Investment value		C	€	€ 16.897.945,17	€ 17.369.295,78	€ 8.138.212,16	€ 22.139.938,69	€ 45.653.653,08	€ 66.570.569,06
E2	Real-estate market potential		G	qualitative	high	medium	high	high	low	high
E3	Development potential		G	cm/sqm	−0,29	0,42	2,30	1,42	0,74	1,63
	Overall score									
	Value/ Income				6%	27 %	11 %	39 %	14 %	35 %
	Classes				**1**	**3**	**2**	**4**	**2**	**4**

	Concluded (training) UDAs								
	A7	A8	A9	A10	A11	A12	A13	A14	A15
A									
A1	low	medium	high	medium	high	medium	low	medium	high
A2	Industrial brownfield	Mixed	Mixed	Greenfield	Urban brownfield	Mixed	Urban brownfield	Urban brownfield	Greenfield
B									
B1	private	private	private	private	private	private	mixed	private	mixed
B2	65 %	100 %	58 %	56 %	68 %	72 %	77 %	100 %	100 %
	35 %	0%	42 %	44 %	32 %	28 %	23 %	0%	0%
C									
C1	1	2	1	2	2	1	1	1	1
C2	2	2	2	2	2	2	2	2	2
D									
D1	0%	0%	0%	5%	80 %	7%	0%	25 %	0%
D2	12533,3	31087,3	26254,0	17733,3	19074,0	28605,0	3725,3	2719,3	109693,3
E									
E1	€ 12.239.411,22	€ 35.815.952,90	€ 30.600.681,12	€ 23.105.833,14	€ 20.779.524,58	€ 41.385.270,65	€ 3.276.970,86	€ 3.428.215,04	€ 121.591.920,83
E2	high	low	low	high	high	high	medium	high	low
E3	0,51	−0,14	0,91	0,65	0,17	0,55	0,72	0,00	0,00
	18 %	5%	9%	19 %	38 %	37 %	29 %	40 %	19 %
	2	**1**	**1**	**2**	**4**	**4**	**3**	**4**	**3**

Table 6
Performance matrix of the on-going (testing) UDAs.

					On-going (testing) UDAs		
	Condition attributes		G/C	U.M.	CS1	CS2	CS3
A	**Urban Context**						
A1	Accessibility		G	qualitative	high	medium	low
A2	Development site		/	nominal	Mixed	Urban brownfield	Industrial brownfield
B	**Institutional context**						
B1	Developer		/	nominal	Private	Private	Private
B2	Ownership	private	G	%	55 %	100 %	100 %
		public	C	%	45 %		
C	**Negotiation context**						
C1	Number of private developers		G	n°	1	1	1
C2	Number of public stakeholders		C	n°	2	2	2
D	**Development context**						
D1	Functional program		G	%	0%	0%	1%
D2	Size		G	sqm	8.430	4.035	41.520
E	**Economic context**						
E1	Investment value		C	€	9.753.175,01 €	4.897.413,92 €	46.651.817,94 €
E2	Real-estate market potential		G	qualitative	high	low	medium
E3	Development potential		G	cm/sqm	0,30	0,19	0,25

	On-going (testing) UDAs						
	CS4	CS5	CS6	CS7	CS8	CS9	CS10
A							
A1	high	high	high	high	low	high	high
A2	Industrial brownfield	Urban brownfield	Urban brownfield	Urban brownfield	Greenfield	Industrial brownfield	Industrial brownfield
B							
B1	Private	Mixed	Private	Private	Private	Private	Private
B2	100 %	85 % 25 %	91 % 9%	22 % 78 %	100 %	100 %	100 %
C							
C1	3	1	1	2	1	1	1
C2	2	2	2	2	2	2	2
D							
D1	13 %	37 %	4%	10 %	5%	6%	93 %
D2	151.725	50.493	828.985	229.693	19.492	20.123	8.950
E							
E1	168.897.320,46 €	65.226.865,02 €	1.044.156.410,52 €	325.039.153,12 €	22.693.203,86 €	31.743.424,30 €	14.592.974,70 €
E2	low	low	low	high	high	high	low
E3	0,57	0,33	0,75	1,00	0,22	0,45	0,46

Table 7
Approximations of the downward and upward class unions.

Class union	Alternatives	Class union	Alternatives
$Cl_2^{\geq}$	A4 A6 A11 A12 A14 A2 A13 A15 A3 A5 A7 A10	$Cl_1^{\leq}$	A1 A8 A9
$Cl_3^{\geq}$	A4 A6 A11 A12 A14 A2 A13 A15	$Cl_2^{\leq}$	A1 A8 A9 A3 A5 A7 A10
$Cl_4^{\geq}$	A4 A6 A11 A12 A14	$Cl_3^{\leq}$	A1 A8 A9 A3 A5 A7 A10 A2 A13 A15

Table 8
All reducts constructed for the information table analyzed within the case study, mean absolute errors and respective weights assigned to the sets of decision rules based on the corresponding reducts.

Reduct	Attributes	MAE	Weight	Reduct	Attributes	MAE	Weight
RED_1	D1 D2 E1	1.167	0.857	RED_{12}	A1 C1 E2 E3 A2	1.011	0.989
RED_2	D2 E1 E2	0.944	1.058	RED_{13}	D1 D2 E3 A2	1.111	0.900
RED_3	A1 D2 E1 E3	1.433	0.697	RED_{14}	D2 E1 E3 B1	1.511	0.661
RED_4	B2 C1 E3 A2	1.467	0.681	RED_{15}	D2 E2 E3 A2	0.800	1.250
RED_5	B2 D2 E1 E3	1.311	0.762	RED_{16}	C1 D1 E1 A2 B1	1.100	0.909
RED_6	A1 C1 D1 E1 A2	1.133	0.882	RED_{17}	B2 C2 D1 D2 E3 B1	1.144	0.873
RED_7	B2 C1 D1 E1 A2	1.167	0.857	RED_{18}	C1 D1 E3 A2 B1	1.044	0.957
RED_8	B2 D2 E3 A2	1.378	0.725	RED_{19}	C1 E1 E2 A2 B1	0.800	1.250
RED_9	A1 C1 D1 E3 A2	1.100	0.909	RED_{20}	D1 D2 E2 E3 B1	0.911	1.097
RED_{10}	A1 C1 E1 E2 A2	0.944	1.370	RED_{21}	C1 E2 E3 A2 B1	1.200	0.833
RED_{11}	B2 C1 E1 E2 A2	1.067	0.937				

concerning both the overall value generated by the urban transformations and sustainable growth objectives, when they enter into negotiation with the private parties.

To support the definition of the agreements, according to the analysis of the UDAs' main features and to the value they generate, four preference ordered classes have been specified. These decision classes have been defined by considering the incidence of each *UDA Value* on the total amount of potential revenues with the value captured by public administrations being a part of it (see Table 4). In particular, the worst (Cl_1) and the best (Cl_4) classes correspond to, respectively, at most 10 % and more than 30 % of the ratio between UDA's value and total potential revenues from the urban development. Such a discretization has been approved by the urban planning experts and regional delegates involved in the case study.

Within the framework of DRSA, the performances of 15 concluded (training) UDAs on the respective attributes or criteria (see Table 5) along with the assigned decision classes formed an input data set. It was analyzed to derive structured knowledge that has been subsequently used for the holistic evaluation (i.e., deriving class assignments) of 10 on-going UDAs, whose performances are given in Table 6.

From the performance matrices for the two groups of UDAs, it is possible to draw some preliminary conclusions. As far as Accessibility (A1) is concerned, both the concluded and on-going UDAs obtained relatively high evaluations. When it comes to the Development site (A2), the concluded UDAs are mainly characterized by industrial brownfields, while the on-going ones by urban brownfields. About the Institutional context, both groups of UDAs involve private developers (B1) and refer to private areas. In general, the negotiation takes place between two private developers (C1) and two public stakeholders (C2). The percentage of commercial function (D1) exceeds, in a few cases, 20 %, whereas the project size (D2) differs significantly from one UDA to another. High variability of performances is also typical for the Investment value (E1), where the minimal and maximal development costs are around 3 million (see A13 and A14) and 1 billion (see CS6), respectively. When it comes to the Real-estate market potential (E2), the concluded UDAs obtained on average high evaluation, while the on-going UDAs attained both low and high performances. As far as the Development potential (E3) is concerned, in the concluded UDAs both positive and negative performances are present, while for the on-going ones rather positive evaluations are observed. Such a high variability of performances for most of the attributes and criteria used to characterize the two groups of UDAs confirms the complexity of negotiations, which are generally run without any decision support system aimed to minimize the discretion of the choices.

4.4. Learning process

The analysis of classification provided for the concluded UDAs starts with the consistency check and computation of the approximations of class unions. Since four preference ordered classes are considered (Cl_1, Cl_2, Cl_3, and Cl_4), the following six unions are relevant: upward unions – $Cl_2^{\geq}$, $Cl_3^{\geq}$, and $Cl_4^{\geq}$, and downward unions – $Cl_1^{\leq}$, $Cl_2^{\leq}$, and $Cl_3^{\leq}$. The consistency check indicates that there is no UDA from a worse class that would dominate in terms of Δ_I^G any UDA from a better class. This, in turn, implies that the lower and upper approximations of all unions are equal, and composed of UDAs assigned to the class contained in a given union (see Table 7). For example, the approximations of union $Cl_2^{\leq}$ contain all reference UDAs assigned to either Cl_1 or Cl_2.

The absence of inconsistencies in the analyzed information table implies the perfect quality of approximation of classification (i.e., $\gamma_G(Cl) = 1$). This means that for all reference UDAs there is no reason to contest the assigned class at least concerning its consistency with Δ_I^G, and the constructed set of criteria and attributes is capable of fully explaining these assignments. Such a unitary quality of approximation can be also treated as a confirmation of the appropriate problem formulation, including construction of both a family of relevant attributes and a database with UDAs' performances.

The perfect quality of approximation of classification is also maintained when reducing a set of criteria in different ways. In Table 8, we present 21 different reducts that were identified for the analyzed information table. For example, when considering only D1, D2, and E1 (denoted as reduct RED_1), the quality of approximation would be preserved, which indicates that these three attributes can be used to distinguish the assignment of all 15 reference UDAs among the classes. The cardinalities of reducts range between 3 and 6.

Each reduct speaks in favor of the discriminatory role played by the attributes contained in it. The statistics concerning the numbers of various attributes appearing in 21 reducts differ vastly. More specifically, these numbers are as follows: A1 − 5, A2 − 14, B1 − 7, B2 − 6, C1 − 11, C2 − 1, D1 − 9, D2 − 10, E1 − 11, E2 − 8, and E3 − 13. Consequently, A2 (development site), E3 (development potential), E1 (investment value) and C1 (number of private developers) can be

Table 9
Decision rules induced with DOMLEM for the case study.

Symbol	Rule syntax / Interpretation of rule in the natural language	Supported alternatives
Rules for class at most Cl_1		
$r_1^{\leq 1}$	***if*** A2 = mixed *and* D1≤5.0 ***then*** $Cl_1^{\leq}$	A8 A9
	if the development site is mixed *and* functional program is at most 5% ***then*** class at most Cl_1	
$r_2^{\leq 1}$	***if*** A2 = industrial brownfield *and* C1≤1.0 *and* E1≥16897945.17 ***then*** $Cl_1^{\leq}$	A1
	if the development site is industrial brownfield *and* the number of private developers is at most 1 *and* investment value is at least 16897945.17 Euro ***then*** class at most Cl_1	
Rules for class at most Cl_2		
$r_1^{\leq 2}$	***if*** A2 = industrial brownfield *and* C1≤1.0 ***then*** $Cl_2^{\leq}$	A1 A3 A7
	if the development site is industrial brownfield *and* the number of private developers is at most 1 ***then*** class at most Cl_2	
$r_2^{\leq 2}$	***if*** A1≤medium *and* A2 = greenfield ***then*** $Cl_2^{\leq}$	A10
	if accessibility is at most medium *and* development site is greenfield ***then*** class at most Cl_2	
$r_3^{\leq 2}$	***if*** *A2* = industrial brownfield *and* C1≤2.0 *and* E1≥23105833.14 ***then*** $Cl_2^{\leq}$	A5
	if the development site is industrial brownfield *and* the number of private developers is at most 2 *and* investment value is at least 23105833.14 Euro ***then*** class at most Cl_2	
$r_4^{\leq 2}$	***if*** A2 = mixed *and* D1≤5.0 *then* $Cl_1^{\leq}$	A8 A9
	if the development site is mixed *and* functional program is at most 5% ***then*** class at most Cl_2	
Rules for class at most Cl_3		
$r_1^{\leq 3}$	***if*** D1≤5.0 ***then*** $Cl_3^{\leq}$	A1 A8 A9 A7 A10 A13 A15
	if functional program is at most 5% ***then*** class at most Cl_3	
$r_2^{\leq 3}$	***if*** C1≤2.0 *and* E1≥45653653.08 ***then*** $Cl_3^{\leq}$	A5 A15
	if the number of private developers is at most 2 *and* investment value is at least 45653653.08 Euro ***then*** class at most Cl_3	
$r_3^{\leq 3}$	***if*** A2 = urban brownfield *and* C1≤1.0 *and* E1≥17369295.78 ***then*** $Cl_3^{\leq}$	A2
	if the development site is urban brownfield *and* the number of private developers is at most 1 *and* investment value is at least 17369295.78 Euro ***then*** class at most Cl_3	
$r_4^{\leq 3}$	***if*** A2 = industrial brownfield *and* C1≤1.0 ***then*** $Cl_3^{\leq}$	A1 A3 A7
	if the development site is industrial brownfield *and* the number of private developers is at most 1 ***then*** class at most Cl_3	
Rules for class at least Cl_2		
$r_1^{\geq 2}$	***if*** D1≥7.0 ***then*** $Cl_2^{\geq}$	A3 A5 A2 A4 A6 A11 A12 A14
	if the functional program is at least 7% ***then*** class at least Cl_2	
$r_2^{\geq 2}$	***if*** E1≤12239411.22 ***then*** $Cl_2^{\geq}$	A3 A7 A13 A14
	if investment value is at most 12239411.22 Euro ***then*** class at least Cl_2	
$r_3^{\geq 2}$	***if*** A2 = greenfield ***then*** $Cl_2^{\geq}$	A10 A15
	if the development site is greenfield ***then*** class at least Cl_2	
Rules for class at least Cl_3		
$r_1^{\geq 3}$	***if*** A2 = urban brownfield ***then*** $Cl_3^{\geq}$	A2 A14 A11 A14
	if the development site is urban brownfield ***then*** class at least Cl_3	
$r_2^{\geq 3}$	***if*** A1≥high *and* A2 = greenfield ***then*** $Cl_3^{\geq}$	A15
	if accessibility is high *and* development site is greenfield ***then*** class at least Cl_3	
$r_3^{\geq 3}$	***if*** C1≥3.0 ***then*** $Cl_3^{\geq}$	A6
	if the number of private developers is at least 3 ***then*** class at least Cl_3	
$r_4^{\geq 3}$	***if*** C1≥2.0 *and* E1≤22139938.69 ***then*** $Cl_3^{\geq}$	A4 A11
	if the number of private developers is at least 2 *and* investment value is at most 22139938.69 Euro ***then*** class at least Cl_3	
$r_5^{\geq 3}$	***if*** A2 = mixed *and* D1≥7.0 ***then*** $Cl_3^{\geq}$	A12
	if the development site is mixed *and* functional program is at least 7% ***then*** class at least Cl_3	
Rules for class at least Cl_4		
$r_1^{\geq 4}$	***if*** C1≥2.0 *and* E1≤22139938.69 ***then*** $Cl_4^{\geq}$	A4 A11
	if the number of private developers is at least 2 *and* investment value is at most 22139938.69 Euro ***then*** class at least Cl_4	
$r_2^{\geq 4}$	***if*** A2 = mixed *and* D1≥7.0 ***then*** $Cl_4^{\geq}$	A12
	if the development site is mixed *and* functional program is at least 7% ***then*** class at least Cl_4	
$r_3^{\geq 4}$	***if*** A1≥medium *and* E1≤3428215.04 ***then*** $Cl_4^{\geq}$	A14
	if accessibility is at least medium *and* investment value is at most 3428215.04 Euro ***then*** class at least Cl_4	
$r_4^{\geq 4}$	***if*** C1≥3.0 ***then*** $Cl_4^{\geq}$	A6
	if the number of private developers is at least 3 ***then*** class at least Cl_4	

considered as the most significant criteria in terms of explaining the assignments for reference UDAs as well as finely discriminating the choice of classes, because they appear in the greatest number of reducts. On the contrary, A1 (accessibility) and C2 (number of public stakeholders) can be considered as the least significant criteria as they appear in the least number of reducts.

The decision rules induced with the DOMLEM algorithm are presented in Table 9. They contain two rules for class at most Cl_1, four for class at most Cl_2, four for class at most Cl_3, two for class at least Cl_2, five for class at least Cl_3, and four for class at least Cl_4. For clarity of presentation, each rule is distinguished with a unique symbol and additionally formulated in the natural language.

The rules are composed of the condition and decision parts. The condition part refers to the conjunction of the values on nominal attributes and/or the requirements concerning minimal or maximal performances on the criteria with preference directions. The number of conditions contained in different rules ranges from one (see, e.g., rule $r_1^{\geq 2}$) to three (see, e.g., rule $r_2^{\leq 1}$). The induced rules are minimal in a sense that they only contain conditions, which are sufficient to explain a decision (hence not containing any redundant condition). In Table 9, we also provide a subset of reference UDAs, which support each rule. Some rules (e.g., $r_2^{\leq 1}$ or $r_3^{\leq 2}$) are rather peculiar, being supported by few

Table 10
Classification results for the on-going (testing) Urban Development Agreements.

Alternative	Net scores				Matching rules	Recommended class
	Cl_1	Cl_2	Cl_3	Cl_4		
CS1	0.0952	−0.416	−0.137	−0.649	$r_2^{\geq 2}, r_1^{\leq 1}, r_4^{\leq 2}, r_1^{\leq 3}$	Cl_1
CS2	−0.071	−0.250	0.272	−0.566	$r_2^{\geq 2}, r_1^{\geq 3}, r_1^{\leq 3}$	Cl_3
CS3	0.333	0.111	−0.423	−0.900	$r_2^{\leq 1}, r_1^{\leq 2}, r_3^{\leq 2}, r_1^{\leq 3}, r_2^{\leq 3}, r_4^{\leq 3}$	Cl_1
CS4	−0.666	0.000	−0.158	0.625	$r_1^{\geq 2}, r_3^{\geq 3}, r_4^{\geq 4}$	Cl_4
CS5	−0.750	−0.388	0.300	0.255	$r_1^{\geq 2}, r_1^{\geq 3}, r_2^{\leq 3}, r_3^{\leq 3}$	Cl_3
CS6	−0.166	−0.250	0.272	−0.700	$r_1^{\geq 3}, r_1^{\leq 3}, r_2^{\leq 3}, r_3^{\leq 3}$	Cl_3
CS7	−0.750	−0.388	0.300	0.355	$r_1^{\geq 2}, r_1^{\geq 3}, r_2^{\leq 3}$	Cl_4
CS8	0.261	0.142	0.047	−0.700	$r_3^{\geq 2}, r_2^{\leq 2}, r_1^{\leq 3}$	Cl_1
CS9	0.083	0.229	−0.571	−0.400	$r_2^{\leq 1}, r_1^{\leq 2}, r_3^{\leq 2}, r_4^{\leq 3}$	Cl_2
CS10	−0.555	0.225	−0.395	0.325	$r_1^{\geq 2}, r_1^{\leq 2}, r_4^{\leq 3}$	Cl_4

Table 11
Results of robustness analysis conducted with 21 different sets of rules.

Alternative	Weighted Class Acceptability Indices (in %)				Robust assignment
	Cl_1	Cl_2	Cl_3	Cl_4	
CS1	53.66	22.27	3.85	20.21	Cl_1
CS2	22.00	12.63	61.79	3.58	Cl_3
CS3	74.43	25.57	0.00	0.00	Cl_1
CS4	0.00	39.81	9.64	50.55	Cl_4
CS5	0.00	4.68	55.78	39.54	Cl_3
CS6	22.00	4.68	52.95	20.37	Cl_3
CS7	0.00	7.80	13.55	78.65	Cl_4
CS8	41.88	39.59	8.89	9.64	Cl_1
CS9	49.58	27.17	0.00	23.26	Cl_1
CS10	37.50	34.66	0.00	27.84	Cl_1

alternatives, whereas other rules (e.g., $r_1^{\geq 3}$ or $r_1^{\leq 3}$) generalize the characteristics of a given class union better, covering numerous reference UDAs. Let us explicitly recall some peculiarities captured with the induced rules covering the greatest number of UDAs induced from each downward class union:

- rule $r_1^{\leq 1}$ shows that for 67 % of UDAs assigned to class Cl_1 the development site was mixed and the functional program was limited to few percent;
- rule $r_1^{\leq 2}$ reveals that for 3 out of 7 UDAs evaluated as at most Cl_2 the development site was industrial brownfield and the number of private developers was not greater than one;
- rule $r_1^{\leq 3}$ indicates that for 70 % of UDAs not assigned to the most preferred class Cl_4 the functional program was limited to at most 5%.

In general, the rules convey useful knowledge as they highlight the conditions that might have implied the success or failure of some UDAs. When applied to a set of 15 reference UDAs, these rules reproduce all pre-defined class assignments, hence attaining perfect classification accuracy on the learning set. Moreover, the inferred set of rules was directly analyzed by the Italian urban planning experts involved in the case study. These experts comprehensively validated the set and confirmed its soundness. In this view, the inferred rules pose a basis for the comprehensive assessment of UDAs, which have not been yet concluded.

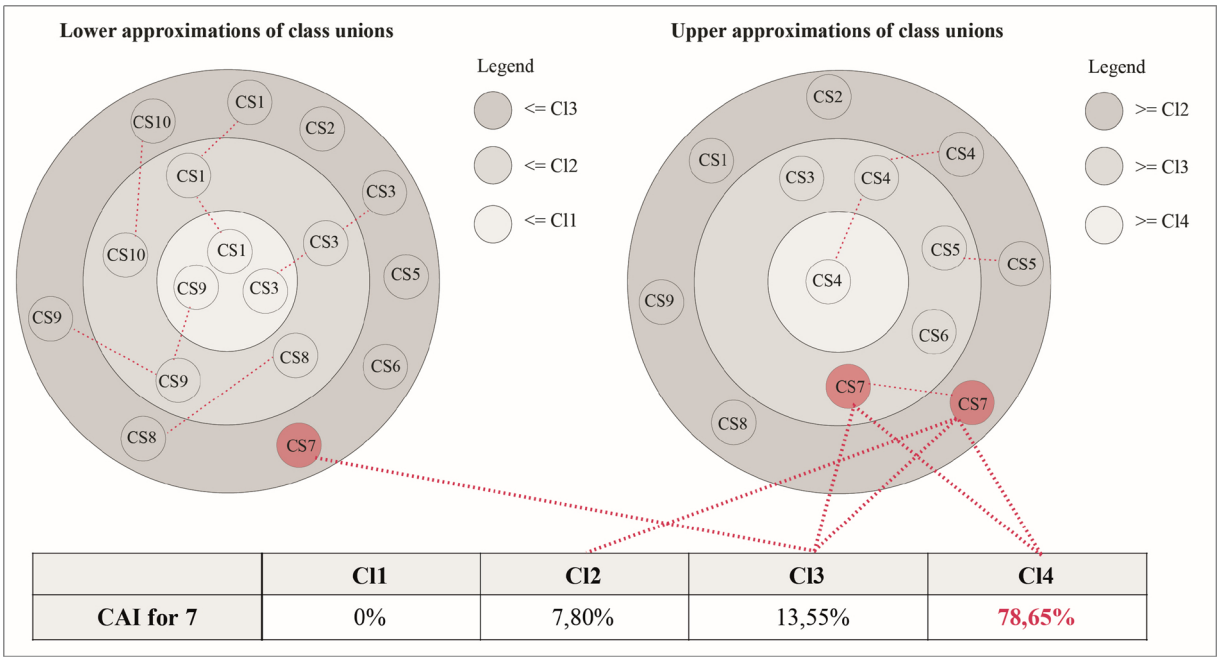


Fig. 4. Robustness of the class assignments.

4.5. Application of decision rules

The decision model formed by the rules presented in Table 9 can support the assessment of on-going UDAs. In Table 10, we present the classification results for 10 non-reference UDAs, including the rules matching a particular UDA, the net scores supporting its assignments to different classes as well as the recommended class. The assignment suggested by the new classification scheme depends on the decision recommended by the matching rules as well as the assignment of reference UDAs that support these rules.

For some non-reference UDAs, there is a prevailing score for a single class (see, e.g., CS2, CS4 or CS6). In particular, CS2 is covered by the rules suggesting the assignment to class at least Cl_2 ($r_2^{\geq 2}$), at least Cl_3 ($r_1^{\geq 3}$), and at most Cl_3 ($r_1^{\leq 3}$), which are mostly supported by the reference UDAs from class Cl_3. In this perspective, both the greatest score for the recommended class Cl_3 and the negative scores for the remaining classes can be easily justified. On the contrary, for some other UDAs, the highly positive scores are observed for more classes. For example, the assignments suggested for CS7 indicate class at least Cl_2 ($r_1^{\geq 2}$), at least Cl_3 ($r_1^{\geq 3}$), and at most Cl_3 ($r_2^{\leq 3}$), but the greatest scores are observed for Cl_3 and Cl_4. The latter can be explained by referring to many reference alternatives from Cl_4 supporting $r_1^{\geq 2}$ and $r_1^{\geq 3}$, i.e., two out of three rules employed for deriving the classification for CS7. Overall, three UDAs (CS1, CS3, and CS8) are assigned to Cl_1, one (CS9) to Cl_2, three (CS2, CS5, and CS6) to Cl_3, and three (CS4, CS7, and CS10) to Cl_4.

To verify the assignment suggested by a set of rules constructed with the DOMLEM algorithm against the recommendation that can be obtained for other sets of rules induced from the same set of reference UDAs, we conducted a robustness analysis. Specifically, we induced 21 minimal sets of rules, each referring to the attributes contained in a unique reduct. Moreover, we validated a predictive ability for the set of rules based on different reducts by performing a leave-one-out cross-validation on a set of 15 reference UDAs (A1-A15). In particular, the results for each reduct were averaged over 15 independent runs. In each run, a subset of 14 UDAs served as a knowledge base for induction of decision rules and the classification obtained with this set for a single UDA was compared with its pre-defined assignment. The *MAE*s for the sets of rules based on different reducts are presented in Table 8. The best results were obtained with the sets of rules based on RED_{15} = {D2, E2, E3, A2} and RED_{19} = {C1, E1, E2, A2, B1}, whereas the worst performance was observed for the sets of rules based on RED_{14} = {D2, E1, E3, B1} and RED_4 = {B2, C1, E3, A2}. Such outcomes can be interpreted in terms of the potential of the sets of rules based on particular reducts to suggest a correct classification.

The 21 minimal sets of rules based on the attributes and criteria contained in all unique reducts are presented in the supplementary material available as an *e-Appendix*. Since the assignments for all 15 reference UDAs were used in the rule induction phase, each of these sets of rules perfectly reproduces all pre-defined classification examples, hence attaining a classification accuracy equal to one when applied on the learning set.

For each non-reference UDA, we determined the classes recommended with different rule sets using the same algorithm and verified the spaces of agreement and disagreement between these recommendations. Moreover, when computing *WCAI*s, we accounted for the weights equal to be the reciprocals of respective *MAE*s (see Table 8). In this way, the sets of rules that proved to perform better in the leave-one-out cross-validation had a greater impact on the final recommendation derived for the non-reference UDAs. For example, the weights assigned to a recommendation suggested by the sets of rules based on RED_{15} and RED_{19} were almost twice as great as the weight associated with the set of rules based on RED_{15} (1.250 vs. 0.661).

Overall, *WCAI*s quantify the strength of considered sets of rules recommending a certain class for a given alternative. These indices are presented in Table 11 along with the class having the greatest support in terms of *WCAI*. Specifically, for CS2 – 11 out of 21 sets of rules indicate Cl_3, 4 sets suggest either Cl_1 or Cl_2, and 2 sets recommend Cl_4. Hence, the non-weighted class acceptability indices for CS2 would be as follows: $CAI(\text{CS2}, 1) = \frac{4}{21} = 19.05\%$, $CAI(\text{CS2}, 2) = \frac{4}{21} = 19.05\%$, $CAI(\text{CS2}, 3) = \frac{11}{21} = 52.38\%$, and $CAI(\text{CS2}, 4) = \frac{2}{21} = 9.52\%$. However, when accounting for the weights associated with different sets of rules, the support given to the assignment of CS2 to Cl_3 ($WCAI(\text{CS2}, 3) = 61.79\%$) or Cl_1 ($WCAI(\text{CS2}, 1) = 22\%$) gets higher at the cost of lowering the support for the assignments of CS2 to Cl_2 ($WCAI(\text{CS2}, 2) = 12.63\%$) or Cl_4 ($WCAI(\text{CS2}, 4) = 3.58\%$). Note that since the profiles of CS1 and CS9 did not match any rule in either one or two sets of rules, the classes recommended in these cases were extrapolated from the rules that were "nearest" to these UDAs in terms of a valued closeness relation.

In general, there is no agreement concerning the recommended assignment between different sets of rules relevant for our analysis (i.e., there is no single *WCAI* equal to 100 %). However, for many UDAs, the suggested recommendation is robust with some class being supported by significantly more considered sets of rules (e.g., Cl_3 for CS2, CS5, and CS6, Cl_1 for CS1 and CS3, Cl_4 for CS4 and CS7) (Fig. 4). On the contrary, for some few UDAs (see, e.g., CS8 and CS10) the recommendation quantified with *WCAI* s is rather conflicting. This means that their characteristics are untypical and have some common aspects with the reference alternatives from different classes. Overall, for 8 out of 10 non-reference UDAs, the most robust assignment (see Table 11) agrees with the class suggested by a set of rules induced with DOMLEM. Only for CS9 and CS10, these assignments differ. Nonetheless, the support given to the assignments of CS9 to Cl_2 and CS10 to Cl_4 by different considered sets of rules is also substantial.

Let us emphasize that *WCAI* s can be used for identifying the classes with the greatest support, but they also convey useful knowledge on the least probable assignments. Specifically, *WCAI* s equal to 0% indicate that the assignment was not supported by any set of rules. For example, the most robust assignment for CS7 is class Cl_4, there is limited support given to the assignment to either Cl_2 or Cl_3, but the assignment to the least preferred class Cl_1 is excluded due to $WCAI(\text{CS7}, 1) = 0\%$.

All results presented in this paper have been computed with the original software implemented by the authors. Its source code written in *Java* is available to other researchers on request. We also envisage a release of the software with an intuitive user interface in the following months.

5. Discussion and conclusions

In this paper, we considered the problem of a holistic evaluation of Urban Development Agreements in the Lombardy Region, with the aim to analyze their most relevant features and to identify the drivers deciding upon the success of UDAs. We accounted for two subsets of UDAs, one concluded and the other still on-going.

Firstly, we defined a set of eleven characteristics that were suitable for describing the complexity of UDAs. They involved five dimensions related to the urban, institutional, negotiation, development and economic contexts. These characteristics included both criteria with preference-ordered scales as well as nominal attributes. In order to apply such diverse scales to the various features at the same time, we suitably adapted the traditional Dominance-based Rough Set Approach. Moreover, we confirmed that the proposed family of criteria and attributes was comprehensive given the case study under investigation. Indeed, the quality of approximation of classification for the analyzed datasets was equal to one, which means that the classification of all concluded UDA could be perfectly explained while referring to their performances on the proposed set of pertinent factors.

Secondly, by considering the training sample composed of 15 concluded UDAs, the proposed approach pointed out some confirmations regarding the factors guiding the processes of urban transformations as well as the drivers able to influence the value generated by such

processes. In this regard, it proved to be successful as an analytical framework able to capture the complexity of the decision context and to describe the most and the least important characteristics. In particular, the results showed that the most significant criteria were the development of the site, the development potentials, the value of the investment and the number of private developers. When considering a specific aspect of the area's land use function, a greenfield was preferred to an industrial field as limited investments are necessary for its reclaim. As far as the development potential is concerned, an agreement that offers an increase in such potential is generally more preferred. When it comes to the number of private promoters, the involvement of a great number of investors reduces the risk of an investment, whereas a large number of public stakeholders could slow down the administrative-bureaucratic process. These findings can be pinned as pertinent characteristics that play a crucial role in determining the fairness and appropriateness of the public-private partnership. Thus, in addition to confirming some empirical evidence already known to the private developers and municipal administrations, the use of DRSA made it possible to identify the factors to which particular attention should be assigned in the preventive evaluation phase of new development projects. This emphasizes the role of the proposed methodological framework in supporting the negotiations between the private and public parties helping them to focus on the essentials.

Thirdly, by analyzing the test sample consisting of the ten on-going negotiation-testing alternatives, we revealed the potential of DRSA in guiding the private developers in the presentation of proposals capable of meeting the requirements that are crucial for the success of the intervention programs under a cooperative negotiation perspective. In this regard, two aspects of the proposed method proved to be of utmost importance. On the one hand, the form of inferred decision rules was essential for capturing structured knowledge and conveying transparent recommendations for supporting complex decision processes. On the other hand, the use of robustness analysis helped to validate the results obtained from the traditional DRSA and to reduce the uncertainty related to the holistic evaluation of on-going or planned UDAs.

The results obtained with our approach can be used to lead the parties involved into negotiations to firstly establish the business as usual scenario that corresponds to the urban development allowed by the local statutory plan and to focus on the economic impact of changes to the functional program and building development potentials. According to this perspective, it could be useful to provide the public parties with standard development costs and market values for a wide range of typologies of intervention and uses, to base the evaluation of the planning gain on robust assumptions rather than on continuously changing and subjective inputs.

The limitations of the current study indicate the most appealing directions for future research. Specifically, we have not included the environmental and social issues in the model although they play a crucial role in the investment decisions and to promote sustainable developments (Rocchi et al., 2019). The future developments should be aimed at integrating these aspects within a triple-bottom-line approach. This requires, however, collecting dedicated data on the environmental and social factors, which were not available for the already concluded UDAs. Furthermore, our learning data set consisted of only several completed UDAs. Since this way of cooperation between the public and the private becomes more and more relevant – which is confirmed by the high cardinality of our testing set – the knowledge base should be systematically updated. The model, that will be based on the analysis of more numerous set of reference alternatives,would reflect both more general regularities supported by a greater number of concluded UDAs and more specific characteristics capturing the divergence between UDAs assigned to the same decision class. Finally, the specificity of the decision rules model consists of seeking the characteristics, which distinguish the options belonging to a given class (union) from the remaining alternatives. As a result, the explanation of a classification of the reference alternatives and the recommendation for the non-reference alternatives can be based on the performances on a subset of few selected criteria or attributes. Although this peculiarity was found appealing for this first case study, a different way of proceeding should apply value or outranking-based MCDA methods. They aggregate the performances on all rather than on only some criteria into an overall measure of alternative's plausibility or its holistic evaluation, which implies that each performance does matter to some degree.

CRediT authorship contribution statement

Alessandra Oppio: Conceptualization, Investigation, Writing - review & editing, Supervision, Project administration. **Marta Dell'Ovo:** Writing - review & editing, Visualization. **Francesca Torrieri:** Data curation. **Grzegorz Miebs:** Methodology, Software, Formal analysis. **Miłosz Kadziński:** Methodology, Software, Validation, Data curation, Writing - review & editing.

Acknowledgements

This study was partially supported by Lombardy Region (Direzione Generale Territorio e Protezione civile) and Polis Lombardia, for which the authors are most grateful. Grzegorz Miebs acknowledges support from the Polish Ministry of Science and Higher Education (Ministerstwo Nauki i Szkolnictwa Wyzszego) under the Diamond Grant project (grant no. DI2018004348). Miłosz Kadziński acknowledges support from the Polish National Science Center (Narodowe Centrum Nauki) under the SONATA BIS project (grant no. DEC-2019/34/E/HS4/00045).

The authors are grateful to the anonymous reviewers for their constructive and detailed comments, which helped us to improve the previous versions of the paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.landusepol.2020.104678.

References

Abastante, F., Bottero, M., Greco, S., Lami, I.M., 2013. Dominance-based rough set approach and analytic network process for assessing urban transformation scenarios. Int. J. Multicriteria Decis. Mak. 3 (2-3), 212–235 73-74.

Alterman, R., 2012. Land use regulations and property values: the' windfalls capture' idea revisited. In: Brooks, Nancy, Donaghy, Kieran, Knaap, Gerrit-Jan (Eds.), Oxford Handbook of Urban Economics and Planning, pp. 755–786.

Bai, L., Li, Y., Du, Q., Xu, Y., 2017. A fuzzy comprehensive evaluation model for sustainability risk evaluation of ppp projects. Sustainability 9 (10), 1890.

Błaszczyński, J., Greco, S., Słowiński, R., 2007. Multi-criteria classification–A new scheme for application of dominance-based decision rules. Eur. J. Oper. Res. 181 (3), 1030–1044.

Błaszczyński, J., Słowiński, R., Szeląg, M., 2011. Sequential covering rule induction algorithm for variable consistency rough set approaches. Inf. Sci. 181 (5), 987–1002.

Bowers, J., 1992. The economics of planning gain: a re-appraisal. Urban Stud. 29 (8), 1329–1339.

Brito, V.T., Ferreira, F.A., Pérez-Gladish, B., Govindan, K., Meidutė-Kavaliauskienė, I., 2019. Developing a green city assessment system using cognitive maps and the Choquet Integral. J. Clean. Prod. 218, 486–497.

Calabrò, F., Della Spina, L., 2012. Stima e ripartizione del plusvalore generato da programmi urbani complessi a Reggio Calabria. In: Stefano Stanghellini, Il negoziato pubblico privato nei progetti urbani, principi metodi e tecniche di valutazione. DEI, Roma.

Camagni, R., 1996. Economia e pianificazione delle città sostenibili. Il Mulino, Bologna. pp. 53–82.

Camagni, R., 2008. Il finanziamento della città pubblica. In: Curti, F. (Ed.), Urbanistica e fiscalità locale, Maggioli, Ravenna, pp. 321–342.

Camagni, R., 2016. Urban development and control on urban land rents. Ann. Reg. Sci. 56 (3), 597–615.

Camagni, R., Modigliani, D., 2013. La rendita fondiaria/immobiliare a Roma: 6 studi di caso. XXVIII Congresso INU-Istituto Nazionale di Urbanistica, Città come motore di sviluppo del Paese, Salerno 24–26.

Chen, Y., Hipel, K.W., Kilgour, D.M., Zhu, Y., 2009. A strategic classification support system for brownfield redevelopment. Environ. Model. Softw. 24 (5), 647–654.

Cilona, T., Granata, M.F., 2015. Multicriteria prioritization for multistage implementation of complex urban renewal projects. In: International Conference on Computational

Science and Its Applications. Springer, Cham. pp. 3–19.
Cinelli, M., Spada, M., Kadziński, M., Miebs, G., Burgherr, P., 2019. Advancing hazard assessment of energy accidents in the natural gas sector with rough set theory and decision rules. Energies 12 (21), 4178.
Copiello, S., 2011. Progetti urbani in partenariato. Studi di fattibilità e piano economico finanziario. Valori e Valutazioni, 7.
Gan, J., Zhang, Y., Hu, Y., Liu, S., 2018. The cooperation partner selection of private sector under public-private-Partnership projects: an improved approach under group decision-making based on FRS, SAW, and integrated Objective/Subjective attributes. Discrete Dyn. Nat. Soc. 2018.
Geisser, S., 1993. Predictive Inference. Chapman and Hall, New York, US.
Greco, S., Matarazzo, B., Słowiński, R., Stefanowski, J., 2000. An algorithm for induction of decision rules consistent with the dominance principle. In: International Conference on Rough Sets and Current Trends in Computing. Springer, Berlin, Heidelberg. pp. 304–313.
Greco, S., Matarazzo, B., Słowiński, R., 2001. Rough sets theory for multicriteria decision analysis. Eur. J. Oper. Res. 129 (1), 1–47.
Greco, S., Matarazzo, B., Słowiński, R., 2016. Decision rule approach. In: In: Greco, S., Ehrgott, M., Figueira, J. (Eds.), Multiple Criteria Decision Analysis. International Series in Operations Research & Management Science, vol 233. Springer, New York, USA, pp. 497–552.
Han, Q., Zhu, Y., Ke, G.Y., Hipel, K.W., 2018. An ordinal classification of brownfield remediation projects in China for the allocation of government funding. Land Use Policy 77, 220–230.
Han, Q., Zhu, Y., Ke, G.Y., Hipel, K.W., 2019. Public private partnership in brownfield remediation projects in China: identification and structure analysis of risks. Land Use Policy 84, 87–104.
Kadziński, M., Greco, S., Słowiński, R., 2014. Robust ordinal regression for dominance-based rough set approach to multiple criteria sorting. Inf. Sci. 283, 211–228.
Kadziński, M., Słowiński, R., Greco, S., 2015. Multiple criteria ranking and choice with all compatible minimal cover sets of decision rules. Knowl. Based Syst. 89, 569–583.
Kadziński, M., Słowiński, R., Greco, S., 2016. Robustness analysis for decision under uncertainty with rule-based preference model. Inf. Sci. 328, 321–339.
Kadziński, M., Rocchi, L., Miebs, G., Grohmann, D., Menconi, M.E., Paolotti, L., 2018. Multiple criteria assessment of insulating materials with a group decision framework incorporating outranking preference model and characteristic class profiles. Group Decis. Negot. 27 (1), 33–59.
Kadziński, M., Badura, J., Figueira, J.R., 2020. Using a segmenting description approach in multiple criteria decision aiding. Expert Syst. Appl. 147, 113186.
Liang, Q., Hu, H., 2018. Study on identification of spurious public–Private partnership projects in China. IEEE Trans. Eng. Manag. https://doi.org/10.1109/TEM.2018.2877261. in press.
Liou, J.J.H., Tzeng, G.-H., 2010. A Dominance-based Rough set Approach to customer behavior in the airline market. Information Sciences 18, 2230–2238.
Malme, J.H., Youngman, J.M. (Eds.), 2001. The Development of Property Taxation in Economies in Transition: Case Studies from Central and Eastern Europe. World Bank Publications.
Mambelli, T., Stanghellini, S., 2003. La valutazione dei programmi di riqualificazione urbana proposti dai soggetti privati. Scienze Regionali.
Micelli, E., 2011. La gestione dei piani urbanistici. Perequazione, accordi, incentivi, Marsilio, Venezia.
Morano, P., Tajani, F., 2017. Decision support methods for public-private partnerships: an application to the territorial context of the Apulia Region (Italy). Appraisal: From Theory to Practice. Springer, Cham, pp. 317–326.
Nijkamp, P., Van Der Burch, M., Vindigni, G., 2002. A comparative institutional evaluation of public-private partnerships in Dutch urban land-use and revitalisation projects. Urban Stud. 39 (10), 1865–1880.
OECD, 2017. The Governance of Land Use in OECD Countries Policy Analysis and Recommendations.
Oppio, A., Torrieri, F., 2018. Public and private benefits in Urban development agreements. In: International Conference on Smart and Sustainable Planning for Cities and Regions. Springer, Cham. pp. 345–356.
Oppio, A., Torrieri, F., Dell'Oca, E., 2018. Land value in Urban Development Agreements: methodological perspectives and operational recommendations. Valori e Valutazioni 21, 87–96.
Oppio, A., Torrieri, F., Bianconi, M., 2019. Land value capture by Urban development agreements: the case of Lombardy region (Italy). In: International Symposium on New Metropolitan Perspectives. Springer, Cham. pp. 346–353.
Ottomano Palmisano, G., Loisi, R.V., Ruggiero, G., Rocchi, L., Boggia, A., Roma, R., Dal Sasso, P., 2016. Using analytic network process and dominance-based rough set approach for sustainable requalification of traditional farm buildings in Southern Italy. Land Use Policy 59, 95–110.
Pawlak, Z., 1991. Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Dordrecht.
PwC and Urban Land Institute, 2018. Emerging Trends in Real Estate.
Raiffa, H., 1982. The Art and Science of Negotiation. Harvard University Press.
Reimer, M., Getimis, P., Blotevogel, H., 2014. Spatial planning systems and practices in Europe: a comparative perspective. Spatial Planning Systems and Practices in Europe. Routledge, pp. 21–40.
Rita, D.I., Ferreira, F.A., Meidutė-Kavaliauskienė, I., Govindan, K., Ferreira, J.J., 2018. Proposal of a green index for small and medium-sized enterprises: a multiple criteria group decision-making approach. J. Clean. Prod. 196, 985–996.
Rocchi, L., Paolotti, L., Rosati, A., Boggia, A., Castellini, C., 2019. Assessing the sustainability of different poultry production systems: a multicriteria approach. J. Clean. Prod. 211, 103–114.
Roy, B., 1996. Multicriteria Methodology for Decision Aiding. Kluwer, Dordrecht, The Netherlands.
Saleh, E., Błaszczyński, J., Moreno, A., Valls, A., Romero-Aroca, R., de la Riva-Fernandez, S., Słowiński, R., 2018. Learning ensemble classifiers for diabetic retinopathy assessment. Artif. Intell. Med. 85, 50–63.
Słowiński, R., Stefanowski, J., et al., 1994. Rough classification with valued closeness relation. In: Diday, E. (Ed.), New Approaches in Classification and Data Analysis. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 482–489.
Słowiński, R., Greco, S., Matarazzo, B., 2015. Rough set methodology for decision aiding. In: Kacprzyk, J., Pedrycz, W. (Eds.), Springer Handbook of Computational Intelligence. Springer Handbooks. Springer, Berlin, Heidelberg, pp. 349–370.
Smajgl, A., Ward, J., 2015. Evaluating participatory research: framework, methods and implementation results. J. Environ. Manage. 157, 311–319.
Stanghellini, S., 2019. Un approccio integrato alla rigenerazione urbana. Urbanistica 69 (160), 8–15.
Stanghellini, S., Copiello, S., 2011. Urban models in Italy: partnership forms, territorial contexts, tools, results. Urban Models and Public-Private Partnership. Springer, Berlin, Heidelberg, pp. 47–130.
Susmaga, R., Słowiński, R., 2015. Generation of rough sets reducts and constructs based on inter-class and intra-class information. Fuzzy Sets Syst. 274, 124–142.
Tánczos, K., Kong, G.S., 2001. A review of appraisal methodologies of feasibility studies done by public private partnership in road project development. Period. Polytech. Transp. Eng. 29 (1-2), 71–81.
Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79–82.
Xiong, W., Chen, B., Wang, H., Zhu, D., 2019. Governing public–private partnerships: a systematic review of case study literature. Aust. J. Public Adm. 78 (1), 95–112.
Youngman, J.M., Malme, J.H., 1994. An International Survey of Taxes on Land and Buildings. Kluwer Law International.
Youngman, J., Malme, J., 2004. The property tax in a new environment: lessons from international tax reform efforts. Andrew Young School's Fourth Annual Conference on Public Finance Issues. Available from:http://isp-aysps.gsu.edu/academics/conferences/conf2004/Youngmannmalme.pdf(Accessed 10th of May 2016).
Zopounidis, C., Doumpos, M., 2002. Multicriteria classification and sorting methods: a literature review. Eur. J. Oper. Res. 138 (2), 229–246.

Publication [P6]

Miebs, G., and Kadziński, M. (2021). Heuristic algorithms for aggregation of incomplete rankings in multiple criteria group decision making. In Information Sciences (Vol. 560, pp. 107–136). Elsevier BV., DOI: 10.1016/j.ins.2021.01.055.

Number of citations[6]:

- according to Web of Science: 9
- according to Google Scholar: 14

[6] as on September 25, 2023

Information Sciences 560 (2021) 107–136

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Heuristic algorithms for aggregation of incomplete rankings in multiple criteria group decision making

Grzegorz Miebs, Miłosz Kadziński *

Institute of Computing Science, Faculty of Computing, Poznan University of Technology, Piotrowo 2, 60-965 Poznań, Poland

ARTICLE INFO

Article history:
Received 7 August 2020
Received in revised form 3 January 2021
Accepted 17 January 2021
Available online 2 February 2021

Keywords:
Group decision
Multiple criteria decision analysis
Ranking aggregation
Incomplete ranking
Metaheuristics
Insulating materials

ABSTRACT

We propose heuristics for constructing a compromise incomplete ranking based on partial rankings admitting incomparability. We consider the utilitarian and egalitarian perspectives oriented toward minimizing an average or a maximal distance from any input ranking. The proposed algorithms incorporate genetic algorithms, simulated annealing, tabu search, local search, and intuitive, dedicated procedures. We demonstrate their efficiency in a real-world case study concerning the ranking of insulating materials based on the conflicting, incomplete preferences of a few tens of Decision Makers (DMs). For each DM, we consider a single representative ranking consistent with his/her preferences or thousands of such rankings following incorporation of the robustness concern. The experimental comparison is generalized to artificially generated problems that differ in terms of the numbers of alternatives and input rankings, and diversity levels. The results are quantified with the quality of obtained rankings and computation time.

1. Introduction

Multiple Criteria Decision Aiding (MCDA) provides methods for comparing alternatives evaluated in terms of at least two, usually conflicting, criteria [7]. These approaches incorporate the preferences of the Decision Maker (DM) to deliver a recommendation consistent with his/her value system. Typically, such a recommendation takes one of the following three forms: indication of a subset of the most preferred alternatives, assignment of alternatives to pre-defined and ordered classes, or ranking of alternatives from the best to the worst [7].

In multiple criteria ranking, the alternatives are compared against each other, and the results depend on relative rather than absolute evaluations [26]. In many real-world problems, one expects a complete ranking to be constructed, hence making all pairs of alternatives comparable. This is often attained by assigning a score to each alternative. However, using a scoring method is often fragile and limiting in terms of relations that can be modeled for a given pair of alternatives. For these reasons, Roy introduced the possibility of representing incomparability [41]. It corresponds to a symmetric and non-reflexive binary relation that is useful if no clear reasons justify either preference or indifference. When some alternatives remain incomparable, the obtained ranking is partial or incomplete.

The incomparability relation is justified for a pair of alternatives when none of them is at least as good as the other [41]. However, an incomplete ranking delivered by the existing MCDA methods results from intersecting the complete rankings

* Corresponding author.
E-mail addresses: grzegorz.miebs@cs.put.poznan.pl (G. Miebs), milosz.kadzinski@cs.put.poznan.pl (M. Kadziński).

https://doi.org/10.1016/j.ins.2021.01.055
0020-0255/

exploiting the DM's preferences or some partial results in slightly different ways. For example, in ELECTRE II and III, the final ranking is obtained by aggregating the two pre-orders derived from the downward and upward distillations [40]. These procedures exploit the matrix of a valued outranking relation by constructing the ranking in a top-down or a bottom-up manner, respectively. Furthermore, a ranking delivered by PROMETHEE I is obtained by aggregating the two orders imposed by the positive and negative flows [5]. These flows quantify for each alternative its comprehensive advantage and disadvantage compared to all remaining alternatives. Thus, a pair of alternatives is incomparable in ELECTRE III or PROMETHEE I if either alternative proves to be more advantageous than the other in one of the two elementary rankings.

The other family of approaches that provide a partial ranking incorporates robustness analysis. For example, Preference Programming [43], UTA^{GMS} [23], and PROMETHEE^{GKS} [26] exploit all (infinitely many) instances of an assumed preference model to construct a robust ranking. These instances may be criteria weights [43], additive value functions [23], or parameters of outranking relations [26] consistent with the imprecise or indirect DM's preference information. The robust ranking is determined by the so-called preference dominance [43] or necessary preference relation [23], which needs to be unanimously confirmed by all compatible model instances. The incomparability in such robust rankings occurs if one alternative is strictly preferred to the other for at least one compatible model instance, whereas for some other compatible instance, the preference is inverse. The various MCDA methods that deliver partial rankings have been applied in multiple domains such as management of municipal solid waste and water infrastructure [49], green supplier selection [1], school quality assessment, value engineering, or performance evaluation of construction plants [45].

The applications mentioned above involved a single DM. However, the increasing complexity of multiple criteria problems implies that a group makes most decisions in firms, organizations, families, agencies, or governments. For example, [36] considered a participative approach involving a few dozen local authorities and representatives from different communities involved in the decision-making process in the District of Turin. Furthermore, [6] faced a problem of revitalizing an urban area in the City of Kitchener, which involved city hall representatives, residents, environmental groups, and consultants collaborating to rank the possible clean-up plans. Moreover, [14] included a group of experts in transportation, environmental engineering, landscape ecology, and operational research as well as a representative of a local government to study the redevelopment projects for disused railways in Italy. Finally, [21] dealt with a case study where a group composed of representatives of the purchase, production, and food safety departments of some Indian food processing company aimed at ranking different options for organizing its supply chain.

Even though the group members face a common problem and share the responsibility, they may differ substantially in terms of their aspirations, beliefs, value systems, interests, or vision on attaining the goals and handling the problem. These differences are often implied by various roles played in the group by the involved stakeholders. In this perspective, Group Decision (GD) is oriented toward supporting Decision Makers (DMs) in transforming their individual preferences into a collective recommendation [28].

There exist two basic paradigms in GD for aggregating group preference, differing in terms of whether the consensus between DMs should be reached at the input or output level [31]. On the one hand, in the input-oriented perspective, the group is assumed to act as a unit [39]. Its members are expected to interact during a consultation process or focus group meeting and reach a consensus on the problem setting (including sets of alternatives and criteria and performances) and preferences. Often, the group members need to agree on the weights, thresholds, and other assumed preference model parameters. Such an approach is applicable when differences between group members regarding their preferences are small, or the group needs to answer collectively [31]. Nevertheless, in many real-world decision scenarios, either such an agreement cannot be reached, and the collective parameters and opinion do not correspond to any individual preferences.

On the other hand, in an output-oriented approach, the group needs to agree only on the considered set of alternatives. However, each DM can define his/her criteria and performances or express individual judgments [22]. An individual ranking is then constructed for each DM, hence avoiding the reference to vague collective parameters or preferences. The joint ranking is obtained by aggregating the personal recommendations according to some pre-defined consensus rules and accounting for priorities assigned to various group members [28]. The supra-DM often determines these rules and priorities. (S)he perfectly knows the interest of an organization to implement the decision in practice and understands both how the group was formed and what roles are played by its particular members [31].

Rank aggregation is a popular problem [48]. However, the vast majority of papers focus only on exploiting complete rankings (see, e.g., [12,32]). In this paper, we focus on the aggregation of partial rankings into a collective one. Some GD systems dealing with a similar scenario incorporate ideas and methods that originated in MCDA. In particular, [33] proposed the GDSS PROMETHEE system, which aggregates the positive and negative flows obtained individually for each DM using PROMETHEE I. The aggregation is performed with a weighted sum model where weights correspond to the DMs' priorities. Furthermore, [31] developed the ELECTRE GD framework that incorporates the individual rankings derived with ELECTRE III, constructs a collective outranking relation, and exploits it to acquire a group compromise ranking. Also, [11] proposed the VIP-G method, which accounted for a setting with imprecise preference information of each DM being modeled with a set of additive value functions. A collective decision (whether a range of scores or preference dominance relation) was deemed acceptable when supported by a sufficient majority of group members. Similarly, [22] adjusted the framework of robustness analysis to verify the stability of the necessary preference relation by introducing the level of certainty related to the support given by the group of DMs. Specifically, a robust relation for a given pair of alternatives was deemed as necessary-necessary or necessary-possible if it was confirmed by all or at least one group member, respectively. A comprehensive GD support system, called ARGOS II, was developed by [9]. It aggregates the individual rankings constructed with

different outranking methods into a collective ordering using various procedures. The latter ones include, e.g., some social choice functions, minimization of the number of ranking deviations, or the Euclidean distance between ranks. However, even if this system accepts partial rankings at the input, it only offers complete rankings at the output. In the same spirit, [19] used interval goal programming within a distance-based framework for deriving consensus ranks based on individual incomplete patterns of preference. Conversely, [3] analyzed the partial rankings of individual DMs obtained with ELECTRE III and selected the one which minimizes the sum of distances from all remaining rankings as the collective one. This idea was advanced in [21] who proposed Mixed-Integer Linear Programming (MILP) for constructing a group compromise ranking that would minimize either the sum of distance or the maximal from the partial ranking of any DM. Its usefulness was demonstrated in the context of incomplete rankings constructed with PROMETHEE I.

The approaches mentioned above for constructing a group compromise ranking based on partial rankings of the individual DMs can be criticized for a few reasons. Firstly, even if they accept partial inputs, some do not accept incomparability at the output, suggesting a complete ranking as the collective recommendation [9,19]. Note that incomparability is an appealing option to be considered at the method's output if at the group level there are no convincing arguments for imposing a weak preference relation for a given pair of alternatives (e.g., if the DMs are equally divided in support of a over b or b over a). Secondly, some algorithms investigate the spaces of consensus and disagreement between DMs in a straightforward way by referring to the support they give as a group to some elementary outcomes [11,22]. Although such approaches are useful for quantifying the results of robustness analysis and supporting negotiation between DMs, they leave constructing a collective, sufficiently decisive ranking far from being solved. Thirdly, the approaches which aim at creating a group compromise partial ranking either exploit a limited subset of candidate solutions [3] or involve a high computational cost related to the use of MILP [21]. Consequently, they may fail to find the optimal solution or cannot be applied to problems involving numerous alternatives or, in the case of an egalitarian problem, many rankings to aggregate. Note that in the vast majority of problems, sets of alternatives in MCDA consist of modestly-sized collections of choices, involving at most several dozens of alternatives. However, the number of scenarios in which the alternatives are more numerous is growing each year (e.g., see the applications in the pharmaceutical market [34] or energy supply security [16]). In the same spirit, group decision-making problems involving a large number of DM are becoming more and more popular, e.g., in health-related experiments, e-commerce, e-marketplaces, and social media platforms [46]. Nevertheless, many rankings to be aggregated can also result from incorporating the robustness analysis into group decision problems with a limited number of DM. Due to accounting for uncertainty and imperfections in the DMs' preferences, tens or even hundreds of rankings can be consistent with the value system of each DM, hence increasing the computational complexity of the ranking aggregation problem.

Another field dealing with ranking aggregation is Computational Social Choice Theory (CSCT) [4]. Its relations with MCDA are very close. Many voting procedures can be applied for both aggregating rankings of various DMs in MCDA and deriving a recommendation for multiple criteria problems when treating alternatives as candidates and criteria as voters. However, the primary focus of CSCT has been on dealing with complete rankings, with all alternatives being fully comparable in both input and output rankings. Recent interest in social choice has been on accounting for partial rankings interpreted as orders of subsets of alternatives (without specifying the rank order for other alternatives) and top-k lists (indicating only the top-ranked options) [37]. However, even for such inputs, the interest of CSCT has remained on computing a complete output ranking or determining only the possible winner(s). Hence, the approach considered in this paper is more general in terms of both accepted inputs and outputs admitting incomparability.

This paper's main contribution consists of proposing various heuristic approaches for constructing a group compromise incomplete ranking based on individual partial rankings. These algorithms incorporate different metaheuristics such as genetic algorithms [18], simulated annealing [29], tabu search [17], and local search as well as intuitive, dedicated procedures. Their heuristic character implies that they are applicable to large-scale problems involving numerous alternatives and thousands of input rankings. We consider different variants of the problem. On the one hand, we aim to construct either a utilitarian or an egalitarian ranking. The utilitarian principle asserts that the best group policy is the one that maximizes the sum of utilities or minimizes the regret for all individuals [35,25]. Hence, the compromise utilitarian ranking minimizes the sum of distances or an average distance from any input ranking. In turn, the egalitarian principle is equivalent to the maxmin (or minmax) principle [35]. In this perspective, the egalitarian ranking minimizes the maximal distance from any input ranking. On the other hand, we generalize the primary problem to account for the weights associated with each elementary ranking. This allows us to use the proposed methods when various priorities are assigned to the DMs or the same input ranking is repeated more than once. The latter is particularly useful in the context of robustness analysis conducted with the Monte Carlo simulations exploiting the rankings compatible with imprecise or indirect preference information [20]. Under such a setting, instead of analyzing a single ranking associated with each DM, one accounts for multiple (possibly hundreds or thousands) rankings following various sets of parameters consistent with the DM's preferences affected by some natural imperfections.

We also report the results of a case study concerning the ranking of thirteen insulating materials. They were evaluated by thirty-eight DMs providing their individual preferences on the importance of six criteria. Such information was used in a twofold way. We derived a precise vector of weights for each DM using the revised Simos (SRF) procedure or used its robust counterpart incorporating the Monte Carlo simulations to sample 1000 weight vectors compatible with the ranking of criteria [20]. We applied the ELECTRE III method for each weight vector to derive a partial ranking [40]. Thus obtained incomplete rankings formed input for the proposed algorithms. Their effectiveness is compared in terms of the quality of

constructed group compromise utilitarian or egalitarian rankings, quantified using the average or maximal distances from the input rankings, respectively. As a secondary metric, we also report the computation time needed for constructing a compromise ranking. This allows for demonstrating a trade-off between quality and time.

We generalize an experimental comparison to artificially generated rankings by studying various settings differing in terms of the numbers of alternatives and rankings and diversity levels between the input rankings. Specifically, we consider the scenarios in which the input rankings are i) similar, being generated from the same order through some mutations, ii) completely random, and iii) an intermediate setting involving related rankings and a single randomly generated one.

The remainder of this paper is organized in the following way. Section 2 introduces the tackled problem and the basic operators that are incorporated into the proposed algorithms. In Section 3, we present the algorithms for constructing an incomplete compromise ranking. Section 4 reports the results for a case study concerning the evaluation of insulating materials. In Section 5, we discuss the outcomes of extensive experimental comparison on artificially generated test problems. The final section concludes the paper and indicates the avenues for future research.

2. Notation and basic concepts

2.1. Problem of constructing a compromise incomplete ranking

The considered problem consists of constructing a partial compromise ranking based on the multiple input incomplete rankings admitting incomparability of some pairs of alternatives. This can be attained by analyzing the pairwise relations in the elementary rankings [21] rather than aggregating individual scores or flows, criteria weights, or alternatives' performances [31,33]. We consider a few variants of the problem, oriented toward constructing either a utlitarian or an egaliterian ranking with respect to the input rankings. Moreover, in the utilitarian case – we can consider the weights or priorities associated with the rankings.

Formally, we account for a set of input rankings $\mathscr{R} = \{R', R'', \ldots, R^r\}$, specifying an incomplete order of n alternatives. They should be aggregated into a single compromise ranking that may be either utilitarian R^U or egalitarian R^E. The compromise utilitarian ranking minimizes the sum of distances or an average distance from any input ranking. In turn, the egalitarian ranking minimizes the maximal distance from any input ranking. To compare the distances between incomplete rankings, we will employ a function defined in [42]. Specifically, such a distance is a sum of elementary divergences between all pairs of alternatives (a_i, a_j) in the two rankings.

For each ordered pair (a_i, a_j) of alternatives, one and only one of the following relations holds in each ranking: a_iPa_j (preference of a_i over a_j), a_jPa_i (preference of a_j over a_i), a_iIa_j (indifference between a_i and a_j), or $a_i?a_j$ (incomparability between a_i and a_j). For the sake of convenience, we shall substitute a_iPa_j and a_jPa_i with, respectively, $a_iP^+a_j$ and $a_iP^-a_j$. Let R'_{ij} and R''_{ij} indicate the relations holding between a_i and a_j in the rankings R' and R'', respectively. Thus, R'_{ij} or R''_{ij} can be instantiated as one of the four relations: P^+, P^-, I, or $?$.

A detailed consideration of the distances $\delta\left(R'_{ij}, R''_{ij}\right)$ between R'_{ij} and R''_{ij} was presented in [42]. Specifically, a distance matrix for the preference, indifference, and incomparability relations was justified by a set of axioms, logical and significance conditions. In what follows, we list these points. For their detailed justification and underlying discussion, the readers are referred to [42]. On the one hand, to be a distance, $\delta\left(R'_{ij}, R''_{ij}\right)$ has to verify the following three axioms:

- the distance is zero if the two relations are the same (i.e., $\delta\left(R'_{ij}, R''_{ij}\right) = 0$ if and only if $R'_{ij} = R''_{ij}$ for $R'_{ij}, R''_{ij} \in \{P^+, P^-, I, ?\}$); otherwise, the distance should be positive i.e., $\delta\left(R'_{ij}, R''_{ij}\right) > 0$ if and only if $R'_{ij} \neq R''_{ij}$);
- the distance matrix is symmetric, i.e., $\delta\left(R'_{ij}, R''_{ij}\right) = \delta\left(R''_{ij}, R'_{ij}\right)$ for $R'_{ij}, R''_{ij} \in \{P^+, P^-, I, ?\}$;
- the triangular inequaility is satisfied, i.e., $\delta\left(R'_{ij}, R''_{ij}\right) + \delta\left(R''_{ij}, R'''_{ij}\right) \geqslant \delta\left(R'_{ij}, R'''_{ij}\right)$.

On the other hand, the following consistency and significance conditions should be satisfied:

- the distance between P^+ or P^- and I or $?$ should be the same, i.e., $\delta(P^+, I) = \delta(P^-, I)$ and $\delta(P^+, ?) = \delta(P^-, ?)$;
- the contradiction between P^+ and P^- in two different preorders is not smaller that the sum of contradictions between P^+ and I as well as I and P^-, i.e., $\delta(P^+, P^-) \geqslant \delta(P^+, I) + \delta(I, P^-)$;
- the contradiction between P^+ and P^- in two different preorders is not smaller that the contradiction between P^+ and $?$, which, in turn, is not smaller than the contradiction between P^+ and I, i.e., $\delta(P^+, P^-) \geqslant \delta(P^+, ?) \geqslant \delta(P^+, I)$.

The joint consideration of all these requirements led to the definition of distances presented in Table 1. A general result in [42] left some domain of variation for $\delta(?, I) = x$ and $\delta(P^+, ?) = \delta(P^-, ?) = y$. Specifically, they can take values in the feasible space presented in Fig. 1. After a detailed consideration and introducing some additional restrictions, it was suggested that a

Table 1
Definition of distances $\delta\left(R'_{ij}, R''_{ij}\right)$ between different pairs of relations [42].

R'_{ij}/R''_{ij}	$a_iP^+a_j$	$a_iP^-a_j$	a_iIa_j	$a_i?a_j$
$a_iP^+a_j$	0	4	2	y (3)
$a_iP^-a_j$	4	0	2	y (3)
a_iIa_j	2	2	0	x (2)
$a_i?a_j$	y (3)	y (3)	x (2)	0

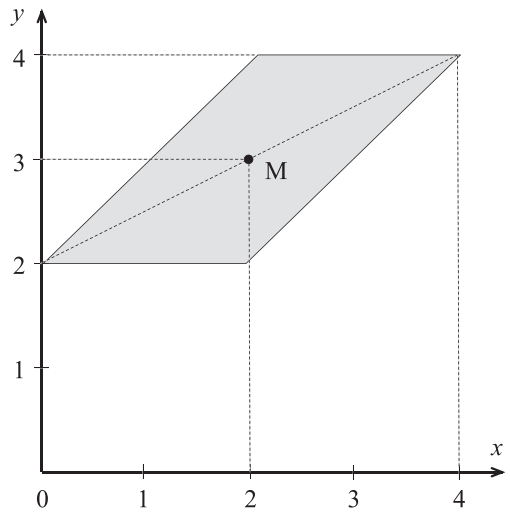


Fig. 1. A domain of variation of distance: $\delta(?, I) = x$ and $\delta(P^+, R) = \delta(P^-, R) = y$ according to the analysis presented in [42].

central point of the feasible domain, denoted by M in Fig. 1, corresponds to the finally adopted solution. Hence, in what follows, we will assume that $x = 2$ and $y = 3$. Nevertheless, for illustrative purposes, we will also consider some other feasible distances (i.e., realizations of x and y) to verify their impact on the final results.

The comprehensive distance between two rankings R' and R'' that can be decomposed into relations R'_{ij} and R''_{ij}, respectively, for all pairs of alternatives $a_i, a_j \in A$, such that $i < j$ (e.g., $R' = \left\{R'_{ij} \text{ for } a_i, a_j \in A, i < j\right\}$), can be computed as follows [42]: $\sum_{i,j\ i<j}\delta\left(R'_{ij}, R''_{ij}\right)$.

When it comes to the aggregation of different rankings, utilitarianism postulates that the collective decision should maximize the sum of individuals' utilities in the same way as it is rational for an individual to maximize his/her utility. It has been advocated by philosophers going back to Bentham and Harsanyi [24], as it allows maximizing the total satisfaction of individuals in the group. This idea can be implemented in the context of ranking aggregation by constructing a compromise utilitarian ranking R^U that minimizes an average distance from the DMs' individual rankings [21]:

$$\sum_{R\in\mathscr{R}}\frac{1}{|\mathscr{R}|}\sum_{i,j\ i<j}\delta\left(R^U_{ij}, R_{ij}\right).$$

Furthermore, in case weight w_R is assigned to each ranking $R \in \mathscr{R}$, we account for the following function:

$$\sum_{R\in\mathscr{R}}\frac{1}{|\mathscr{R}|}w_R\sum_{i,j\ i<j}\delta\left(R^U_{ij}, R_{ij}\right).$$

Let us emphasize that weights $w_R, R \in \mathscr{R}$, do not have to sum up to one. For example, they can reflect the numbers of DMs supporting the same ranking. Overall, the utilitarian compromise will expose the elements supported by a large share of DMs, being close to the center mass of all DMs' rankings. Thus, it potentially leads to a solution that is highly acceptable for the majority of DMs, possibly neglecting the opinion of radical or outlying ones.

In turn, egalitarianism is one of the essential concepts of justice in social and political thought [35]. It postulates that a collective decision should distribute the available "good" in such a way that all the individuals end up enjoying equal benefits. Although there exist various theories of egalitarianism that differ in terms of how the "good" is to be equalized, we implement the "Difference Principle" postulated by Rawls [38]. According to it, the priority should be given to the worst-off individuals, which can be implemented by maximizing the most unfortunate individuals' utility. In the context of ranking

aggregation, this leads to the minimax principle, i.e., constructing a compromise egalitarian ranking R^E minimizing the maximal distance from any DM's ranking, i.e. [21]:

$$max_{R\in\mathscr{R}}\left\{\sum_{i,j\ i<j}\delta\left(R^E_{ij},R_{ij}\right)\right\}.$$

According to [10], the egalitarian perspective is the most useful when groups are heavily polarised, ensuring that the distribution of satisfaction is not too uneven. In this perspective, the egalitarian ranking does not neglect anyone's viewpoints, potentially leading to a solution that is acceptable, to some degree, by everyone, while not entirely satisfying any DM.

To quantify the quality of a given ranking, the values of the above objective functions will be interpreted as a fitness of the ranking. Hence whenever referring to the concept of fitness in the remainder of the paper, we mean either an average or a maximal distance to any input ranking in $\mathscr{R}$ when dealing with the construction of the utilitarian or egalitarian ranking, respectively.

2.2. Representation of rankings

In all algorithms, ranking $R \in \mathscr{R}$ concerning n alternatives is represented as matrix M^R of size $n \times n$. For each ordered pair of alternatives (a_i, a_j), $1 \leqslant i,j \leqslant n$, one of the following relations $(P^+, P^-, I, ?)$ is stored in the respective cell. In addition, since P^+ and P^- are inverse, we have $M^R_{ij} = P^+$ when $M^R_{ji} = P^-$. Furthermore, an indifference relation is reflexive and symmetric, and thus $M^R_{ij} = I$ implies $M^R_{ji} = I$, and $M^R_{ii} = I$. Finally, since incomparability is also symmetric, $M^R_{ij} = ?$ implies $M^R_{ji} = ?$.

2.3. Repairing the incomplete ranking

When considering an incomplete ranking, a weak preference relation needs to be transitive. That is, for all triples of alternatives (a_i, a_j, a_k), if a_i is at least as good as a_j and a_j is not worse than a_k, then also a_i needs to be weakly preferred to a_k. The incorporated matrix representation does not guarantee such a property. The proposed heuristic approaches for constructing a compromise ranking will iteratively improve some (pool of) solutions. For this reason, the solutions stored in the form of a matrix of relations that do not correspond to a valid ranking need to be fixed to satisfy the transitivity requirement.

The repairing method should be efficient to not significantly affect the algorithms' computation time. Thus, we use a Net Flow Score (NFS) procedure [5]. On the one hand, the positive flow of a given alternative is equal to the number of alternatives over which it is preferred. On the other hand, its negative flow corresponds to the number of alternatives that are preferred to it. When considering a pair of alternatives (a_i, a_j) and their positive (a_i^+ and a_j^+) and negative (a_i^- and a_j^-) flows, the relation between a_i and a_j in the repaired ranking can be established in the following way:

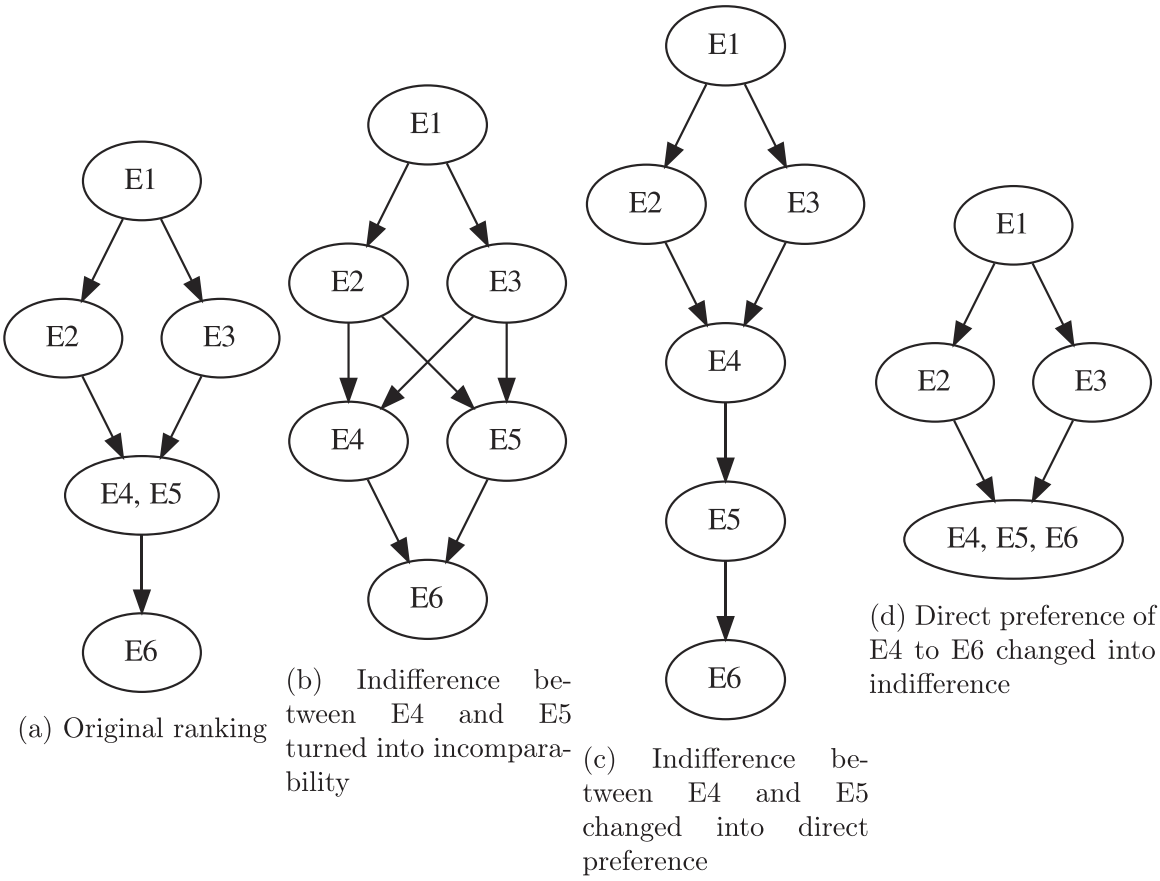
- $a_i P^+ a_j \iff \left(a_i^+ \geqslant a_j^+ \wedge a_i^- < a_j^-\right) \vee \left(a_i^+ > a_j^+ \wedge a_i^- \leqslant a_j^-\right)$,
- $a_i P^- a_j \iff \left(a_i^+ \leqslant a_j^+ \wedge a_i^- > a_j^-\right) \vee \left(a_i^+ < a_j^+ \wedge a_i^- \geqslant a_j^-\right)$,
- $a_i I a_j \iff a_i^+ = a_j^+ \wedge a_i^- = a_j^-$,
- $a_i ? a_j$, *otherwise*.

Whenever a random ranking needs to be created, we fill a matrix of a desired size with the randomly selected relations. We then apply the above NFS procedure to fix it, hence forming a proper incomplete ranking adhering to the weak preference relation's transitivity.

2.4. Definition of neighborhood

To enable searching the space around a given ranking (e.g., through mutating it within the genetic algorithm) or to generate some test instances, we need to define a neighborhood of a given incomplete ranking. For this purpose, we tested the following three definitions of the neighborhood:

[N-I.] Changing a relation for a pair of alternatives a_i and a_j.
[N-II.] Swapping two alternatives a_i and a_j if they are not indifferent.
[N-III.] Swapping two alternatives a_i and a_j as in N-II or changing the relation between them in the following way:
– Set two alternatives incomparable only if initially, they were indifferent (compare Fig. 2a and b). To maintain the weak preference relation's transitivity, all alternatives formerly indifferent with a_j remain indifferent with a_j, thus becoming incomparable with a_i.



(a) Original ranking

(b) Indifference between E4 and E5 turned into incomparability

(c) Indifference between E4 and E5 changed into direct preference

(d) Direct preference of E4 to E6 changed into indifference

Fig. 2. Possible transformations of an original ranking into the neighboring rankings according to three different definitions of neighborhood.

- Set one alternative preferred over the other only if they were indifferent (compare Fig. 2a and c). All alternatives indifferent with a_j remain indifferent with a_j. Thus, a preference or an inverse preference relation is imposed between these alternatives and a_i.
- Set two alternatives a_i and a_j indifferent only if one of them was directly preferred over the other (i.e., there was no alternative a_k such that $a_i R a_k \wedge a_k R a_j$, $R \in \{P^+, P^-\}$ (compare Fig. 2a and d). In addition, a_i will acquire the relation when compared with all other alternatives indifferent to a_j.

Three possible definitions of the neighborhood have been tested to measure an average distance between neighbors, an average distance in fitnesses between neighbors, and an average neighborhood size. To select the most favorable definition for the purpose of further analysis, we employed the following testing procedure:

- To simulate the input rankings to be aggregated, 10 sets were picked randomly, each consisting of 20 rankings.
- 500 rankings were generated randomly, and their full neighborhood was generated to find an average distance between neighbors and an average difference in fitnesses between them, where fitness was defined as an average distance of a given ranking and each pre-order from an input set of rankings.
- This procedure was repeated for different numbers of alternatives n ranging from 15 to 35 with a step of 5.

The results are shown in Table 2. Application of neighborhood N-I may result in invalid solutions, which have to be repaired. This may make them very different from the original ranking, increasing both an average distance between neighbors and an average difference in fitness. However, these measures should be relatively small for a favorable definition of neighborhood. The remaining two definitions (N-II and N-III) attain acceptable results in these regards. In N-II, the structure of a ranking remains the same. Thus when starting from a given point, only a subset of all possible solutions can be achieved. It is impossible to attain ranking with, e.g., a larger number of preferences at the cost of incomparability. Because of that N-III will be applied in further studies. In what follows, mutating ranking R, i.e., picking a solution from the neighborhood N-III of R, will be denoted as an application of function $mutate(R)$.

Table 2
Experimental results concerning the properties of different definitions of neighborhood.

Measure	Neighborhood	$n = 15$	$n = 20$	$n = 25$	$n = 30$	$n = 35$
Average distance between neighbors	$N - I$	406.50	737.14	1159.03	1692.3	2292.06
	$N - II$	80.95	110.92	140.64	170.46	201.21
	$N - III$	61.73	84.85	107.95	131.17	155.2
Average difference in fitnesses	$N - I$	744.69	1348.17	2250.66	3473.39	5361.11
	$N - II$	73.82	91.81	116.17	130.32	148.92
	$N - III$	111.33	135.96	167.67	196.38	220.04
Size	$N - I$	315	570	900	1305	1785
	$N - II$	102.86	187.94	298.0	432.93	593
	$N - III$	184.83	329.56	514.42	738.2	1000.3

2.5. Recombination of rankings

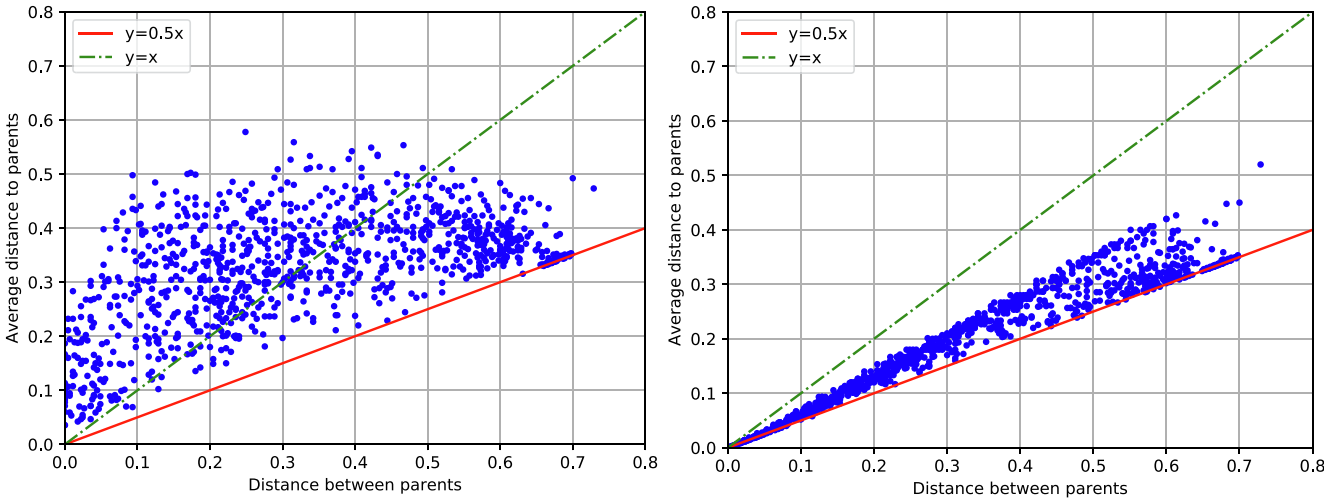
Genetic algorithms make use of the recombination operator. We tested various cross-over schemes, and in this section, we discuss two of them:

[R-I.] For each ordered pair of alternatives, pick a relation from one of the parents randomly.

[R-II.] Get rid of incomparability from both parent rankings by replacing it with a preference relation respecting the ascending order of indices in one parent and the descending order of indices in the other parent (see Fig. 4). Then, find an intersection of the two matrices R' and R'' representing the combined rankings in the following way:

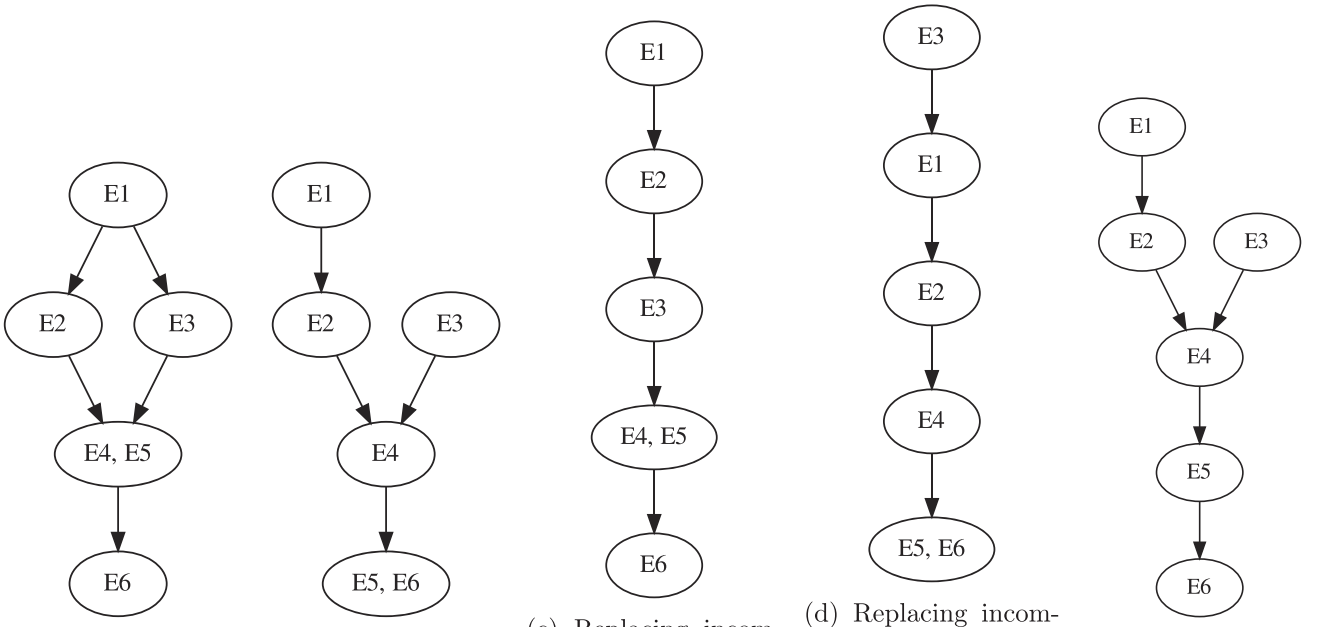
- $a_i P^+ a_j \iff \left(a_i P^+_{R'} a_j \vee a_i I_{R'} a_j\right) \wedge \left(a_i P^+_{R''} a_j \vee a_i I_{R''} a_j\right) \wedge \left(a_i P^+_{R'} a_j \vee a_i P^+_{R''} a_j\right)$,
- $a_i P^- a_j \iff \left(a_j P^+_{R'} a_i \vee a_j I_{R'} a_i\right) \wedge \left(a_j P^+_{R''} a_i \vee a_j I_{R''} a_i\right) \wedge \left(a_j P^+_{R'} a_i \vee a_j P^+_{R''} a_i\right)$,
- $a_i I a_j \iff \left(a_i I_{R'} a_j \wedge a_i I_{R''} a_j\right)$,
- $a_i ? a_j$, otherwise.

These approaches were tested by generating 1000 random rankings, which played the role of a single parent. The other parent was obtained by applying a mutation operation an arbitrary number of times to this first parent. This allowed us to attain a diversity of similarities between these two parents. The child was generated by using a cross-over operation, and an average distance between child and parents was calculated. Recombination R-I was very likely to generate invalid solutions, which had to be repaired using the procedure introduced in Section 2.3. This often resulted in rankings that were distant from the original ones. In turn, recombination R-II always returned a valid ranking. Analysis of the obtained results, which are presented in Fig. 3, leads to the conclusion that the child obtained through the application of R-II is usually nearly a linear combination of its parents. Due to these favorable properties, we decided to use it in the proposed algorithms. In what follows, recombining solutions R' and R'', i.e., applying recombination operator R-II to rankings R' and R'', will be denoted as an application of function $recombination(R', R'')$.



(a) Recombination R-I (b) Recombination R-II

Fig. 3. Distance between parents and their child as a function of a distance between parents for a pair of tested recombination operators.



(a) First parent (b) Second parent (c) Replacing incomparability in (a) between E2 and E3 with preference respecting the ascending order of indices (d) Replacing incomparability in (b) between pairs E1, E3 and E2, E3 with preference respecting the descending order of indices (e) Ranking after intersecting rankings from (c) and (d)

Fig. 4. Steps of recombination operator R-II.

3. Methods for constructing a compromise ranking

In this section, we present ten algorithms for constructing a compromise incomplete ranking based on a set of partial input rankings. These approaches either implement some dedicated heuristic procedures or adapt some metaheuristics to the tackled problem.

3.1. Heuristics based on Net Flow Score

The Net Flow Score procedure introduced in Section 2.3 can serve as the basis for heuristics constructing a compromise ranking. For this purpose, one needs to find support for different relations while accounting for weights w_R associated with the input rankings. The first method denoted as NFS-I, iterates over all input ranking and quantifies the positive flow a_i^+ of a_i as the number of alternatives which a_i is either preferred to or indifferent with, multiplied by the weight w_R of a given ranking:

$$a_i^+ = \sum_{R\in\mathcal{R}} \sum_{j=1,\ldots,n; j\neq i: (a_i P^+ a_j)\vee(a_i I a_j)} w_R.$$

In turn, the negative flow a_i^- of a_i is computed analogously by taking into account the number of alternatives which are preferred to a_i or indifferent with a_i:

$$a_i^- = \sum_{R\in\mathcal{R}} \sum_{j=1,\ldots,n; j\neq i: (a_j P^+ a_i)\vee(a_i I a_j)} w_R.$$

The other variant of the NFS procedure, denoted as NFS-II, computes the scores similarly by assigning different strengths to the observed relations. In particular, in case a_i is preferred to a_j, the score is multiplied by the number $f(i,j,R)$ of alternatives that a_i is preferred to and which are preferred to a_j. They can be seen as the number of alternatives "between" a_i and a_j in ranking R. When a_i and a_j are incomparable or indifference, the score is multiplied by weights $w_?$ and w_I, respectively. The positive and negative flows are computed in the following way:

$$a_i^+ = \sum_{R\in\mathscr{R}}\left[\sum_{j=1,\ldots,n;j\neq i:\left(a_iP^+a_j\right)} w_R\cdot f(i,j,R) + \sum_{j=1,\ldots,n;j\neq i:\left(a_i?a_j\right)} w_R\cdot w_? + \sum_{j=1,\ldots,n;j\neq i:\left(a_iIa_j\right)} w_R\cdot w_I\right];$$

$$a_i^- = \sum_{R\in\mathscr{R}}\left[\sum_{j=1,\ldots,n;j\neq i:\left(a_jP^+a_i\right)} w_R\cdot f(j,i,R) + \sum_{j=1,\ldots,n;j\neq i:\left(a_i?a_j\right)} w_R\cdot w_? + \sum_{j=1,\ldots,n;j\neq i:\left(a_iIa_j\right)} w_R\cdot w_I\right].$$

Having experimentally studied different values of $w_?$ and w_I, we concluded that the best results are obtained when $w_?$ is very large and w_I is very small. For each weight, we tested the following values $[0, 1, n/3, n/2, n]$. Based on the performed experiments, we determined the best parameter values: $w_? = n/2$ and $w_I = 0$. The relations in the compromise ranking constructed with both NFS-I and NFS-II are established based on the positive and negative flows, as discussed in Section 2.3.

3.2. Genetic algorithm

The genetic algorithms evolve a population of solutions in a way that mimics the process of natural evolution. The solutions (in our case, rankings) are modified using operators such as mutation and recombination, and the fittest ones survive to the next generation [18]. Such algorithms become more and more popular in MCDA, mainly to solve Multiple Objective Optimization problems and learn preferences from large sets of decision examples [13]. For the tackled problem of constructing a compromise ranking, a general scheme of the implemented genetic approach is presented as Algorithm 1. In what follows, we discuss its main steps.

Algorithm 1. General scheme of the implemented genetic algorithm

```
1: iter ← 0; populationSize ← 150
2: population ← startPopulation(populationSize) 3: while iter ⩽ 250 do
4:    iter + +
5:    count = findModeFrequency(population)
6:    p ← (count / populationSize)^0.5
7:    for i = 0; i < populationSize; i + + do
8:       if random() < p then
9:          r ← mutate(tournament(population, 2))
10:      else
11:         r ← recombination(tournament(population, 2), tournament(population, 2))
12:      end if
13:      population[reverseTournament(population, 2)] ← r
14:      if r.fitness < best.fitness then
15:         iter ← 0; best ← r 16:      end if
17:   end for
18: end while
```

A starting population (see the *startPopulation* function) is generated by including the input rankings, rankings derived from the NFS-based methods, and rankings where all alternatives are incomparable. Also, some random rankings are added to obtain a population with the desired size. By default, we use a size of 150, but if the number of rankings to be aggregated was larger, this size would be suitably increased.

We use a steady state version of the genetic algorithm. Thus, in every generation, only a single new solution is created by a mutation of one solution (see function *mutate*) or cross-over of two solutions (see function *recombination*) [44]. The probability of mutation is equal to the square root of a ratio of solutions with the most frequent fitness. The latter – determined with the *findModeFrequency* function – is used as a measure of the diversity of a population, which helps avoid premature convergence. For example, if the fitnesses of solutions are $\{4, 5, 3, 4, 5, 4\}$, then the most frequent fitness is 4, and it appears 3 times. Thus, this ratio is equal to $\frac{3}{6} = 0.5$, so the chances of mutation are equal to $0.5^{0.5} \approx 0.707$. This value is calculated once in *populationSize* steps not to generate high computational costs.

Tournament selection is used to pick the solutions to generate a new one and choose the solution that will be replaced by a new one. In the first case, the best solution in a tournament is selected, while the worst one is chosen in the second case. Let us remind that fitness of solution r, denoted by $r.fitness$, is interpreted as either an average or a maximal distance to any input ranking in $\mathscr{R}$, and its value is to be minimized. Note that the $tournament(pop, x)$ function performs a tournament selection from a set of solutions *pop* with the size of a tournament equal to x. During the subsequent tests, it turned out that many solutions contained in the population – despite representing different rankings – share the same best fitness. Thus, to

increase the diversity of solutions, we use the minimal size of a tournament equal to two. The algorithm stops after 250 iterations without improving the best solution. This value was selected based on the tests, which indicated that in more than 99% cases, the best result found so far was improved before 150 iterations.

3.3. Local search algorithms

The main idea underlying the local search method is to constantly improve a current solution until no further improvement can be made. The approaches implementing this principle incorporate the notion of a neighborhood. Specifically, in each step, they move from a current solution to its neighbor with better fitness. In the case this is not possible (i.e., no solution in the neighborhood has better fitness), the algorithm terminates, having identified some Local Optimum (LO). We consider two variants of the algorithm:

- *Steepest*, where in each step, an entire neighborhood is analyzed, and the best neighbor is picked as the base for the next step.
- *Naive*, where the neighborhood is searched only until a better solution than the current one is found, becoming the base in the following step.

The local search algorithms do not guarantee to find the optimal solution to the tackled problem since LO does not need to correspond to the global optimum. However, to increase the chances of finding the best possible solution, the algorithm can be run multiple times, starting from different solutions. Note that the deterministic procedures of exploiting the neighborhood by the steepest and naive variants always result in the same LO when starting from a given point. In the proposed implementation, the starting points were generated in the same way as an initial population in the genetic algorithm.

3.4. Genetic algorithm combined with local search

The genetic algorithms can be combined with local search, which can potentially improve the obtained results [47]. Two possibilities were considered:

- *Lamarckian* variant, where during a process of local search optimization, both fitness and genotype of a ranking are updated.
- *Baldwin* variant, where only fitness is updated, while the genotype remains the same.

For this particular problem, the Lamarckian approach performed better, and hence we will focus on it. This method introduces some minor changes to the genetic algorithm described in Section 3.2. In the tournament procedure, rankings are firstly optimized by local search algorithms, and their new fitnesses are used to determine which is the best among them. This algorithm is time-consuming. Thus such parameters as *populationSize* and the maximum number of iterations need to be decreased. Our tests confirmed that with the lower values of these parameters, the computation time is decreased but at the cost of deteriorating the results. The choice was performed by analyzing the size of testing data, time, and results of other approaches. Specifically, the size of a population was set to 50, and the algorithm was terminated after 15 iterations without an improvement.

3.5. Tabu search

Tabu search was proposed to answer a problem revealed by the local search algorithms [17]. The method continues the search after finding a LO by going to the best neighbor even if it is worse than the current base solution. This will likely result in a cycle composed of two solutions: LO, its best neighbor, LO, its best neighbor, etc. A tabu list of a fixed size containing the last N movements is employed to prevent such moves. In the next step, the opposite direction moves to those included on the tabu list are forbidden. As a result, the same move can be performed again after at least N steps. When applying the standard tabu list for the considered problem, the algorithm got stuck in cycles very quickly with the sequences of moves performed on the same pair of alternatives, e.g., set $a_1 I a_2$, set $a_1 ? a_2$, swap a_2 and a_3, set $a_1 I a_3$. Note that due to the swap between a_2 and a_3, in all operations a_1 is compared against the same alternative (a_2).

To avoid this problem, the adopted tabu list does not contain complete information about the previous moves. In turn, we store just the pair of alternatives and type of performed action: swap or change relation. Using such a tabu list makes it impossible to change the relation for a given pair of alternatives more frequently than once per N steps. Even with such kind of forbidden actions, applying a too small value of N (< 7) resulted in frequent cycles. Thus, the size of a tabu list was set to the number of alternatives n. The search was stopped after 100 steps without improving the best solution found so far.

3.6. Simulated annealing

Similarly to tabu search, simulated annealing overcomes the main weakness of local search, i.e., terminating the search in a LO [29]. Hence in this method, it is possible to accept a neighbor of the current solution with worse fitness. The probability

of such an action is proportional to the difference in fitnesses of these solutions divided by parameter T, called temperature. The latter is initialized using the formula $T_0 = -\frac{\overline{\Delta f}}{ln(p)}$, where $\overline{\Delta f}$ is an estimation of the fitness decrease of strictly positive transitions, and $p = 0.8$ is the probability of accepting such ranking. Such estimation is conducted by generating some pairs of neighbor solutions randomly [2]. Then, T is slowly decreased by multiplying it by parameter $\alpha = 0.95$ after $iter_{max} = n^2$ steps, where each step corresponds to choosing a random neighbor and deciding if it should replace the current solution. Specifically, current solution S is replaced by worse solution S' with a probability $p = exp\left(\frac{f(S)-f(S')}{T}\right)$, where $f(S)$ returns the fitness of solution S.

However, running the simulated annealing's basic variant was not sufficient for finding a satisfactory solution to the tackled problem. The algorithm stuck in a LO when the value of T was so low that the probability of accepting a worse solution a few times in a row was close to zero, hence making it impossible to escape from the LO. To enable visiting a greater number of LOs, the process of reheating was incorporated. Specifically, when the algorithm did not improve, T was increased to enrich the exploration potential. The method was considered stuck when the probability of accepting an average positive transition, calculated in the same way as for the initial temperature, is lower than 0.01, and for the last 10 iterations, a ranking with different fitness was not attained. Then, the temperature T was set to half of the temperature used before (i.e., the temperature with which the process was initiated or set after the previous reheating). The algorithm terminated after three reheating processes without improving the best solution.

The pseudocode of simulated annealing used to construct a compromise ranking is shown in Algorithm 2. Function *setT* returns an initial temperature and a temperature at which reheating should be called. Function *randomNeighbor* returns a random neighbor of a solution passed as an argument. Further, *iter* counts a number of iterations without any changes to check if the reheating needs to be imposed, and *bigIter* stores a number of reheating processes without any improvement to verify if the algorithm should be terminated.

Algorithm 2. General scheme of the simulated annealing.

```
1: iter ← 0; bigIter ← 0; t, t_min ← setT()
2: while bigIter < 3 do
3:    bigIter++; iter ← 0; startT ← t
4:    while t > t_min ∨ iter < 10 do
5:       iter++
6:       for i = 0; i < iter_max; i++ do
7:          r ← randomNeighbor(R)
8:          if r.distance ⩽ R.distance then
9:             R ← r
10:            if r.distance < R.distance then
11:               iter ← 0
12:            end if
13:            if r.distance < best.distance then
14:               best ← r; bigIter ← 0
15:            end if
16:         else if random() < exp((R.distance − r.distance)/t) then
17:            R ← r; iter ← 0
18:         end if
19:      end for
20:      t ← t · α
21:   end while
22:   t ← startT · 0.5
23: end while
```

3.7. Benchmark approaches

To compare the results obtained with metaheuristic algorithms with some elementary solutions, we introduced some simple procedures for constructing an incomplete compromise ranking. On the one hand, the "pick-one" approach constructs a set of rankings in the same way as the initial population is formed by the genetic algorithm (see Section 3.2) and selects the best ranking in such a set as the final result. On the other hand, a random search creates a number of random rankings, out of which the best is returned. If different is not explicitly stated, we generated 50000 random rankings.

4. Results for a case study concerning ranking of insulating materials

In this section, we consider a case study operating on real-world data. It deals with a comprehensive evaluation of 13 insulating materials [27]. Specifically, we account for coating materials with a thickness of 15 cm used in buildings retrofitting. They were placed internally on the roof of a model building typical for central Italy and evaluated from the socio-economic and environmental viewpoints (see Table 3). The six criteria capturing overall sustainability of the materials were: g_1 – Hour of Discomfort (an overall time during which the temperature falls outside the comfort category); g_2 – CO_2 avoidance (energy saved during the building life by using a particular material); g_3 – Net Present Value (NPV; the difference between the present values of cash outflows and inflows related to the use of the material); g_4 – Human Health (environmental Eco-indicator built on the following normalized impact categories: carcinogens, respiratory organics and inorganics, climate change, radiation, and ozone layer); g_5 – Ecosystem Quality (environmental Eco-indicator involving ecotoxicity, acidification/eutrophication, and land use), and g_6 – Resources (environmental Eco-indicator considering minerals and fossil fuels). For a more detailed explanation of these criteria, see [27]. Their importance was assessed by 38 DM. Each of them provided a ranking of criteria along with some blank cards inserted in-between the criteria, as required by the SRF procedure [15]. For some example preference information of this kind, see Table 4.

We will implement two approaches for exploiting the provided rankings of criteria. On the one hand, we will conduct robustness analysis using the Monte Carlo simulation extracting 1000 compatible weights vectors for each DM (for some example compatible weights and mean weights, see Table 5) [20]. On the other hand, we will directly apply the SRF procedure to derive a single weight vector for each DM (such example weight vectors are given in Table 4) [15]. Each compatible weight vector was incorporated into Electre III [40]. The method constructed a credibility matrix and exploited it through the downward and upward distillation procedures to build, respectively, the descending and ascending pre-orders of materials. Finally, these two pre-orders were intersected to obtain an incomplete ranking of insulating materials.

4.1. Results for robustness analysis conducted with the Monte Carlo simulation

When conducting the robustness analysis, for each DM, we derived 1000 weight vectors with the Monte Carlo simulation. Thus, overall 38000 weight vectors were analyzed, and for each of them, we constructed an incomplete ranking with Electre III. The entire process resulted in 488 different rankings. In Fig. 5, we present four of them that repeated the greatest number of times and hence were assigned the highest weights within the procedure, constructing a utilitarian ranking.

The results attained when minimizing an average distance by running every algorithm 10 times are shown in Table 6. For clarity of the presentation, the algorithms are ordered in all tables according to the quality of constructed rankings. The steepest local search, as well as the Lamarckian and traditional genetic algorithms, achieved the same best result (see Fig. 6a), hence outperforming the remaining algorithms in terms of quality. The average distance of this ranking from all input rankings is 26.641. The naive local search turned out to be just slightly worse (26.773). All these algorithms use input rankings as initial solutions, which may give them an advantage when some of these rankings are close to the optimal solution. Moreover, the application of the Monte Carlo simulation leads to a vast number of rankings, which are all contained in the initial population of solutions that are further improved or evolved by these methods. Note that the steepest local search's deterministic character implies that the same best ranking was returned in every iteration, whereas the genetic algorithms were, on average, slightly worse.

When it comes to the tabu search, it proved to be the fastest (average time of 0.367 seconds), but its performance was very unstable (standard deviation more than 7 times greater than for the genetic algorithm). It may suggest that for the tackled problem, this algorithm cannot find solutions that are much different from an initial point, and it exploits a relatively small hyperspace around such a point. Therefore, the overall performance of this method strongly depends on where it

Table 3
Performance matrix of 13 insulating materials evaluated in terms of 6 criteria (g_1 – Hour of Discomfort; g_2 – CO_2 avoid; g_3 – NPV; g_4 – Human Health; g_5 – Ecosystem Quality; g_6 – Resources)

Alternative	g_1	g_2	g_3	g_4	g_5	g_6
Autoclaved aerated concrete	4889.339	158.63	283.41	0.009703	0.000636	0.015876
Corkslab	3974.451	178.49	282.01	0.022122	0.018376	0.040660
Expanded perlite	3893.646	179.11	326.26	0.006451	0.000759	0.043280
Fibreboard hard	3657.799	185.29	243.45	0.039111	0.014516	0.136345
Glass woll	3681.898	187.35	316.92	0.010608	0.001307	0.033364
Gypsum fibre board	7051.231	103.24	135.88	0.047131	0.003916	0.070469
Hemp fibres	3921.449	182.59	334.10	0.002336	0.003079	0.008207
Kenaf fibres	3685.51	186.82	341.79	0.004760	0.015137	0.003079
Mineralized wood	4392.808	167.63	245.45	0.042932	0.004548	0.083149
Plywood	7636.502	87.58	71.26	0.095717	0.201332	0.126167
Polystyrene foam	3750.482	187.13	322.02	0.002801	0.000217	0.016521
Polyurethane	3357.309	194.18	330.35	0.013225	0.000564	0.043280
Rock wool	3659.441	188.45	346.14	0.019183	0.000825	0.009846

Table 4
The order of criteria and inserted blank cards provided by DM_1, DM_2, DM_5, DM_6, DM_8, and DM_{10} in the SRF procedure ($l(i)$ – rank (the higher, the better), e_s – the number of inserted blank cards); and the weights w_i derived with the SRF procedure.

g_i	$l(i)$	e_s	w_i
DM_1			
g_1	1		0.024
		1	
g_3	2		0.085
		2	
g_2	3		0.177
		1	
g_4, g_5, g_6	4		0.238
DM_2			
g_3	1		0.05
		0	
g_1	2		0.088
		1	
g_4	3		0.167
		0	
g_2	4		0.206
		0	
g_5, g_6	5		0.245
DM_5			
g_2, g_3	1		0.063
		1	
g_1, g_4	2		0.188
		0	
g_5, g_6	3		0.25
DM_6			
g_4	1		0.027
		0	
g_2	2		0.077
		1	
g_1, g_3	3		0.175
		1	
g_5, g_6	4		0.273
DM_8			
g_3	1		0.028
		2	
g_1	2		0.124
		1	
g_4	3		0.188
		0	
g_2, g_5, g_6	4		0.22
DM_{10}			
g_1, g_3	1		0.045
		0	
g_2, g_4, g_5, g_6	2		0.227

starts. The average results attained by the simulated annealing and the tabu search were the worst ones among all metaheuristics. They were even worse than for the pick-one method. Such an advantageous performance of the latter algorithm was implied by a very high number of input rankings. The NFS-based methods performed rather poorly, but their computation time was low, e.g., 120 shorter than for the Lamarckian genetic algorithm. In any case, the quality of rankings attained by these approaches was over twice better than for the random search.

In e-Appendix 1 (supplementary material available online), we discuss a detailed performance of all metaheuristics in view of constructing a utilitarian ranking based on 38000 rankings consistent with the weights derived from the Monte Carlo simulation.

When minimizing the maximal distance, all metaheuristics (steepest and naive local search, genetic algorithms, simulated annealing, and tabu search) outperformed the remaining methods (see Table 7). Moreover, these approaches, except the traditional genetic method, attained the same quality of the best results. For the ranking constructed by these methods, the maximal distance from all input rankings is equal to 65. However, the constructed rankings were not the same. A pair of example rankings obtained with the local search and the tabu search with the same maximum distance from the 38000 input rankings are presented in Fig. 6b and c. Again, the steepest local search and the Lamarckian genetic algorithm were the least favorable in terms of computation time but attained the most advantageous average results. Although the genetic algorithm was the worst among the metaheuristic approaches, it still proved to be better than the pick-one method or ran-

Table 5
Four example weight vectors, mean weights (μ) and standard deviations (σ) based on 1000 weight vectors obtained with the Monte Carlo simulation for various Decision Makers.

Sample	g_1	g_2	g_3	g_4	g_5	g_6	Sample	g_1	g_2	g_3	g_4	g_5	g_6
			DM_1							DM_2			
1	0.022	0.213	0.095	0.223	0.223	0.223	1	0.061	0.219	0.047	0.191	0.239	0.239
2	0.024	0.161	0.069	0.247	0.247	0.247	2	0.065	0.196	0.052	0.163	0.261	0.261
3	0.022	0.194	0.092	0.229	0.229	0.229	3	0.099	0.205	0.048	0.160	0.242	0.242
4	0.023	0.204	0.068	0.234	0.234	0.234	4	0.054	0.222	0.049	0.173	0.249	0.249
μ	0.022	0.213	0.095	0.223	0.223	0.223	μ	0.076	0.209	0.049	0.176	0.245	0.245
σ	0.001	0.0206	0.0239	0.0104	0.0104	0.0104	σ	0.0168	0.0152	0.0026	0.0177	0.0129	0.0129
			DM_5							DM_6			
1	0.215	0.057	0.057	0.215	0.228	0.228	1	0.205	0.030	0.205	0.026	0.266	0.266
2	0.2	0.06	0.06	0.2	0.24	0.24	2	0.198	0.051	0.198	0.026	0.262	0.262
3	0.171	0.066	0.066	0.171	0.263	0.263	3	0.214	0.039	0.214	0.025	0.253	0.253
4	0.185	0.063	0.063	0.185	0.252	0.252	4	0.105	0.049	0.105	0.035	0.352	0.352
μ	0.194	0.061	0.061	0.194	0.245	0.245	μ	0.152	0.056	0.152	0.030	0.305	0.305
σ	0.0158	0.0032	0.0032	0.0158	0.0126	0.0126	σ	0.0412	0.0167	0.0412	0.004	0.0403	0.0403
			DM_8							DM_{10}			
1	0.109	0.226	0.028	0.182	0.226	0.226	1	0.045	0.227	0.045	0.227	0.227	0.227
2	0.115	0.223	0.027	0.186	0.223	0.223	2	0.045	0.227	0.045	0.227	0.227	0.227
3	0.119	0.217	0.027	0.199	0.217	0.217	3	0.045	0.227	0.045	0.227	0.227	0.227
4	0.147	0.207	0.025	0.204	0.207	0.207	4	0.045	0.227	0.045	0.227	0.227	0.227
μ	0.137	0.215	0.027	0.192	0.215	0.215	μ	0.045	0.227	0.045	0.227	0.227	0.227
σ	0.0204	0.0082	0.001	0.0099	0.0082	0.0082	σ	0.00	0.00	0.00	0.00	0.00	0.00

dom search. Tabu search was the fastest metaheuristic, attaining equally good results as the local search algorithms. The difference in outcomes between constructing the utilitarian and egalitarian rankings suggests that these two types of problems are substantially different and require dedicated treatment.

To confirm that the problem of constructing a compromise ranking is non-trivial, let us present the distribution of fitnesses from points obtained by randomly sampling the space of solutions 50000 times for the problem of constructing a utilitarian problem is presented in Fig. 7. It resembles a normal distribution with $\mu \approx 188$ and $\sigma \approx 20.2$. For constructing an egalitarian ranking, the histogram of random results is very similar; just the parameters are different ($\mu \approx 211, \sigma \approx 19$). These observations may be used to compare the results attained for different datasets or problems.

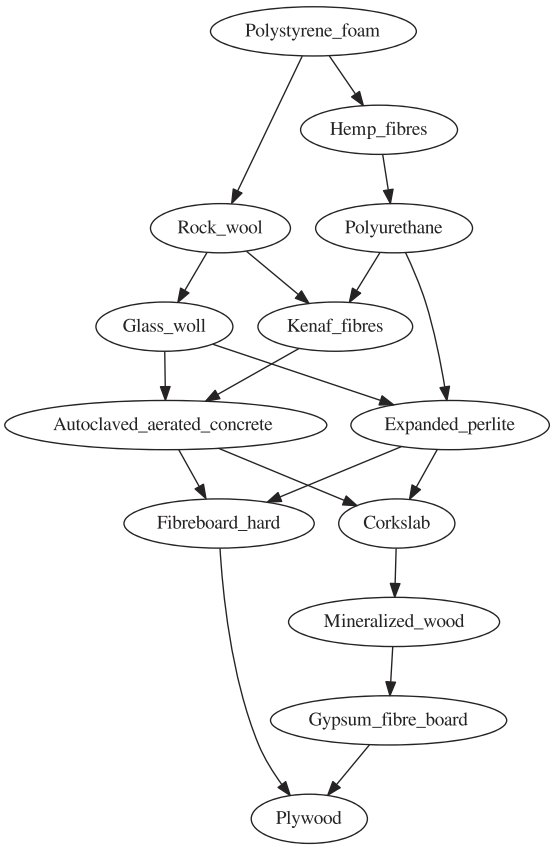
4.2. Results for rankings obtained for weights derived with the SRF procedure

The SRF procedure derives a single weight vector for each DM. Table 4 presents example rankings of criteria provided by various DMs. The respective rankings obtained for some of these weights are presented in Fig. 8. Overall, 38 rankings of criteria led to 27 unique rankings obtained with Electre III, which form the input for the proposed algorithms.

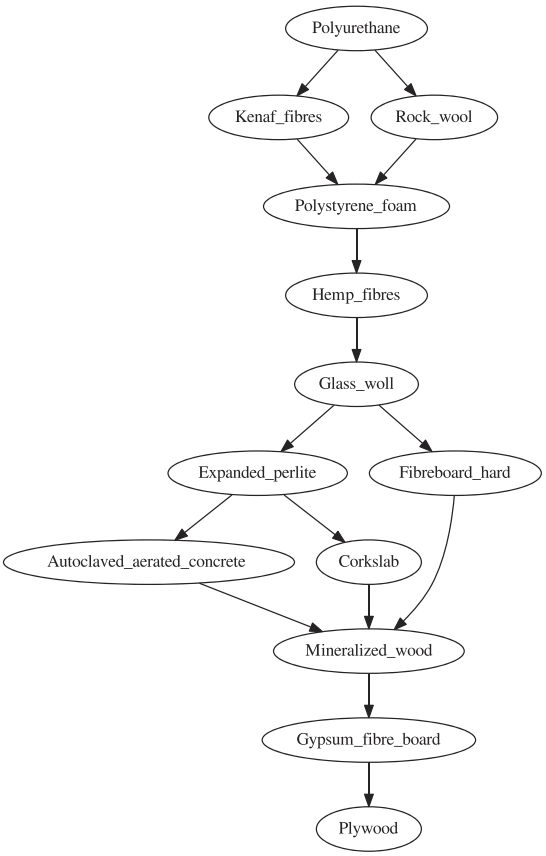
When minimizing an average distance, the local search algorithms and the genetic approaches again proved to be the most advantageous (see Table 8 and Fig. 9a). The average distance of the ranking they constructed from all input rankings was 25.579. Due to a smaller number of input rankings, the best-performing methods' advantage was not that significant. For example, for the naive local search and the genetic algorithm, the average distance of the built ranking was lesser by 0.263, and the ranking constructed by simulated annealing was worse by 0.974. Since 27 input rankings do not fill the admissible input pool of rankings for the proposed metaheuristic algorithms, some randomly generated rankings were included in the initial population. The computation time of the genetic algorithms and simulated annealing was the longest. This was mainly due to the stopping condition adapted by these approaches, which implied that they spent more time after finding their best solution than before attaining this point. On the contrary, the naive local search and tabu search computations took, respectively, 0.134 and 0.244 seconds.

Table 9 presents the results for the scenario oriented toward minimizing a maximum distance from the input ranking. The best solutions were found by the Lamarckian genetic algorithm, tabu search, and simulated annealing (see Fig. 9b). The maximal distance from any input ranking was 42. Among these best-performing methods, the first required significantly more computation time (over 2 minutes) than the remaining two approaches (0.582 and 1.489 seconds). However, an increase in computation time resulted in the best results attained in the average run (the mean quality of the ranking equal to 42.167). The best rankings attained by the genetic and local search algorithms had the same quality with the maximal distance of 43, but the local search proved to be slightly better in the average case. In the egalitarian scenario, the genetic algorithm attained the worst result among all considered metaheuristics. However, it still outperformed the pick-one approach.

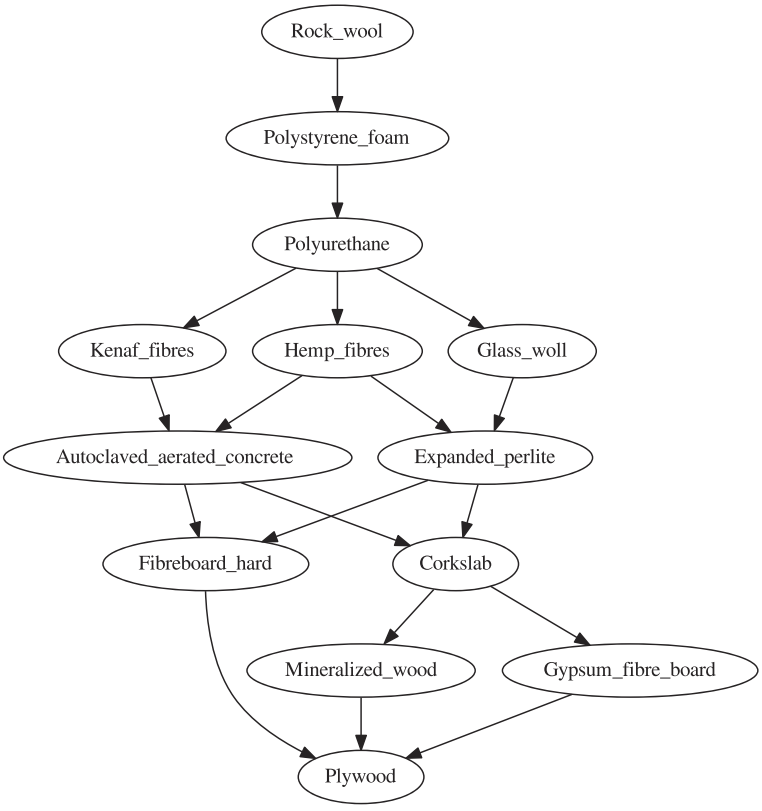
Fig. 10 presents the relation between input and compromise rankings. It is quantified with a distance matrix for each pair and then scaled down with the Multi-Dimensional Scaling (MDS) [30] procedure to a two-dimensional space. The distances from each input ranking to the best egalitarian and utilitarian rankings are presented in Table 10. These results reveal that



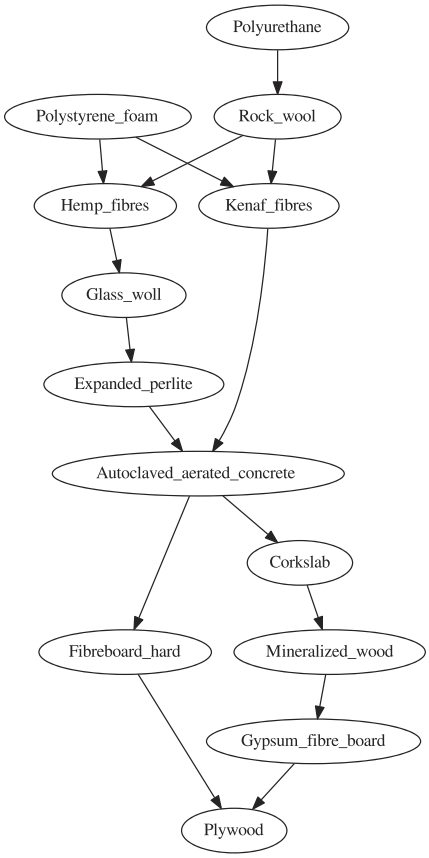
(a) weight $w_R = 6988$



(b) weight $w_R = 1882$



(c) weight $w_R = 1257$



(d) weight $w_R = 1183$

Fig. 5. Input rankings with the greatest weights, consistent with the weight vectors derived from the Monte Carlo simulations.

Table 6
The best and average fitness along with the standard deviation attained by different algorithms in view of constructing a utilitarian compromise ranking based on the 38000 inputs rankings consistent with the weight vectors derived with the Monte Carlo simulations.

Algorithm	Time [ms]	Best	Mean	St. dev.
Steepest Local Search	6854.2	26.641	26.641	0.0
Genetic Lamarckian	11287.0	26.641	26.663	0.054
Genetic Algorithm	3168.2	26.641	26.778	0.217
Naive Local Search	4142.7	26.773	26.773	0.0
Simulated Annealing	1264.5	27.288	27.62	0.257
Tabu Search	367.8	27.288	28.768	1.578
Pick-One	106.0	27.398	27.398	0.0
NFS-I	85.2	33.153	33.153	0.0
NFS-II	92.0	36.04	36.04	0.0
Random Search	1969.5	79.569	89.59	6.447

the best rankings obtained in the considered scenarios (see Fig. 9) are very different. Thus, both settings should be analyzed separately and require dedicated algorithmic solutions.

Let us note that for all scenarios related to the problem of ranking the insulating materials, we have compared the rankings constructed with the proposed metaheuristic approaches with the optimal rankings obtained by solving the MILP model proposed in [21]. When considering the utilitarian and egalitarian settings given the input rankings compatible with weights derived from the SRF procedure or the Monte Carlo simulation, the best-constructed rankings corresponded to the optimal preorders identified with MILP.

Finally, in e-Appendix 2 (supplementary material available online), we analyze the utilitarian rankings obtained for the case study when using different distance matrices. In the main paper, we assumed $x = \delta(?, I) = 2$ and $y = \delta(P^+, ?) = \delta(P^-, ?) = 3$. The additional experiments concern a grid search of eleven different settings for (x, y), ranging from $(0, 2)$ to $(4, 4)$. We discuss how much the rankings are affected by using different distance values. We also report the distribution of the preference, indifference, and incomparability relations.

5. Experimental comparison on artificially generated problems

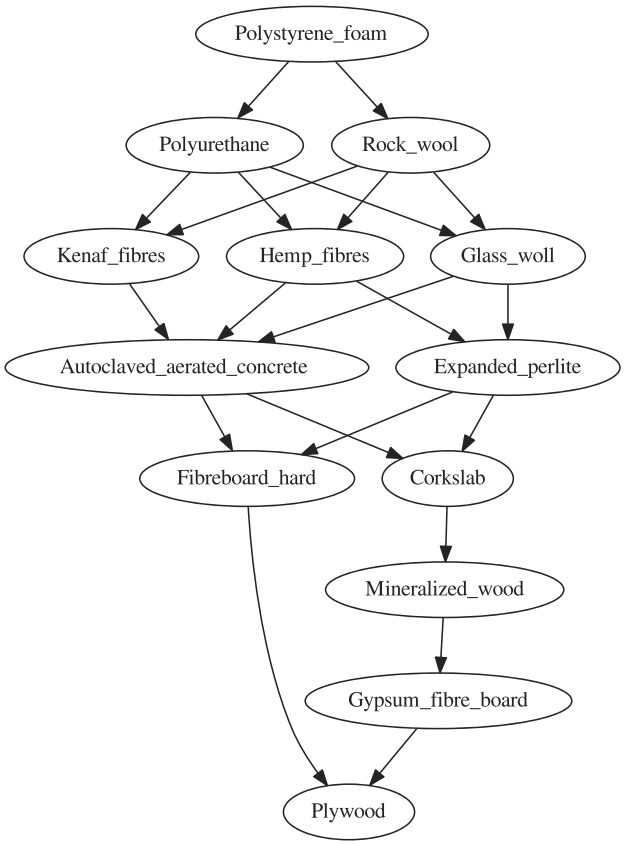
In this section, we discuss an extensive experimental evaluation of the proposed approaches from constructing an incomplete compromise ranking. We account for the results on some artificial data sets with various characteristics. In particular, we generated tests with different diversity levels between the input rankings and various sizes and numbers of input rankings. We considered four numbers of alternatives (n): $15, 20, 25$ and 30, and three numbers of input rankings (s): $15, 20$ and 25. When it comes to the diversity levels, they are discussed in the following subsections.

For every set of parameters, the test case was generated 10 times. For each particular scenario, we sampled 5000 random rankings from which approximate values of the mean (μ) and standard deviation (σ) were computed. These parameters were used to standardize the fitnesses (x) of output compromise rankings in the following way: $\frac{\mu - x}{\sigma}$. In this way, the outcomes for different instances of the problem could be compared and aggregated. Thus, the values above zero mean that a result is better than an average from the randomly generated rankings. Moreover, the higher these values are, the better. For clarity of presentation, in what follows, we will call such a standardized score – a fitness.

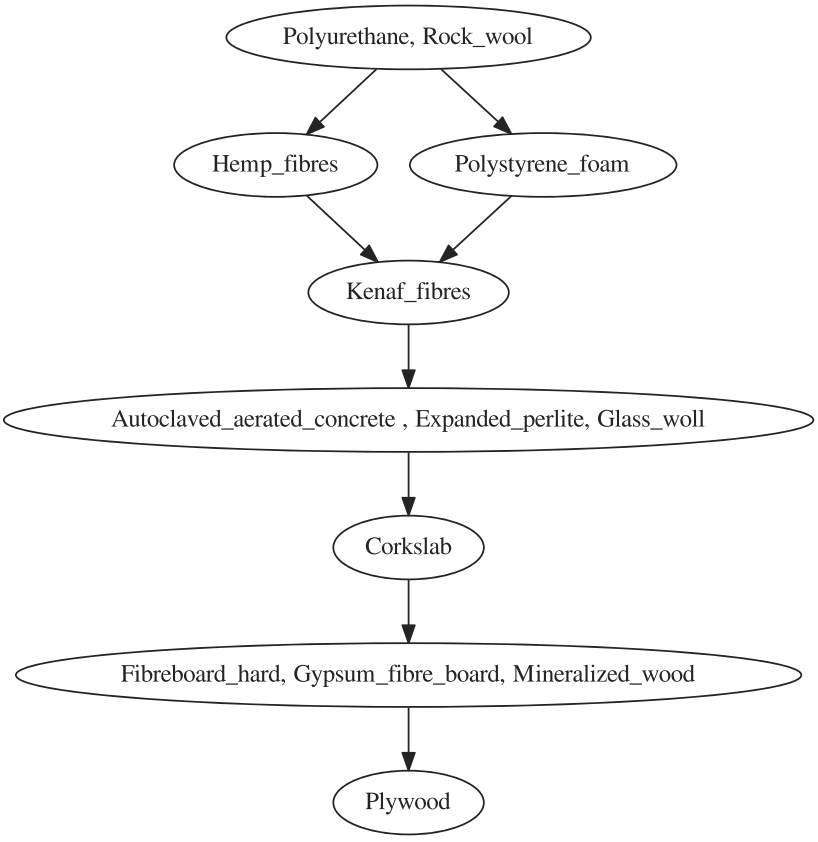
5.1. Related input rankings

The first diversity level concerns aggregation of input rankings, which are related to some degree, being generated from the same original ranking. We averaged the results over nine various levels of diversity (d). Each level was controlled by replacing a given ranking with its random neighbor and repeating this process sequentially d times. The computation times and qualities of rankings obtained for the utilitarian and egalitarian scenarios are presented in Tables 11 and 13, respectively. The respective frequencies of ranks attained by the considered algorithms in all considered scenarios are given in Tables 12 and 14.

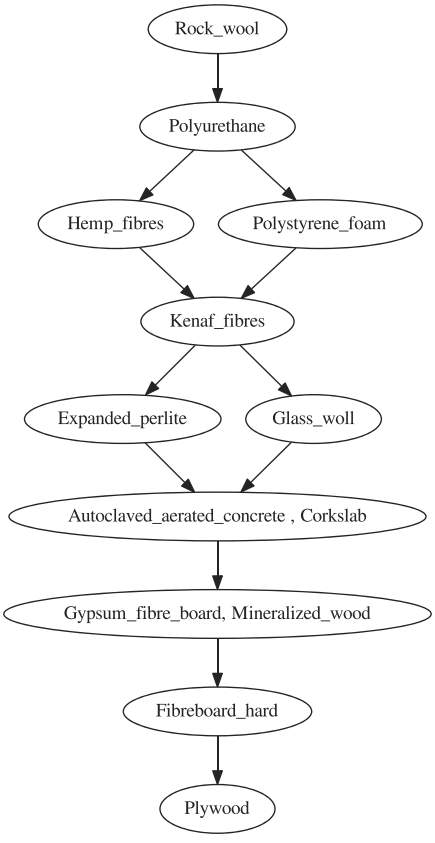
When it comes to constructing a utilitarian ranking, the best results are attained by the genetic Lamarckian approach and both variants of local search (see Table 11). These algorithms were ranked in the top three for, respectively, 70%, 84%, and 61% of considered settings (see Table 12). Even if the steepest local search attained the best average rank, the most advantageous mean quality and superiority in the greatest number of scenarios could be attributed to the genetic Lamarckian approach. These algorithms use input rankings as the initial points, which offers them a competitive advantage under scenarios with similar rankings provided at the input. Also, tabu search proved to be the best in 29% of considered scenarios, though its performance was relatively worse for other settings. All proposed metaheuristics achieved better results than heuristics designed particularly for dealing with this problem (pick-one, NFS-I and II). However, their computational costs were much higher. In this regard, the genetic Lamarckian approach's computations took significantly longer than for other methods (on average, over 2 minutes). In contrast, naive local search, tabu search, and simulated annealing found their best



(a) The best utilitarian ranking



(b) One of the best egalitarian rankings



(c) One of the best egalitarian rankings

Fig. 6. The best compromise rankings based on the 38000 inputs rankings consistent with the weight vectors derived with the Monte Carlo simulations.

Table 7
The best and average fitness along with the standard deviation attained by different algorithms in view of constructing an egalitarian ranking based on the 38000 inputs rankings consistent with the weight vectors derived with the Monte Carlo simulations.

Algorithm	Time [ms]	Best	Mean	St. dev.
Steepest Local Search	139281.3	65	65.0	0.0
Naive Local Search	68531.3	65	65.167	0.408
Genetic Lamarckian	195312.3	65	65.167	0.408
Simulated Annealing	19141.3	65	65.833	0.983
Tabu Search	9781.3	65	66.0	0.894
Genetic Algorithm	24608.2	66	66.167	0.408
Pick-One	357.5	71	71.0	0.0
NFS-I	92.7	73	73.0	0.0
NFS-II	102.3	74	74.0	0.0
Random Search	24933.2	115	121.167	5.707

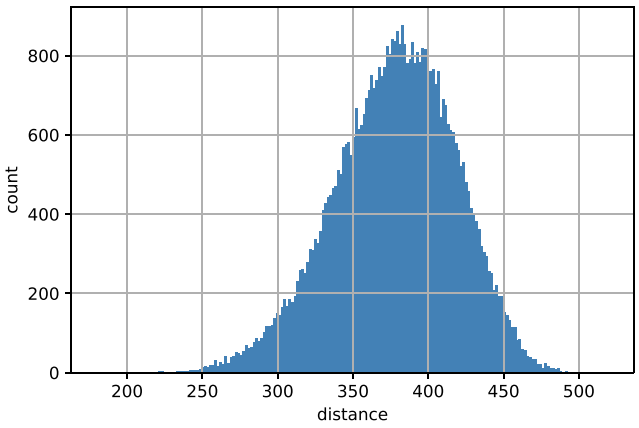


Fig. 7. The distribution of distances based on 500000 randomly generated rankings for the problem of constructing a utilitarian ranking based on the input ranking attained with the Monte Carlo simulations.

solutions in a much shorter time. The random search and NFS-based approaches attained the worst results. In all scenarios, selecting one of the input rankings proved to be more advantageous than for these three methods.

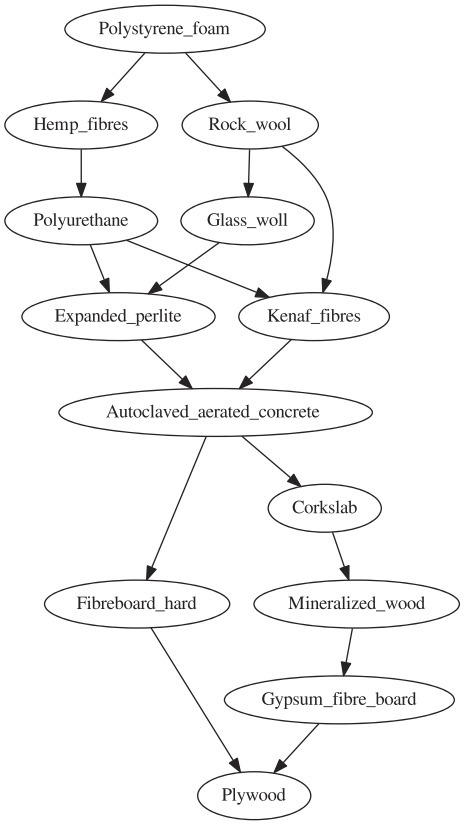
As far as the construction of an egalitarian ranking is concerned, the genetic Lamarckian approach confirmed its superiority over the remaining methods in terms of quality of the attained compromise ranking (see Table 13). Specifically, it attained the best results for 62% of considered settings (see Table 14). The performance of tabu search and simulated annealing was only slightly worse. However, these approaches found their solutions faster than genetic algorithms. The performance of both variants of the local search was relatively worse (the average ranks for the steepest and naive variants are, respectively, 4.51 and 4.65). The worst results were attained by the NFS-based methods, which were outperformed even by the random search in most scenarios.

Let us also emphasize that, in general, the standardized results obtained when minimizing an average distance were much better than when dealing with the optimization of a maximal distance. For example, the average quality of the utilitarian ranking constructed by the genetic Lamarckian algorithm is 5.546, whereas for the egalitarian case – it is 4.042. It may be implied by the higher diversity in the results when accounting for the egalitarian setting. Specifically, the observed standard deviation, in relation to the mean fitness, was more than 1.5 times higher when minimizing a maximal distance as compared to its average counterpart.

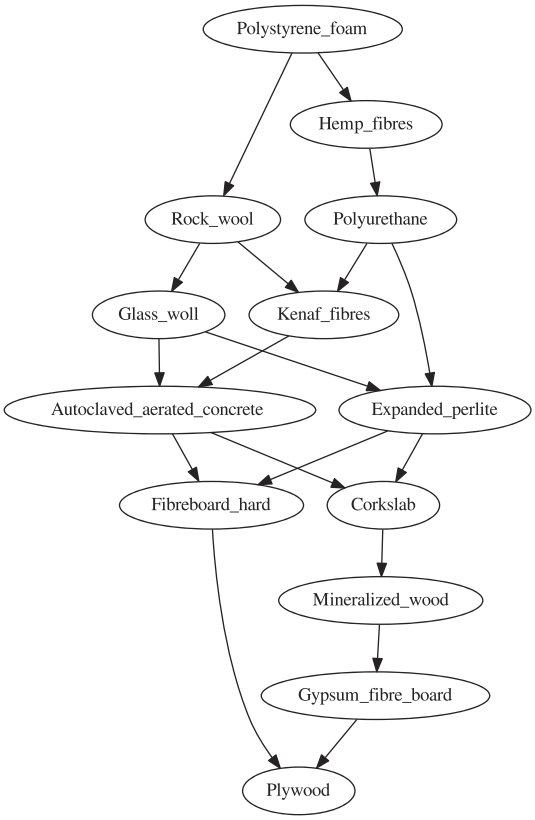
5.2. Random input rankings

The second considered setting related to the diversity of input data involved random input rankings. Thus, the diversity level was significantly higher than in the previous test, and the problem of constructing a compromise ranking is very challenging. Other parameters (i.e., the numbers of alternatives, input rankings, and iterations) remained the same.

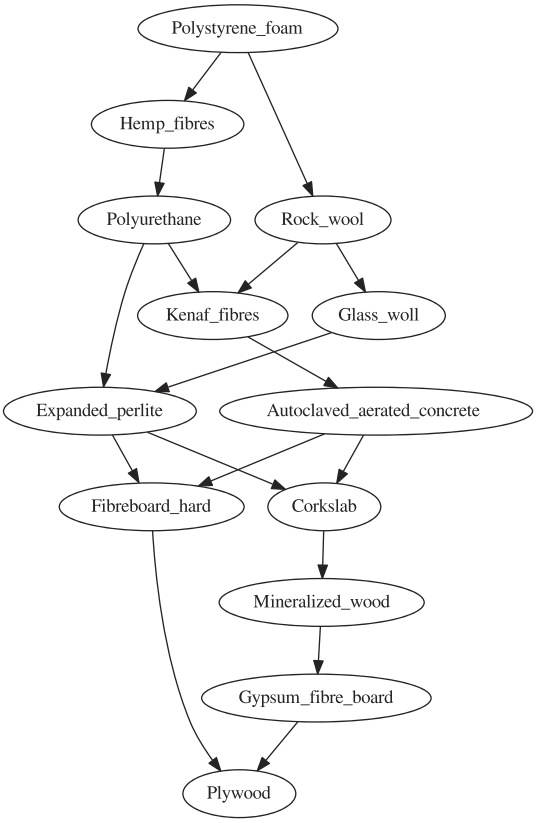
The computation time, quality of constructed ranking, and frequency of ranks attained by all algorithms in all runs for the case of creating a utilitarian ranking are presented in Tables 15 and 16. The qualities of rankings constructed by the metaheuristics are very similar. For example, the best and fifth mean fitness values (equal to 5.022 and 5.018, respectively) are attributed to the genetic Lamarckian approach and simulated annealing. However, the most favorable results in the highest number of runs were attained by tabu search and steepest local search (54% and 22%, respectively). In terms of the average ranks, they were followed by the simulated annealing and naive local search. The latter two approaches found their best solutions in the shortest time. The genetic algorithm attained the worst results among the metaheuristics, requiring, in addi-



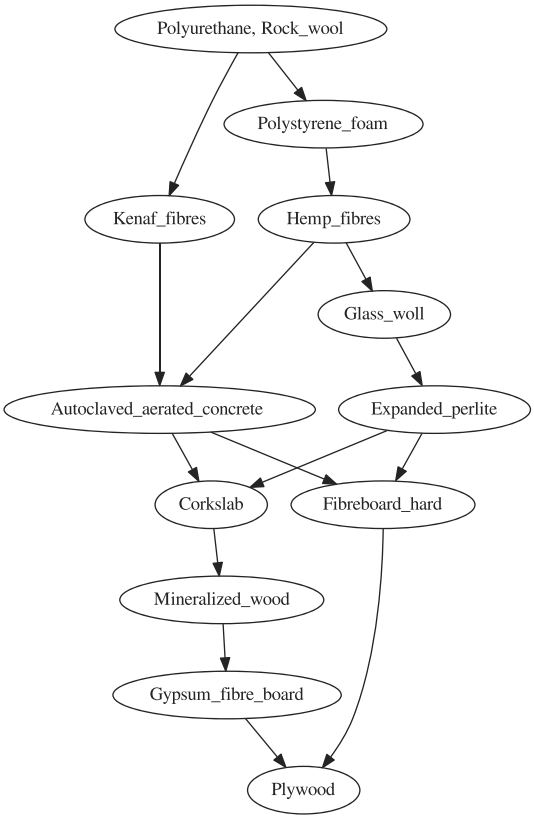
(a) Ranking for DM_1



(b) Ranking for DM_2 and DM_8



(c) Ranking for DM_{10}

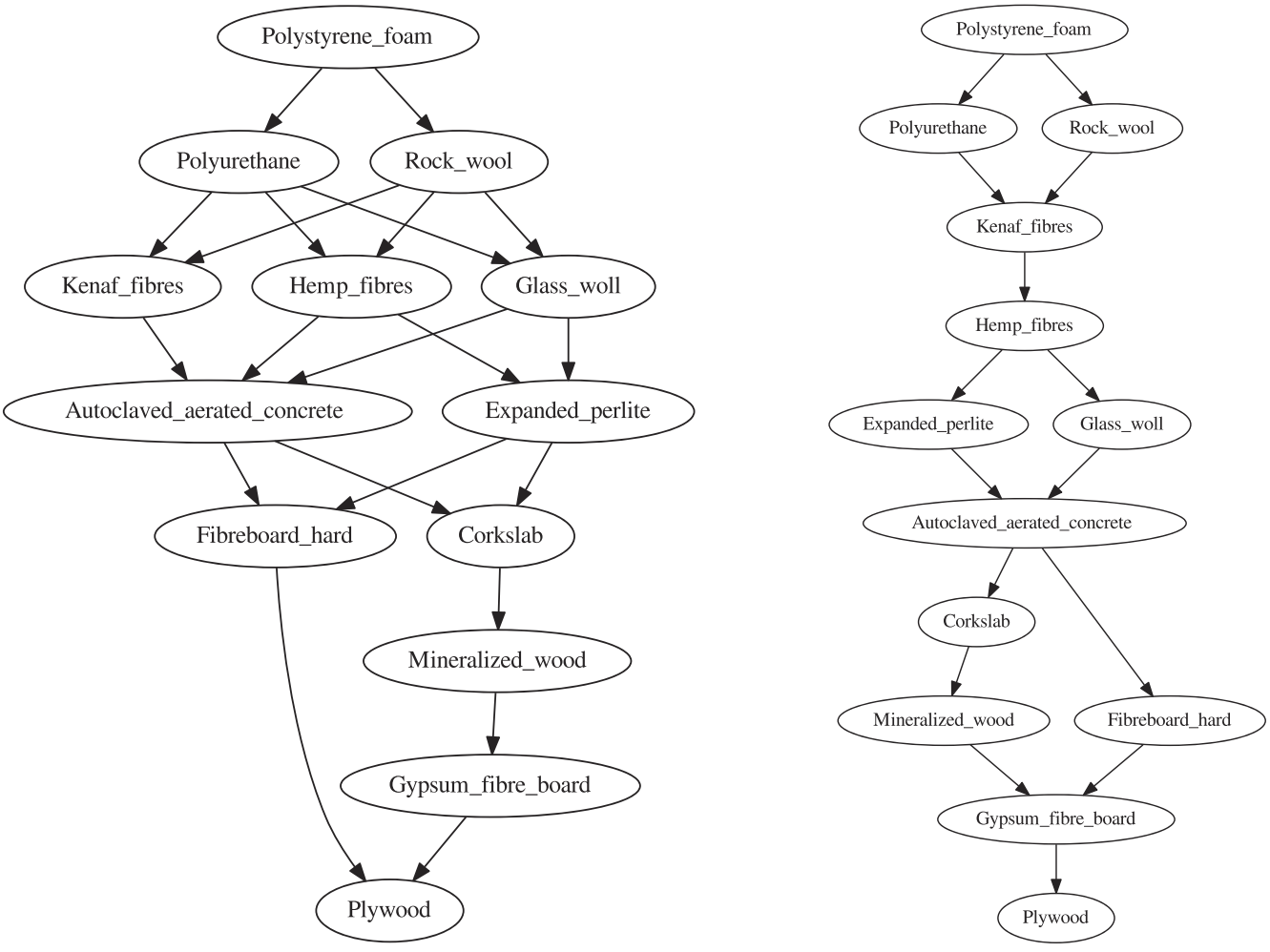


(d) Ranking for DM_6

Fig. 8. Rankings constructed for various Decision Makers for weights derived with the SRF procedure.

Table 8
The best and average fitness along with the standard deviation attained by different algorithms in view of constructing a utilitarian compromise ranking based on the 38 inputs rankings consistent with the weight vectors derived with the SRF procedure.

Algorithm	Time [ms]	Best	Mean	St. dev.
Steepest Local Search	306.0	25.579	25.579	0.0
Genetic Lamarckian	46627.5	25.579	25.798	0.107
Naive Local Search	134.0	25.842	25.842	0.0
Genetic Algorithm	2851.3	25.842	25.842	0.0
Simulated Annealing	1260.5	26.553	26.553	0.0
Pick-One	26.5	26.711	26.711	0.0
Tabu Search	244.5	26.842	28.583	1.369
NFS-I	5.7	33.184	33.184	0.0
NFS-II	6.0	34.289	34.289	0.0
Random Search	1961.5	89.105	92.395	2.674



(a) The best utilitarian ranking (b) The best egalitarian ranking

Fig. 9. The best compromise rankings for the 38 inputs rankings consistent with the weight vectors derived with the SRF procedure.

tion, almost 10 times more computation time than some of them. On the other extreme, the pick-one approach turned out to be better than the random search, which, in turn, outperformed the NFS-based approaches.

The lesser differences among the best performing methods are implied by the increased difficulty of the considered problem. When looking for the best compromise between random input rankings, the best solution is usually very different from the considered rankings. In this perspective, considering the input rankings as the starting points or the initial population does not offer a competitive advantage.

The detailed results for the scenarios oriented toward constructing an egalitarian ranking are given in Tables 17 and 18. The relative advantage of the best performing genetic Lamarckian approach is higher than in the previous case (the mean

Table 9
The best and average fitness along with the standard deviation attained by different algorithms in view of constructing an egalitarian compromise ranking based on the 38 inputs rankings consistent with the weight vectors derived with the SRF procedure.

Algorithm	Time [ms]	Best	Mean	St. dev.
Genetic Lamarckian	122964.8	42	42.167	0.408
Tabu Search	582.7	42	42.5	0.548
Simulated Annealing	1489.0	42	42.667	1.211
Naive Local Search	262.0	43	43.5	0.548
Steepest Local Search	549.8	43	43.833	0.408
Genetic Algorithm	3498.2	43	44.0	0.894
NFS-II	5.8	49	49.0	0.0
Pick-One	46.0	49	49.0	0.0
NFS-I	6.0	50	50.0	0.0
Random Search	2594.8	102	108.333	4.502

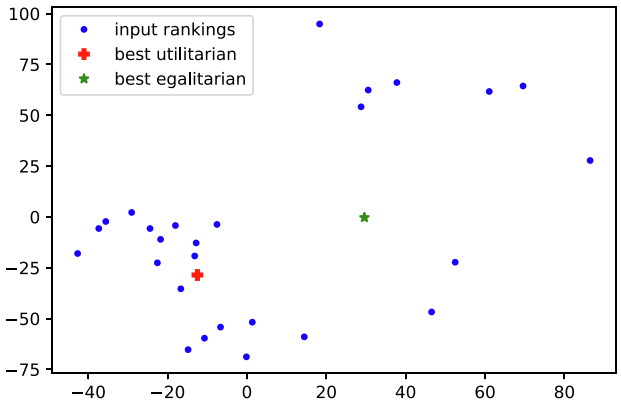


Fig. 10. The best compromise rankings for the 38 inputs rankings consistent with the weight vectors derived with the SRF procedure after Multi-Dimensional Scaling.

value of 3.703 as compared with 3.607 for simulated annealing and 3.575 for the genetic algorithm). This method identified the best solution in 64% of runs. It is followed by simulated annealing and tabu search, which proved to be the most advantageous for 12% and 24% of runs, respectively. Although simulated annealing attains the second-best average fitness of constructed ranking, tabu search performs poorly under some scenarios. As a result, the average quality of an egalitarian ranking constructed by the genetic algorithm is slightly better, though at the cost of twice as high computation time. When minimizing the maximal distance from the random input rankings, the local search algorithms proved to be less favorable. Both the average quality of the rankings and the attained ranks were the worst among the metaheuristic approaches. Nonetheless, it is worth noting that the naive variant was faster and, on average better than its steepest counterpart. This confirms that the latter is not always more advantageous than the former. Interestingly, when the considered problem involves so much randomness in the provided inputs, in over 30% of runs, the random search offered a better solution than picking the best among the input rankings.

Table 10
Distance of each input ranking from (U) the best utilitarian ranking and (E) the best egalitarian ranking.

Rank	DM_1	DM_2	DM_3	DM_4	DM_5	DM_6	DM_7	DM_8	DM_9
U	16	13	22	16	24	19	63	13	63
E	29	32	41	35	37	34	42	32	42
Rank	DM_{10}	DM_{11}	DM_{12}	DM_{13}	DM_{14}	DM_{15}	DM_{16}	DM_{17}	
U	16	16	52	14	14	10	11	55	
E	35	35	37	29	35	25	29	34	
Rank	DM_{18}	DM_{19}	DM_{20}	DM_{21}	DM_{22}	DM_{23}	DM_{24}		
U	15	48	51	7	16	60	22		
E	24	33	42	28	29	39	41		
Rank	DM_{25}	DM_{26}	DM_{27}	DM_{28}	DM_{29}	DM_{30}	DM_{31}		
U	14	22	32	36	10	48	6		
E	35	37	29	27	25	33	27		
Rank	DM_{32}	DM_{33}	DM_{34}	DM_{35}	DM_{36}	DM_{37}	DM_{38}		
U	19	16	12	17	52	16	16		
E	38	35	27	32	37	35	35		

Table 11
Results (computation time, mean fitness along with standard deviation) obtained when constructing a utilitarian ranking based on the related input rankings.

Algorithm	Time [ms]	Mean	St. dev.
Genetic Lamarckian	123141.5	5.546	1.3
Steepest Local Search	13124.3	5.539	1.322
Naive Local Search	2473.7	5.52	1.294
Genetic Algorithm	18688.8	5.45	1.24
Tabu Search	2815.9	5.416	1.14
Simulated Annealing	1934.7	5.371	1.075
Pick-One	58.2	4.942	1.159
NFS-II	0.7	3.549	1.037
Random Search	5638.3	3.356	0.312
NFS-I	0.4	0.337	1.591

Table 12
Frequency of ranks attained by different algorithms when constructing a utilitarian ranking based on the related input rankings.

Algorithm	1	2	3	4	5	6	7	8	9	10	Mean
Steepest Local Search	0.23	0.27	0.34	0.10	0.05	0.01	0.00	0.00	0.00	0.00	2.48
Genetic Lamarckian	0.37	0.21	0.12	0.06	0.08	0.14	0.02	0.00	0.00	0.00	2.78
Naive Local Search	0.06	0.20	0.35	0.31	0.08	0.01	0.00	0.00	0.00	0.00	3.17
Tabu Search	0.29	0.06	0.08	0.19	0.25	0.12	0.01	0.00	0.00	0.00	3.43
Simulated Annealing	0.03	0.23	0.07	0.12	0.25	0.26	0.04	0.00	0.00	0.00	4.27
Genetic Algorithm	0.01	0.02	0.04	0.22	0.27	0.44	0.00	0.00	0.00	0.00	5.04
Pick-One	0.00	0.02	0.00	0.00	0.02	0.04	0.92	0.00	0.00	0.00	6.84
NFS-II	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.57	0.42	0.01	8.44
Random Search	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.53	0.05	8.62
NFS-I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.94	9.93

Table 13
Results (computation time, mean fitness along with standard deviation) obtained when constructing an egalitarian ranking based on the related input rankings.

Algorithm	Time [ms]	Mean	St. dev.
Genetic Lamarckian	419696.4	4.042	0.862
Tabu Search	8602.9	3.871	0.767
Simulated Annealing	3008.8	3.851	0.72
Genetic Algorithm	16992.1	3.847	0.813
Steepest Local Search	11901.1	3.814	0.889
Naive Local Search	3427.8	3.809	0.863
Pick-One	69.9	2.706	0.865
Random Search	6871.0	2.439	0.225
NFS-II	0.8	2.23	0.836
NFS-I	0.2	0.062	0.995

Table 14
Frequency of ranks attained by different algorithms when constructing an egalitarian ranking based on the related input rankings.

Algorithm	1	2	3	4	5	6	7	8	9	10	Mean
Genetic Lamarckian	0.62	0.27	0.08	0.02	0.00	0.00	0.00	0.00	0.00	0.00	1.51
Tabu Search	0.22	0.27	0.16	0.14	0.10	0.11	0.00	0.00	0.00	0.00	2.99
Simulated Annealing	0.05	0.20	0.33	0.21	0.10	0.09	0.01	0.00	0.00	0.00	3.43
Genetic Algorithm	0.04	0.13	0.20	0.27	0.18	0.17	0.00	0.00	0.00	0.00	3.92
Steepest Local Search	0.05	0.07	0.10	0.16	0.30	0.31	0.00	0.00	0.00	0.00	4.51
Naive Local Search	0.01	0.05	0.12	0.20	0.32	0.30	0.00	0.00	0.00	0.00	4.65
Pick-One	0.00	0.00	0.00	0.00	0.00	0.01	0.56	0.43	0.00	0.00	7.41
Random Search	0.00	0.00	0.00	0.00	0.00	0.00	0.43	0.19	0.37	0.01	7.97
NFS-II	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.60	0.02	8.65
NFS-I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.96	9.96

Table 15
Results (computation time, mean fitness along with standard deviation) obtained when constructing a utilitarian ranking based on the random input rankings.

Algorithm	Time [ms]	Mean	St. dev.
Genetic Lamarckian	83770.9	5.022	0.934
Steepest Local Search	21723.9	5.021	0.933
Naive Local Search	2873.7	5.020	0.933
Tabu Search	3169.4	5.019	0.933
Simulated Annealing	2526.5	5.018	0.933
Genetic Algorithm	23040.3	5.009	0.93
Pick-One	68.9	4.977	0.927
Random Search	6565.5	3.28	0.304
NFS-II	0.3	2.43	0.7
NFS-I	0.2	−1.71	0.683

Table 16
Frequency of ranks attained by different algorithms when constructing a utilitarian ranking based on the random input rankings.

Algorithm	1	2	3	4	5	6	7	8	9	10	Mean
Tabu Search	0.54	0.18	0.08	0.10	0.10	0.00	0.00	0.00	0.00	0.0	2.03
Steepest Local Search	0.22	0.14	0.36	0.21	0.06	0.01	0.00	0.00	0.00	0.0	2.78
Simulated Annealing	0.04	0.35	0.29	0.08	0.15	0.08	0.00	0.00	0.00	0.0	3.21
Naive Local Search	0.04	0.15	0.17	0.42	0.19	0.03	0.00	0.00	0.00	0.0	3.65
Genetic Lamarckian	0.15	0.11	0.10	0.12	0.33	0.11	0.07	0.00	0.00	0.0	3.99
Genetic Algorithm	0.00	0.00	0.00	0.07	0.14	0.67	0.12	0.00	0.00	0.0	5.85
Pick0One	0.00	0.07	0.00	0.00	0.03	0.10	0.81	0.00	0.00	0.0	6.50
Random Search	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.08	0.0	8.08
NFS-II	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.92	0.0	8.92
NFS-I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.0	10.00

Table 17
Results (computation time, mean fitness along with standard deviation) obtained when constructing an egalitarian ranking based on the random input rankings.

Algorithm	Time [ms]	Mean	St. dev.
Genetic Lamarckian	524177.2	3.703	0.569
Simulated Annealing	4195.8	3.607	0.524
Genetic Algorithm	22686.6	3.575	0.53
Tabu Search	12690.4	3.535	0.529
Naive Local Search	4938.4	3.47	0.515
Steepest Local Search	19054.1	3.441	0.508
Pick-One	86.1	2.604	0.541
Random Search	8805.4	2.387	0.182
NFS-II	0.8	1.611	0.604
NFS-I	0.1	−1.12	0.63

Table 18
Frequency of ranks attained by different algorithms when constructing an egalitarian ranking based on the random input rankings.

Algorithm	1	2	3	4	5	6	7	8	9	10	Mean
Genetic Lamarckian	0.64	0.22	0.11	0.03	0.00	0.00	0.00	0.00	0.00	0.0	1.53
Simulated Annealing	0.12	0.29	0.42	0.15	0.01	0.00	0.00	0.00	0.00	0.0	2.64
Tabu Search	0.24	0.25	0.11	0.12	0.10	0.17	0.00	0.01	0.00	0.0	3.17
Genetic Algorithm	0.00	0.21	0.24	0.35	0.17	0.04	0.00	0.00	0.00	0.0	3.60
Naive Local Search	0.00	0.01	0.06	0.24	0.36	0.33	0.00	0.00	0.00	0.0	4.94
Steepest Local Search	0.00	0.01	0.07	0.11	0.36	0.44	0.00	0.00	0.00	0.0	5.15
Pick-One	0.00	0.00	0.00	0.00	0.00	0.01	0.62	0.36	0.00	0.0	7.35
Random Search	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.54	0.08	0.0	7.71
NFS-II	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.92	0.0	8.92
NFS-I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.0	10.00

5.3. Related input rankings involving a single random ranking

The third setting concerning a diversity of the input rankings can be seen as a combination of the previous two. The considered scheme and parameters were the same as in the case of dealing with the related input rankings except for including a single randomly generated ranking in each scenario. The latter makes the problem of finding a compromise ranking more challenging.

When it comes to the construction of a utilitarian ranking, the results in terms of the quality of constructed compromise ranking, computation times, and attained ranks (see Tables 19 and 20) are very similar to those observed for the setting when all rankings are related. Hence, the average quality of the ranking constructed by the genetic Lamarckian approach is the best at the cost of very high computation time. It is followed by the local search steepest and naive, with the former attaining the best average rank across all runs. Other metaheuristic approaches outperform tabu search and simulated annealing in terms of the quality of identified rankings. However, their computation cost is low, which confirms that they can find a relatively good solution quickly.

As far as the identification of an egalitarian ranking is concerned, the outcomes (see Tables 21 and 22) are very similar to those reported for the setting with all input rankings generated randomly. Hence, the Lamarckian genetic approach attains the best results. Simulated annealing finds a very good solution quickly, and the local search algorithms are the worst among the metaheuristics. The slight differences are as follows. First, the Lamarckian genetic approach's advantage is greater as it outperforms other methods in 70% of runs. Second, tabu search compares to the genetic algorithm better in terms of the mean quality and the simulated annealing given an average rank across all runs. Third, the differences in the performance of the two variants of local search are lesser.

5.4. Summary of results for all settings considered in the experimental evaluation

In this section, we summarize the experimental results on artificial data with various characteristics. In Table 23, we present the mean fitness attained by ten considered methods for different settings. These results are associated with the Hasse diagrams illustrating the rankings of approaches concerning the statistically significant differences in their performances (see Fig. 11). Specifically, to confirm the observed differences in performance, we first conducted the Friedman test. Its outcomes allowed us to reject the hypothesis about no differences in various algorithms' performance for all considered settings. Then, the Wilcoxon signed-rank test was performed for each pair of methods under the six considered settings. For most pairs, we confirmed a statistically significant difference in their performances, assuming a significance level $\alpha = 0.01$. Precisely, whenever an arc connects some pair of algorithms in Fig. 11, this means that a difference in their performances is statistically significant in terms of the quality of constructed group compromise rankings. This way of presenting the results of statistical tests has been adopted after [8].

Overall, the Lamarckian variant of the genetic algorithm proved to be the best for all scenarios, attaining the most advantageous mean fitness. Its advantage in performance over all remaining methods is statistically significant for all considered setting except the construction of a utilitarian ranking for random input rankings (see Fig. 11c). Moreover, its competitive advantage over other algorithms is more evident for the egalitarian setting, with differences in the maximal distance being more remarkable than when the average distance is optimized in the utilitarian approach.

The Lamarckian genetic algorithm's best competitors differ depending on whether we minimize an average or a maximal distance from the input rankings. When it comes to the utilitarian approach, the second and third ranks are occupied by the steepest and naive local search variants, respectively. As far as the egalitarian setting is considered, tabu search and simulated annealing proved to more advantageous. Moreover, when considering the random input rankings or related ranking with one random ranking, the differences between performances of the latter two approaches are not statistically significant. The advantage of all six metaheuristic algorithms over the four simple procedures for constructing an incomplete compro-

Table 19
Results (computation time, mean fitness along with standard deviation) obtained when constructing a utilitarian ranking based on the related rankings involving one random ranking.

Algorithm	Time [ms]	Mean	St. dev.
Genetic Lamarckian	125952.2	5.531	1.285
Steepest Local Search	15225.3	5.527	1.308
Naive Local Search	2805.2	5.508	1.279
Genetic Algorithm	20706.9	5.439	1.222
Tabu Search	3113.2	5.397	1.111
Simulated Annealing	2240.7	5.368	1.066
Pick-One	61.9	4.981	1.147
NFS-II	0.5	3.481	1.027
Random Search	5934.2	3.343	0.283
NFS-I	0.4	0.239	1.586

Table 20
Frequency of ranks attained by different algorithms when constructing a utilitarian ranking based on the related rankings involving one random ranking.

Name	1	2	3	4	5	6	7	8	9	10	Mean
Steepest Local Search	0.26	0.25	0.33	0.12	0.03	0.00	0.00	0.00	0.00	0.00	2.42
Genetic Lamarckian	0.34	0.20	0.13	0.08	0.09	0.16	0.01	0.00	0.00	0.00	2.91
Naive Local Search	0.05	0.20	0.33	0.34	0.07	0.00	0.00	0.00	0.00	0.00	3.18
Tabu Search	0.30	0.07	0.08	0.17	0.22	0.14	0.03	0.00	0.00	0.00	3.49
Simulated Annealing	0.05	0.23	0.07	0.11	0.30	0.20	0.04	0.00	0.00	0.00	4.13
Genetic Algorithm	0.00	0.03	0.06	0.19	0.27	0.45	0.00	0.00	0.00	0.00	5.04
Pick-One	0.00	0.01	0.00	0.00	0.03	0.03	0.92	0.00	0.00	0.00	6.84
NFS-II	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.57	0.43	0.01	8.44
Random Search	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.53	0.04	8.60
NFS-I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.96	9.95

Table 21
Results (computation time, mean fitness along with standard deviation) obtained when constructing an egalitarian ranking based on the related rankings involving one random ranking.

Algorithm	Time [ms]	Mean	St. dev.
Genetic Lamarckian	420151.3	3.897	0.671
Simulated Annealing	3407.1	3.743	0.615
Tabu Search	10192.7	3.736	0.628
Genetic Algorithm	19478.5	3.711	0.62
Naive Local Search	4018.1	3.653	0.628
Steepest Local Search	14640.0	3.641	0.617
Pick-One	75.1	2.522	0.637
Random Search	7448.2	2.415	0.226
NFS-II	0.8	2.049	0.678
NFS-I	0.4	−0.407	0.91

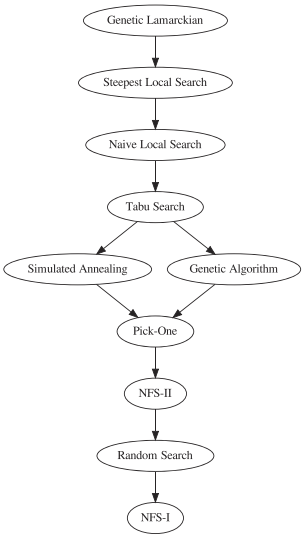
Table 22
Frequency of ranks attained by different algorithms when constructing an egalitarian ranking based on the related rankings involving one random ranking.

Algorithm	1	2	3	4	5	6	7	8	9	10	Mean
Genetic Lamarckian	0.70	0.20	0.09	0.01	0.01	0.00	0.00	0.00	0.00	0.00	1.44
Tabu Search	0.19	0.31	0.17	0.13	0.08	0.12	0.00	0.00	0.00	0.00	2.96
Simulated Annealing	0.07	0.26	0.33	0.20	0.09	0.05	0.00	0.00	0.00	0.00	3.14
Genetic Algorithm	0.03	0.15	0.21	0.27	0.16	0.17	0.00	0.00	0.00	0.00	3.91
Steepest Local Search	0.01	0.06	0.09	0.18	0.34	0.33	0.00	0.00	0.00	0.00	4.76
Naive Local Search	0.00	0.03	0.11	0.20	0.31	0.33	0.00	0.00	0.00	0.00	4.79
Pick-One	0.00	0.00	0.00	0.00	0.00	0.00	0.52	0.48	0.00	0.00	7.48
Random Search	0.00	0.00	0.00	0.00	0.00	0.00	0.48	0.25	0.27	0.00	7.78
NFS-II	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.26	0.72	0.01	8.75
NFS-I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.99	9.98

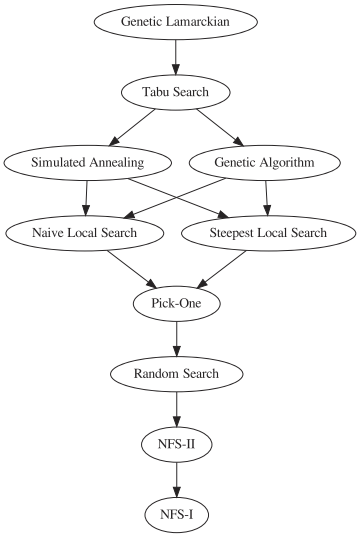
Table 23
Mean fitness attained by different algorithms when constructing a utilitarian or an egalitarian ranking under different settings (the respective ranks implied by these values are provided in the round brackets).

	Related rankings		Random rankings		Single random ranking	
Algorithm	Utilitarian	Egalitarian	Utilitarian	Egalitarian	Utilitarian	Egalitarian
Genetic Algorithm	5.450 (4)	3.847 (4)	5.009 (6)	3.575 (3)	5.439 (4)	3.711 (4)
Genetic Lamarckian	5.546 (1)	4.042 (1)	5.022 (1)	3.703 (1)	5.531 (1)	3.897 (1)
Naive Local Search	5.520 (3)	3.809 (6)	5.020 (3)	3.470 (5)	5.508 (3)	3.653 (5)
NFS-I	0.337 (10)	0.062 (10)	−1.71 (10)	−1.120 (10)	0.239 (10)	−0.407 (10)
NFS-II	3.549 (8)	2.230 (9)	2.430 (9)	1.611 (9)	3.481 (8)	2.049 (9)
Pick-One	4.942 (7)	2.706 (7)	4.977 (7)	2.604 (7)	4.981 (7)	2.522 (7)
Random Search	3.356 (9)	2.439 (8)	3.280 (8)	2.387 (8)	3.343 (9)	2.415 (8)
Simulated Annealing	5.371 (6)	3.851 (3)	5.018 (5)	3.607 (2)	5.368 (6)	3.743 (2)
Steepest Local Search	5.539 (2)	3.814 (5)	5.021 (2)	3.441 (6)	5.527 (2)	3.641 (6)
Tabu Search	5.416 (5)	3.871 (2)	5.019 (4)	3.535 (4)	5.397 (5)	3.736 (3)

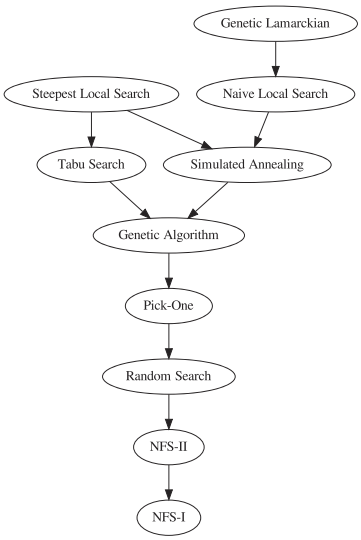
mise ranking is statistically significant. Among the worst-performing methods, NFS-I is always ranked at the bottom. In contrast, pick-one is the best (i.e., seventh overall), proving its superiority over Random Search and NFS-based approaches.



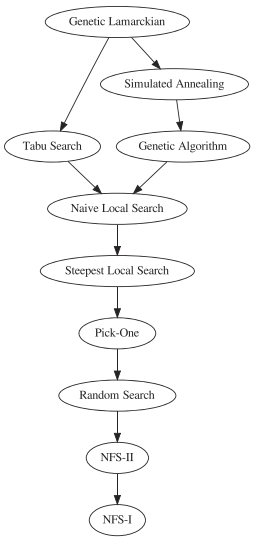
(a) Utilitarian approach with related rankings



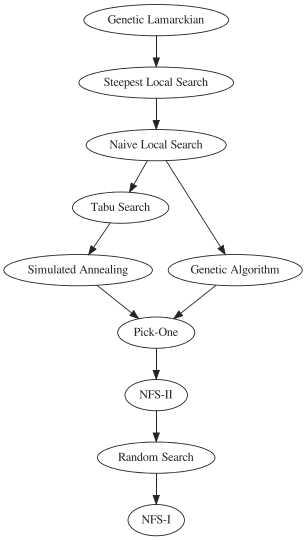
(b) Egalitarian approach with related rankings



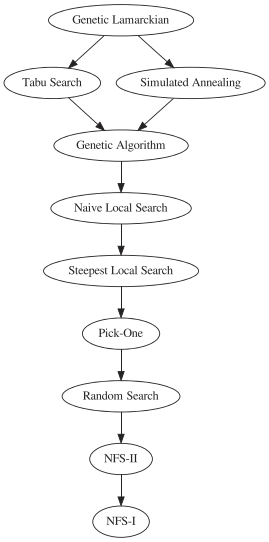
(c) Utilitarian approach with random rankings



(d) Egalitarian approach with random rankings



(e) Utilitarian approach with one random ranking



(f) Egalitarian approach with one random ranking

Fig. 11. Hasse diagrams illustrating the rankings of algorithms with respect to the statistically significant differences in their performances for $\alpha = 0.01$.

6. Conclusions and future research

We considered the problem of constructing a compromise incomplete ranking based on individual partial rankings admitting incomparability. The compromise is interpreted in terms of minimizing the maximal or an average distance from the elementary rankings. The tackled problem is computationally complex, and for large instances – it cannot be solved precisely due to the limitations of the contemporary solvers. Thus, we proposed various heuristics that either adjusted metaheuristic methods, such as genetic algorithms, simulated annealing, tabu search, or local search, to the considered setting or incorporated some intuitive procedures for finding a compromise ranking.

We verified the practical usefulness of these methods in a twofold way. On the one hand, we applied them for constructing the compromise rankings for real-world data concerning insulating materials. We considered the preferences of thirty-eight Decision Makers, taking into account either a single representative ranking for each DM or a multiplicity of such rankings derived from the robustness analysis. On the other hand, we conducted extensive experimental comparison on artificially generated problems. These problems differed in terms of the numbers of alternatives and input rankings and the diversity level.

The performance of algorithms depends on the characteristics of the tackled problem. There are, however, some general conclusions that follow the experimental outcomes. The Lamarckian genetic algorithm constructs, on average, the best compromise rankings at the cost of high computation time. Similar outcomes in the context of minimizing an average distance can be obtained with the local search algorithms. They find the compromise rankings faster than the genetic algorithms, though still requiring significant time. When the latter is a paramount concern, it is recommended to use simulated annealing, which proves to be the fastest metaheuristic. This method was one of the best approaches when minimizing the maximal distance, mainly when the input rankings differed vastly. In both utilitarian and egalitarian settings, tabu search attained the best outcomes for many considered instances. However, its computations took longer than for simulated annealing, and for some other instances, it was outperformed by all remaining metaheuristics.

The proposed algorithms work, without any additional adjustments, for any distance matrix for the preference, indifference, and incomparability relations. In this paper, we used the distances that were justified by a set of axioms as well as logical and significance conditions. However, by setting them to some arbitrarily selected values, one may provide some specific guidelines for constructing a group compromise ranking. For example, if the distances between incomparability and other relations would be relatively great, the algorithm might avoid imposing incomparability in the final preorder. Moreover, even though symmetry is one of the basic axioms underlying the definition of a distance, one can opt to use an asymmetric matrix. In particular, one can promote the preference relation over incomparability in the output ranking. Assuming that in an ordered pair of relations, the first one corresponds to the input ranking and the other one to the output ranking, this can be attained, e.g., by requiring that a distance for $(?, P^+)$ is lesser than for $(P^+, ?)$. Such adaptations should account for a trade-off between using a rigorous mathematical concept and adequately representing a particular group decision problem.

We envisage the following directions for future research. First, when executing the algorithms, we assumed that the values of some parameters (e.g., the number of iterations not yielding any improvement, the probability of mutation in the genetic algorithms, or the pace of temperature decrease in simulated annealing) were fixed. These values were selected following the results of some internal experiments. However, in general, a set of parameter values leading to the best results could be different for each problem. Hence, one should investigate how to select them for a particular instance. Second, to compare various algorithms fairly, they shared the same operators for modifying the rankings. Even though we tested a variety of options for these operators (the worse performing were not reported in the paper), it would be beneficial to verify even more variants. The most promising direction concerns the elaboration of the neighborhood operators based on fuzzy logic. In the same spirit, one could work on novel procedures for restoring the ranking's transitivity based on, e.g., cutting some arcs that would break the potential cycles. Third, other metaheuristics or combinations of the already used ones can be tested. For example, a local search method can be used to find some local optima that can be subsequently employed as an initial population for the genetic algorithms.

Finally, let us emphasize that our contribution was mainly methodological and consisted of introducing various approaches for constructing an egalitarian or a utilitarian compromise ranking. In most real-world decision problems, only one approach would be adopted. An exciting and practically relevant direction for future research consists of verifying the suggested ranking in terms of the DMs' commitment to the agreement, sense of justice, and satisfaction [50]. The latter two variables can be measured on a Likert-type scale of, e.g., five levels. According to [50], the chances that the sense of justice and satisfaction are greater when compromise is supported by some mathematical or decision support method. However, such verification can be conducted only in the context of a concrete real-world problem involving real-world DMs. Therefore, it cannot be part of the experimental verification where preferences or individual rankings are simulated. Testing and employing the proposed algorithms to real-world group decision problems is the most appealing direction for future works.

CRediT authorship contribution statement

Grzegorz Miebs: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing - original draft, Visualization. **Miłosz Kadziński:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Grzegorz Miebs acknowledges support from the Polish Ministry of Science and Higher Education under the Diamond Grant project (Grant No. DI2018 004348). Miłosz Kadziński acknowledges support from the Polish National Science Center under the SONATA BIS project (Grant No. DEC-2019/34/E/HS4/00045). The authors thank the editors and three anonymous reviewers whose remark allowed us to improve the paper significantly.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ins.2021.01.055.

References

[1] L. Abdullah, W. Chan, A. Afshari, Application of PROMETHEE method for green supplier selection: a comparative result based on preference functions, Journal of Industrial Engineering International 15 (2) (2019) 271–285.
[2] W. Ben-Ameur, Computing the initial temperature of simulated annealing, Computational Optimization and Applications 29 (3) (2004) 369–385.
[3] D. Bollinger, L. Maystre, R. Słowiński, Un cas concret de la contribution de l'aide multicritére á la décision á la coordination inter-cantonale pour l'incinération des déchets urbains. In: 45th Meeting of the EURO Working Group on Multiple Criteria Decision Aiding. Celàkovice, Czech Republic, 1997.
[4] D. Bouyssou, T. Marchant, P. Perny, Social Choice Theory and Multicriteria Decision Aiding, John Wiley & Sons, Ltd, Ch. 19, 2010, pp. 779–810.
[5] J. Brans, P. Vincke, A preference ranking organisation method, Management Science 31 (6) (1985) 647–656.
[6] Y. Chen, J. Witmer, K. Hipel, D. Kilgour, Strategic decision support for brownfield redevelopment, in: IEEE International Conference on Systems, Man and Cybernetics, 2007, pp. 1860–1865.
[7] M. Cinelli, M. Kadziński, M. Gonzalez, R. Słowiński, How to support the application of multiple criteria decision analysis? Let us start with a comprehensive taxonomy, Omega 96 (2020) 102261.
[8] K. Ciomek, M. Kadziński, T. Tervonen, Heuristics for selecting pair-wise elicitation questions in multiple criteria choice problems, European Journal of Operational Research 262 (2) (2017) 693–707.
[9] G. Colson, The OR's prize winner and the software ARGOS: how a multijudge and multicriteria ranking GDSS helps a jury to attribute a scientific award, Computers & Operations Research 27 (7) (2000) 741–755.
[10] P. Contucci, A. Sibu, Egalitarianism vs. Utilitarianism in Preferential Voting: An Interdisciplinary Approach, 2019, pp. 24–37.
[11] L. Dias, J. Climaco, Dealing with imprecise information in group multicriteria decisions: a methodology and a GDSS architecture, European Journal of Operational Research 160 (2) (2005) 291–307.
[12] J. Ding, D. Han, Y. Yang, Iterative ranking aggregation using quality improvement of subgroup ranking, European Journal of Operational Research 268 (2) (2018) 596–612.
[13] M. Doumpos, Y. Marinakis, M. Marinaki, C. Zopounidis, An evolutionary approach to construction of outranking models for multicriteria classification: The case of the ELECTRE TRI method, European Journal of Operational Research 199 (2009) (2008) 496–505.
[14] V. Ferretti, A. Degioanni, How to support the design and evaluation of redevelopment projects for disused railways? A methodological proposal and key lessons learned, Transportation Research Part D: Transport and Environment 52 (2017) 29–48.
[15] J. Figueira, B,R., Determining the weights of criteria in the ELECTRE type methods with a revised Simos' procedure, European Journal of Operational Research 139 (2) (2002) 317–326.
[16] P. Gasser, J. Suter, M. Cinelli, M. Spada, P. Burgherr, S. Hirschberg, M. Kadziński, B. Stojadinović, Comprehensive resilience assessment of electricity supply security for 140 countries, Ecological Indicators 110 (2020) 105731.
[17] F. Glover, Tabu Search – Part I, ORSA Journal on Computing 1 (3) (1989) 190–206.
[18] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, first ed., Addison-Wesley Longman Publishing Co., Inc., USA, 1989.
[19] J. Gonzalez-Pachon, C. Romero, Aggregation of partial ordinal rankings: an interval goal programming approach, Computers & Operations Research 28 (8) (2001) 827–834.
[20] K. Govindan, M. Kadziński, R. Ehling, G. Miebs, Selection of a sustainable third-party reverse logistics provider based on the robustness analysis of an outranking graph kernel conducted with ELECTRE I and SMAA, Omega 85 (2019) 1–15.
[21] K. Govindan, M. Kadziński, R. Sivakumar, Application of a novel PROMETHEE-based method for construction of a group compromise ranking to prioritization of green suppliers in food supply chain, Omega 71 (2017) 129–145.
[22] S. Greco, M. Kadziński, V. Mousseau, R. Słowiński, Robust ordinal regression for multiple criteria group decision: UTA-GMS-GROUP and UTADIS-GMS-GROUP, Decision Support Systems 52 (3) (2012) 549–561.
[23] S. Greco, V. Mousseau, R. Słowiński, Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions, European Journal of Operational Research 191 (2) (2008) 416–436.
[24] J.C. Harsanyi, Nonlinear social welfare functions, Theory and Decision 6 (3) (1975) 311–332.
[25] J.C. Harsanyi, Bayesian decision theory and utilitarian ethics, The American Economic Review 68 (2) (1978) 223–228.
[26] M. Kadziński, S. Greco, R. Słowiński, Extreme ranking analysis in robust ordinal regression, Omega 40 (4) (2012) 488–501.
[27] M. Kadziński, L. Rocchi, G. Miebs, D. Grohmann, M.E. Menconi, L. Paolotti, Multiple criteria assessment of insulating materials with a group decision framework incorporating outranking preference model and characteristic class profiles, Group Decision and Negotiation 27 (1) (2018) 33–59.
[28] D.M. Kilgour, Y. Chen, K.W. Hipel, Multiple Criteria Approaches to Group Decision and Negotiation, Springer, Boston, USA, 2010, pp. 317–338.
[29] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671–680.
[30] J.B. Kruskal, Nonmetric multidimensional scaling: A numerical method, Psychometrika 29 (2) (1964) 115–129.
[31] J. Leyva-Lopez, E. Fernandez-Gonzalez, A new method for group decision support based on ELECTRE III methodology, European Journal of Operational Research 148 (1) (2003) 14–27.
[32] S. Lin, Rank aggregation methods, WIREs Computational Statistics 2 (5) (2010) 555–570.
[33] B. Mareschal, J.P. Brans, C. Macharis, The GDSS PROMETHEE procedure: a PROMETHEE-GAIA based procedure for group decision support, Journal of Decision Systems 7 (1998) 283–307.

[34] K. Mastorakis, E. Siskos, Value focused pharmaceutical strategy determination with multicriteria decision analysis techniques, Omega 59 (2016) 84–96.
[35] R.B. Myerson, Utilitarianism, egalitarianism, and the timing effect in social choice problems, Econometrica 49 (4) (1981) 883–897.
[36] M. Norese, ELECTRE III as a support for participatory decision-making on the localisation of waste-treatment plants, Land Use Policy 23 (1) (2006) 76–85.
[37] D. Parkes, S. Seuken, Social Choice and Rank Aggregation, Cambridge University Press, Ch. 13, 2013, pp. 323–367.
[38] J. Rawls, Some reasons for the maximin criterion, American Economic Review 64 (2) (1974) 141–146.
[39] M. Rogers, M. Bruen, L.-Y. Maystre, Electre and Decision Support, Springer, NY, USA, 2000.
[40] B. Roy, ELECTRE III: Un algorithme de classements fondé sur une représentation floue des préférences en présence de critères multiples, Cahiers du CERO 20 (1) (1978) 3–24.
[41] B. Roy, The outranking approach and the foundations of ELECTRE methods, Theory and Decision 31 (1) (1991) 49–73.
[42] B. Roy, R. Słowiński, Criterion of distance between technical programming and socio-economic priority, RAIRO Recherche Opérationnelle 27 (1) (1993) 45–60.
[43] A. Salo, R. Hamalainen, Preference programming through approximate ratio comparisons, European Journal of Operational Research 82 (3) (1995) 458–475.
[44] G. Syswerda, A study of reproduction in generational and steady-state genetic algorithms, vol. 1 of Foundations of Genetic Algorithms, Elsevier, 1991, pp. 94–101.
[45] C. Tam, T. Tong, C. Lau, ELECTRE III in evaluating performance of construction plants: case study on concrete vibrators, Construction Innovation 3 (2003) 45–61.
[46] M. Tang, H. Liao, From conventional group decision making to large-scale group decision making: What are the challenges and how to meet them in big data era? A state-of-the-art survey, Omega 102141 (2019).
[47] D. Whitley, V.S. Gordon, K. Mathias, Lamarckian evolution, the Baldwin effect and function optimization, in: Y. Davidor, H.-P. Schwefel, R. Männer (Eds.), Parallel Problem Solving from Nature – PPSN III, Springer Berlin Heidelberg, Berlin, Heidelberg, 1994, pp. 5–15.
[48] Z. Xue, Z. Lin, H. Wang, S. McClean, Quantifying consensus of rankings based on q-support patterns, Information Sciences 518 (2020) 396–412.
[49] J. Zheng, J. Lienert, Stakeholder interviews with two MAVT preference elicitation philosophies in a S]wiss water infrastructure decision: Aggregation using SWING-weighting and disaggregation using UTA-GMS, European Journal of Operational Research 267 (1) (2018) 273–287.
[50] V.C. Ziotti, A.B. Leoneti, Improving commitment to agreements: The role of group decision-making methods in the perception of sense of justice and satisfaction as commitment predictors, Pesquisa Operacional 40 (2020) e230300.

Publication [P7]

Cinelli, M., Kadziński, M., Miebs, G., Gonzalez, M., and Słowiński, R. (2022). Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system. In European Journal of Operational Research (Vol. 302, Issue 2, pp. 633–651). Elsevier BV., DOI: 10.1016/j.ejor.2022.01.011.

Number of citations[7]:

- according to Web of Science: 31
- according to Google Scholar: 62

Contribution of co-authors (excluding the author and the supervisor of this dissertation):

- Marco Cinelli:
  - Co-authorship of the idea underlying the paper
  - Co-authorship of the database with the MCDA methods' descriptions
  - Co-authorship of the text of the publication
- Michael Gonzalez:
  - Consultation of the system and methods' descriptions
  - Editing of the manuscript
- Roman Słowiński:
  - Co-authorship of the idea underlying the paper
  - Consultation of the system's and methods' descriptions, including the characteristics and questions
  - Editing of the manuscript

[7] as on September 25, 2023

European Journal of Operational Research 302 (2022) 633–651

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Decision Support

Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system

Marco Cinelli [a,*], Miłosz Kadziński [a], Grzegorz Miebs [a], Michael Gonzalez [b], Roman Słowiński [a,c]

[a] Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznań, Poland
[b] Environmental Decision Analytics Branch, Land Remediation and Technology Division, Center for Environmental Solutions and Emergency Response, Office of Research and Development, U.S. Environmental Protection Agency, 26 West Martin Luther King Dr., Cincinnati, 45268, OH, USA
[c] Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland

ARTICLE INFO

Article history:
Received 9 July 2021
Accepted 5 January 2022
Available online 21 January 2022

Keywords:
Decision analysis
Multiple criteria
Taxonomy
Decision support system
Method recommendation

ABSTRACT

We present a new methodology to lead the selection of Multiple Criteria Decision Analysis (MCDA) methods. It is implemented in the Multiple Criteria Decision Analysis Methods Selection Software (MCDA-MSS), a decision support system that helps analysts answer a recurring question in decision science: "*Which is the most suitable Multiple Criteria Decision Analysis method (or a subset of MCDA methods) that should be used for a given Decision-Making Problem (DMP)?*". The MCDA-MSS provides guidance to lead decision-making processes and choose among an extensive collection (>200) of MCDA methods. These are assessed according to an original comprehensive set of problem characteristics. The accounted features concern problem formulation, preference elicitation and types of preference information, desired features of a preference model, and construction of the decision recommendation. The applicability of the MCDA-MSS has been tested on several case studies. The MCDA-MSS includes the capabilities of (i) covering from very simple to very complex DMPs, (ii) offering recommendations for DMPs that do not match any method from the collection, (iii) helping analysts prioritize efforts for reducing gaps in the description of the DMPs, and (iv) unveiling methodological mistakes that occur in the selection of the methods. A community-wide initiative involving experts in MCDA methodology, analysts using these methods, and decision-makers receiving decision recommendations will contribute to the expansion of the MCDA-MSS.

1. Introduction

Multiple Criteria Decision Analysis (MCDA) is a scientific process aiming to frame Decision-Making Problems (DMPs) and develop a comprehensive assessment of alternatives (Cinelli, 2017; Roy, 1990; Tsoukiàs, 2007). It has two key potentials. First, it allows identifying, with structured and traceable protocols, the alternatives to be considered and the criteria to evaluate them. Second, it enables conveying a wealth of information that describes each alternative in a synthetic fashion, like a ranking from the best to the worst, a sorting of good, medium, and bad classes, or the choice of a subset of the most preferred alternatives.

There are several complexities involved in conducting the MCDA process. These include (i) the framing of the decision situation to be studied (Ley-Borrás, 2015), (ii) generation and characterization of alternatives to be considered (Keeney, 1996), (iii) development and identification of evaluation criteria (Keeney & Gregory, 2005), and (iv) selection of MCDA methods for each case study (Wątróbski, Jankowski, Ziemba, Karczmarczyk & Zioło, 2019). Due to these complexities, tools are needed to aid the MCDA-based research. This paper proposes a novel methodology to select an MCDA method, or a subset of these methods, relevant for a particular DMP. In order to enable its use by decision analysts, it is implemented in a Decision Support System (DSS) called the MCDA Methods Selection Software (MCDA-MSS).

Over the last few decades, the number of MCDA methods has grown steadily (Alinezhad & Khalili, 2019; Greco, Ehrgott & Figueira, 2016a; Hwang & Yoon, 1981), and an analyst can find it challenging to select the most suitable method. The main issue that a decision analyst is faced with is summarized by this question:

* Corresponding author.
E-mail addresses: m.cinelli@luc.leidenuniv.nl (M. Cinelli), milosz.kadzinski@cs.put.poznan.pl (M. Kadziński), gregorz.miebs@cs.put.poznan.pl (G. Miebs), gonzalez.michael@epa.gov (M. Gonzalez), roman.slowinski@cs.put.poznan.pl (R. Słowiński).

https://doi.org/10.1016/j.ejor.2022.01.011

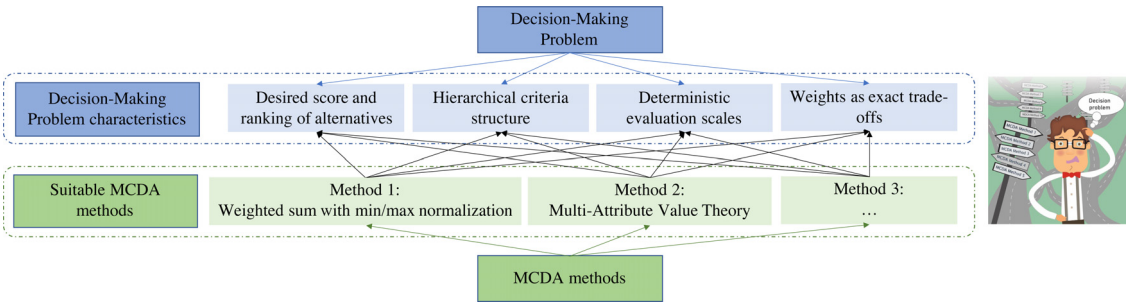


Fig. 1. The challenge faced by decision analysts when the DMP is structured. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

"Which is the most appropriate MCDA method (or a subset of methods) that should be used for a given DMP?"

This dilemma and one possible solution are presented in Fig. 1. The analyst is usually in a situation where the characteristics of the DMP must be accounted for in order to lead to selecting the MCDA method(s) that fit best with the DMP.

These characteristics are the features or elements that define the MCDA process. As seen in Fig. 1, the elements (light blue boxes) are: (i) desired recommendation is a ranking of the alternatives, (ii) criteria are structured hierarchically, (iii) the evaluation scales of the criteria are deterministic, and (iv) weights of the criteria are exact trade-offs. Thus, the challenge consists in finding the MCDA method(s) that can support these characteristics, which with this simple example could be a weighted sum with normalization (Itsubo, 2015) or MAVT (von Winterfeldt & Edwards, 1986). In other words, the challenge is to obtain a match between the DMP characteristics and the MCDA method(s) that can address such characteristics or at least satisfy as many of them as possible.

To avoid overwhelming the analysts and Decision Makers (DMs) with a wide pool of methods and/or prevent wrongly choosing the method(s) (or at least not selecting the most suitable one), tools are needed to assist analysts in choosing an MCDA method(s) for each specific application. It is essential to also add that the consequences of choosing an inadequate method among the plethora of those available are substantial. This can lead to neglecting some critical aspects of the problem, undesired compromises, and ultimately lead to a recommendation not aligned with the actual problem's characteristics and preferences of the involved stakeholders (Ebert & Welsch, 2004). An additional warning sign is the fact that popularity, simplicity, and intuitiveness are among the main reasons for selecting MCDA methods (Cajot, Mirakyan, Koch & Maréchal, 2017). These are some key reasons that drove the emergence of DSSs to aid in selecting MCDA method(s), which are briefly reviewed in the next section.

1.1. Review of DSSs for MCDA method(s) recommendation

MCDA methods selection is an MCDA problem in its own right (Gershon, 1981; Guarini, Battisti & Chiovitti, 2018). It involves the MCDA methods as the alternatives and their decision support capabilities as the evaluation criteria. The most up-to-date comprehensive review of the available DSSs for MCDA method(s) recommendation has been recently presented by Cinelli, Kadziński, Gonzalez and Słowiński (2020). They analyzed 23 of these DSSs and clustered them into three groups: (i) rule-based, (ii) algorithm-based, and (iii) artificial neural network-based. It was found that each DSS is structured on a formal representation of the methods, which can be called a taxonomy. In these DSSs, the taxonomy contains a set of characteristics (i.e., features) that describe the MCDA methods and the types of decision-making challenges they can support solving. For example, Celik and Topcu (2009) investigated the type of supported problem statement, measurement scale, weights, and thresholds. Benoit and Rousseaux (2003) emphasized the effect of the level of compensation in each method and the sensitivity to thresholds. Gershon and Duckstein (1983) highlighted the qualitative features, like the easiness of use, processing time needed to compile the data required for the method, and alternatives and/or criteria the MCDA method can work with. Salinesi and Kornyshova (2006) focused on the possible dynamic character of the DMP, the structure of the family of criteria (i.e., flat or hierarchical), and the type of preference models.

1.2. Motivation: What's missing in the existing DSSs for MCDA method(s) recommendation?

DSSs for MCDA method(s) recommendation require their own criteria to lead the selection process. As indicated by Cajot et al. (2017), a system capable of describing the MCDA methods and the practical implications of using one method rather than another is pivotal for these DSSs. This requires the systematic axiomatization of MCDA methods, which was an issue that already emerged in the 1990s' (French, 1993), but has only been addressed partially thus far. An initial contribution to achieving this ambitious target was presented in the recent review by Cinelli et al. (2020), which analyzed 56 peer-reviewed publications that consider the features (called characteristics) that should be accounted for when conducting an MCDA process and leading to the selection of an MCDA method or a subset of these methods. This work resulted in a comprehensive taxonomy to describe the MCDA process as well as its methods. The taxonomy is composed of 10 main characteristics (and sub-characteristics) clustered into three phases. The first is the problem formulation phase, which examines problem typology, the structure of the criteria, and evaluating the performance of alternatives. The second phase is focused on how the decision recommendation is developed with different strategies to elicit preferences of the DMs, among which the type of weights, thresholds, aggregation functions, and indirect elicitation approaches. The third phase is focused on the technical support to handle the problem with an MCDA software and the qualitative features of the DMP, describing how complex and flexible the method is, its data preparation requirements, and its reported use in the relevant literature.

The available DSSs include a limited set of taxonomy characteristics to describe the MCDA methods and their decision support capabilities (Cinelli et al., 2020). This means that they do not account for several key problem characteristics that analysts regularly deal with, making it difficult to identify the relevant MCDA method(s) in real case studies. These include, among others, the presence of hierarchies in the set of criteria, the type of preferences provided by the decision makers (i.e., direct or indirect), a

wide array of uncertainty analyses, the mathematical foundations of the underlying algorithms, and the capacity to handle inconsistent and/or dynamic datasets and preferences. Consequently, the type of MCDA problems that can be tackled and the way MCDA methods can be described with the available DSSs are limited to a subset of the proposed features in Cinelli et al. (2020). In addition, the available DSSs consider a rather limited group of methods, ranging from five (Li, Weston & Mavris, 2008) to 56 (Wątróbski et al., 2019). This means a large share of the most recent and advanced methods is not included in the available DSSs. What is more, the available DSSs do not face critical issues that can emerge when searching for an MCDA method for a particular case study. These include the cases when (i) no method is recommended or (ii) many methods are recommended. In the first case, the analyst finds himself in a situation where no decision support is provided at all, while in the second one, the too general advice is given.

1.3. Main contributions of the paper

This paper proposes the MCDA-MSS, a DSS that tackles the research gaps listed above by integrating a novel methodology developed to lead a comprehensive, dynamic and transparent selection process of the most suitable MCDA method(s) for every DMP. This DSS has two aims: (i) to allow describing complex DMPs and distinguish many methods proposed for MCDA by means of a set of relevant features (characteristics), and (ii) to guide an analyst assisting a DM in choosing the most appropriate method(s) for a given MCDA problem. The MCDA-MSS is now available for free at the following link http://mcdamss.com. The type of DMP we consider involves a one-stage decision by a single decision maker, concerning a finite set of alternatives with known consequences, evaluated by a finite set of conflicting criteria.

The users of the MCDA-MSS are envisioned to be of three main types:

- Type 1: Experienced MCDA users (including teachers) who want to use the tool for educational purposes. These educators can use the software to train future decision analysts (including students) by catering a comprehensive overview of the features that characterize MCDA processes and the methods that can be used to provide decision recommendations to the DMs;
- Type 2: MCDA methods developers and researchers who want to test, compare, and comprehend the characteristics of the existing MCDA methods and/or to develop new and more advanced ones;
- Type 3: MCDA methods developers and practitioners (including consultants and analysts) who want to support actual DMs in applying these methods in real-life DMPs.

Compared to the existing DSSs, the MCDA-MSS provides several unique contributions. First, it uses a comprehensive taxonomy of 156 characteristics to describe MCDA methods within its library, updating the taxonomy presented in Cinelli et al. (2020), which included “only” 66. The taxonomy in the MCDA-MSS thus represents the most detailed vocabulary that can be used to describe and develop MCDA methods and guide complex DMPs. Second, it includes a large set of MCDA methods, more than 200, which represent different approaches, schools, tendencies, and methodological streams that have evolved in the field of MCDA over the past sixty years. As a result, we offer the most comprehensive database of MCDA methods. Their number is almost four times greater than in the state-of-the-art decision support system elaborated by Wątróbski et al. in 2019. Third, it provides solutions to cases where no method matches all the requirements set by the analysts. These solutions involve a dialog with the DM concerning demands that must be fulfilled by the recommended methods and aiming at maximizing the share of requirements that can be satisfied. Fourth, it offers a strategy to reduce a large set of relevant methods when there is much uncertainty in the description of the DMP. It suggests the questions that maximize the potential information gain from the users’ answers by minimizing the number of recommended methods once a desired feature is specified. Fifth, it can be used as an identifier of errors in MCDA method selection. This capacity is illustrated by studying ten case studies reported in the literature and discussing the reasons for mismatches of MCDA methods.

The paper is structured as follows. Section 2 presents the methodology that was used to develop MCDA-MSS and test it with a set of case studies. Section 3 describes the application of MCDA-MSS to the case studies and the contributions that it provides. Section 4 concludes the paper and provides some avenues for future research.

2. MCDA-MSS development

The methodology formulated to develop and test the MCDA-MSS included three stages (see Fig. 2). The first stage concentrates on shaping the taxonomy used within the MCDA-MSS to describe each MCDA method in its library. The second stage is tailored to the development of the MCDA-MSS in a web-software form, while the last stage focuses on testing the MCDA-MSS on a series of case studies to assess its usability and performance. Each stage is described in detail in the following sections.

2.1. Stage 1: Develop the taxonomy and the database of MCDA methods used in the MCDA-MSS

Stage 1 was focused on (i) the development of the methodological backbone of the MCDA-MSS, i.e., the taxonomy of characteristics (i.e., features) used to describe the MCDA methods, and (ii) the database of MCDA methods themselves. This was achieved in two steps. The first one was the study of the taxonomy introduced in Cinelli et al. (2020), called here *taxonomy v.1*. The second one consisted in its application to a large set of MCDA methods. The rationale for the selected methods, together with their brief description, is given in Section 2.1.2 and Appendix A in the Electronic Supplementary Information (ESI). An important remark on the database of methods in the MCDA-MSS is that the objective was not to include all the available MCDA methods but rather propose an initially wide list, which can be complemented in the future. Also, the current focus has been on methods for a single DM, while those for group DMs can be a topic of further inclusion (see Section 4.2 for future research avenues).

2.1.1. The taxonomy used in the MCDA-MSS

The application of taxonomy v.1 from Cinelli et al. (2020) to the MCDA methods led to its refinement and revision, resulting in an elaborated form, called *taxonomy v.2*, that has been implemented in the MCDA-MSS. Its structure is shown in Table S1 in Appendix B of the ESI. This one expanded and also restructured the first two phases, while the third one was not included. The reason for such exclusion is that the MCDA-MSS aims at describing the MCDA methods according to the features that can be evaluated as objectively as possible. Being that the third phase of the taxonomy in Cinelli et al. (2020) focused on qualitative features, it did not fit with the purpose of the MCDA-MSS. Two examples of these qualitative features are easiness of use of the method and the time required to compile the needed data to use the method. Their assessment depends on the knowledge and expertise of the analysts who lead the MCDA process, and so they cannot be characterized objectively.

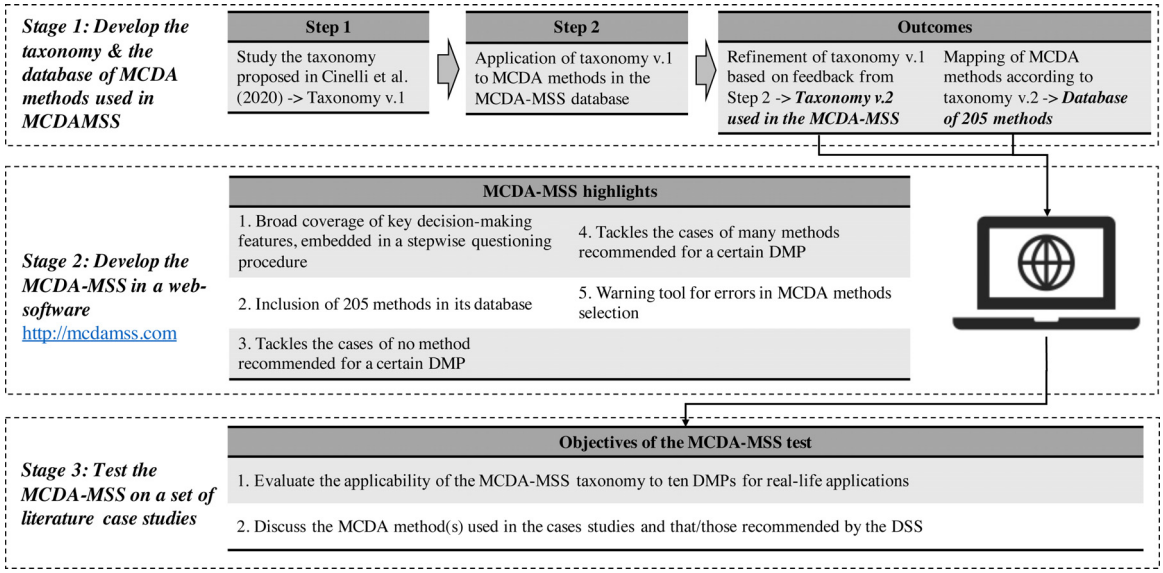


Fig. 2. The methodology used to develop and test the MCDA-MSS.

The differences between taxonomy v.1 and v.2 are substantial, starting from the number of considered objective features. Those in taxonomy v.1 of (Cinelli et al., 2020) are 66, while those in taxonomy v.2 of the MCDA-MSS are 156. As a matter of context regarding the most recent software to recommend MCDA methods, the number of features they account for is 9 in Wątróbski et al. (2019) and 39 in Guarini et al. (2018).

The taxonomy in the MCDA-MSS is structured in four main sections:

1. Problem typology: Defines the type and structure of the DMP;
2. Preference model: Defines what type of model the analyst would like to apply;
3. Elicitation of preferences: Defines the type, modality, and frequency of elicited preferences;
4. Exploitation of the preference relation induced by the preference model: Defines the strategy used to derive and enrich the decision recommendation.

2.1.1.1. MCDA-MSS section 1: Problem typology (c.1). In the problem typology section, the analyst can define how the problem is framed by (i) choosing the type of DMP under consideration and (ii) describing the criteria used to assess the alternatives.

As far as the type of decision-making challenge is concerned, the problem statement (c.1.1), in other words, the kind of desired decision recommendation, can be of four types (Belton & Stewart, 2002; Cailloux, Lamboray & Nemery, 2007). These include ranking (i.e., order the alternatives from the most to the least preferred), sorting/ordinal classification (i.e., assign the alternatives to predefined preference-ordered decision classes), clustering (i.e., divide alternatives into groups according to some similarity measure or preference relation), and choice (i.e., select the most preferred subset of alternatives). Except for choice problems, it is also possible to distinguish the type of order of the alternatives/classes/clusters, as either partial or complete (Roy, 2016b). Partial ordering admits incomparability, and it does not necessarily lead to a univocal ordering. This does not hold for the complete order, where all the alternatives/classes/clusters are ordered from the most to the least preferred. In addition, the scale leading the recommendation for ranking and sorting problems can also be chosen between ordinal and cardinal (Roy, 2016b). Ordinal recommendations are based on binary relations, where only the position of the alternatives is meaningful. In contrast, cardinal recommendations are driven by a score, where the distance between alternatives is meaningful in quantitative terms. As far as sorting and choice are concerned, it is possible to set the cardinality of the DMP, either with or without constraints (Kadziński & Słowiński, 2013). Cardinality with constraints consists in the DMP where a pre-defined number of alternatives is either chosen or assigned to each class. On the contrary, cardinality without constraints does not restrict the number of chosen alternatives or assignments to each decision class. The type of set of alternatives (c.1.2) can then also be chosen, being either stable (i.e., no new alternatives are foreseen and added to the set) or incremental (i.e., new alternatives keep arriving as the decision context evolves) (Siebert & Keeney, 2015). The assumption that we applied in the development of the MCDA-MSS database is that methods that handle an incremental set of alternatives can also handle one with a stable set. In addition, in case the user chooses an incremental set of alternatives and the problem statements of ranking, choice, sorting with cardinality constraints, and clustering, the MCDA-MSS notes that the recommendation provided by the suitable methods might change when new alternatives are added.

Four features characterize the description of the criteria used to assess the alternatives. The first one is the structure of the set of criteria (c.1.3), being either flat (i.e., the criteria are all at the same level) or hierarchical (i.e., the criteria are organized in levels, hierarchically, from general to detailed ones) (Marttunen, Belton & Lienert, 2018). The second is the type of performance of the criteria (c.1.4), which can be either deterministic (i.e., exact input) or uncertain Keeney and Gregory (2005). A further differentiation here is on whether the performance of an alternative is provided on a criterion per se, or whether it refers to the performance of an alternative on a criterion with respect to the performance of another alternative on the same criterion, like in the Analytical Hierarchy Process (AHP) (Saaty, 1980), for example. Uncertain performance includes multiple options, among which the lack of information, intervals, probability distributions, and fuzzy (Dias, Antunes & Insua, 2012). The third feature describing the criteria is the preference direction of the criteria (c.1.5), which defines whether the order of preference for the values of the criteria is known or has to be discovered (Błaszczyński, Greco & Słowiński, 2012). In case it is known, two options are available. The first one is for monotonic criteria (Nardo et al., 2008), whose order of preference is de-

fined as constantly non-decreasing or non-increasing with respect to the performances on the criterion. The second one is for non-monotonic criteria (Pap, Marichal, Grabisch & Mesiar, 2009), whose order of preference is defined as non-decreasing or non-increasing in different regions of the evaluation scale with respect to the performances on the criterion. The taxonomy does not include attributes without preference-ordered scales, as the objective of the MCDA-MSS is to focus on problems that are characterized in classical MCDA terms, where for each criterion, a preference-order can be assumed, even if the positive or negative monotonicity relationships do not hold in the whole evaluation space. The last feature in this section looks at whether the criteria set is complete or not (c.1.6) (Roy, 2016b). The set is complete if all criteria relevant for the problem and having an impact on the recommendation were identified at the problem modeling phase and included in the set. In contrast, it is incomplete when at least one criterion relevant for the problem and having a potential impact on the recommendation is not included in the set. As far as modeling assumptions used in the MCDA-MSS database are concerned, all methods that deal with a hierarchical structure of the set of criteria, accept uncertain criteria performances and/or incomplete criteria sets, are also assumed to handle a flat structure of the set of criteria, work with deterministic and/or complete criteria sets, respectively. Except for rule-based methods, when the user chooses an incomplete set of criteria, the MCDA-MSS points out that for the recommended methods, due to the aggregation of the performances on all the criteria in a single measure of quality, they should only be used when all relevant criteria are accounted for.

2.1.1.2. MCDA-MSS section 2: Preference model (c.2). The definition of the preference model is the second section of the taxonomy in the MCDA-MSS. It consists of nine main features, starting with looking at the measurement scale used by the method (c.2.1). This can be of three types, ordinal (qualitative scale carrying order information), cardinal (quantitative scale carrying order information), and relative (Pap et al., 2009). The first one only considers the order of performances, the second accounts for the differences between performances, and the third is grounded on comparisons between alternatives to express preference intensity. The methods that use scales in a quantitative matter are further divided according to how they assess the performance of the alternatives, into performance-based and pairwise comparison-based. In the former case, the methods evaluate the desirability of each alternative's performance individually (e.g., in MAVT (Keeney & Raiffa, 1976)), while in the latter, the methods compare the alternatives pairwise to define if one performs at least as well as (or better than) another one (e.g., in ELECTRE methods (Figueira, Mousseau & Roy, 2016)). Performance-based methods that employ quantitative scales are moreover split according to how they treat the raw information before the aggregation step, distinguishing between those applying linear, piecewise linear, and non-linear transformation strategies (Cinelli, Spada, Kim, Zhang & Burgherr, 2021; Nardo et al., 2008). A subdivision is also applied to the pairwise comparison-based methods, discerning those that use comparisons of performance differences with (e.g., PROMETHEE methods (Brans & De Smet, 2016)) and without thresholds (e.g., EVAMIX (Voogd, 1982)). Lastly, the taxonomy distinguishes between the methods that use relative comparisons of performances to express the intensity of preference either in ordinal (e.g., MACBETH (Bana E Costa & Vansnick, 1999)) or in ratio terms. The latter group further differentiates the methods according to whether they are performance-based (e.g., weighted means without transformation of criteria performances (Itsubo, 2015)) or pairwise comparison-based (e.g., AHP (Saaty, 1980)).

The second feature of the preference model is the one that considers how the comparison of the performances on the criteria is performed by the method. (c.2.2). Four options are provided with this feature:

1. Performances are transformed with a data-driven normalization approach and then compared (e.g., Cinelli et al. (2021); Nardo et al. (2008));
2. Performances are compared by the DM with respect to the graded intensity of preference: The comparisons are performed by the DM who has to choose one value from a scale that is pre-defined (e.g., Bana E Costa and Vansnick (1999); Saaty (1980));
3. Performances are compared by the DM with respect to the non-graded intensity of preference: The comparisons are performed by the DM whose intensity of preference is not pre-defined in a set of values (e.g., Morton (2018), , Siskos, Grigoroudis and Matsatsinis (2016));
4. Raw performances are compared directly (e.g., Figueira et al. (2016), Greco, Matarazzo and Słowiński (2016b)).

The four features that follow define binarily whether standard components of MCDA methods are part of or not part of the DMP, being (i) the weights of the criteria (c.2.3), (ii) the per-criterion pairwise comparison thresholds (c.2.4), (iii) the interactions between criteria (c.2.5), and (iv) the multi-criteria profiles (c.2.6). Criteria weights are used to differentiate the role of criteria in the aggregation procedure (Greco, Ishizaka, Tasiou & Torrisi, 2019); per-criterion pairwise comparison thresholds characterize the preference sensitivity of the DM when comparing two alternatives on a single criterion (Granata, 2017), interactions denote interdependencies between the criteria (Grabisch & Labreuche, 2008), and multi-criteria profiles (Dias & Mousseau, 2018) - not corresponding to the considered alternatives - serve as the basis for deriving the decision recommendation by comparing the performances of the alternatives with them.

Compensation level between criteria (c.2.7) is the seventh characteristic of the preference model, looking at how much the good performance on a criterion can compensate for the poor performance on another criterion (Rowley, Peters, Lundie & Moore, 2012). Methods are assigned to one or more of the three available compensatory levels, which are null, partial, and full (Langhans, Reichert & Schuwirth, 2014).

The eighth feature is focused on determining whether and how the aggregation of the performances on multiple criteria should be performed by the method (c.2.8). In case of no aggregation, the respective method will develop the recommendation by considering the evaluation of the alternatives on a criterion-by-criterion basis (Bouyssou, Marchant, Pirlot, Tsoukiàs & Vincke, 2006a). In case the aggregation takes place, three options are available, being scoring functions, binary relations, and decision rules (Słowiński, Greco & Matarazzo, 2002). Scoring functions aggregate the individual criteria performances (usually normalized) to define the overall quality of each alternative (Nardo et al., 2008). Methods using binary relations employ pairwise comparisons of alternatives, which lead to a comprehensive assessment of each alternative (Roy, 2016b). Lastly, rule-based methods aggregate the performances on different criteria using information connectors in the form of "if ..., then ..." statements (Greco et al., 2016b). These aggregation options are not mutually exclusive, and the analyst can request a method that employs one or more of these aggregation modes.

The last feature for the definition of the preference model is the capacity to deal with inconsistent preference information (c.2.9). Three types of inconsistencies can be considered, including (i) only violation of dominance relation (e.g., DRSA-based method (Greco et al., 2016b)), (ii) only other types of inconsistency, not including dominance (e.g., Best Worst Method (Rezaei, 2015) and AHP (Saaty, 1980)), and (iii) violation of dominance and other types of inconsistency (e.g., NAROR-Choquet (Angilella, Greco & Matarazzo,

2010)). The case of violation of dominance only is further split between a strict and a relaxed treatment of the violation. The former captures the inconsistency of all objects (individual alternatives or pairs) violating the dominance principle (Greco, Matarazzo & Słowiński, 2001a). In contrast, the latter captures the inconsistency of all objects (individual alternatives or pairs) insufficiently consistent (defined by a user-specified threshold on the consistency measure) with respect to the dominance principle (Greco, Matarazzo, Słowiński & Stefanowski, 2001b). The category including only other types of inconsistency and not including dominance accepts two sub-categories. The first one consists in handling inconsistency with respect to relative comparisons expressed on a cardinal scale for different pairs of alternatives and/or criteria (i.e., cardinal inconsistency) (Saaty, 1980). The second refers to the capacity of handling inconsistency with respect to relative comparisons expressed on a cardinal scale for the same pair of criteria (i.e., inconsistency concerning preference comparison of criteria) (Rezaei, 2015). The most comprehensive set of inconsistencies is the third one, as it captures the inconsistency of all pieces of preference information that violate the dominance principle and/or cannot be reproduced with an assumed (score- or relation-based) preference model.

2.1.1.3. MCDA-MSS section 3: Elicitation of preferences (c.3). The type (c.3.1), frequency (c.3.2), and confidence (c.3.3) of preferences provision constitute the third section of the taxonomy. The first main distinction is between the type of preferences, which can either be direct or indirect (Dias, Morton & Quigley, 2018). In case they are direct (c.3.1.1), the model parameters are defined directly, while in the indirect mode (c.3.1.2), local or holistic judgments of the experts/DMs on some reference alternatives are used to elicit them. With respect to direct preferences, the analyst can define whether the features are either specified directly by the decision maker or not. The latter includes DMPs where the model parameters are missing, they cannot be specified, or the decision maker requires a method that does not use them.

Direct preferences (c.3.1.1) include four characteristics, starting from the weights of the criteria (c.3.1.1.1), which can be used to set the difference of importance between them (Greco et al., 2019). Two distinctive overarching categories can be defined for the type of criteria weights, namely precise and imprecise. As part of the precise ones, trade-offs and importance coefficients are distinguished (Munda & Nardo, 2005). Trade-offs indicate the exchange rate accepted between the criteria to compensate each other performance (Munda, 2008). Relative importance coefficients specify the strength of one criterion in comparison with others in a voting-line procedure (Munda & Nardo, 2005). Weights as precise relative importance coefficients are further split in those defined per-criterion (Riabacke, Danielson & Ekenberg, 2012) and with pairwise comparison ratios (e.g., the AHP method (Saaty, 1980)).

In the case of imprecise weights, where no exact values are defined and they are driven by constraints of a different type, multiple options can be chosen, including missing input (Lahdelma, Hokkanen & Salminen, 1998), fuzzy numbers (Nădăban, Dzitac & Dzitac, 2016), ordering of some or all criteria with or without intensity of preferences (Dias & Climaco, 2000; Punkka & Salo, 2013), pairwise comparison-based difference of importance (Angilella et al., 2010), ratios (Salo & Hämäläinen, 1992), distributions (Pelissari, Oliveira, Amor, Kandakoglu & Helleno, 2020), intervals (Ahn, 2017), and rank requirements (Salo & Punkka, 2005). The distinction between absolute comparisons (i.e., per-criterion) and pairwise comparison ratios is also applied. With respect to the assumptions made in the development of the MCDA-MSS database, the methods that accept imprecise weights are also assumed to work with the weights in their precise form (e.g., SMAA-PROMETHEE II (Corrente, Figueira & Greco, 2014) can work with imprecise weights defined with an ordering of some of the criteria, but it can also accept precise relative importance coefficients).

Pairwise comparison thresholds (c.3.1.1.2) are the second characteristic of direct preferences. They can be used to characterize the preference sensitivity of the DM when comparing two alternatives. The three most common types are included, being indifference, preference, and veto thresholds (Figueira et al., 2016), each distinguished according to whether they are precise or imprecise. Indifference threshold is the maximum difference between performances of two alternatives under which they are considered indifferent (e.g., if the prices of two cars differ by not more than $100, then they are indifferent on the price) (Dias & Mousseau, 2018). Preference threshold sets the minimum difference between performances of two alternatives above which a strict preference can be defined (e.g., if the price of two cars differs by at least $2000, then one strictly prefers the cheaper option) (Dias & Mousseau, 2018). Veto threshold determines the minimum difference in performances of two alternatives which, when exceeded, invalids the preference of the worse alternative over the better one, irrespective of their performances on the remaining criteria (e.g., if one car costs more than $10,000 in comparison to another car, then the former cannot be preferred to it at the comprehensive level despite its advantages on the remaining criteria) (Dias & Mousseau, 2006). It was assumed that methods using these imprecise pairwise comparison thresholds could also accept precise ones. Even though we explicitly consider only three types of thresholds, more do exist (e.g., discordance, reinforced preference, or counter-veto) (Roy & Słowiński, 2008). When the use of these others in MCDA becomes more common, the taxonomy can be extended to account for each of them explicitly.

The third feature of direct preferences is the type of interactions between the criteria (c.3.1.1.3), which are divided into two groups, (i) positive and negative and (ii) antagonistic. Positive interaction means that the comprehensive impact of the criteria on the quality of an alternative is greater than the impact of these criteria taken separately (Grabisch & Labreuche, 2008). Negative interaction implies that the comprehensive impact of the criteria on the quality of an alternative is smaller than the impact of these criteria taken separately (Grabisch & Labreuche, 2008). The antagonistic effect operates by lowering the influence of other criteria in the case the performance on one criterion is very low (Figueira, Greco & Roy, 2009a). All these interactions can be provided as precise and imprecise, according to how much knowledge is available on the DMP.

The last feature of the direct preferences is the definition of the discriminatory profiles (c.3.1.1.4) when sorting, choice, and ranking problems are chosen. This can be attained either by using characteristic profiles (i.e., representative/most typical for each class) (Almeida-Dias, Figueira & Roy, 2010) or boundary profiles (i.e., DM-specified frontiers/boundaries which steer their assessment) (Figueira et al., 2016). The methods may incorporate single or multiple profiles to characterize each class or form the basis for comparing the existing alternatives (Fernández, Figueira, Navarro & Roy, 2017). What is more, profiles for sorting are differentiated between precise and imprecise (Fernández, Figueira & Navarro, 2020), using the modeling assumption that methods accepting the imprecise type also accept the precise one.

There exist other types of parameters whose values need to be specified to run a specific method and derive a recommendation. Such example parameters include a number of characteristic points of marginal value functions deciding upon their flexibility and ability for representing various decision-making policies (Jacquet-Lagreze & Siskos, 1982), a credibility threshold indicating the minimal value of a valued outranking relation justifying the truth of a crisp relation (Figueira et al., 2016), or a compensation

parameter between different types of aggregation functions that are jointly used to evaluate decision alternatives (Brauers & Zavadskas, 2010). However, they cannot be categorized under any more general type of characteristic.

Direct specification of model parameters is sometimes a challenging task for the DM, who might have difficulties in understanding their meaning and/or not have the time to devote to providing input on each of them. For these reasons, indirect elicitation techniques (c.3.1.2) that use local or holistic judgments of the experts/DMs on some reference alternatives have become popular in the last decades (Doumpos & Zopounidis, 2011; Jacquet-Lagrèze & Siskos, 2001). The reasons are that experts/DMs find themselves more comfortable in exerting choices rather than explaining them (Buchanan & O'Connell, 2006).

As far as sorting problems are concerned, six types of indirect preferences are offered. The first is the assignment of reference alternatives to decision classes (Özpeynirci, Özpeynirci & Mousseau, 2018), while the second one accounts for such assignment but with a variable level of certainty (Liu, Kadziński, Liao, Mao & Wang, 2020). The third type is the assignment-based pairwise comparisons of alternatives (e.g., alternative A should be assigned to a decision class at least as good as alternative B) (Özpeynirci et al., 2018). And the fourth is the desired comprehensive value of alternatives assigned to a given class or class range (e.g., alternatives assigned to a class at most medium should have a value not greater than 0.4) (Kadziński, Ciomek & Słowiński, 2015). The last two options include the comparisons of reference alternatives with respect to intensity of preference expressed on an ordinal (Ishizaka & Gordon, 2017) or ratio scale (Ishizaka, Pearman & Nemery, 2012), respectively.

Regarding ranking and choice problems, five types of indirect preferences are considered. The pairwise comparisons of the alternatives (e.g., alternative A is preferred to alternative B) represents the first type. The second consists in the comparisons of reference alternatives with respect to the intensity of preference expressed on an ordinal scale (e.g., alternative A is very strongly preferred to alternative B) (Argyris, Morton & Figueira, 2014), followed by the comparisons of reference alternatives with respect to the intensity of preference expressed on a ratio scale (e.g., alternative A is preferred to alternative B by a factor of 2 on criterion 1) (Saaty, 1980). The set is completed by the rank-related requirement (e.g., alternative A is in the top 3) (Kadziński, Greco & Słowiński, 2013) and by the desired comprehensive value of alternatives (e.g., the value of the alternative with at least the fifth rank is at least 0.6) (Salo & Hamalainen, 2001).

The second to last feature of the preference model is the frequency of preference input (c.3.2). It distinguishes whether the preferences are provided at the start of the elicitation process or whether they are given successively in different iterations (Tsoukiàs, 2007), thus encouraging the user to observe the evolution or convergence of the recommendation.

The elicitation of preferences is completed by accounting for the level of confidence in the provided preferences (c.3.3), offering the option of distinguishing between preferences with or without a confidence level, with such levels that can be, e.g., characterized by the statements "absolutely sure", "sure", or "mildly sure" (Greco, Mousseau & Słowiński, 2010). The modeling assumption used here for the MCDA-MSS database development is that methods working with a level of confidence can also work without it.

2.1.1.4. MCDA-MSS section 4: Exploitation of the preference relation induced by the preference model (c.4). The last section of the taxonomy aims to delineate how the preference relation induced by the preference model should be exploited to derive or enhance the decision recommendation (Perny & Roubens, 1998; Pirlot, 1995). Two main options are available, with a univocal recommendation (c.4.1) on one side, and the output variability analysis (c.4.2) on the other side. The first consists in offering a univocal outcome for the chosen problem statement. The second focuses on showing how variable the recommendation can be when there is uncertainty with respect to the performances of alternatives and/or representation of the DM's preferences by an assumed preference model (Dias, 2007).

Univocal recommendations can be derived with or without output variability analysis. When no output variability analysis is used (c.4.1.1), the recommendation can be provided either by a single or multiple contingent models (Kadziński, Ghaderi & Dąbrowski, 2020). The option of a single model is further differentiated in deterministic and representative. The deterministic model is a single preference model instance with precise parameter values directly specified by the DM (Nardo et al., 2008). The representative model is a single preference model instance with precise parameter values that are either selected by the method (i.e., algorithmic) (Jacquet-Lagreze & Siskos, 1982) or directly chosen by the DM (i.e., direct involvement) (in both cases, there exist other possible preference model instances) (Kostkowski & Słowiński, 1996).

When output variability analysis is embedded in the development of the univocal recommendation (c.4.1.2), the taxonomy (and the MCDA-MSS) offers two options. The first is compromise exploitation, where the univocal recommendation is constructed by aggregating or building on the outcomes of output variability analysis without selecting a representative preference model instance (Vetschera, 2017). The second is representative exploitation, which derives the univocal recommendation from an application of a representative preference model instance selected based on the outcomes of output variability analysis (Greco, Kadziński & Słowiński, 2011; Kadziński, Greco & Słowiński, 2012b).

The last part of the taxonomy is devoted to defining how to conduct the output variability analysis (c.4.2.1) and which results to focus on for the output variability analysis (c.4.2.2). Two types of analysis (c.4.2.1) can be performed, with the first providing the extreme results with all compatible models (Corrente, Greco, Kadziński & Słowiński, 2013) and the second supplying the distribution of results with a sample of compatible models (Lahdelma & Salminen, 2010). Regarding what results to focus on when performing the output variability analysis (c.4.2.2), a distinction is made between choice, ranking, and sorting problems. For choice and ranking problems, the exploitation of the preference relation induced by the preference model can be tailored to the selection of the alternatives (e.g., with robustness analysis on the kernel and/or preference relations (Govindan, Kadziński, Ehling & Miebs, 2019)), the ranking (e.g., with extreme ranking analysis (Kadziński, Greco & Słowiński, 2012a)), the score (e.g., with score variability (Dias & Climaco, 2000)), the preference relations and also preference intensities (Figueira, Greco & Słowiński, 2009b). For sorting problems, the exploitation of the preference relation induced by the preference model is available for class assignments (e.g., alternative A is assigned to class 2 with all the models (Greco et al., 2010)), assignment based pairwise relations (e.g., alternative A is assigned to a class at least as good as alternative B for all models) (Kadziński et al., 2015), and class cardinalities (e.g., at least ten and at most 15 alternatives can be assigned to class medium (Kadziński et al., 2015)).

2.1.2. The database of MCDA methods used in the MCDA-MSS

The database of 205 MCDA methods in the MCDA-MSS is, to the best of the authors' knowledge, the widest available of the DSSs that have been proposed to recommend such methods. It comprises representatives of the three main families of MCDA methods, namely approaches incorporating scoring functions, binary relations, and decision rules. As far as the methods employing scoring functions are concerned, there are widely

used representatives like AHP (Saaty, 1980), ANP (Saaty, 2016), MAVT (Keeney & Raiffa, 1976), TOPSIS (Hwang & Yoon, 1981), VIKOR (Opricovic & Tzeng, 2004), and additive weighted average (Razmak & Aouni, 2014). Less common methods like EVAMIX (Voogd, 1982) and REMBRANDT (Van den Honert & Lootsma, 2000) complement this set. A similar trend is visible for the binary relation methods, where several original options of the most used from this family were included, namely those from ELECTRE and PROMETHEE ones (Behzadian, Kazemzadeh, Albadvi & Aghdasi, 2010; Govindan & Jepsen, 2016). In this case, too, less frequently used methods are incorporated, such as ARGUS (De Keyser & Peeters, 1994), NAIADE (Munda, 1995), QUALIFLEX (Paelinck, 1976), and REGIME (Hinloopen, Nijkamp & Rietveld, 1983). Regarding methods based on decision rules, this led to the inclusion of Dominance-based Rough Set Approach (DRSA) methods (Greco et al., 2016b), which have never been accounted for in this type of DSSs.

A broad set of methods that include dated as well as recent developments in the MCDA area was then selected to complement the dataset. The rationale for their selection was as follows:

1. Expansion of ELECTRE and PROMETHEE methods to account for the advancement of these well-known methods to deal with more elaborate DMPs, like ELECTRE TRI-nB (Fernández et al., 2017), ELECTRE TRI-nC (Almeida-Dias, Figueira & Roy, 2012), ELECTRE SORT (Ishizaka & Nemery, 2014), PROMETHEE^{GKS} (Kadziński et al., 2012a), and PROMSORT (Araz & Ozkarahan, 2007);
2. Methods dealing with less common problem types including multiple criteria clustering (e.g., P2CLUST (Smet, 2013)), multiple criteria sorting with unknown decision classes (Rocha, Dias & Dimas, 2013)), sorting with partially ordered decision classes (e.g., ELECTRE-SORT (Ishizaka & Nemery, 2014)), and problems with decision classes size constraints (e.g., sorting with constrained decision classes (Özpeynirci et al., 2018), constrained choice (Podinovski, 2010);
3. Methods that deal with a hierarchical structure of a family of criteria such as ELECTRE III-H (Del Vasto-Terrientes, Valls, Slowinski & Zielniewicz, 2015) and MCHP-UTA (Corrente, Greco & Słowiński, 2012);
4. MCDA methods that work with imprecise and indirect preferences, in order to include a very relevant and increasing branch in the MCDA domain Corrente et al. (2013); Doumpos and Zopounidis (2011), which most DSSs have not accounted for, e.g., ACUTA (Bous, Fortemps, Glineur & Pirlot, 2010), Preference programming with incomplete ordinal information (Punkka & Salo, 2013), PRIME (Salo & Hamalainen, 2001), DIS-CARD (Kadziński & Słowiński, 2013), GRIP (Figueira et al., 2009b), UTA^{GMS} (Greco, Mousseau & Słowiński, 2008), UTA^{GMS} with imprecise evaluations (Corrente, Greco & Słowiński, 2017);
5. Single score methods tailored to problem formulations with, e.g., interactions (Grabisch & Labreuche, 2016), adjustable compensation levels (Cabello, Navarro-Jurado, Rodríguez, Thiel-Ellul & Ruiz, 2019), hierarchical criteria, uncertainty, and criteria interactions (Angilella et al., 2018);
6. Methods that account for variability in the evaluation of performances on the criteria and/or preferences of the stakeholders while accounting for a sample of the compatible preference models, e.g., SMAA-ELECTRE-I (Govindan et al., 2019), SMAA-AHP (Durbach, Lahdelma & Salminen, 2014), or while using the consequences of applying all compatible models with mathematical programming techniques, e.g., ARIADNE (Sage & White, 1984), CUT (Argyris et al., 2014), or multiple criteria majority-rule sorting (Meyer & Olteanu, 2019);
7. Less compensatory weighted sums that have been commonly used, especially in sustainability-related research (Langhans et al., 2014), i.e., geometric and harmonic.

Overall, 205 MCDA methods are included in the MCDA-MSS, mapped according to the taxonomy presented in Section 2.1.1. "Mapped" is interpreted in this research as the evaluation of which characteristics of the taxonomy are supported by the MCDA methods (see also Section 2.2 for more details). The complete database is available in Appendix C in the ESI.

2.2. Stage 2: Development of the MCDA-MSS in a web-software

The MCDA-MSS has been developed using rule-based modeling, which belongs to the first group of DSSs presented in Cinelli et al. (2020). It is the most used modeling approach to develop this type of DSSs. Its added advantages and justifications for its selection for this DSS, are the objectivity, traceability, and understandability of the reasoning system empowered by the decision rules (Słowiński, Greco & Matarazzo, 2009). This modeling uses rules in the form of *"if the conjunction of conditions on some features is true, then the decision is …"*. The database of MCDA methods already represents an information table shaped with this modeling approach, where the characteristics are the conditions to describe the methods and the methods themselves are the decisions. Table 1 shows a sample of MCDA methods in the MCDA-MSS database, mapped according to a subset of features of the taxonomy for each type of problem statement. Once applied to the taxonomy of the MCDA-MSS, the rules syntax reads as "If characteristics a, b, … are activated/met by the DMP, then MCDA method I, II, … fits with the problem and can be recommended". Four examples of decision rules from Table 1 are as follows, with the **1s** in the table showing those that are used to activate the rules:

1. "If the DMP is one requiring a partial and cardinal ranking, then GRIP method (Figueira et al., 2009b) is a suitable one";
2. "If the DMP is one requiring a complete and cardinal order of decisions classes, with constraints on the number of alternatives assigned to the decision classes, then ROR-UTADIS (Kadziński et al., 2015) is a suitable method";
3. "If the DMP is one requiring a partial clustering, then MCUC–CSA method (Rocha et al., 2013) is a suitable one";
4. "If the DMP is one requiring a choice without any constraints on the number of alternatives to be recommended, then ELECTRE I method (Figueira et al., 2016) is a suitable one".

The DSS presented in this paper can be defined as a supporting tool for choosing the most appropriate MCDA method(s) for a given multiple criteria problem. It can substantially enhance, while selecting the relevant MCDA method(s), the interaction between the analyst and the DM in several ways.

First, the DMP is partitioned into four manageable sections. In each of these sections, a wide amount of information on the MCDA methods is neatly structured according to the taxonomy of characteristics, using a sequenced stepwise questioning process (*DSS highlight 1 in* Fig. 2). Each answer in the software is coupled with an information box that appears when the user moves the cursor on such an answer. This box provides a succinct definition of each answer (i.e., characteristic) so that the user can clearly understand the meaning of each concept. These explanations provided for each question within the information boxes constitute knowledge transfer and a learning exercise for the DMs and inexperienced MCDA practitioners. These boxes are solutions to the black box effect that many MCDA methods have been criticized for. In fact, MCDA methods require several parameters to be defined to use them, and providing a brief description of each of them can strengthen the confidence of the DM in their answers. An example for the information

Table 1
Example of MCDA methods in the MCDA-MSS database, mapped according to a subset of features of the taxonomy for each type of problem statement. Bold "1" are those that are used to trigger the rules in the main text.

MCDA methods		Section 1 - Problem typology																		Section 2 - ...
		c.1.1 - Problem statement																	Other features	Other features
		Ranking				Sorting						Clustering			Choice				...	...
		Order		Scale leading the recommendation		Order of decision classes		Scale leading the recommendation		Cardinality		Non-relational	Relational: Partial tournament	Relational: Complete tournament	Order		Cardinality			
		Partial	Complete	Ordinal	Cardinal	Partial	Complete	Ordinal	Cardinal	With constraints	Without constraints				Partial	Complete	With constraints	Without constraints		
For ranking problems																				
1	AHP	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	...	...
2	GRIP	**1**	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	...	...
3	PROMETHEE II	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	...	...
For sorting problems																				
4	DRSA - sorting	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	...	...
5	ELECTRE SORT	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	...	...
6	ROR-UTADIS	0	0	0	0	0	**1**	0	**1**	**1**	1	0	0	0	0	0	0	0	...	...
For clustering problems																				
7	MCUC-CSA	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	...	...
8	P2CLUST	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	...	...
For choice problems																				
9	Constrained choice	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	...	...
10	ELECTRE I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	...	...
...	Method ...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

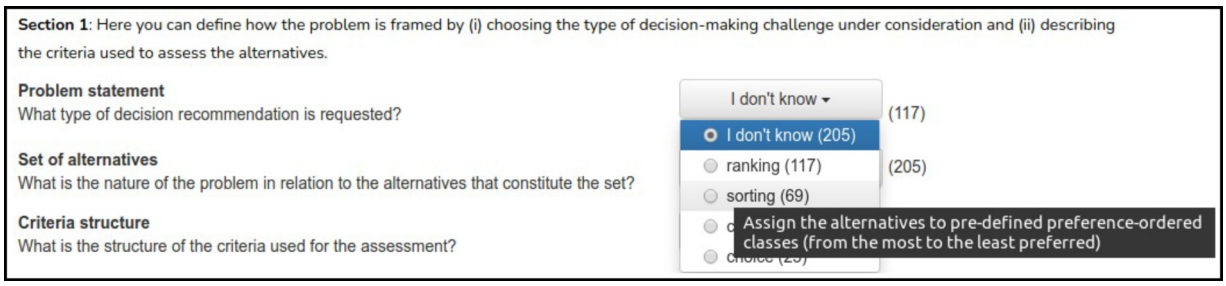


Fig. 3. Example of information box showing the definition of "sorting" as a problem statement.

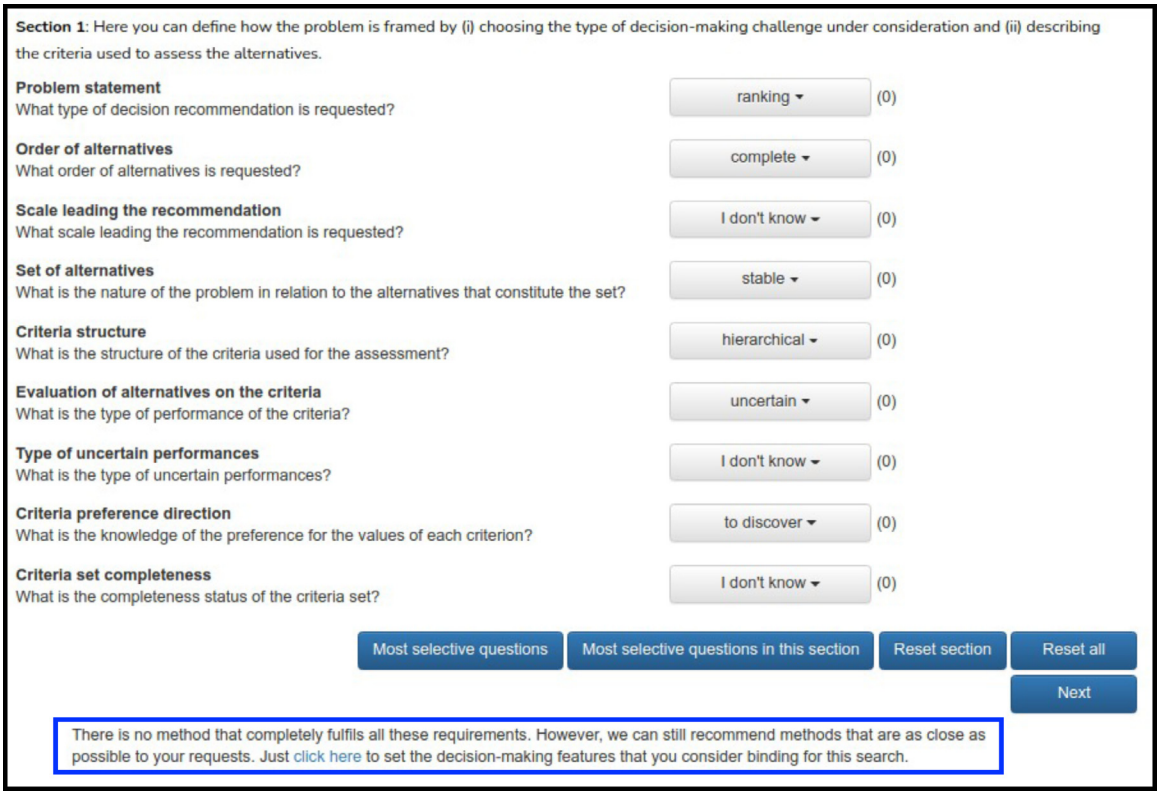


Fig. 4. Example of a DMP where no method completely matches the description of the DMP. The bright blue box highlights the information that is shown when this happens, and the hyperlinked "click here" leads to the window in Fig. 5. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

box showing the definition of "sorting" as a problem statement is presented in Fig. 3.

Second, more than 200 methods are included in the dataset of the MCDA-MSS, which represents a notable library of well-established, as well as more recent methods (*DSS highlight 2 in* Fig. 2). While an initial, although advanced step, towards creating a consistently organized repository of these methods, it can be of use and benefit for all operational researchers and beyond (see also Section 4.1).

Third, even if the dataset of MCDA methods in the MCDA-MSS is wide and a broad range of DMPs can be covered (*DSS highlight 3 in* Fig. 2), there can still be cases where no method completely matches the description of a DMP (see Fig. 4). In this case, the MCDA-MSS is equipped with a recommender function that allows the user defining the binding features that must be satisfied (in this example, the method must have the capability of discovering the preference direction of the criteria performances), as shown in Fig. 5. This strategy caters a solution to the lack of a fully matching method to the DMP by offering one (or more) that is (or are) as close as possible to the DMP.[1] Figure 6 shows a sample of the methods that can still be recommended for the DMP in Fig. 4, together with the feature(s) that are not supported by each method. In this case, UTA-NM and UTA-NM-GRA are shown as suitable methods for the DMP, though they are not tailored for hierarchical criteria with uncertain performances (see column "Missed" in Fig. 6). It is important to note that the recommended method(s) in this scenario (i.e., no complete match with the DMP) should not be blindly applied. Instead, the analyst should discuss the feature(s) that are missed and define the course of action. This can imply that (i) the DMP is re-discussed to assess whether a different formulation of the problem can fully fit with a method, or (ii) a new method is developed that completely satisfies the requirements of the DMP under consideration.

[1] In case resources and competencies are available, another solution to this challenge is that the analyst develops a tailored method for the specific DMP.

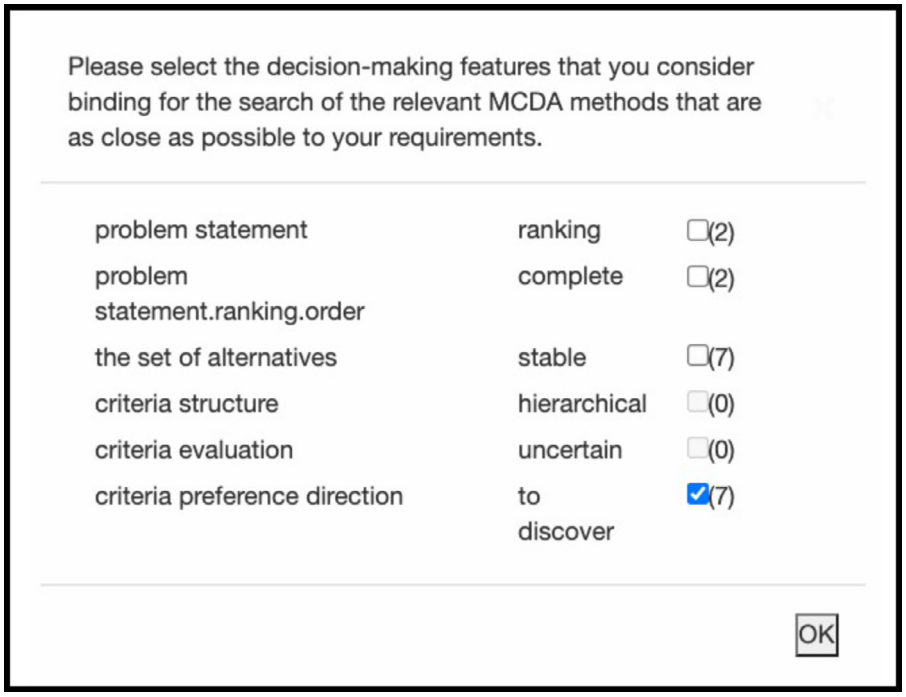


Fig. 5. Choice box to select the features that must be satisfied by the method(s) that will be as close as possible to the DMP.

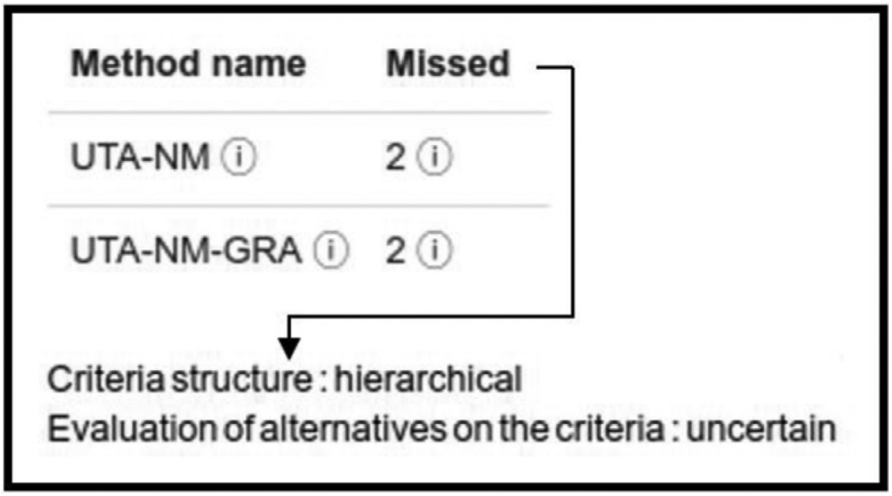


Fig. 6. A sample of the recommended set of methods for the DMP presented in Fig. 4, showing the features that are not satisfied by the UTA-NM and UTA-NM-GRA methods.

Fourth, an opposite situation to the one described above can emerge when the user cannot answer all the questions, and a large set of methods is recommended. This might be caused by a lack of information on the DMP. If there is the possibility of extending the interaction with the DM, the analyst must decide which questions to focus on to reduce the methods set. The button "Most selective questions" shows the questions that (increasingly) minimize the maximal number of recommended MCDA methods irrespective of the answer provided by the user. An example is presented in Fig. 7, which shows that for a DMP where one would only select 'ranking' as a problem statement and 'hierarchical' as a criteria structure, the question on the comparison of performances is the first one that leads to as few methods as possible. In case the user cannot answer this question, the next most selective is the one on the scale used by the method. This capability of the MCDA-MSS of showing the reduction of methods according to each answer provides an immediate understanding of how the available methods are reduced according to the user's choices (*DSS highlight 4 in* Fig. 2). The analyst can, in fact, look at this piece of information to tailor the efforts of interaction with the DM. For example, assuming two questions could be answered from the DM after some discussion, it is more efficient to select the question that leads to the lowest number of methods once answered. This objective is eventually achieved with the presented feature.

Even if the MCDA-MSS recommends methods based on a large set of objective features, it is then still the responsibility of the MCDA practitioners/researchers who are going to use the method to verify which axioms have to be satisfied to meaningfully apply it (Cinelli et al., 2020; Roy & Słowiński, 2013).

The DSS is available freely at http://mcdamss.com. To access it, users need to create an account by providing their email address and password. These are stored in the MySQL database (www.

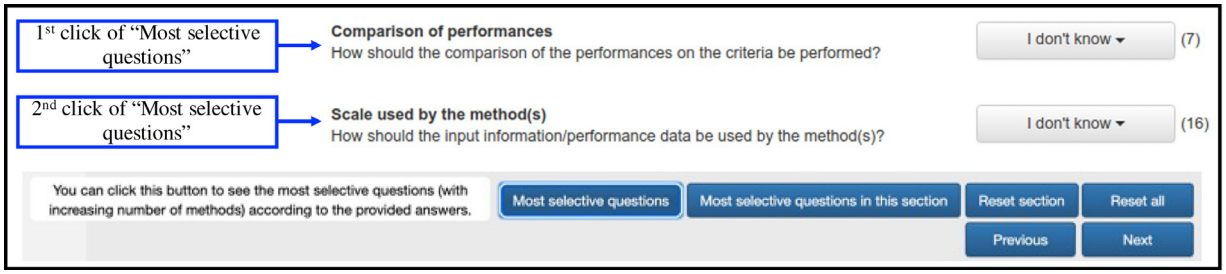


Fig. 7. Activation of the "Most selective questions" button, showing that for this DMP the question on the comparison of performances is the most selective, followed by the one scale used by the method.

mysql.com), with passwords hashed using the bcrypt function. The software's frontend was implemented using bootstrap v4.3.1 (www.getbootstrap.com) and jQuery v3.4.1 (www.jquery.com). It is responsible for the user interface layout. The backend was written in PHP7 (www.php.net). Each time a user changes input information, an appropriate SQL query is formulated to select suitable methods stored in the MySQL database along with their descriptions and features. Metadata such as the lists of questions and methods, the hierarchical structure of questions, or their explanations are supplied as csv files. They are subsequently processed using Python scripts.

3. MCDA-MSS in action

This section describes stage 3 of the methodology in Fig. 2, presenting the test of the MCDA-MSS on a set of case studies selected from the literature. This was performed to (i) evaluate the applicability of the MCDA-MSS to multiple DMPs for different real-life applications, and (ii) discuss the MCDA method(s) used in the case studies and that/those recommended by the MCDA-MSS. The set-up of the test of MCDA-MSS with the literature case studies is firstly given in Section 3.1. Then, the results of its application are presented and discussed in Section 3.2.

3.1. Set-up for the MCDA-MSS test with literature case studies

The case studies selected for the test of the MCDA-MSS included a set of MCDA applications from the peer-reviewed literature based on two conditions. First, they should cover a varied set of application areas given that MCDA methods have been used in a wide variety of domains (Cajot et al., 2017; Razmak & Aouni, 2014). Second, they should include some case studies where the authors of this paper think that mistakes were performed in selecting the MCDA method. Based on these conditions, ten case studies were chosen, which dealt with airline service quality (Liou, 2011), land remediation (Sparrevik, Barton, Bates & Linkov, 2012), biofuels energy chains (Dias, Passeira, Malça & Freire, 2016), urban regeneration (Ferretti & Degioanni, 2017), inventory management (Partovi & Burton, 1993), land use suitability analysis (Qiu, Zhu, Pan, Hu & Amable, 2017), and energy production planning (Haurant, Oberti & Muselli, 2011; Jun, Tian-tian, Yi-sheng & Yu, 2014; Maxim, 2014).

It must be noted that a case study is defined as "a DMP that is solved with an MCDA method". Thus, there can be more than one case study in a single journal paper, as is the case for Dias et al. (2016), in this test of the MCDA-MSS, where two case studies are identified. It must also be stressed that the objective of this test was not to have a large set of studies from the same application area to apply the MCDA-MSS to derive some general conclusions on its applicability and performance (see Section 4.2 on future research). Rather, it was intended as an initial evaluation of the MCDA-MSS in terms of its capability to transparently lead the analyst through the selection of MCDA methods. In addition, it provides preliminary considerations from the choices made by analysts in past research on the chosen MCDA methods.

The taxonomy features are implemented in the MCDA-MSS as a set of questions that evaluate whether those activated in each DMP are matched with that/those of a (set of) MCDA method(s). Table 2 presents a few example questions that can be used to describe the MCDA methods and the case studies. The taxonomy can, in fact, be used to describe the case studies in the same manner as the MCDA methods.

Table 3 shows a simplified example of the description of the MCDA methods, the case studies, the chosen method(s) in the case studies, and that/those recommended by the MCDA-MSS. It can be noticed that as far as literature case study 1 is concerned, the method chosen by the authors of the study is the same as the one recommended by MCDA-MSS, while this is not true for literature case study 2.

3.2. Results of the application of the MCDA-MSS to literature case studies

Nine case studies were selected from the peer-reviewed literature to study the applicability of the MCDA-MSS and its decision support capabilities, using its questioning procedure presented in Section 3.1. The results are summarized in Table 4, which shows (i) the MCDA method that was chosen by the authors of the case study, (ii) the decision-making features that have been missed by the MCDA method chosen in the case study with respect to the case study description by the authors of this paper, and (iii) the recommended MCDA method(s) by the MCDA-MSS with a complete match of the decision-making features and (iv) with missed features.

The taxonomy of the MCDA-MSS has been capable of describing completely all the case studies (for the mapping of each case study, see Appendix D in the ESI). The selected MCDA methods in six case studies are part of those recommended by the MCDA-MSS, as shown in Table 4. Based on the framing of the MCDA-MSS, this is an indication of the same description of the MCDA methods and the case studies, like in literature case study 1 in Table 3. Specific details on the choices driving the selection of the MCDA methods are provided in these cases. Ferretti and Degioanni (2017), who use a multi-attribute value function model, emphasize both the meaning of weights as tradeoffs and the need to create value functions to "harmonize" the different measurement scales. Dias et al. (2016) account for the robustness of the decision recommendation by adopting a value-based model with partial information on the input weights, showing the synergistic benefit of stochastic and exact robustness analysis. In the former (i.e., stochastic), the probability of receiving a certain rank provides a perspective on the trend of the results, while in the latter (i.e., exact), the visualization of how much one alternative can outperform the other highlights how different the alternatives do actually score. Liou (2011) devised a transparent and straightforward

Table 2
Example questions used to describe the MCDA methods and a case study (Adapted from Cinelli et al. (2020)).

Section	Question context	Questions used to describe the MCDA method	Questions used to describe the case study
1: Problem typology	Problem statement	What type of decision recommendation does the method provide?	What type of decision recommendation is requested?
	Order of alternatives, if ranking as a problem statement was chosen	Does the method provide a partial or complete order of alternatives as a final decision recommendation?	Does the DM require a complete order of alternatives as a final decision recommendation, or a partial one would be enough?
	Structure of the set of criteria	Can the method accept a flat and/or hierarchical structure of the criteria?	Is the structure of the criteria flat or hierarchical?
2: Preference model	Comparison of the performances on all the criteria	How does the method perform the comparison of the performances on the criteria?	How should the comparison of the performances on the criteria be performed?
	Compensation between criteria	What is the level of compensation between the criteria performances that the method implements?	How much can the good performance on a criterion compensate for the bad performance on another criterion?
3: Elicitation of preferences	Type of weights, if the user wants criteria weights to be used	Does the method accept precise and/or imprecise weights?	Should precise or imprecise weights be used in this case study?
	Type of pairwise comparison thresholds if the user wants pairwise comparison thresholds to be used	Which type, if any, of pairwise comparison thresholds does the method accept?	What type of pairwise comparison thresholds should be used?
	Interactions between criteria	What type, if any, of interactions can the method handle?	What type, if any, of interactions should be handled?
4: Exploitation of the preference relation induced by the preference model	Type of exploitation of the preference relation induced by the preference model	Which type of exploitation of the preference relation induced by the preference model does the method support?	What type of exploitation of the preference relation induced by the preference model should be applied for this case study?
	Type of output variability analysis, if the user indicated that output variability analysis be performed	What type, if any, of output variability analysis can the method perform?	How should the output variability analysis be conducted?

sorting system for airline service evaluation using the Dominance-based Rough Set Approach (Greco et al., 2016b), which has the additional noteworthy capacity of accepting inconsistencies in the input preferences. Haurant et al. (2011) stress the limited compensation, variable preference thresholds, and incommensurable measurement scales as the main reasons for the choice model they developed using ELECTRE IS (Figueira et al., 2016). Lastly, the added value of accounting for stochastic input was particularly emphasized in the work of Sparrevik et al. (2012), who used a ranking method (i.e., SMAA-PROMETHEE (Corrente et al., 2014)) to tackle a DMP that lacks deterministic performances, and provided a clear indication of robustness for the final recommendations.

However, the trend presented in the previous paragraph was not found in four case studies since the methods chosen by the authors of the studies were not included in those recommended by the MCDA-MSS. This implies that the description of the MCDA methods and the case studies are not the same, resulting in a situation like in literature case study 2 in Table 3. This set of case studies with a mismatch of MCDA methods shows how the MCDA-MSS can be a warning tool for errors committed in the selection of an MCDA method selection. The reason(s) for the mismatches are presented in the column "Missed features between the MCDA method chosen in the case study and the case study description by the authors of this paper" in Table 4.

Maxim (2014) proposed a composite indicator in the form of an index, obtained by combining min-max normalization and additive weighted mean, and it is the first mismatch found by the MCDA-MSS. The first reason for this mismatch is using an MCDA method (i.e., weighted mean) that is not suited for criteria whose measurement scale is ordinal, such as the ability to respond to demand and social acceptability. In these criteria, the numbers coding an order have qualitative meaning. Any operator that assumes that these numbers have a quantitative meaning implements compensations that are not mathematically justifiable, leading to results that are not scientifically meaningful (Ebert & Welsch, 2004; Munda & Nardo, 2009). The assumption of the quantitative meaning of these numbers can result in the following. The arbitrary choice of the coding in the input data can lead to different results, solely due to the discretionary numbering for the qualitative measurement scale (e.g., low = 1, medium = 2, high = 3 could also be represented as low = 2, medium = 4, high = 7 as far as the increase is preserved) (Pollesch & Dale, 2016). The requirement of choosing an aggregation algorithm that can properly handle the input information is an important consideration that has been made repeatedly in the MCDA literature (in, e.g., Guitouni and Martel (1998), Roy (2016a)). The second reason for the mismatch is that the weights used in this case study were elicited with an elicitation protocol that provided importance coefficients as an outcome. Unfortunately, the weighted mean, which was the algorithm used by these authors, would have required weights in the form of trade-offs (Munda, 2008). The subjective assignment of importance weights based on expert judgment, irrespective of the measurement scale and with no reference to the acceptable trade-offs between the criteria, is a common mistake reported in the MCDA literature when a single score model is selected (Cinelli, Coles & Kirwan, 2014; Munda & Nardo, 2005). The MCDA-MSS recommends QUALIFLEX (Paelinck, 1976) as one of the methods as close as possible to this DMP, as summarized in the last column in Table 4. This is a method that uses weights as set in the study (i.e., importance coefficients). However, it implies that the stakeholder accepts that the way the information is used by the method consists in pairwise comparisons between the alternatives and not in taking into account the individual performance on each criterion.

The second mismatch was found in the case study of Qiu et al. (2017), who used the weighted additive mean (Langhans et al., 2014) to sort land according to its suitability for livestock development. The first reason for the mismatch is the use of number-coded qualitative criteria in a method interpreting all scales of cri-

Table 3
Example of a description of MCDA methods and case studies according to a sample of features of the taxonomy. √ in "Description of the MCDA methods" = the method supports this feature; √ in "Description of the case studies" = the case study requests this feature. The last three columns show the method(s) chosen by the authors of the case study, the recommendation of the MCDA-MSS with the relevant method(s) for each case study with complete match, and the one(s) with the missed features, respectively
(Adapted from Cinelli et al. (2020) and Wątróbski et al. (2019)).

	c.1 - Problem typology							Other features ...	Method(s) chosen in the case study	MCDA method(s) recommended by the MCDA-MSS, complete match	MCDA method(s) recommended by the MCDA-MSS, with missed features
	c.1.1 - Problem statement			c.1.2 – Set of alternatives		c.1.3 – Criteria structure					
	Ranking	Sorting	...	Stable	Incre-mental	Flat	Hierarchical				
Description of the MCDA methods											
Method 1		√		√	√	√					
Method 2	√			√		√	√				
Method ...			√		√		√	√			
Description of the case studies											
Literature case study 1		√			√	√			Method 1	Method 1	/
Literature case study 2	√			√		√		√	Method 2	/	Method 3 & 4

Table 4
Summary of results of the MCDA-MSS test with the literature case studies.

Case study	Reference	MCDA method chosen in the case study	Missed features between the MCDA method chosen in the case study and the case study description by the authors of this paper	MCDA method(s) recommended by the MCDA-MSS with complete match of features	Some of the MCDA method(s) recommended by the MCDA-MSS with fewest missed features, some shown in '[]' brackets
1	Ferretti & Degioanni, 2017	MAVT	None	MAVT	/
2	Dias et al., 2016	SMAA-WAM	None	SMAA-WAM, HSMAA	/
3	Dias et al., 2016	VIP	None	VIP	/
4	Liou, 2011	DRSA Variable Consistency (VC) - sorting	None	DRSA Variable Consistency (VC) - sorting	/
5	Haurant et al., 2011	ELECTRE IS	None	ELECTRE IS, MCHP-ELECTRE, RUBIS	/
6	Sparrevik et al., 2012	SMAA-PROMETHEE	None	SMAA-PROMETHEE	/
7	Maxim, 2014	WAM with performances transformation	1. Use of number-coded qualitative-ordinal criteria 2. Use of weights meaning importance coefficients	/	QUALIFLEX [works with pairwise comparisons between the alternatives]
8	Qiu et al., 2017	WAM with performances transformation	1. Use of number-coded qualitative-ordinal criteria 2. Use of AHP for weighting combined with other methods 3. Problem statement sorting	/	FlowSort, PROMSORT [criteria weights are used as importance coefficients; works with pairwise comparisons between the alternatives]
9	Partovi & Burton, 1993	AHP	1. Problem statement sorting 2. Use of number-coded qualitative-ordinal criteria	/	AHP Sort - Boundary profiles [use of number-coded qualitative-ordinal criteria]
10	Jun et al., 2014	ELECTRE II	1. Data-driven transformation of criteria performances	/	ELECTRE II [performances are transformed with a data-driven normalization approach and then compared]

teria as quantitative, which has been presented above. The second one is related to the process of weighting (see also Table 4), as the authors use weights obtained with the AHP (Saaty, 1980) in a method (the weighted additive mean) that requires a different type of weights (i.e., trade-offs). AHP weights have a specific meaning assigned to them, the one of a relative priority of an element of the hierarchy which can be a (sub-)criterion or a performance of an alternative on a bottom-level (also called elementary) criterion. They serve to calculate the global score of each alternative as a sum of products of these weights along all paths of the hierarchy tree from the alternative to the goal (Salo & Hämäläinen, 1997). Each product includes the weight assigned to the performance of the alternative on a given criterion and the weights of all upper-level criteria on the path to the goal. Consequently, these are not trade-off weights, because they are not used in a weighted sum where the performances of alternatives on the bottom-level criteria are multiplied by the weights, and they are also not relative importance coefficients that are used in a voting-like procedure. Their use in a method like the weighted mean (Langhans et al., 2014) is thus incorrect. The third aspect that causes the mismatch, as also summarized in Table 4, is novel in this test, and it relates to the formulation of the DMP. It shows the implication of a non-optimal formulation of the problem statement, a delicate and key step of the initial phase of any MCDA process (Bouyssou, Marchant, Pirlot, Tsoukiàs & Vincke, 2006b). A weighted average is a method tailored to ranking problems, providing a score that is used to rank the alternatives. To obtain a sorting out of this score, arbitrary and abstract cut-off levels need to be defined to lead the assignments to different decision classes (e.g., 0.3 could be the threshold between a poor and medium overall performance). The MCDA-MSS suggests more suitable methods for this case study that involves a sorting problem statement, namely FlowSort (Nemery & Lamboray, 2008) and PROMSORT (Araz & Ozkarahan, 2005), assuming weights are elicited as importance coefficients, and it is agreed to use pairwise comparisons-based modeling.

The third mismatch was found by the MCDA-MSS for the case study of Partovi and Burton (1993), who used the original AHP method to sort inventory items in three classes (very important, important, and least important). There are two explanations for this mismatch. The first is the same as the latter presented in the previous study, namely the selection of a method for ranking when the DMP is clearly a sorting one. This forced the authors to define the cut-off levels for the final score to assign the inventory items to the different classes. The second reason is the treatment of the criteria as qualitative in a method (AHP) that requires ratio criteria to provide meaningful results. In fact, AHP operates by comparing alternatives and criteria on a ratio scale to define how many times one alternative is better than another and how many times one criterion is preferred over another one (Ishizaka & Labib, 2011; Wedley, 2010). These comparisons can be meaningfully made only with criteria that are measured in a ratio scale, which is not the approach used in this case study. The MCDA-MSS allows tackling these shortcomings of the case study and proposes a solution that still belongs to the AHP family. This is the AHPSort method (Ishizaka et al., 2012), a sorting method that implements that AHP modeling and which nicely fits with all but one of the requirements of this case study. Thus, the only modification that is required consists in using criteria that are on a ratio scale. This could be realistic for this research as three out of four selected criteria are measurable in such a way, being price (in $), lead time (in days), and demand (in units/year).

Lastly, the MCDA-MSS flagged another mismatch for the case study by Jun et al. (2014). These authors applied a composite indicators framework, which requires, among other steps, the normalization of the data to convert the evaluation of the performances from different scales to a common one. The MCDA-MSS signals this

choice as an error, because these authors selected ELECTRE II (Roy, 1991), which is from the outranking methods family. One key feature of the methods that belong to this family is that the independence from normalization was actually one of the main reasons that justified their conception, since the developers wanted these methods to be "as close as possible" to the original dataset. The MCDA-MSS accounts for this requirement in its recommendation, reported in Table 4, which is to reapply the same method, but without any normalization of criteria performances.

4. Conclusions

4.1. Key contributions of the MCDA-MSS

This paper presents a DSS, called the MCDA-MSS, that helps decision analysts describe complex decision-making processes and choose the MCDA method(s) relevant for each case study. It provides five main contributions.

First, it is a tool for decision analysts to transparently and comprehensively describe and reproducibly shape case studies while interacting with the DM. It can thus guarantee consistent and homogeneous communication among different stakeholders. The MCDA-MSS can streamline the complexity involved in MCDA methods selection, while being aware of the possibilities at each stage, as well as their implications. If this DSS would be used in future case studies, this could guarantee accountability and trust of the research, as well as enable its comparability. In fact, the MCDA-MSS provides a recordable description of the choices that led to selecting the MCDA method(s). It uses an accessible vocabulary to describe each step that needs to be taken while moving towards the method(s) selection. It also divides the selection process into four user-friendly sections, where a wide amount of information on the case studies is elicited and linked to the database of MCDA methods, using a stepwise questioning procedure *(DSS highlight 1)*.

Second, thanks to its extensive database of 205 MCDA methods and the large set of characteristics (156 in total), it is capable of dealing with DMPs that span from very simple to very complex, leading to the respective identification of those methods which are most appropriate *(DSS highlight 2)*. Suppose we adopt a knowledge transfer perspective among decision analysts. In that case, the taxonomy of the MCDA-MSS operationalized in its questioning procedure can also be seen as an educational and training tool for inexperienced decision analysts. They can first learn the components and steps necessary to lead an MCDA problem and then guarantee a sound description of each case study, resulting in an informed choice of the suitable MCDA method(s). The information boxes used to describe each DMP feature serve the main purpose of enabling inexperienced MCDA practitioners to learn how MCDA processes can be structured. Furthermore, they transparently show DMs why certain choices are made based on their requirements, lifting the black-box effect that some technical and succinct terminology from the MCDA domain might cause.

Third, it offers MCDA method(s) recommendations also for DMPs where there is no perfectly matching method. This feature ensures the analyst is never "left alone" in the selection process by making the most of the available information on the DMP and the knowledge in the MCDA-MSS *(DSS highlight 3)*.

Fourth, a refinement strategy for directing the user to the most discriminatory questions is proposed in cases where not all the questions can be answered. In these situations, the MCDA-MSS aids the decision analyst to choose where to focus efforts to reduce the set of available options as efficiently as possible *(DSS highlight 4)*.

Fifth, the MCDA-MSS has the potentials to unveil methodological mistakes that analysts have made in a selection of the methods *(DSS highlight 5)*. These include the erroneous use of weights as importance coefficients instead of as trade-offs, using ordinal measurement scales in a method that requires quantitative ones, employing a ranking method to deal with a sorting problem, or performing redundant normalization of performances. The use of the MCDA-MSS has the capacity to avoid these mistakes being committed in future case studies. Thus, making sure that the decision recommendation is aligned with the problem's characteristics and preferences of the involved stakeholders. The MCDA-MSS was tested with a set of ten case studies from the peer-reviewed literature. It confirmed its capability to describe DMPs for different real-life applications, find suitable MCDA methods for each DMP, and unveil some methodological flaws in the selection of methods.

4.2. Future research

The MCDA-MSS and its underlying structure, called taxonomy, can represent an evolving repository of knowledge in the MCDA domain. In this respect, several extensions of this research can be envisioned. The first one consists in expanding the test of the MCDA-MSS by including a large set of case studies from a chosen application area. This could lead to the identification of trends in the use of MCDA methods, together with its misuses. This research could thus start providing an initial list of the most frequent methodological mistakes when choosing an MCDA method. Another extension could involve developing the (already substantial) pool of single DM-focused MCDA methods included in the MCDA-MSS, considering their ever-increasing availability. A further extension could be devoted to the inclusion of methods that support multiple DMs, an avenue of operational research that is crucial due to the recognition that many decisions are made not by individuals, but in groups, such as committees and boards (Lahdelma & Salminen, 2001). As far as software support is concerned, the presence of user-friendly software implementing MCDA methods can be of notable advantage in the development of the MCDA process and the interaction with the DM, making it a further very valuable feature for the extension of the MCDA-MSS.

The MCDA-MSS could also lead to recommending more than one method given its wide repository. This can lead to a notable challenge for the analysts, who might be asked to still propose one method for a certain project. This can be driven by the limited time availability or the DM's commitment to providing one solution to the problem at hand. The final choice can then be influenced by some of the qualitative features of the methods, like the reported easiness of using it (Polatidis, Haralambopoulos, Munda & Vreeker, 2006), its use in certain application areas (Moghaddam, Nasiri & Mousavi, 2011), or the availability of software implementing it (Weistroffer & Li, 2016). One solution provided in this regard is to develop suitability indices to aggregate the characteristics of the DMP and provide a synthetic measure of the suitability of different methods, which are respectively ranked according to it (e.g., Li et al. (2008) and Guarini et al. (2018)). This, however, includes subjectivity due to the inherent nature of the qualitative features as well as their aggregation, making the final result less based on objective facts.

Overall, the MCDA-MSS is intended to help an analyst facing a multiple criteria decision problem to choose consciously an MCDA method that will respond positively to the needs of the DM and satisfy all the constraints characterizing the decision situation. The authors of this paper are eager to engage in a community-wide initiative involving experts in MCDA methods, decision analysts using these methods, and decision makers receiving decision recommendations. This combined action can result in an expansion of the methods repository as well as the tests on case studies, coupled with the inclusion of additional decision aiding features in the web software. The outcome of this initiative can be a sustained contribution to the relevant and transparent use of MCDA methods to

solve real-world problems. All these activities can be monitored on the MCDA-MSS dedicated webpage http://mcdamss.com.

5. Disclaimer

The U.S. Environmental Protection Agency, through its Office of Research and Development, collaborated in the research described herein as part of the Collaborative Research and Development Agreement #1040–18 with Poznan University of Technology. It has been subjected to the Agency's administrative review and has been approved for external publication. Any opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Agency. Therefore, no official endorsement should be inferred. Any mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Each method in the database of the MCDA-MSS required the assignment of binary values (i.e., either 0 or 1) to each possible answer considered in the taxonomy. Given that the selected possible answers are 156, this means that the database of the MCDA-MSS is composed of 205 rows and 156 columns, requiring the manual assignment of 31,980 binary values (i.e., 1 or 0). Given the very large number of manual inputs, the authors acknowledge that there might be some mappings of MCDA methods that other authors/researchers might not agree with. Rather than this being cause for arguments between the developers of the MCDA-MSS and those authors, we encourage them to contact us to improve the database and contribute to this long-lasting initiative, since this is just the beginning of the MCDA-MSS.

Lastly, the developers of the MCDA-MSS do not take responsibility for the choices a user makes based on the recommendations of the MCDA-MSS.

Funding

Marco Cinelli acknowledges financial support from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 743553. Miłosz Kadziński acknowledges financial support from the Polish National Science Center under the SONATA BIS project (grant no. DEC-2019/34/E/HS4/00045). Grzegorz Miebs acknowledges support from the Polish Ministry of Science and Higher Education under the Diamond Grant project (grant no. DI2018 004348).

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank MSc Samson Tu, MSc Josef Hardi, Dr Matthew Horridge, Dr Rafael Gonçalves, and Professor Mark Musen from the Stanford Center for Biomedical Informatics Research (BMIR) at Stanford University for the constructive feedback provided during the conceptualization of our taxonomy.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2022.01.011.

References

Ahn, B. S. (2017). The analytic hierarchy process with interval preference statements. Omega, *67*, 177–185.
Alinezhad, A., & Khalili, J. (2019). *New Methods and Applications in Multiple Attribute Decision Making (MADM)*. Cham: Springer.
Almeida-Dias, J., Figueira, J. R., & Roy, B. (2010). ELECTRE TRI-C: A multiple criteria sorting method based on characteristic reference actions. *European Journal of Operational Research, 204*, 565–580.
Almeida-Dias, J., Figueira, J. R., & Roy, B. (2012). A multiple criteria sorting method where each category is characterized by several reference actions: The ELECTRE TRI-nC method. *European Journal of Operational Research, 217*, 567–579.
Angilella, S., Catalfo, P., Corrente, S., Giarlotta, A., Greco, S., & Rizzo, M. (2018). Robust sustainable development assessment with composite indices aggregating interacting dimensions: The hierarchical-SMAA-Choquet integral approach. *Knowledge-Based Systems, 158*, 136–153.
Angilella, S., Greco, S., & Matarazzo, B. (2010). Non-additive robust ordinal regression: A multiple criteria decision model based on the Choquet integral. *European Journal of Operational Research, 201*, 277–288.
Araz, C., & Ozkarahan, I. (2005). A Multicriteria sorting procedure for financial classification problems: The case of business failure risk assessment. *Intelligent Data Engineering and Automated Learning - IDEAL 2005* (pp. 563–570). Berlin, Heidelberg: Springer Berlin Heidelberg.
Araz, C., & Ozkarahan, I. (2007). Supplier evaluation and management system for strategic sourcing based on a new multicriteria sorting procedure. *International Journal of Production Economics, 106*, 585–606.
Argyris, N., Morton, A., & Figueira, J. (2014). CUT: A Multicriteria approach for Concavifiable preferences. *Operations Research, 62*, 633–642.
Bana E Costa, C. A., & Vansnick, J.-. C. (1999). The MACBETH approach: Basic ideas, software, and an application. In N. Meskens, & M. Roubens (Eds.), *Advances in decision analysis* (pp. 131–157) (Eds.). Dordrecht: Springer Netherlands.
Behzadian, M., Kazemzadeh, R. B., Albadvi, A., & Aghdasi, M. (2010). PROMETHEE: A comprehensive literature review on methodologies and applications. *European Journal of Operational Research, 200*, 198–215.
Belton, V., & Stewart, T. J. (2002). The multiple criteria problem. pp. *13-34 Multiple Criteria Decision Analysis; an Integrated Approach*. Kluwer Academic Publisher.
Benoit, V., & Rousseaux, P. (2003). Aid for aggregating the impacts in life cycle assessment. *International Journal of Life Cycle Assessment, 8*, 74–82.
Błaszczyński, J., Greco, S., & Słowiński, R. (2012). Inductive discovery of laws using monotonic rules. *Engineering Applications of Artificial Intelligence, 25*, 284–294.
Bous, G., Fortemps, P., Glineur, F., & Pirlot, M. (2010). ACUTA: A novel method for eliciting additive value functions on the basis of holistic preference statements. *European Journal of Operational Research, 206*, 435–444.
Bouyssou, D., Marchant, T., Pirlot, M., Tsoukiàs, A., & Vincke, P. (2006a). Aggregation Procedures. In D. Bouyssou, T. Marchant, M. Pirlot, A. Tsoukiàs, & P. Vincke (Eds.), *Evaluation and Decision Models with Multiple Criteria: Stepping Stones for the Analyst* (pp. 169–235) (Eds.). Boston, MA: Springer US.
Bouyssou, D., Marchant, T., Pirlot, M., Tsoukiàs, A., & Vincke, P. (2006b). Problem formulation and structuring: The decision aiding process. In D. Bouyssou, T. Marchant, M. Pirlot, A. Tsoukiàs, & P. Vincke (Eds.), *Evaluation and Decision Models with Multiple Criteria: Stepping Stones for the Analyst* (pp. 19–65) (Eds.). Boston, MA: Springer US.
Brans, J.-. P., & De Smet, Y. (2016). PROMETHEE Methods. In S. Greco, M. Ehrgott, & J. R. Figueira (Eds.), *Multiple Criteria Decision Analysis: State of the Art Surveys* (pp. 187–219) (Eds.). New York, NY: Springer New York.
Brauers, W. K. M., & Zavadskas, E. K. (2010). Project management by multimoora as an instrument for transition economies. *Technological and Economic Development of Economy, 16*, 5–24.
Buchanan, L., & O'Connell, A. (2006). A brief history of decision making. *Harvard Business Review*. https://hbr.org/2006/01/a-brief-history-of-decision-making.
Cabello, J. M., Navarro-Jurado, E., Rodríguez, B., Thiel-Ellul, D., & Ruiz, F. (2019). Dual weak–strong sustainability synthetic indicators using a double reference point scheme: The case of Andalucía. *Spain. Operational Research, 19*, 757–782.
Cailloux, O., Lamboray, C., & Nemery, P. (2007). A taxonomy of clustering procedures. In *66th Meeting of the European Working Group on MCDA*.
Cajot, S., Mirakyan, A., Koch, A., & Maréchal, F. (2017). Multicriteria decisions in urban energy system planning: A review. *Frontiers in Energy Research, 5*, 10.
Celik, M., & Topcu, Y. I. (2009). Analytical modelling of shipping business processes based on MCDM methods. *Maritime Policy & Management, 36*, 469–479.
Cinelli, M. (2017). The art of supporting decision-making. *Exchanges: The Warwick Research Journal, 4*, 298–308.
Cinelli, M., Coles, S. R., & Kirwan, K. (2014). Analysis of the potentials of multi criteria decision analysis methods to conduct sustainability assessment. *Ecological Indicators, 46*, 138–148.
Cinelli, M., Kadziński, M., Gonzalez, M., & Słowiński, R. (2020). How to support the application of multiple criteria decision analysis? Let us start with a comprehensive taxonomy. Omega, *96*, Article 102261.
Cinelli, M., Spada, M., Kim, W., Zhang, Y., & Burgherr, P. (2021). MCDA Index Tool: An interactive software to develop indices and rankings. *Environment Systems and Decisions, 41*, 82–109.
Corrente, S., Figueira, J. R., & Greco, S. (2014). The SMAA-PROMETHEE method. *European Journal of Operational Research, 239*, 514–522.
Corrente, S., Greco, S., Kadziński, M., & Słowiński, R. (2013). Robust ordinal regression in preference learning and ranking. *Machine Learning, 93*, 381–422.

Corrente, S., Greco, S., & Słowiński, R. (2012). Multiple criteria hierarchy process in robust ordinal regression. *Decision Support Systems, 53*, 660–674.
Corrente, S., Greco, S., & Słowiński, R. (2017). Handling imprecise evaluations in multiple criteria decision aiding and robust ordinal regression by n-point intervals. *Fuzzy Optimization and Decision Making, 16*, 127–157.
De Keyser, W. S. M., & Peeters, P. H. M. (1994). ARGUS — A new multiple criteria method based on the general idea of outranking (1994). (Ed.) M. Paruccini (Ed.), *Applying Multiple Criteria AID for Decision to Environmental Management*. Kluwer Academic Publishers.
Del Vasto-Terrientes, L., Valls, A., Slowinski, R., & Zielniewicz, P. (2015). ELECTRE-II-H: An outranking-based decision aiding method for hierarchically structured criteria. *Expert Systems with Applications, 42*, 4910–4926.
Dias, L. C. (2007). *A Note on the Role of Robustness Analysis in Decision-Aiding Processes* (pp. 1872–4981). Paris: LAMSADE.
Dias, L. C., Antunes, C. H., & Insua, D. R. (2012). Dealing with uncertainty in decision support systems: Recent trends 2000–2011. *Intelligent Decision Technologies, 6*, 245–264.
Dias, L. C., & Climaco, J. N. (2000). Additive aggregation with variable interdependent parameters: The VIP analysis software. *The Journal of the Operational Research Society, 51*, 1070–1082.
Dias, L. C., Morton, A., & Quigley, J. (2018). Elicitation: State of the art and science. In L. C. Dias, A. Morton, & J. Quigley (Eds.), *Elicitation: The Science and Art of Structuring Judgement* (pp. 1–14) (Eds.). Cham: Springer International Publishing.
Dias, L. C., & Mousseau, V. (2006). Inferring ELECTRE's veto-related parameters from outranking examples. *European Journal of Operational Research, 170*, 172–191.
Dias, L. C., & Mousseau, V. (2018). Eliciting multi-criteria preferences: ELECTRE models. In L. C. Dias, A. Morton, & J. Quigley (Eds.), *Elicitation: The science and art of Structuring Judgement* (pp. 349–375) (Eds.). Cham: Springer International Publishing.
Dias, L. C., Passeira, C., Malça, J., & Freire, F. (2016). Integrating life-cycle assessment and multi-criteria decision analysis to compare alternative biodiesel chains. *Annals of Operations Research*. https://doi.org/10.1007/s10479-10016-12329-10477.
Doumpos, M., & Zopounidis, C. (2011). Preference disaggregation and statistical learning for multicriteria decision support: A review. *European Journal of Operational Research, 209*, 203–214.
Durbach, I., Lahdelma, R., & Salminen, P. (2014). The analytic hierarchy process with stochastic judgements. *European Journal of Operational Research, 238*, 552–559.
Ebert, U., & Welsch, H. (2004). Meaningful environmental indices: A social choice approach. *Journal of Environmental Economics and Management, 47*, 270–283.
Fernández, E., Figueira, J. R., & Navarro, J. (2020). Interval-based extensions of two outranking methods for multi-criteria ordinal classification. Omega, *95*, Article 102065.
Fernández, E., Figueira, J. R., Navarro, J., & Roy, B. (2017). ELECTRE TRI-nB: A new multiple criteria ordinal classification method. *European Journal of Operational Research, 263*, 214–224.
Ferretti, V., & Degioanni, A. (2017). How to support the design and evaluation of redevelopment projects for disused railways? A methodological proposal and key lessons learned. *Transportation Research Part D: Transport and Environment, 52*, 29–48 **Part A**.
Figueira, J. R., Greco, S., & Roy, B. (2009a). ELECTRE methods with interaction between criteria: An extension of the concordance index. *European Journal of Operational Research, 199*, 478–495.
Figueira, J. R., Greco, S., & Słowiński, R. (2009b). Building a set of additive value functions representing a reference preorder and intensities of preference: GRIP method. *European Journal of Operational Research, 195*, 460–486.
Figueira, J. R., Mousseau, V., & Roy, B. (2016). ELECTRE Methods. In S. Greco, M. Ehrgott, & J. R. Figueira (Eds.), *Multiple Criteria Decision Analysis: State of the Art Surveys* (pp. 155–185) (Eds.). New York, NY: Springer New York.
French, S. (1993). A manifesto for the new MCDA era. *Journal of Multi-Criteria Decision Analysis, 2*, 125–127.
Gershon, M. E. (1981). *Model Choice in Multiobjective Decision-Making in Natural Resources Systems*. University of Arizona Ph.D. thesis. Accessed on 6 Nov 2018 at: https//repository.arizona.edu/handle/10150/281947.
Gershon, M., & Duckstein, L. (1983). An algorithm for choosing of a Multiobjective technique. *Essays and Surveys on Multiple Criteria Decision makIng* (pp. 53–62). Berlin, Heidelberg: Springer Berlin Heidelberg.
Govindan, K., & Jepsen, M. B. (2016). ELECTRE: A comprehensive literature review on methodologies and applications. *European Journal of Operational Research, 250*, 1–29.
Govindan, K., Kadziński, M., Ehling, R., & Miebs, G. (2019). Selection of a sustainable third-party reverse logistics provider based on the robustness analysis of an outranking graph kernel conducted with ELECTRE I and SMAA. Omega, *85*, 1–15.
Grabisch, M., & Labreuche, C. (2008). A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. *4OR, 6*, 1–44.
Grabisch, M., & Labreuche, C. (2016). Fuzzy measures and integrals in MCDA. In S. Greco, M. Ehrgott, & J. R. Figueira (Eds.), *Multiple Criteria Decision Analysis: State of the Art Surveys* (pp. 553–603) (Eds.). New York, NY: Springer New York.
Granata, M. F. (2017). Tentative reflections on construction of assessment models for Buildings' sustainability certification. *Computational Science and Its Applications – ICCSA 2017* (pp. 501–515). Cham: Springer International Publishing.
Greco, S., Ehrgott, M., & Figueira, J. (2016a). *Multiple Criteria Decision Analysis: State of the Art Surveys*. New York: Springer-Verlag.
Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research, 141*, 61–94.
Greco, S., Kadziński, M., & Słowiński, R. (2011). Selection of a representative value function in robust multiple criteria sorting. *Computers & Operations Research, 38*, 1620–1637.
Greco, S., Matarazzo, B., & Słowiński, R. (2001a). Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research, 129*, 1–47.
Greco, S., Matarazzo, B., & Słowiński, R. (2016b). Decision Rule Approach. In S. Greco, M. Ehrgott, & J. R. Figueira (Eds.), *Multiple Criteria Decision Analysis: State of the Art Surveys* (pp. 497–552) (Eds.). New York, NY: Springer New York.
Greco, S., Matarazzo, B., Słowiński, R., & Stefanowski, J. (2001b). Variable consistency model of dominance-based rough sets approach. *Rough sets and current trends in computing* (pp. 170–181). Berlin, Heidelberg: Springer Berlin Heidelberg.
Greco, S., Mousseau, V., & Słowiński, R. (2008). Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions. *European Journal of Operational Research, 191*, 416–436.
Greco, S., Mousseau, V., & Słowiński, R. (2010). Multiple criteria sorting with a set of additive value functions. *European Journal of Operational Research, 207*, 1455–1470.
Guarini, M. R., Battisti, F., & Chiovitti, A. (2018). Public Initiatives of Settlement Transformation: A Theoretical-Methodological Approach to Selecting Tools of Multi-Criteria Decision Analysis. *Buildings, 8*. https://doi.org/10.3390/buildings8010001.
Guitouni, A., & Martel, J. M. (1998). Tentative guidelines to help choosing an appropriate MCDA method. *European Journal of Operational Research, 109*, 501–521.
Haurant, P., Oberti, P., & Muselli, M. (2011). Multicriteria selection aiding related to photovoltaic plants on farming fields on Corsica Island: A real case study using the ELECTRE outranking framework. Energy Policy, *39*, 676–688.
Hinloopen, E., Nijkamp, P., & Rietveld, P. (1983). Qualitative discrete multiple criteria choice models in regional planning. *Regional Science and Urban Economics, 13*, 77–102.
Hwang, C. L., & Yoon, K. (1981). *Multiple Attribute Decision Making: Methods and Applications*. New York: Springer-Verlag.
Ishizaka, A., & Gordon, M. (2017). MACBETHSort: A multiple criteria decision aid procedure for sorting strategic products. *Journal of the Operational Research Society, 68*, 53–61.
Ishizaka, A., & Labib, A. (2011). Review of the main developments in the analytic hierarchy process. *Expert Systems with Applications, 38*, 14336–14345.
Ishizaka, A., & Nemery, P. (2014). Assigning machines to incomparable maintenance strategies with ELECTRE-SORT. Omega, *47*, 45–59.
Ishizaka, A., Pearman, C., & Nemery, P. (2012). AHPSort: An AHP-based method for sorting problems. *International Journal of Production Research, 50*, 4767–4784.
Itsubo, N. (2015). Weighting. In M. Z. Hauschild, & M. A. J. Huijbregts (Eds.), *Life Cycle Impact Assessment* (pp. 301–330) (Eds.). Dordrecht: Springer Netherlands.
Jacquet-Lagreze, E., & Siskos, Y. (1982). Assessing a set of additive utility functions for multicriteria decision making: The UTA method. *European Journal of Operational Research, 10*, 151–164.
Jacquet-Lagrèze, E., & Siskos, Y. (2001). Preference disaggregation: 20 years of MCDA experience. *European Journal of Operational Research, 130*, 233–245.
Jun, D., Tian-tian, F., Yi-sheng, Y., & Yu, M. (2014). Macro-site selection of wind/solar hybrid power station based on ELECTRE-II. *Renewable and Sustainable Energy Reviews, 35*, 194–204.
Kadziński, M., Ciomek, K., & Słowiński, R. (2015). Modeling assignment-based pairwise comparisons within integrated framework for value-driven multiple criteria sorting. *European Journal of Operational Research, 241*, 830–841.
Kadziński, M., Ghaderi, M., & Dąbrowski, M. (2020). Contingent preference disaggregation model for multiple criteria sorting problem. *European Journal of Operational Research, 281*, 369–387.
Kadziński, M., Greco, S., & Słowiński, R. (2012a). Extreme ranking analysis in robust ordinal regression. Omega, *40*, 488–501.
Kadziński, M., Greco, S., & Słowiński, R. (2012b). Selection of a representative value function in robust multiple criteria ranking and choice. *European Journal of Operational Research, 217*, 541–553.
Kadziński, M., Greco, S., & Słowiński, R. (2013). RUTA: A framework for assessing and selecting additive value functions on the basis of rank related requirements. Omega, *41*, 735–751.
Kadziński, M., & Słowiński, R. (2013). DIS-CARD: A new method of multiple criteria sorting to classes with desired cardinality. *Journal of Global Optimization, 56*, 1143–1166.
Keeney, L. R., & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: Wiley.
Keeney, R. L. (1996). Value-focused thinking: Identifying decision opportunities and creating alternatives. *European Journal of Operational Research, 92*, 537–549.
Keeney, R. L., & Gregory, R. S. (2005). Selecting attributes to measure the achievement of objectives. *Operations Research, 53*, 1–11.
Kostkowski, M., & Słowiński, K. (1996). *UTA+ Application (v. 1.20) – User's manual*. Paris: Université de Paris-Dauphine Technical report Documents du LAMSADE 95.
Lahdelma, R., Hokkanen, J., & Salminen, P. (1998). SMAA - Stochastic multiobjective acceptability analysis. *European Journal of Operational Research, 106*, 137–143.
Lahdelma, R., & Salminen, P. (2001). SMAA-2: Stochastic Multicriteria acceptability analysis for group decision making. *Operations Research, 49*, 444–454.
Lahdelma, R., & Salminen, P. (2010). Stochastic Multicriteria Acceptability Analysis (SMAA). In M. Ehrgott, J. R. Figueira, & S. Greco (Eds.), *Trends in Multiple Criteria Decision Analysis* (pp. 285–315) (Eds.). Boston, MA: Springer US.
Langhans, S. D., Reichert, P., & Schuwirth, N. (2014). The method matters: A guide for indicator aggregation in ecological assessments. *Ecological Indicators, 45*, 494–507.

Ley-Borrás, R. (2015). Deciding on the decision situation to analyze: The critical first step of a decision analysis. *Decision Analysis, 12*, 46–58.
Li, Y., Weston, N., & Mavris, D. (2008). An approach for Multi-criteria decision making method selection and development. In *26th international Congress of the Aeronautical Sciences* Accessed on 4 Dec 2018 at:http://www.icas.org/ICAS_ARCHIVE/ICAS2008/PAPERS/597.PDF.
Liou, J. J. H. (2011). Variable consistency dominance-based rough set approach to formulate airline service strategies. *Applied Soft Computing, 11*, 4011–4020.
Liu, J., Kadziński, M., Liao, X., Mao, X., & Wang, Y. (2020). A preference learning framework for multiple criteria sorting with diverse additive value models and valued assignment examples. *European Journal of Operational Research, 286*, 963–985.
Marttunen, M., Belton, V., & Lienert, J. (2018). Are objectives hierarchy related biases observed in practice? A meta-analysis of environmental and energy applications of Multi-Criteria Decision Analysis. *European Journal of Operational Research, 265*, 178–194.
Maxim, A. (2014). Sustainability assessment of electricity generation technologies using weighted multi-criteria decision analysis. Energy Policy, *65*, 284–297.
Meyer, P., & Olteanu, A.-. L. (2019). Handling imprecise and missing evaluations in multi-criteria majority-rule sorting. *Computers & Operations Research, 110*, 135–147.
Moghaddam, N. B., Nasiri, M., & Mousavi, S. M. (2011). An appropriate multiple criteria decision making method for solving electricity planning problems, addressing sustainability issue. *International Journal of Environmental Science & Technology, 8*, 605–620.
Morton, A. (2018). Multiattribute value elicitation. In L. C. Dias, A. Morton, & J. Quigley (Eds.), *Elicitation: The Science and Art of Structuring Judgement* (pp. 287–311) (Eds.). Cham: Springer International Publishing.
Munda, G. (1995). Multicriteria evaluation in a fuzzy environment: The naiade method. In G. Munda (Ed.), *Multicriteria Evaluation in a Fuzzy Environment: Theory and Applications in Ecological Economics* (pp. 131–148) (Ed.). Heidelberg: Physica-Verlag HD.
Munda, G. (2008). The issue of consistency: Basic discrete multi-criteria "Methods". In G. Munda (Ed.), *Social Multi-Criteria Evaluation for a Sustainable Economy* (pp. 85–109) (Ed.). Berlin: Springer Berlin Heidelberg.
Munda, G., & Nardo, M. (2005). *Constructing Consistent Composite Indicators: The Issue of Weights European Commission*. Joint Research Centre EUR 21834 EN.
Munda, G., & Nardo, M. (2009). Noncompensatory/nonlinear composite indicators for ranking countries: A defensible setting. *Applied Economics, 41*, 1513–1523.
Nădăban, S., Dzitac, S., & Dzitac, I. (2016). Fuzzy TOPSIS: A general view. *Procedia Computer Science, 91*, 823–831.
Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2008). *Handbook on Constructing Composite Indicators. Methodology and user Guide*. Paris: OECD.
Nemery, P., & Lamboray, C. (2008). FlowSortt: A flow-based sorting method with limiting or central profiles. *TOP, 16*, 90–113.
Opricovic, S., & Tzeng, G.-. H. (2004). Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research, 156*, 445–455.
Özpeynirci, S., Özpeynirci, Ö., & Mousseau, V. (2018). An interactive algorithm for multiple criteria constrained sorting problem. *Annals of Operations Research, 267*, 447–466.
Paelinck, J. H. P. (1976). Qualitative multiple criteria analysis, environmental protection and multiregional development. *Papers in Regional Science, 36*, 59–76.
Pap, E., Marichal, J.-. L., Grabisch, M., & Mesiar, R. (2009). Properties for aggregation. In E. Pap, J.-L. Marichal, M. Grabisch, & R. Mesiar (Eds.), *Aggregation functions* (pp. 11–55) (Eds.). Cambridge: Cambridge University Press.
Partovi, F. Y., & Burton, J. (1993). Using the analytic hierarchy process for ABC analysis. *International Journal of Operations & Production Management, 13*, 29–44.
Pelissari, R., Oliveira, M. C., Amor, S. B., Kandakoglu, A., & Helleno, A. L. (2020). SMAA methods and their applications: A literature review and future research directions. *Annals of Operations Research, 293*, 433–493.
Pages Perny, P., & Roubens, M. (1998). Fuzzy Preference Modeling. editor. *Fuzzy Sets in Decision Analysis, Operations Research and Statistics*, 3–30.
Pirlot, M. (1995). A characterization of 'min' as a procedure for exploiting valued preference relations and related results. *Journal of Multi-Criteria Decision Analysis, 4*, 37–56.
Podinovski, V. V. (2010). Set choice problems with incomplete information about the preferences of the decision maker. *European Journal of Operational Research, 207*, 371–379.
Polatidis, H., Haralambopoulos, D. A., Munda, G., & Vreeker, R. (2006). Selecting an appropriate multi-criteria decision analysis technique for renewable energy planning. *Energy Sources, Part B: Economics, Planning, and Policy, 1*, 181–193.
Pollesch, N. L., & Dale, V. H. (2016). Normalization in sustainability assessment: Methods and implications. *Ecological Economics, 130*, 195–208.
Punkka, A., & Salo, A. (2013). Preference Programming with incomplete ordinal information. *European Journal of Operational Research, 231*, 141–150.
Qiu, L., Zhu, J., Pan, Y., Hu, W., & Amable, G. S. (2017). Multi-criteria land use suitability analysis for livestock development planning in Hangzhou metropolitan area. *China. Journal of Cleaner Production, 161*, 1011–1019.
Razmak, J., & Aouni, B. (2014). Decision support system and multi-criteria decision aid: A state of the art and perspectives. *Journal of Multi-Criteria Decision Analysis, 22*, 101–117.
Rezaei, J. (2015). Best-worst multi-criteria decision-making method. Omega, *53*, 49–57.
Riabacke, M., Danielson, M., & Ekenberg, L. (2012). State-of-the-Art prescriptive criteria weight elicitation. *Advances in Decision Sciences, 2012*, 24.
Rocha, C., Dias, L. C., & Dimas, I. (2013). Multicriteria classification with unknown categories: A clustering–sorting approach and an application to conflict management. *Journal of Multi-Criteria Decision Analysis, 20*, 13–27.
Rowley, H. V., Peters, G. M., Lundie, S., & Moore, S. J. (2012). Aggregating sustainability indicators: Beyond the weighted sum. *Journal of Environmental Management, 111*, 24–33.
Roy, B. (2016b). Paradigms and Challenges. In S. Greco, M. Ehrgott, & J. R. Figueira (Eds.), *Multiple Criteria Decision Analysis: State of the art Surveys* (pp. 19–39) (Eds.). New York, NY: Springer New York.
Roy, B. (1990). Decision-aid and decision-making. *European Journal of Operational Research, 45*, 324–331.
Roy, B. (1991). The outranking approach and the foundations of electre methods. *Theory and Decision*, 49–73.
Roy, B. (2016a). Do not confuse. *International Journal of Multicriteria Decision Making, 6*, 112–117.
Roy, B., & Słowiński, R. (2008). Handling effects of reinforced preference and counter-veto in credibility of outranking. *European Journal of Operational Research, 188*, 185–190.
Roy, B., & Słowiński, R. (2013). Questions guiding the choice of a multicriteria decision aiding method. *EURO Journal on Decision Processes, 1*, 69–97.
Saaty, T. L. (1980). *The analytic hierarchy process*. New York: McGraw-Hill.
Saaty, T. L. (2016). The analytic hierarchy and analytic network processes for the measurement of intangible criteria and for decision-making. In S. Greco, M. Ehrgott, & J. R. Figueira (Eds.), *Multiple criteria decision analysis: State of the art surveys* (pp. 363–419) (Eds.). New York, NY: Springer New York.
Sage, A. P., & White, C. C. (1984). ARIADNE: A knowledge-based interactive system for planning and decision support, *IEEE Transactions on Systems, Man, and Cybernetics SMC* (14, pp. 35–47).
Salinesi, C., & Kornyshova, E. (2006). Choosing a prioritization method – Case of IS security improvement. In *The 18th Conference on Advanced Information Systems Engineering (CAiSE '06)*.
Salo, A. A., & Hämäläinen, R. P. (1992). Preference assessment by imprecise ratio statements. *Operations Research, 40*, 1053–1061.
Salo, A. A., & Hamalainen, R. P. (2001). Preference ratios in multiattribute evaluation (PRIME)-elicitation and decision procedures under incomplete information. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 31*, 533–545.
Salo, A., & Hämäläinen, R. P. (1997). On the measurement of preferences in the analytic hierarchy process. *Journal of Multi-Criteria Decision Analysis, 6*, 309–319.
Salo, A., & Punkka, A. (2005). Rank inclusion in criteria hierarchies. *European Journal of Operational Research, 163*, 338–356.
Siebert, J., & Keeney, R. L. (2015). Creating more and better alternatives for decisions using objectives. *Operations Research, 63*, 1144–1158.
Siskos, Y., Grigoroudis, E., & Matsatsinis, N. F. (2016). UTA Methods. In S. Greco, M. Ehrgott, & J. R. Figueira (Eds.), *Multiple Criteria Decision Analysis: State of the Art Surveys* (pp. 315–362) (Eds.). New York, NY: Springer New York.
Słowiński, R., Greco, S., & Matarazzo, B. (2002). Axiomatization of utility, outranking and decision preference models for multiple-criteria classification problems under partial inconsistency with the dominance principle. *Control and Cybernetics, 31*, 1005–1035.
Słowiński, R., Greco, S., & Matarazzo, B. (2009). Rough sets in decision making. In A. R. Meyers (Ed.), *Encyclopedia of Complexity and Systems Science* (pp. 7753–7787) (Ed.). New York, NY: Springer New York.
Smet, Y. D. (2013). P2CLUST: An extension of PROMETHEE II for multicriteria ordered clustering. In *2013 IEEE International Conference on Industrial Engineering and Engineering Management* (pp. 848–851).
Sparrevik, M., Barton, D. N., Bates, M. E., & Linkov, I. (2012). Use of stochastic multi-criteria decision analysis to support sustainable management of contaminated sediments. *Environmental Science & Technology, 46*, 1326–1334.
Tsoukiàs, A. (2007). On the concept of decision aiding process: An operational perspective. *Annals of Operations Research, 154*, 3–27.
Van den Honert, R. C., & Lootsma, F. A. (2000). Assessing the quality of negotiated proposals using the REMBRANDT system. *European Journal of Operational Research, 120*, 162–173.
Vetschera, R. (2017). Deriving rankings from incomplete preference information: A comparison of different approaches. *European Journal of Operational Research, 258*, 244–253.
von Winterfeldt, D., & Edwards, W. (1986). *Decision Analysis and Behavioral Research*. London: Cambridge University Press.
Voogd, H. (1982). Multicriteria evaluation with mixed qualitative and quantitative data. *Environment and Planning B: Planning and Design, 9*, 221–236.
Wątróbski, J., Jankowski, J., Ziemba, P., Karczmarczyk, A., & Zioło, M. (2019). Generalised framework for multi-criteria method selection. Omega, *86*, 107–124.
Wedley, W. C. (2010). Issues in aggregating AHP/ANP Scales. *Advances in Intelligent Decision Technologies* (pp. 29–42). Berlin, Heidelberg: Springer Berlin Heidelberg.
Weistroffer, H. R., & Li, Y. (2016). Multiple criteria decision analysis software. In S. Greco, M. Ehrgott, & J. R. Figueira (Eds.), *Multiple criteria decision analysis: State of the Art Surveys* (pp. 1301–1341) (Eds.). New York, NY: Springer New York.

Publication [P8]

Jurczak, M., Miebs, G., and Bachorz, R. A. (2022). Multi-criteria human resources planning optimisation using genetic algorithms enhanced with MCDA. In Operations Research and Decisions (Vol. 32, Issue 4). Politechnika Wroclawska Oficyna Wydawnicza., DOI: 10.37190/ord220404.

Contribution of co-authors (excluding the author and the supervisor of this dissertation):

- Marcin Jurczak:
  - Co-authorship of the idea underlying the paper
  - Consultation of the context of case study and interpretation of its results
  - Co-authorship of the text of the publication
- Rafał Bachorz:
  - Consultation of the algorithm and results analysis
  - Corrections to the first draft of the manuscript

Vol. 32, No. 4 (2022) | DOI: 10.37190/ord220404

OPEN ACCESS

Operations Research and Decisions

www.ord.pwr.edu.pl

OPERATIONS RESEARCH AND DECISIONS QUARTERLY

Multi-criteria human resources planning optimisation using genetic algorithms enhanced with MCDA

Marcin Jurczak[1] Grzegorz Miebs[2,3] Rafał A. Bachorz[2,4]

[1]*Department of Logistics, Poznań University of Economics and Business, Poznań, Poland*

[2]*Advanced Analytics Team, PSI Poland, Poznań, Poland*

[3]*Institute of Computing Science, Faculty of Computing, Poznan University of Technology, Poznań, Poland*

[4]*Institute of Medical Biology of Polish Academy of Sciences, Łódź, Poland*

[*]*Corresponding author: marcin.jurczak@ue.poznan.pl*

Abstract

The main objective of this paper is to present an example of the IT system implementation with advanced mathematical optimisation for job scheduling. The proposed genetic procedure leads to the Pareto front, and the application of the multiple criteria decision aiding (MCDA) approach allows extraction of the final solution. Definition of the key performance indicator (KPI), reflecting relevant features of the solutions, and the efficiency of the genetic procedure provide the Pareto front comprising the representative set of feasible solutions. The application of chosen MCDA, namely elimination et choix traduisant la réalité (ELECTRE) method, allows for the elicitation of the decision maker (DM) preferences and subsequently leads to the final solution. This solution fulfils all of the DM expectations and constitutes the best trade-off between considered KPIs. The proposed method is an efficient combination of genetic optimisation and the MCDA method.

Keywords: *mathematical optimisation, multi-criteria optimisation, scheduling, job-shop problem, MCDA*

1. Introduction

Employee scheduling has been a common problem in the literature since the 1950s [51]. This can be a crucial process due to at least two reasons: high labour costs, which can be reduced by proper scheduling, and a labour shortage in which loss of profit can be minimised by optimising human resources planning [62].

Beaker [3] proposed three classes of personnel scheduling problems: shift scheduling, day-off scheduling, and rotating scheduling. Shift scheduling is the simplest one where the daily planning horizon is considered with either overlapping or non-overlapping shifts. In day-off scheduling employee's workweek is of a different length than the operation week. The most common example is the five days workweek

Received 17 December 2021, accepted 12 December 2022, published online 8 February 2023
ISSN 2391-6060 (Online)/

The costs of publishing this issue have been cofinansed by the program *Development of Academic Journals* of the Polish Ministry of Education and Science under agreement No. RCN/SP/0241/2021/1

and two free days with an operational week, altogether lasting for seven days. Rotating scheduling is a combination of both previous models. A company works seven days a week and each day consists of more than one shift. Employees do not have a fixed schedule, thus workweek can be of different lengths but the required daily and weekly breaks have to be satisfied.

Different types of methods can be used to solve a personnel scheduling problem. Yakoob et al. [1] proposed a classification with ten classes: (1) manual solution, (2) integer programming, (3) implicit modelling, (4) decomposition, (5) goal programming, (6) working set generation, (7) linear programming-based solution, (8) construction/improvement, (9) metaheuristics, and (10) other methods.

Personnel scheduling is computationally complex, in a general case, NP-complex problem. Due to its non-polynomial complexity, only small instances can be treated with systematic approaches like linear programming or mixed-integer programming [9]. For larger problems, which are more common in real-world applications, heuristics that introduce a trade-off between the time of computations and the quality of results have to be used [53].

Real-world personnel scheduling is a multi-objective problem where criteria such as length of the schedule, utilisation of resources, the satisfaction of people's preferences, and compliance with regulations have to be considered [59]. To address the multi-criteria nature of this problem optimisation algorithms designed for a multi-objective goal function were used [45], including the multi-objective genetic algorithms [14]. Proposed here non-dominated sorting genetic algorithm II (NSGAII) methodology is broadly applied for scheduling problems [4, 38, 44] as well as in many other areas [2, 17, 36, 57, 64, 67]. The method is designed for efficient handling of the multi-objective optimisation problems providing high-quality, uniformly distributed approximated Pareto frontier.

Within the current study, we propose a novel hybrid approach which combines the NSGAII method with the ELECTRE approach. This combined methodology provides a tool which is capable of determining a Pareto-optimal solution fulfilling the current quality expectations. For an in-depth review of the applications of the NSGAII framework in the area of scheduling see [53]. The problem addressed in the presented research is a standard and frequently encountered business scheduling problem. Within this business case, the task of scheduling algorithms is to find the best possible match between transportation tasks and workers. In operations research, this type of problem is called the job-shop problem. From a formal point of view, it means that there is a finite number of jobs, a set denoted by J, and a finite number of resources, a set denoted by M. The mathematical goal is to find the best solution being the best match between the task and resources. It is an NP-complex problem, its complexity is non-polynomial and grows very fast with the number of tasks and the number of resources. Thus, it is necessary to design and use efficient heuristics to find approximate, possibly good solutions.

The remaining of this paper is organised as follows: Section 2 introduces the general business context of the problem, Section 3 provides the discussion about the business rules and constraints as well as the considered here KPIs are defined there. In Section 4 the NSGAII and ELECTRE are thoroughly discussed with an emphasis on the synergy between these two. In Section 5, the attention of the reader is focused on the numerical experiments and the results and finally, in Section 6, the conclusions are formulated.

2. Business context of labour planning

Resource planning remains one of the fundamental problems in economic science. Planning is a specific element of corporate decision-making that relates to the future of an organisation. It is one of the four key elements of the management process. Generally, it is continuous and is also a reflection of the changes taking place in the organisation [61]. Planning is a strategic function and produces results in the long term [65]. Planning allows one to properly approach the implemented activities and to determine its priorities or to maximally use the available competencies [43]. Resource planning is often equated with work planning in manufacturing (production plant). Therefore, much of the literature on job-shop refers to the optimisation of production planning. Some authors, especially those interested in computer scheduling, refer also to these machines as processors because each activity means an operation for a machine [23]. In turn, analysing the literature on multi-criteria scheduling, it can be seen that de facto individual job execution scenarios imply the execution of different job parameters – and the value of each parameter is related to the job and the specific scenario. This makes scenario analysis a complex problem [28]. The choice of mathematical methods also remains an interesting problem, addressed by researchers both directly in relation to urban transport and more broadly, for example in the context of resource planning in supply chains [27]. In essence, the problems of the two mentioned areas are similar. Resource planning and task scheduling also remain an interesting issue wherever these processes are carried out in a decentralised manner – there too, the potential for supporting these processes with appropriate software is seen today [26]. More and more often various simulation and machine learning techniques are also being used to support planning in manufacturing processes and beyond [58]. At the same time, planning on defined metrics allows for being more tailored to the specific business [66].

3. The economical efficiency of labour planning

The discussed case of scheduling tasks is a real business problem for a public transport operator providing public transport services with the use of buses and trams. Miejskie Przedsiębiorstwo Komunikacyjne w Poznaniu (Municipal Public Transport Company) operates a fleet of approximately 600 vehicles. Clearly, it faces the necessity of planning the work of approx. 800 bus drivers and 600 tram drivers. Due to the continuous development of the department responsible for task scheduling and increasingly complex scheduling issues, a strategic decision was made to implement a new IT solution for the job scheduling of these 1400 employees. The mathematical algorithms are part of an IT system implemented for this purpose. The implemented IT system was related to the creation and development of comprehensive tools which allow maximisation of the efficiency of working time management: implementation of a system for job planning, supervision of job planning implementation and clearance of workers' jobs.

In the context of this issue, the basic data set are transport tasks (shifts). They define specific tasks to be realised by specific vehicles – separated vehicles on each bus line or tram line. Each shift characterise the parameters: start time, end time, duration, driving time or rolling stock type, and others.

On the other hand, within the second set of data, there is a group of all available employees. Each of them has certain, individual characteristics like type of contract, nominal working hours, holiday volume or specific days that can be planned for work or days off.

The idea of scheduling is to pair shifts and workers. This has to be a 1:1 relationship – for obvious reasons a vacant vehicle would not leave the depot, with few exceptions the presence of two drivers in one vehicle would not make sense either. In practice, transport tasks also have features that act as business constraints – for example in terms of the planned type of rolling stock. Employees also have defined qualifications. So if, for example, a specific job defines a type of rolling stock, it is necessary at the planning stage to take into account the qualification for that type of rolling stock that the specific employee is equipped with. This clearly increases the complexity of the planning problem, especially in tram operations where the number of rolling stock types can reach a dozen.

In the discussed enterprise, much attention is paid to the issue of work planning, among others due to the difficult labour market. The continuous shortage of bus and tram drivers makes planning work one of the most difficult tasks in operational management. This is an effect of a phenomenon called employee market, i.e., permanent shortage of employees supply. This phenomenon, in relation to professional drivers, is observed practically in all larger Polish cities.

One of the main challenges at the implementation stage of the optimisation algorithms was to take into account all the requirements: different criteria and constraints. The resource planning business process itself has four stages. The layout of the stages is shown in Figure 1. It should be noted that this process has the nature of feedback – the stage of implementation of the carriage is also a source of data for subsequent stages of the planning process.

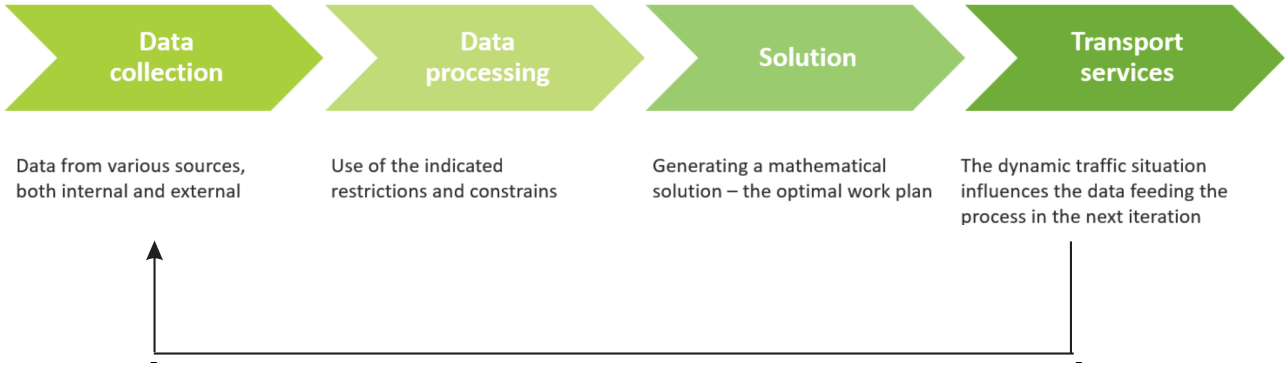


Figure 1. Resource planning business process

From the organisation's point of view, all these stages focus on the successive collection and entering information into the system about:

- previously known absences of employees (e.g., long-term absences),
- restrictions on employees (e.g., their availability schedules),
- assignments to employee groups (in case of group planning or over a planning period longer than one month),
- availability on specific dates (e.g., for employment/ dismissal).

This data is entered manually using dedicated views, and in some cases imported from external systems (e.g., ERP systems).

Three levels of data are used in the planning process (as well as in the entire planning information system): the monthly plan, the updated monthly plan (as a daily plan, called "shifts list") and the execution (the a posteriori information about the executed schedule). Planning with the optimisation engine is

done at the first level, but taking into account the plan update and execution data as far as it is possible and data are available.

3.1. Business rules and constraints

The essential task of mathematical algorithms is to create a monthly work schedule. This schedule must take into account basic business-like constraints. Planning constraints can be divided into several basic groups and according to different criteria, namely:

- mandatory and optional restrictions (preferences, business rules),
- external and internal constraints,
- restrictions concerning the employee, working hours and days or holidays.

The primary pieces of legislation outlining work planning requirements are:

- labour Code,
- drivers' Hours of Work Act,
- internal regulations (work regulations, company collective agreement and others),
- provisions specifying the requirements for medical, psychological and other health and vocational qualifications.

In the case of bus drivers and tram drivers, the so-called equivalent working time is usually used as the basic working time system. According to art. 129. §1 of the Labour Code Act, "working time cannot exceed 8 hours per day and an average of 40 hours in an average five-day working week". This provision is complemented by article 135. §1 of the Act: "if it is justified by the type of work or its organisation, the system of equivalent working time may be used, in which it is permissible to extend the daily working time, but not more than to 12 hours, in the settlement period not exceeding 1 month. The extended daily working time is balanced by a shorter daily working time on certain days or by days off". These provisions allow for more flexible scheduling wherever it is necessary to fill jobs in a two- or three-shift system or on holidays.

In practice, this means that bus and tram drivers, in accordance with the regulations of the Labour Code, can perform shifts up to 12 hours on selected days of the month (not necessarily only from Monday to Friday, but also on Saturdays, Sundays and holidays). The limit for the number of working days in a month is therefore 8 hours multiplied by the number of working days, which usually means a range of 152–184 hours for a full-time employee. Of course, these are just a few of the many rules, resulting not only from the Labour Code but also from other legal regulations. Additional restrictions (for example up to 10 hours of working time) result from the Act about drivers working time. In order to stay according to the law, the planning algorithms have taken into account the limitations of both Acts.

Selected constraint items from the implemented algorithms are presented in Table 1. Due to the complexity of the planning problem, authors have limited themselves to presenting the key constraints for planning processes. However, it should be noted that there are several dozen of these constraints in the system and all of them are taken into account by the planning algorithms. This is the reason why the task-employee matching process is the most time-demanding step of the algorithm.

Table 1 contains both formal and legal restrictions, resulting from existing legal regulations and those developed as good practices over the years of the company's operation. Both of these groups are inter-

esting. In practice, the second optional group may also be considered obligated, when they have a key and clear influence on the quality of planning or subsequent execution of transport tasks.

Table 1. Selected limitations of work planning

Limitation	Type	Source
Daily rest (11h), weekly rest (24/35h)	obligatory	legislation; act about drivers working time
Maximum working time limits (10 hours per day, 60 hours per week)	obligatory	legislation; act about drivers working time
Number of days off not less than the calendar	obligatory	legislation; labour code
Validity of periodic and psychological tests	obligatory	law
Validity of authorisations for a specific type of rolling stock	obligatory	internal regulations
Staff availability preferences (hours from/to, selected days)	optional	information from employees

To discuss an example, it will be a value to look at the parameter concerning the maximum number of working activities day by day. Although the law allows for planning cycles of up to six days, due to the nature of the work performed (shift work, high level of stress, changing climate conditions) it is desirable to eliminate the sixth day of work. This allows the reduction of the engagement of employees, and the distribution of the workload more evenly over the month. On the other hand, scheduling of single days of work (a day off, a day of work and again a day off), although also not prohibited by law, is eliminated in order to increase the comfort of employees. It also allows for a better grouping of days off, and, statistically, employees more often receive a sequence of two days off (which corresponds to a typical weekend). This constraint is an example of both obligatory constraint (1–6 days of continuous work) and optional constraint (2–5 days of work on consecutive days). To reconcile these two perspectives, the types of constraints take the form of configurable parameters in the system in most cases.

Some restrictions are relative in nature – they depend on other restrictions and the specific tasks performed by the employee. For example, there is a concept of a working day which limits the working day for an employee. For employees which work on the 1st and 2nd shifts, a day is contained in the time period between 3.00 AM and 3.00 AM, and for those who perform also night duties – between 8.00 AM and 8.00 AM. Since the second group is on duty only during the 2nd and 3rd shifts (without the 1st, which means morning shifts), the next task of optimisation algorithms is to appropriately balance task scheduling between workers on the 3rd shift. Planning night tasks (due to the rigid group of employees handling them) has a higher priority. The remaining tasks must be balanced – completing the nominal hours of night workers and between the remaining workers.

Proper distribution of free days (days off) remains a separate, important planning topic. An employee is entitled to have the days off for Sunday (Wn – free for Sunday), for holidays (Ws – free for holiday) and for ensuring a five-day working week (Wd – additional days off for Saturday). Each type of day off may have defined criteria of occurrence, which is an additional limitation for scheduling algorithms. The scheduling algorithm allocates the appropriate amount of time off according to the defined rules. The most important of these rules are:

- every fourth Sunday off (free Sunday may occur more often),
- days off not less than the nominal number of days off (Saturdays, Sundays, holidays) in the given planning period,

- day off for Sunday must be scheduled within six days before or six days after this Sunday,
- holiday and Wd's day off may be scheduled on any day within the planning period (provided it is not a Sunday or a holiday),
- if Sunday and the public holiday fall on a Sunday or public holiday, it may not be taken on another day.

The scheduling algorithm also takes into account all additional activities that are delivered as input data – absences (illnesses, holidays according to the dictionary of types) or training (including defined types of training in specific hours, e.g., OSH training). For each type of activity, there are defined assumptions for later settlement – whether it is a paid or unpaid day, whether it is included in the employee's nominal value or generates overtime etc.

The input data about employees is collected in the "Employee file". This is all the information that refers to individual employees, and from the perspective of scheduling algorithms, the most important is the information that may limit an employee's availability, or otherwise specify limitations or preferences for scheduling.

3.2. Key performance indicators (KPIs)

Algorithms responsible for assigning workers to tasks follow certain optimisation directions based on defined KPIs. Each indicator has a numerical parameter associated with it, called the goal function priority, which reflects to what extent the optimiser should take into account the given goal function. This allows the DM to influence the optimisation process.

A total of over a dozen key indicators were defined to assess the quality of the plan and to match the plan to DM preferences. Key indicators are these, which relate to the main directions of multi-criteria optimisation. These indicators are:

- General task planning – determines the percentage of shifts to which an employee is assigned in relation to all shifts from a given planning period (that means actually: task planning index) in the considered planning horizon. The indicator is applied to morning, afternoon and night shifts (1st, 2nd and 3rd shift).
- Equality of deficiencies – the purpose of the KPI is to determine the uniform distribution of unassigned tasks over a given planning period.
- Home depot indicator – specifies the number of shifts assigned within the home depot to the total number of shifts (working days). This indicator characterises the plan, the optimiser should strive to make this value as high as possible. Additionally, the distribution of assignments from outside the home depot should be equal among the employees.
- Equality of 2nd shift scheduling – assigning second shifts of afternoon services to employees assigned to the night schedule. This is the standard deviation of the number of hours resulting from afternoon shifts assigned to night workers.
- Number of switches between the 1st and 2nd shift – if such transitions are allowed, then one of the criteria for schedule quality is a number of them.

Such indicators are complemented by KPIs of qualitative nature. For instance, a special mechanism was implemented to ensure the implementation of the "diversity of reserves" factor. The diversity of the

length of the generated reserves should be as high as possible, which allows for matching the reserves with the shortfalls in the transport tasks in a given month. The length of the generated reserves is adjusted to the current transport tasks.

Among the KPIs there are also informative indicators which form the basis for plan evaluation for the planner responsible for creating work schedules dedicated to a group of employees (from a given depot or traction). These include indicators related to planning execution (measured at the stage of subsequent plan implementation), service list variability (number of reserves per employee planned and executed), the average length of service (average duration of service divided into shifts and types of days per depot) or the length of non-issued shifts (calculated according to defined time intervals).

3.3. Financial aspects

Scheduling optimisation's main aim, among other issues, is maximising the efficient use of labour. In public transport, shifts for drivers and co-drivers have different lengths, so it is usually impossible to plan the entire monthly hourly quota down to the minute. As a result, some activities are not desirable from the employer's point of view, but necessary to fill the employee's calendar nomination. They are called "plus reserves" (so-called, because they are an addition to the employee's shifts), and in fact, they are lost working time. The employee is formally at work, but usually, it is a short time, which does not allow any use of the employee's availability. Minimising plus reserves is an important planning goal.

A second important objective of planning is to minimise the level of overtime. Proper scheduling – as balanced as possible and with maximum utilisation of the available working time – makes it possible to reduce the amount of overtime – which, in turn, represents financial value for the company. The better the scheduling, the less overtime is worked. Of course, to a certain level, due to the insufficient number of employees, overtime naturally occurs anyway. In practice, the algorithm responsible for scheduling is therefore designed to create a plan that minimises additional costs (e.g., overtime). Obviously, in the case of a shortage of workers, it is impossible to ensure full staffing for the implemented tasks – the task of the algorithms is then to maximise this staffing, as well as to ensure the possibility of employing workers in additional time.

4. Methods

A single-objective optimisation (SOO) is a set of procedures routinely applied in different areas of industry and science. As the name says, the SOO methods focus on problems where only one goal function is considered, and the decision vector is a subject of optimisation according to this goal function. A variety of algorithms solving that task were proposed, e.g., simulated annealing [41], local search, Tabu search [29], or genetic algorithm [30]. On the other hand, the multi-objective optimisation (MOO) problems, are designed to tackle qualitatively different problems, where multiple objective functions exist and must be respected simultaneously. Here, instead of finding the optimum with respect to a single goal, one needs to cope with multiple objectives, often conflicting with each other. The interactions between goals result in multiple solutions for a particular problem, usually called trade-offs, non-dominated, non-inferior or Pareto-optimal solutions. The multi-objective optimisation problems are successfully addressed by many approaches [40], one of them is the family of multi-objective genetic al-

gorithm (MOGA), which was proposed as a generalization of the single-objective genetic algorithm [48]. There are multiple variants of MOGA methods such as strength Pareto evolutionary algorithm (SPEA and SPEAII) [68], Pareto archived evolution strategy (PAES) [42], Pareto envelope-based selection algorithm (PESA, PESAII) [13], Niched Pareto genetic algorithm [35] and many others [21]. The presented research focuses on the application of the NSGAII [16] in a scheduling problem.

4.1. The NSGAII method

The NSGAII method, similar to the majority of the genetic algorithms (GA) approaches, has the usual structure involving population creation, selection and genetic operations. In particular, the key elements of the NSGAII methodology can be sketched in the following way:

1. Create an initial population.
2. Carry out the non-domination sort.
3. Calculate the crowding distance.
4. Select a new population using crowded tournament selection, where solutions are compared based on front ranking and in case of a tie by a crowding distance.
5. Apply genetic operators, i.e., create offspring.
6. If optimisation is not finished, return to Step 2.

Due to the fact that the crowding distance is taken into account explicitly, the NSGAII method tends to return an evenly distributed Pareto front. In principle, this is a desirable feature, because it assures that the solution space is spanned over a relatively large range of objective values, thus considered population covers diverse cases. This leads to a high-quality, representative, Pareto front, which ultimately delivers the final solution of desired properties, i.e., where the trade-offs between goals are at the expected level.

At the general level of consideration, the main optimisation task is, for a defined set of workers (M) and a set of tasks (J), to create a schedule by assigning a single worker m_i to a particular task j_i. Due to obvious reasons, if the availability of workers is not sufficient, some tasks remain unassigned. This problem was proven to be NP-hard [24], thus to provide a solution in a feasible time, a heuristic approach has to be used. In the presented approach the NSGAII procedure was chosen. Each chromosome, i.e., the member of the population, is represented as a presorted vector of workers (M). Such a vector processed within a deterministic algorithm leads unequivocally to an assignment of employees to tasks. The whole procedure is presented in Algorithm 1. At this level of generality, this procedure seems to be very simple. In practice, however, the crucial point is the procedure responsible for the worker-task assignment. It is a complex set of operations producing feasible solutions. Checking if m matches j, i.e., the verification of all the business constraints, requires a thorough check of all the conditions defined in Section 3.1. All the constraints defined either by Polish law or being an internal rule of the company must be explicitly verified here.

These steps are executed to transform a chromosome into a feasible solution within the NSGAII algorithm. The solution space exploration is reduced here to the exploration of potential workers' presorted vectors. This presorted vector of workers should be understood as the optimisation decision vector, fully determining the resulting schedule. The decision vectors of this form, being the chromosomes in the GA language, are the subject of all genetic operations like selection, cross-over or mutation. The resulting population is then the subject of the quality estimation by means of the KPIs defined earlier (see

Section 3.2). The assignment of the tasks to the workers can be considered as the mathematical transformation of the decision vector into the space of the KPIs. The chosen MOO method, i.e., the NSGAII approach, takes care of non-domination. This means that the genetic population approximates the Pareto frontier and each next generation of the genetic procedure improves the Pareto frontier towards the exact solution. In the end, the algorithms return the final, the best, Pareto frontier of the solutions. The final step, discussed in the next Section, is to extract a single solution which reflects the best current preferences.

Input: Array of workers $M = [m_1, \ldots, m_n]$, Array of tasks $J = [j_1, \ldots, j_m]$
Create vector M
for $j \in J$ **do**
  for $m \in M$ **do**
    if m matches j **then**
      Assign m to j
      proceed to the next j
    end
  end
end

Algorithm 1. Procedure of matching tasks with workers

4.2. The application of the MCDA to the extraction of final solution

In contrast to SOO, which returns a single best solution, MOO provides a set of non-dominated solutions called a Pareto frontier. Without any additional preference information from a DM, these solutions are incomparable and represent a trade-off between metrics. The MCDA algorithms are designed in such a way that they support the DM with preference elicitation and ultimately lead to the extraction of a single solution with expected properties [11]. Within these approaches, firstly, preference information reflecting the value system of the DM is collected. Then it is applied to the dataset to provide recommendations. Some of the most popular MCDA algorithms are analytic hierarchy process (AHP) [56], dominance-based rough set approach (DRSA) [34], and family of ELECTRE methods [55]. Such algorithms are widely applied in use cases from different areas including energy [12, 25], finance [49, 60], military [15, 19, 37] or urban development [50, 52]. In particular AHP method has been applied to many real-world problems [32, 54]. However, due to DM's preferences, this method is not suitable for this use case. The preferences of the DM were as follows.

- Poor performance on one criterion cannot be fully compensated by good performance on others.
- Algorithm should directly use per-criterion pairwise comparison thresholds, in particular indifference and veto thresholds.
- Criteria expressed on different quantitative scales must be accepted

There is a variety of the MCDA approaches available, here in order to select an appropriate one we have applied the methods selection system [10]. Among them the ELECTRE [22, 55] method seems to be the most suitable for this problem while being also widely applied to real-world problems [20, 31]. The methods selection system suggested a number of methods satisfying the DM's preferences as well as meeting the problem description, e.g., PROMETHEE [7], TACTIC [63], RUBIS [6], and ELECTRE

[22, 55]. The latter was found to be preferred by the DM due to multiple real-world applications [5, 31, 46] and a variety of extensions of that approach [33, 39].

4.2.1. The ELECTRE method

The family of ELECTRE methods are based on an outranking relation S, which can be interpreted as equal or better. If one alternative outranks the other, it means that it is at least as good, based on the DM's value system, and there are no significant reasons to refute this relation.

In what follows the following notation was used:

- $A = \{a_1, a_2, \ldots, a_n\}$ – a set of decision alternatives (schedules coming from NSGAII).
- $G = \{g_1, g_2 \ldots, g_m\}$ – a family of evaluation criteria.
- $g_j(a_i)$ – the performance of alternative a_i with respect to criterion g_j. For clarity of the presentation in what follows we assume that all criteria are of gain type meaning the higher value the better.
- q_j, p_j, v_j – values of indifference, preference, and veto thresholds on criterion g_j.
- w_j – weight of criterion g_j.

4.2.2. Preference information

The DM provides two types of preference information: weights associated with each criterion and threshold values. Weights w_j represent the strength of a given criterion and should be rather associated with the number of votes than a numeric weight itself. In ELECTRE methods three thresholds are used: indifference q_j, preference p_j, and veto v_j. The first two are called intracriteria, and the last one is intercriteria. The intracriteria thresholds impact only evaluation on that criterion, while intercriteria affect general evaluation. The indifference threshold indicates a maximal difference on a given criterion that is negligible. It can help reduce noise impact when dealing with imperfect knowledge [18]. The preference threshold is a minimal difference denoting strict preference, whereas the veto threshold represents minimal difference which is so significant that it invalidates preference relation.

4.2.3. The model

In this section, we present a variant of ELECTRE that was used for the selection problem. The method constructs a matrix with the credibility of outranking relations for each pair of alternatives. That matrix is later exploited with a net flow score (NFS) procedure to calculate a score for each alternative. Finally, an alternative with the highest score is recommended.

To calculate the credibility of outranking relation, for a given ordered pair of alternatives (a, b), the following procedure is applied:

1. For each criterion g_j, calculate the marginal concordance function $c_j(a,\, b)$ presenting the strength of an outranking b on g_j. Value of $c_j(a,\, b)$ depending on $g_j(a)$ and $g_j(b)$ is presented in Figure 2 and mathematically can be expressed as:

$$c_j(a,\, b) = \begin{cases} 1 & g_j(a) - g_j(b) \geq -q_j \\ 0 & g_j(a) - g_j(b) < -p_j \\ \dfrac{g_j(a) - g_j(b) + p_j}{p_j - q_j} & \text{otherwise} \end{cases} \tag{1}$$

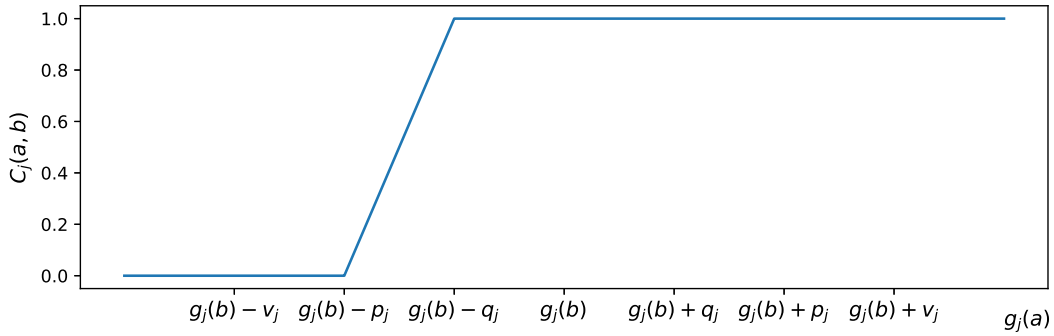


Figure 2. Dependence of the marginal concordance function on the difference between two alternatives on one criterion

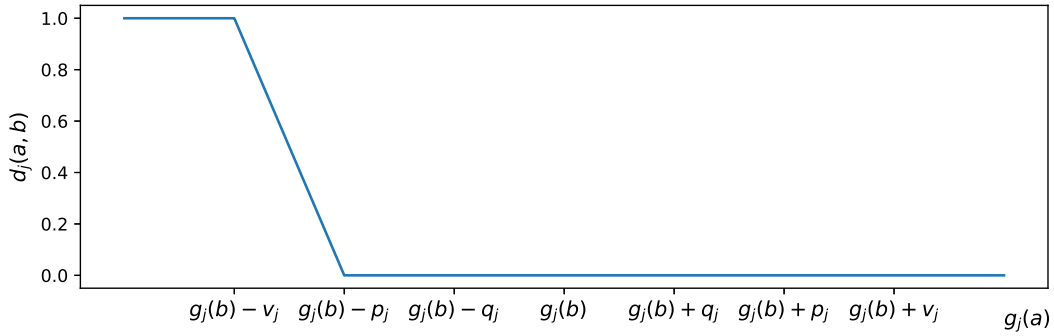


Figure 3. Dependence of marginal discordance function on the difference between two alternatives on one criterion

2. Calculate the comprehensive concordance index $C(a,\, b)$ which denotes the strength of an outranking b on all criteria

$$C(a,\, b) = \frac{\sum_{j=1}^{m} w_j c_j(a,\, b)}{\sum_{j=1}^{m} w_j} \tag{2}$$

3. For each criterion g_j, calculate the marginal discordance function $d_j(a,\, b)$ presenting the strength of negation of an outranking b on g_j. Value of $d_j(a,\, b)$ depending on $g_j(a)$ and $g_j(b)$ is presented in Figure 3 and mathematically can be expressed as:

$$d_j(a,\, b) = \begin{cases} 1 & g_j(b) - g_j(a) \geq v_j \\ 0 & g_j(b) - g_j(a) < p_j \\ \dfrac{g_j(b) - g_j(a) - p_j}{v_j - p_j} & \text{otherwise} \end{cases} \tag{3}$$

4. Calculate the credibility of an outranking relation defines as follows [47]:

$$\sigma(a,b) = C(a,b)\left(1 - \left[\max_{0<j\leq m} d_j(a,b)\right]\right) \tag{4}$$

The procedure is applied for each pair of alternatives resulting in matrix M, where $M[i,j] = \sigma(a_i, a_j)$. This matrix is then exploited using the NFS [8] procedure to calculate a score (s_i) for each alternative, which is defined as follows:

$$s_i = \varphi_i^+ - \varphi_i^- = \sum_{j=0}^{n} M\left[i,j\right] - \sum_{j=0}^{n} M\left[j,i\right] \tag{5}$$

Finally, the alternatives are ranked according to the score s_i, and the one with the highest score is returned as the final recommendation. Such an approach is more robust to the imperfect perspective of the DM. This is in opposition to approaches like ELECTRE 1s, where a subset of solutions can be selected [31], and still, the DM is obliged to make the final choice when only one alternative must be selected.

5. The results of optimisation

This section presents a discussion of the results obtained within a particular application of the optimisation approach to real data. The set of tasks contained 9003 items, and the optimisation problem was to assign them to 573 workers/employees. In the first phase, the NSGAII method was applied to obtain the Pareto front of the solutions. As already mentioned, without the additional preference information from the DM, it is impossible to point out the final solution. Therefore, within the second phase, the preference information is incorporated and utilised within the MCDA method.

The exemplary set of solutions is presented in Figure 4. The blue dots represent the feasible solutions; the orange dots reflect 14 particular solutions that form the Pareto front. The performance of these solutions is shown in Table 2.

Table 2. Pareto front of generated solutions[1]

Name	KPI1↑	KPI2↑	KPI3↑	KPI4↓	KPI5↑	KPI6↓	KPI7↓	KPI8↓
a_1	0.8683	0.8726	0.8000	74.7052	0.6296	5.5364	24.8440	0
a_2	0.8690	0.8724	0.8000	74.6285	0.6298	5.5369	24.8440	0
a_3	0.8692	0.8702	0.8000	81.8977	0.6286	5.5729	25.9767	0
a_4	0.8697	0.8699	0.7905	77.7294	0.6234	5.4849	27.8632	1
a_5	0.8713	0.8695	0.7905	67.6030	0.6218	5.4981	26.6789	0
a_6	0.8724	0.8690	0.7905	68.5985	0.6190	5.4990	28.1957	1
a_7	0.8729	0.8677	0.7905	71.5745	0.6273	5.4997	25.4752	1
a_8	0.8756	0.8672	0.7905	79.5618	0.6234	5.5155	27.8992	1
a_9	0.8770	0.8666	0.7905	63.8953	0.6281	5.5250	25.2300	1
a_{10}	0.8790	0.8652	0.7905	66.1173	0.6288	5.5110	25.6365	0
a_{11}	0.8792	0.8598	0.7905	70.3695	0.6271	5.4850	25.4279	1
a_{12}	0.8797	0.8594	0.8000	68.6371	0.6293	5.5134	25.5040	0
a_{13}	0.8822	0.8587	0.7905	68.4240	0.6222	5.4758	25.1981	0
a_{14}	0.8824	0.8542	0.7810	80.6420	0.6292	5.5265	25.4271	1

[1] Gain-type criteria are denoted by ↑, while cost-type criteria are denoted by ↓.

To illustrate the calculation phase, each step of the calculation of the credibility of the outranking relation between solutions a_{13} and a_7 is discussed in detail. The preference information used for this study is presented in Table 3. The alternative a_{13} is equal to or better than a_7 on KPIs 1, 3, 4, 6, 7, 8, thus the marginal concordance functions in these criteria are equal to one $c_j(a_{13}, a_7) = 1, j \in \{1, 3, 4, 6, 7, 8\}$.

Table 3. Indifference, preference, and veto thresholds along with the weights used for this study

Type	KPI1↑	KPI2↑	KPI3↑	KPI4↓	KPI5↑	KPI6↓	KPI7↓	KPI8↓
Indifference	0.005	0.005	0.01	1	0.002	0.01	0.5	2
Preference	0.02	0.02	0.02	3	0.005	0.03	1	3
Veto	0.05	0.05	0.05	10	0.02	0.1	5	10
Weights	5	5	3	2	2	3	2	1

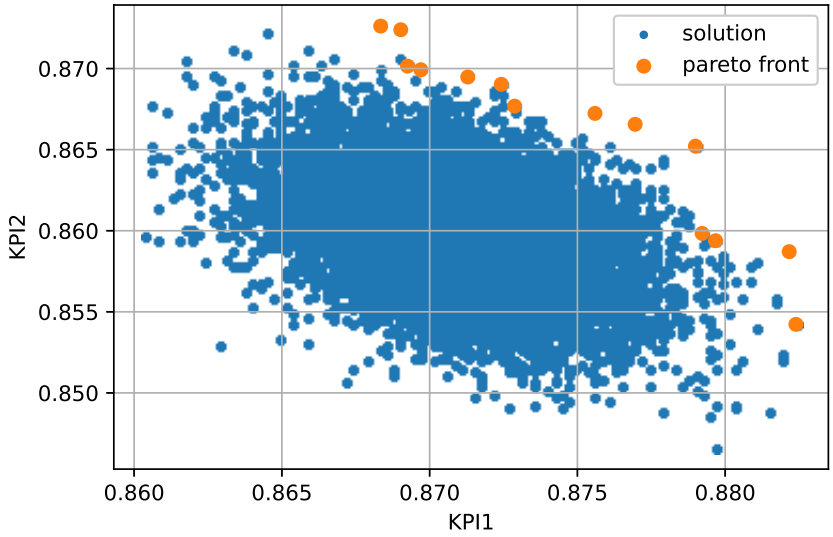


Figure 4. The distribution of the performance of the generated solutions on the first two KPIs with the selection of the Pareto front

On KPI2 a_{13} is worse than a_7 by more than the indifference threshold, however, less than the preference threshold; thus, the marginal concordance function shall be calculated as follows:

$$c_2(a_{13}, a_7) = \frac{(0.8587 - 0.8677 + 0.02)}{(0.02 - 0.005)} = \frac{0.011}{0.15} = 0.7(3) \tag{6}$$

On KPI5, a_{13} is worse than a_7 by more than the preference threshold; therefore, the marginal concordance function is equal to zero. The marginal discordance function shall be calculated in the following way.

$$d_5(a_{13}, a_7) = \frac{0.6273 - 0.6222 - 0.005}{0.02 - 0.005} = \frac{1}{150} = 0.00(6) \tag{7}$$

Finally, one can calculate the comprehensive concordance index as

$$C(a_{13}, a_7) = \frac{(1 \times 5 + 0.7(3) \times 5 + 1 \times 3 + 1 \times 2 + 0 \times 2 + 1 \times 3 + 1 \times 2 + 1 \times 1)}{(5 + 5 + 3 + 2 + 2 + 3 + 2 + 1)} \approx 0.855 \tag{8}$$

After including the marginal discordance function, one can obtain the credibility of an outranking relation for that pair of alternatives.

$$\sigma(a_{13}, a_7) = 0.855\,(1 - 0.00(6)) = 0.8493 \tag{9}$$

After applying this procedure to each ordered pair of alternatives, a matrix that contained the credibilities of an outranking relation was created. Then, within the NFS procedure, the final score can be calculated individually for each alternative. The obtained scores are presented in Table 4. Alternative a_9

attained the highest score, thus it was recommended as the best schedule generated with the genetic approach. Two main factors have decided to recommend the solution a_9. First, it is placed in the middle of the Pareto front with respect to the two most important criteria, and hence, when compared with other alternatives in most comparisons, it is not worse by more than the indifference threshold. It results in a high number of outranked alternatives which causes a relatively high value of a positive flow. Secondly, the alternative has an outstanding performance on criterion g_4. The difference in evaluations on this criterion between a_9 and other alternatives is nearly always higher than the preference threshold and sometimes even above the veto threshold. This fact impacts the value of the marginal discordance function which later invalidates the outranking relation over a_9, thus the negative flow of this alternative is the lowest. Relatively high positive flow with the lowest negative flow results in the highest overall score.

Table 4. Final scores for Pareto front

Name	φ_i^+	φ_i^-	s_i
a_1	8.23	9.57	–1.35
a_2	8.28	9.55	–1.27
a_3	2.73	11.92	–9.2
a_4	4.78	10.8	–6.02
a_5	10.78	7.75	3.03
a_6	7.832	8.52	–0.69
a_7	11.99	9.01	2.98
a_8	3.97	12.95	–8.98
a_9	12.73	5.0	7.73
a_{10}	13.33	6.38	6.95
a_{11}	12.36	7.83	4.52
a_{12}	12.34	7.65	4.69
a_{13}	11.43	6.71	4.72
a_{14}	3.89	11.02	–7.13

6. Conclusions

The main aim of this publication was to provide an example of the implementation of an IT system for human work planning using advanced mathematical algorithms. The business case is based on data collected from different sources and levels of detail, which are used together to create work plans. The introduction of a mathematical optimisation tool has helped to improve the quality of job scheduling.

Firstly, by implementing sophisticated mathematical algorithms, it was possible to expand the list of criteria and constraints that are taken into account during planning. Until now, due to the mathematical complexity of the problem, this was not always possible. In addition, the introduction of KPIs made it possible to adapt planning criteria to individual needs – thanks to the preferences of a DM, it is now possible to plan based on the prioritisation of selected goal functions.

The result of the optimisation mechanisms is the extracted Pareto front. In the course of genetic optimisation, the Pareto front is constantly improved, but at some point, the algorithm finishes due to the fulfilment of certain conditions. The Pareto front can be expressed as a table with the non-dominated alternatives. Each of the alternatives is expressed as a vector of KPIs which reflect the business value. Thus, a quite complex solution to the task assignment problem is expressed here in a very elegant and concise way. The proper definition of the KPI space allows the extraction of the essential information,

which forms the basis for the subsequent application of the MCDA method. On one hand, we cope with the multi-criteria optimisation that the solution is a set of non-dominated alternatives. The NSGAII method takes care of the quality of the Pareto front, especially in the context of the uniform distribution of the solutions along the KPIs directions. On the other hand, the carefully chosen MCDA method is applied to the Pareto front to select a single solution, because, in the end, this is relevant from the practical perspective.

The article explains the basic principles of the planning process, the selected mathematical methods and tools, and the reasons for their choice. The proposed here workflow is capable of efficient treatment of the incoming planning data, turning them into suitable forms for further processing and ultimately applying the proposed here methods. The end-user can introduce the preferences in a very straightforward way and obtains a single solution that can be the subject of manual correction, or can be directly launched into the production environment. The appropriate definition of the KPIs opens the possibility of the business-oriented interpretation of the results, and thus builds trust and confidence. The presented approach provides the necessary in business applications consistency and efficiency. It combines the well-established and commonly accepted multi-objective genetic approach with the application of the MCDA method for the extraction of the final solution.

References

[1] Al-Yakoob, S. M., and Sherali, H. D. Multiple shift scheduling of hierarchical workforce with multiple work centers. *Informatica 18*, 3 (2007), 325–342.

[2] Atiquzzaman, M., Liong, S.-Y., and Yu, X. Alternative decision making in water distribution network with NSGA-II. *Journal of Water Resources Planning and Management 132*, 2 2006, 122–126.

[3] Baker, K. R. Workforce allocation in cyclical scheduling problems: A survey. *Operational Research Quarterly (1970-1977) 27*, 1 (1976), 155–167.

[4] Bandyopadhyay, S. Modified NSGA-II for a bi-objective job sequencing problem. *Intelligent Information Management 4*, 6 (2012), 319–329.

[5] Bari, F., and Leung, V. Application of ELECTRE to network selection in a hetereogeneous wireless network environment. In *2007 IEEE Wireless Communications and Networking Conference*, IEEE, (2007), pp. 3810-3815.

[6] Bisdorff, R., Meyer, P., and Roubens, M. R UBIS: a bipolar-valued outranking method for the choice problem. *4OR 6*, 2 (2007), 143–165.

[7] Brans, J.-P., and Mareschal, B. Promethee methods. In Multiple Criteria Decision Analysis: State of the Art Surveys, J. Figueira, S. Greco and M. Ehrogott, Eds., vol. 78 of *International Series in Operations Research & Management Science*. Springer-Verlag, New York, NY, 2005, pp. 163–186.

[8] Brans, J. P., and Vincke, P. Note—A preference ranking organisation method. (The PROMETHEE Method for Multiple Criteria Decision-Making). *Management Science 31*, 6 (1985), 647–656.

[9] Brucker, P., Qu, R., and Burke, E. Personnel scheduling: Models and complexity. *European Journal of Operational Research 210*, 3 (2011), 467–473.

[10] Cinelli, M., Kadziński, M., Miebs, G., Gonzalez, M., and Słowiński, R. Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system. *European Journal of Operational Research 302*, 2 (2022), 633–651.

[11] Cinelli, M., Kadziński, M., Gonzalez, M., and Słowiński, R. How to support the application of multiple criteria decision analysis? Let us start with a comprehensive taxonomy. *Omega 96*, (2020), 102261.

[12] Cinelli, M., Spada, M., Kadziński, M., Miebs, G., and Burgherr, P. Advancing hazard assessment of energy accidents in the natural gas sector with rough set theory and decision rules. *Energies 12*, 21 (2019), 4178.

[13] Corne, D. W., Knowles, J. D., and Oates, M. J. The Pareto envelope-based selection algorithm for multiobjective optimization. In *Parallel problem solving from nature PPSN VI*, M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J. J. Merelo, and H.-P. Schwefel, Eds., vol. 1917 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2000, pp. 839–848.

[14] Cowling, P., Colledge, N., Dahal, K., and Remde, S. The trade off between diversity and quality for multi-objective workforce scheduling. In *Evolutionary Computation in Combinatorial Optimization*, J. Gottlieb, G. R. Raidl, Eds., vol. 3906 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2006, pp. 13–24.

[15] de Almeida, I. D. P., de Pina Corriça, J. V., de Araújo Costa, A. P., de Araújo Costa, I. P., do Nascimento Maêda, S. M., Gomes, C. F. S., and dos Santos, M. Study of the location of a second fleet for the Brazilian navy:

Structuring and mathematical modeling using SAPEVO-M and VIKOR methods. In *Production Research*, D. A. Rossit, F. Tohmé, G. M. Delgadillo, Eds., vol. 1408 of *Communications in Computer and Information Science* Springer International Publishing, Cham, 2021, pp. 113–124.

[16] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation 6*, 2 (2002), 182–197.

[17] Deb, K., Rao N. U. B., and Karthik, S. Dynamic multi-objective optimization and decision-making using modified NSGA-II: A case study on hydro-thermal power scheduling. In *Evolutionary Multi-Criterion Optimization, 4th International Conference, EMO 2007, Matsushima, Japan, March 5-8, 2007*, S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, T. Murata, Eds., vol. 4403 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2007, pp. 803–817.

[18] Dias, L. C., Antunes, C. H., Dantas, G., de Castro, N., and Zamboni, L. A multi-criteria approach to sort and rank policies based on Delphi qualitative assessments and ELECTRE TRI: The case of smart grids in Brazil. *Omega 76* (2018), 100–111.

[19] dos Santos, M., de Araújo Costa, I. P., and Gomes, C. F. S. Multicriteria decision-making in the selection of warships: A new approach to the AHP method, *International Journal of the Analytic Hierarchy Process 13*, 1 (2021), 147–169.

[20] Drumond, P., Basílio, M. P., de Araújo Costa, I. P., de Moura Pereira, D. A., Gomes, C. F. S., and dos Santos, M. Multicriteria analysis in additive manufacturing: An ELECTRE-MOr based approach. In *Modern management based on big data II and machine learning and intelligent systems III*, A. J. Tallón-Ballesteros, Ed., IOS Press: Amsterdam, The Netherlands, 2021, pp. 126–132.

[21] Durillo, J. J., and Nebro, A. J. jMetal: A Java framework for multi-objective optimization. *Advances in Engineering Software 42*, 10 (2011), 760–771.

[22] Figueira, J. R., Greco, S., Roy, B., and Słowiński, R. An overview of ELECTRE methods and their recent extensions. *Journal of Multi-Criteria Decision Analysis 20*, 1-2 (2013), 61–85.

[23] French, S. *Sequencing and scheduling. An introduction to the mathematics on the job-shop*. Ellis Horwood Ltd, 1981.

[24] Garey, M. R., Johnson, D. S., and Sethi, R. The complexity of flowshop and jobshop scheduling. *Mathematics of Operations Research 1*, 2 (1976), 117–129.

[25] Gasser, P., Suter, J., Cinelli, M., Spada, M., Burgherr, P., Hirschberg, S., Kadziński, M., and Stojadinović, B. Comprehensive resilience assessment of electricity supply security for 140 countries. *Ecological Indicators 110*, (2020), 105731.

[26] Gezer, V., and Wagner, A. Real-time edge framework (RTEF): task scheduling and realisation. *Journal of Intelligent Manufacturing 32*, (2021), 2301–2317.

[27] Gharaei, A., and Jolai, F. A Pareto approach for the multi-factory supply chain scheduling and distribution problem. *Operational Research 21*, (2021), 2333–2364.

[28] Gilenson, M., Shabtay, D., Yedidsion, L., and Rohit, M. Scheduling in multi-scenario environment with an agreeable condition on job processing times. *Annals of Operations Research 307* (2021), 153–173.

[29] Glover, F. Tabu search—part I. *ORSA Journal on Computing 1*, 3 (1989), 190–206.

[30] Goldberg, D. E., and Holland, J. H. Genetic algorithms and machine learning. *Machine Learning 3* (1988), 95–99.

[31] Govindan, K., Kadziński, M., Ehling, R., and Miebs, G. Selection of a sustainable third-party reverse logistics provider based on the robustness analysis of an outranking graph kernel conducted with ELECTRE I and SMAA. *Omega 85* (2019), 1–15.

[32] Goyal, S., Luthra, S., and Garg, D. Shifting systematically towards sustainable consumption and production: A solution framework to overcome the impacts of Covid-19. *International Journal of Information Technology & Decision Making 21*, 03 (2022), 933–968.

[33] Greco, S., Kadziński, M., Mousseau, V., and Słowiński, R. ELECTRE$^{\mathrm{GKMS}}$: Robust ordinal regression for outranking methods. *European Journal of Operational Research 214*, 1 (2011), 118–135.

[34] Greco, S., Matarazzo, B., and Słowiński, R. Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research 129*, 1 (2001), 1–47.

[35] Horn, J., Nafpliotis, N., and Goldberg, D. E. A niched Pareto genetic algorithm for multiobjective optimization. In *Proceedings of the First IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence*, 1994, pp. 82–87.

[36] Huang, B., Buckley, B., and Kechadi, T.-M. Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications. *Expert Systems with Applications 37*, 5 (2010), 3638–3646.

[37] Jardim, R. R. J., Santos, M., Neto, E., da Silva, E., and de Barros, F. Integration of the waterfall model with ISO/IEC/IEEE 29148:2018 for the development of military defense system. *IEEE Latin America Transactions 18*, 12 (2020), 2096–2103.

[38] Jiang, X., and Li, Y. Improved NSGA-II for the job-shop multi-objective scheduling problem. *International Journal of Performability Engineering 14*, 5 (2018), 891–898 .

[39] Kadziński, M., Tervonen, T., and Figueira, J. R. Robust multi-criteria sorting with the outranking preference model and characteristic profiles. *Omega 55*, (2015), 126–140.

[40] Kadziński, M., Tervonen, T., Tomczyk, M. K., and Dekker, R. Evaluation of multi-objective optimization approaches for solving green supply chain design problems. *Omega 68*, (2017), 168–184.

[41] Kirkpatrick, S., Gelatt Jr., C. D., and Vecchi, M. P. Optimization by simulated annealing. *Science 220*, 4598 (1983), 671–680.

[42] Knowles, J. D., and Corne, D. W. Approximating the nondominated front using the Pareto archived evolution strategy. *Evolutionary Computation 8*, 2 (2000), 149–172.

[43] Krzakiewicz, K., Ed., *Theoretical foundations of organization and management.*. Wydawnictwo Akademii Ekonomicznej w Poznaniu, 2006 (in Polish).

[44] Li, C. P., Cui, H. Y., and Wang, G. C. The optimization of flexible job-shop scheduling problem based on NSGA-II. *Advanced Materials Research 651*, (2013), 684–687.

[45] Li, J., Burke, E. K., Curtois, T., Petrovic, S., and Qu, R. The falling tide algorithm: A new multi-objective approach for complex workforce scheduling. *Omega 40*, 3 (2012), 283–293.

[46] Liu, Z., Wang, W., Wang, D., and Liu, P. A modified ELECTRE II method with double attitude parameters based on linguistic z-number and its application for third-party reverse logistics provider selection. *Applied Intelligence 52*, 13 (2022), 14964–14987.

[47] Mousseau, V., and Dias, L. Valued outranking relations in ELECTRE providing manageable disaggregation procedures. *European Journal of Operational Research 156*, 2 (2004), 467–482.

[48] Murata, T., and Ishibuchi, H. Moga: multi-objective genetic algorithms. In *Proceedings of 1995 IEEE International Conference on Evolutionary Computation, 1995, IEEE, Perth, WA, Australia*, vol. 1, Piscataway, NJ, USA, IEEE, 1995, pp. 289–294.

[49] Oliveira, M. D. N. T., Ferreira, F. A. F., Ilander, G. O. P.-B., and Jalali, M. S. Integrating cognitive mapping and MCDA for bankruptcy prediction in small- and medium-sized enterprises. *Journal of the Operational Research Society 68*, 9 (2017), 985–997.

[50] Oppio, A., Dell'Ovo, M., Torrieri, F., Miebs, G., and Kadziński, M. Understanding the drivers of Urban Development Agreements with the rough set approach and robust decision rules. *Land Use Policy 96* (2020), 104678.

[51] Ott, L. E. *Labor utilization in independent Indiana supermarkets*. Ph. D thesis, Purdue University, 1959.

[52] Radziszewska-Zielina, E., Adamkiewicz, D., Szewczyk, B., and Kania, O. Decision-making support for housing projects in post-industrial areas. *Sustainability 14*, 6 (2022), 3573.

[53] Rahimi, I., Gandomi, A. H., Deb, K., Chen, F., and Nikoo, M. R. Scheduling by NSGA-II: Review and bibliometric analysis. *Processes 10*, 1 (2022), 98.

[54] Reisi, M., Afzali, A., and Aye, L. Applications of analytical hierarchy process (AHP) and analytical network process (ANP) for industrial site selections in Isfahan, Iran. *Environmental Earth Sciences 77*, 14 (2018), 537.

[55] Roy, B. Classement et choix en présence de points de vue multiples. *R. I. R. O. 2*, 8 (1968), 57–75.

[56] Saaty, T. L. What is the analytic hierarchy process? In *Mathematical Models for Decision Support*, Mitra G., Ed., Springer, Berlin, Heidelberg, 1988, pp. 109–121.

[57] Sadeghi, J., Sadeghi, S., and Niaki, S. T. A. A hybrid vendor managed inventory and redundancy allocation optimization problem in supply chain management: An NSGA-II with tuned parameters. *Computers & Operations Research 41* (2014), 53–64.

[58] Sakr, A. H., Aboelhassan, A., Yacout, S., and Bassetto, S. Simulation and deep reinforcement learning for adaptive dispatching in semiconductor manufacturing systems. *Journal of Intelligent Manufacturing* (2021), 1–13.

[59] Silva, J. D. L., Burke, E. K., and Petrovic, S. An introduction to multiobjective metaheuristics for scheduling and timetabling. In *Metaheuristics for Multiobjective Optimisation*. X. Gandibleux, M. Sevaux, K. Sörensen, V. T'kindt, Eds., vol. 535 of *Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin, Heidelberg, 2004, pp. 91–129.

[60] Spronk, J., Steuer, R. E., and Zopounidis, C. Multicriteria decision aid/analysis in finance. In *Multiple Criteria Decision Analysis*. S. Greco, M. Ehrgott, J. Figueira, Eds., vol. 233 of *International Series in Operations Research & Management Science*, Springer, New York, 2016, pp. 1011–1065.

[61] Stoner, J. A. F., Freeman, R. E., and Gilbert jr., D. R. *Management*. Polskie Wydawnictwo Ekonomiczne, 1997 (in Polish).

[62] Van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E., and De Boeck, L. Personnel scheduling: A literature review. *European Journal of Operational Research 226*, 3 (2013), 367–385.

[63] Vansnick, J.-C. On the problem of weights in multiple criteria decision making (the noncompensatory approach). *European Journal of Operational Research 24*, 2 (1986), 288–294.

[64] Wang, S., Zhao, D., Yuan, J., Li, H., and Gao, Y. Application of NSGA-II algorithm for fault diagnosis in power system. *Electric Power Systems Research 175* (2019), 105893.

[65] Waters, D. *Operations management: Producing goods and services*. Wydawnictwo Naukowe PWN, 2001 (in Polish).

[66] West, S., Gaiardelli, P., and Saccani, N. Overcoming the barriers to service excellence. In *Modern Industrial Services*. Book series: *Springer Texts in Business and Economics*. Springer, Cham, 2022, pp. 19–174.

[67] Xu, W., Xu, J., He, D., and Tan, K. C. A combined differential evolution and NSGA-II approach for parametric optimization of a cancer immunotherapy model. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (2017), IEEE, Honolulu, HI, USA, pp. 1–8.

[68] Zitzler, E., and Lothar, T. *An evolutionary algorithm for multiobjective optimization: the strength Pareto approach*. vol. 43 of TIK-report, ETH Zurich, Computer Engineering and Networks Laboratory, 1998.

Publication [P9]

Miebs, G., Wójcik, M., Karaszewski, A., Mochol-Grzelak, M., Wawdysz, P., and Bachorz, R. A. (2022). Predicting a Time-Dependent Quantity Using Recursive Generative Query Network. In International Journal of Neural Systems (Vol. 32, Issue 11). World Scientific Pub Co Pte Ltd, DOI: 10.1142/S0129065722500563.

Contribution of co-authors (excluding the author and the supervisor of this dissertation):

- Michał Wójcik:
  - Consultation of the neural network architecture and interpretation of its results
  - Corrections to the first draft of the manuscript
- Adam Karaszewski:
  - Consultation of the context of case study and interpretation of its results
  - Corrections to the first draft of the manuscript
- Małgorzata Mochol-Grzelak:
  - Consultation of the neural network architecture and interpretation of its results
  - Co-authorship of the text of the publication
- Paulina Wawdysz:
  - Consultation of the neural network architecture and interpretation of its results
  - Corrections to the first draft of the manuscript
- Rafał Bachorz:
  - Consultation of the neural network architecture and interpretation of its results
  - Consultation of the context of case study and interpretation of its results
  - Co-authorship of the text of the publication

International Journal of Neural Systems, Vol. 32, No. 11 (2022) 2250056 (15 pages)

DOI: 10.1142/S0129065722500563

World Scientific
www.worldscientific.com

Predicting a Time-Dependent Quantity Using Recursive Generative Query Network

Grzegorz Miebs[*,†], Michał Wójcik[*,†], Adam Karaszewski[*],
Małgorzata Mochol-Grzelak[*], Paulina Wawdysz[*]
and Rafał A. Bachorz[*,‡,§]
[*]*PSI Poland, Advanced Analytics Team*
Towarowa 37, 61-896 Poznań, Poland
[†]*Institute of Computing Science*
Poznań University of Technology, Piotrowo 2
60-965 Poznań, Poland
[‡]*Institute of Medical Biology*
Polish Academy of Sciences
Lodowa 103, 93-232 Łódź, Poland
[§]*rbachorz@psi.pl; rafal@bachorz.eu*

Accepted 26 August 2022
Published Online 29 October 2022

We propose here a novel neural architecture dedicated to the prediction of time series. It can be considered as an adaptation of the idea of Generative Query Network (GQN) to the data which is of a sequence nature. The new approach, dubbed here as the Recursive Generative Query Network (RGQN), allows for efficient prediction of time series. The predictor information (i.e. the independent variable) is one or more of the other time series which are in some relationship with the predicted sequence. Each time series is accompanied by additional meta-information reflecting its selected properties. This meta-information, together with the standard dynamic component, is provided simultaneously in Recurrent Neural Network (RNN). During the inference phase, meta-information becomes a query reflecting the expected properties of the predicted time series. The proposed idea is illustrated with use cases of strong practical relevance. In particular, we discuss the example of an industrial pipeline that transports liquid media. The trained RGQN model is applied to predict pressure signals, assuming that the training was carried out during routine operational conditions. The subsequent comparison of the prediction with the actual data gathered under extraordinary circumstances, e.g. during the leakage, leads to a specific residual distribution of the prediction. This information can be applied directly within the data-driven Leak Detection and Location framework. The RGQN approach can be applied not only to pressure time series but also in many other use cases where the quantity of sequence nature is accompanied by a meta-descriptor.

Keywords: Time-series; neural networks; generative model.

[§]Corresponding author.

1. Introduction

Time series reflect numerous real-world phenomena such as traffic,[1] temperature, wind strength and direction,[2,3] disease spread,[4] energy consumption,[5] hippocampal rhythm modulation[6] and pipeline pressure. They are often augmented with additional static metadata describing the time series, e.g. location and road type in the case of traffic, country size and wealth in disease spread modeling, location along the pipeline and transported medium in the case of pressure time series or time. Generally, despite their continuous nature, measurements are only available at a few specific, discrete locations for fundamental reasons such as sensor placement limitations or simply economic considerations. For these circumstances, metadata can be used to interpolate between data points and generate time series associated with any given description.[5]

Time series generation and forecasting are well-known concepts with many approaches and applications. Markov chain is one of the best examples, widely used for modeling natural processes,[7] for example, synthetically generating wind speed time series,[3] forecasting electricity production for a grid-connected photovoltaic system,[8] EEG analysis,[9] EMG analysis,[10] health monitoring,[11,12] vibration control,[13,14] traffic engineering[15–17] or generating synthetic time series of irradiance.[18] In addition, there are applications of neural networks to multivariate time series,[19] and many other scientific efforts involving sophisticated neural architectures.[20–28] An excellent review on these topics can be found in Refs. 29 and 30. It is worth mentioning that the approach proposed here is not the standard time series forecasting. The result is indeed a time series, but the predictors are other time series that are in some relation to the predicted one. The later use cases presented involve the hydraulic behavior of the pipeline, but, in general, any relation could be reflected in both the static and dynamic components of the data. This makes a comparison of our method with standard time series prediction approaches rather difficult.

The introduction of neural network models to solve the generation problem by proposing Generative Adversarial Networks (GAN)[31] was a noticeable step forward. This approach is widely used in image generation tasks.[24,32,33] The basic concept has evolved into different variants suitable for specific problems, such as Progressive GAN,[34] Conditional GAN,[35] Wasserstein GAN[36] and CycleGAN.[37] GANs have also been applied to generate time series.[38] Yoon[39] proposed a time series GAN suitable for learning the temporal dynamics of time series and reproducing it in the generated ones.

Another approach suitable for generating images considering static metadata is called Generative Query Network GQN[40] where the model learns a representation of the scene based on different camera positions. In the prediction phase, the model is able to generate an appropriate image representing the scene from a given camera orientation that was not provided directly in the training. We took inspiration from the GQN approach and adapted it for the generation of time series. Within the proposed framework the graphical scenes are replaced with time series and the camera position is substituted with the corresponding meta-information associated with the given time series. Technically, to incorporate the information contained in the time series, we used the method proposed by us where both dynamic and static features are captured by Recurrent Neural Network RNN.[5] Such an approach is especially desired in situations when some of the measurement points are temporarily not available, so their outputs have to be recovered based on values from other sensors. The productive use case is analyzed, where pressure measurements are predicted for common scenarios of fluid pipeline operations. However, it is worth noting that the presented framework is more general and in principle can be applied to many other real-live use cases. Moreover, it is particularly useful for phenomena propagating both in time and space domains.

The main focus of this paper is the application of presented method in the context of Leak Detection System (LDS), especially for oil and gas pipelines. Regarding preventive measures for mitigating pipeline accidents and resulting leaks, the well-known, established results from industrial research[41,42] are already implemented both in norms and in practice. In assessing the importance of the measures, the requirement for LDS immediately follows the enforcement of strict regulations, implementation of safety and maintenance programs, and most importantly, extensive cyclic training of pipeline technical operators.[43] The deployment of LDS,

utilizing adequate pipeline monitoring equipment capable of supporting leak detection functions, can provide early warnings for the development of seepage, immediate rupture alarms, and also includes the capability to locate leaks. These significantly limit the pipeline downtime and most importantly, the accumulation of environmental damage due to prolonged leak. Ultimately, the well-performing LDS allows for the effective dispatch of maintenance crews for repair and cleanup.

Ideally, suppliers of LDS should determine exactly how their systems will perform on a considered pipeline configuration before the installation. In practice, however, performance is highly dependent on characteristics of the pipeline medium such as compressibility, thermal expansion, density and viscosity. Other important factors are pipeline parameters such as diameter, length, elevation profile, pipeline flexibility and roughness, and variables measured by process instruments such as flow, temperature or pressure. There are many methods and approaches that can be used to detect pipeline leaks and ruptures. For some of them, the critical information about the pipeline integrity is brought with the pressure signals, in particular, this is the case of the Negative Pressure Wave method.[44–46] There are also other approaches in which these signals are of secondary importance or simply accompany many other quantities, e.g. various mass-volume methods.[47,48] However, in all these cases, the pressure signal plays an important role and from a general perspective, can be considered as a primary measurement.

The gold standards of computational LDS, explicit hydraulic-based simulations of Real Time Transient Model (RTTM) and complementary Negative Pressure Wave (NPW), are improved first with statistical approaches, then with data-driven approaches, including neural networks.[49] These were utilized to interpret running output, perform input parameters tuning, and even provide independent implicit models that share common voting in ensemble architectures. Earlier surveys, including one for gas LDS,[50] focused on the differentiation between pure hardware-based solutions and established computational explicit methods or on their fusions. Today, hardware solutions incorporate advanced software, and the distinction is drawn between explicit models and data-driven approaches or hybrids, see Refs. 41, 51 and 52 for oil, LNG and water use cases, respectively. The roles change, for example, as in the[53] stack of data-driven methodologies, both inputs are processed and inherent knowledge accumulates, while the explainable decision rules on top are used to interpret the results. Traditional methods continue to evolve, such as analysis of pressure points and statistical balance using the Kantorovitch distance[54] or adaptation of time difference NPW to acoustic sensor data by cross-correlating in domains of time or frequency.[55] Unlike Deeppipe[56] with its GAN use of simulation as an additional input, this paper proposes a generative approach that completely replaces the generative hydraulic calculations of RTTM with RGQN for the pressure time series. As an application example, the residues of RGQN generated predicted measurements and sensor data are used to enhance input to implicit attenuation-based NPW leak location, as opposed to the explicit attenuation NPW approaches.[57]

The main goals of the presented research are: an adaptation of the GQN framework to data of sequence nature, and providing the elaborated use-case of LDS relevance. Within this use case, the main responsibility of the RGQN architecture is to generate time series related to a certain position on a pipeline based on an arbitrary number of other time series positioned at different locations.

The remainder of this paper is organized as follows. Section 2 introduces the novel approach in comparison with the traditional GQN. Section 3 describes the results of a case study on pressures in the industrial pipeline, as well as some details related to the computational environment. Section 4 concludes the paper and indicates potential avenues for future research.

2. Methods

In this section, we introduce a variant of GQN, which we dubbed RGQN, adapted for time series. Furthermore, the novel approach takes advantage of proper techniques to incorporate static metadata during the processing of time series by initiating the hidden states of the recurrent layers. In what follows, we discuss the original GQN (see Sec. 2.1), and then point out the adaptation to the nature of time series, which is the key point of this novel approach (see Sec. 2.2).

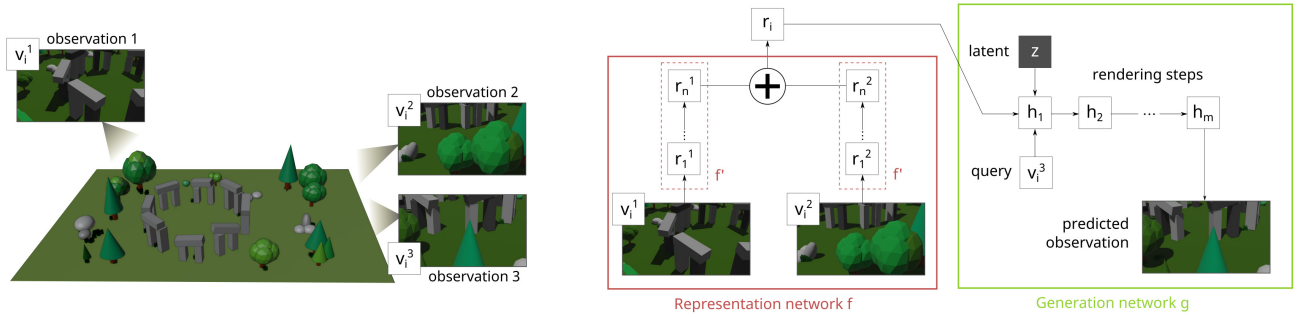


Fig. 1. Schematic diagram of the original generative query network framework.

2.1. *Generative Query Network*

GQN is a deep learning framework suitable for generating an image, representing a scene view associated with the provided metadata, from images of the same scene seen from other views. In other words, the GQN model learns to create a given scene by exploring its surroundings and taking different views of it. In the following paragraphs, we introduce the notation and discuss the main steps to create the GQN framework. An in-depth discussion of this framework can be found in Eslami's paper *et al.*,[40] here we just outline the idea along with the main equations that we will refer to later.

2.1.1. *Notation and data*

The data fed into the representation network consist of a set (o_i) of observations (o_i^k)

$$o_i = \{o_i^k\}_{k=1\ldots K}. \tag{1}$$

Each observation includes K images $(\mathbf{x}_i^k)$ and their associated parameterization $(\mathbf{v}_i^k)$

$$o_i^k = (\mathbf{x}_i^k, \mathbf{v}_i^k). \tag{2}$$

These observations are transformed into the latent representation $\mathbf{r}_i$, which is the final result of the representation network denoted here as f:

$$\mathbf{r}_i = f(o_i) = \sum_{k=1}^{K} f'((\mathbf{x}_i^k, \mathbf{v}_i^k)), \tag{3}$$

where f' is the transformation of single observation, see Figs. 1 and 2 for the pictorial representation of f'. Finally, both latent representation $\mathbf{r}_i$ together with a query $\mathbf{v}_i^{K+1}$ are used to generate an image

$$\mathbf{x}_i^{K+1} = g(\mathbf{r}_i, \mathbf{v}_i^{K+1}). \tag{4}$$

Exemplary workflow is presented in Fig. 1, where two images are used to create latent representation $\mathbf{r}_i$, which is then combined with the metadata $\mathbf{v}_i^3$ corresponding to the generated image.

2.1.2. *Architecture*

The GQN model consists of two neural networks: (1) a representation network f and (2) a generation network g. The former is used to convey information from input images (denoted as vectors $\mathbf{x}_i^k$) and metadata (denoted as vectors $\mathbf{v}_i^k$) to the latent representation $\mathbf{r}_i$, while the latter takes $\mathbf{r}_i$ along with the query metadata $\mathbf{v}_i^{K+1}$ and generates the output image $\mathbf{x}_i^{K+1}$. In addition, a stochastic variable $\mathbf{z}$ is introduced as an auxiliary input to the network g. It allows for the production of different outputs in the case of uncertainty. Since the main subject of the processing is an image, mostly Convolutional Neural Networks (CNNs) are applied within both architectures.

The representation network takes an image together with its corresponding metadata as input. The former is processed by convolutional layers with stride, thus the size of such a layer's output is smaller than its input. The metadata is first projected onto the matching dimensionality and then attached to the processed image as an additional layer. Each image pair with metadata is processed separately by the representation network, and then the latent variables $\mathbf{r}_i^k$, calculated for each of them are summed into a final latent representation $\mathbf{r}_i$ covering all contributions of the individual scene observations.

The generating network utilizes the output of the representation network together with metadata forming the query to generate the final image.

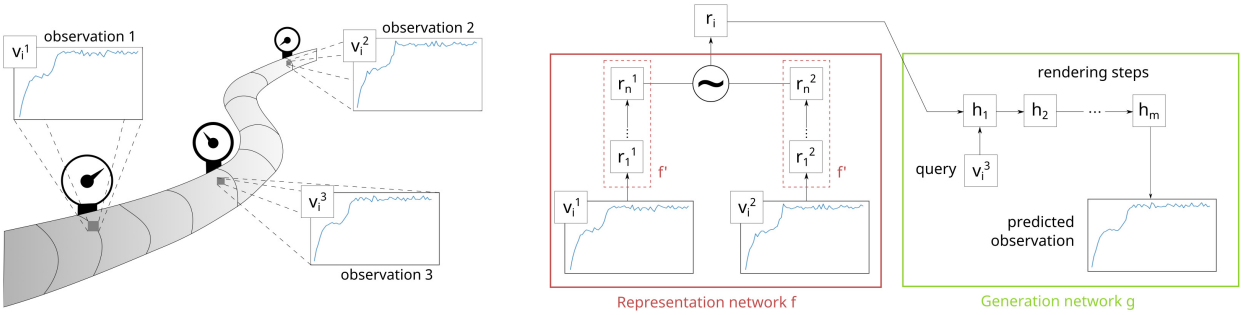


Fig. 2. Schematic diagram of the novel recursive generative query network.

2.2. *Recursive Generative Query Network*

The GQN approach based on convolutional neural networks allows generating a new image from representative images and static meta-data. Analogous to GQN, an example workflow of the novel approach is depicted in Fig. 2. At a general level, the overall network architecture is the same as in the original approach. However, due to the sequential nature of the time series, different neural layers are used. The basic unit of the final model is a recurrent neural network enriched with static metadata, similar to the one presented in Miebs' work *et al.*[5] The first recurrent layer has its states initialized with static metadata processed using dense layers. The recurrent output (all hidden states) is passed as input to the next recurrent layer, while the last states are concatenated with the initial states of the previous layer, processed with a dense layer, and used to initialize states of the next recurrent layer. The LSTM units have both hidden and cell states (contrary to GRU units where only the former one exists). Therefore, besides the dynamic input, the LSTM units are also supplied with the transformed concatenation of current and previous hidden and cell states, resulting in two additional inputs. In standard application of the LSTM it is not necessary, but here this leads to the incorporation of static meta information. The dense layers have the same number of units as the recurrent layers whose initial states they initialize. The architectural diagram of the neural network discussed above is shown in Fig. 3. The RGQN model can be conceptualized as a meta-architecture composed of the sub-architectures being the RNNs with static features incorporated.

Neural models built on this architecture are involved in both the representation and generation parts of the network. The representation component accepts two inputs that together form a single observation o_i (see Eq. (2)): (1) dynamic denoting pressure time series and (2) static metadata describing the location and time related to the dynamic data. The output, or latent representation (see Eq. (3)), consists of time series equal to the number of units in the last Long-Short Term Memory Unit (LSTM) layer. This latent representation is used as a dynamic input to a generation network along with the query consisting of static data describing the location and time for the generation of pressure series (see Eq. (4)). The output of the representation network, i.e. the latent state, can be thought of as a compressed and well-structured representation resulting from any number of incoming observations. This abstract quantity is then passed to the generation network as a feature, and the task of the generation network is to transform the latent state into the expected time series according to a query, which is also passed to the generation network. We have decided not to use the latent stochastic variable $\mathbf{z}$, since the uncertainty is not expected in the considered task. Nevertheless, it can easily be added as an additional input to the generation network g, if needed.

The architecture of the neural network was the subject of optimization. The final model was selected after testing multiple architectures. It consists of the encoder part with 5 layers having 128, 64, 64, 32 units, and 16 for the latent representation, respectively. The decoder part also consists of 5 layers with 128, 64, 32, 8 and 1 unit. It is worth noting that smaller networks do not perform significantly worse and can still be used successfully in a wide range of approaches. However, we have decided to use a larger network in this study.

The architecture described above, being an adaptation of the GQN model to the sequence data,

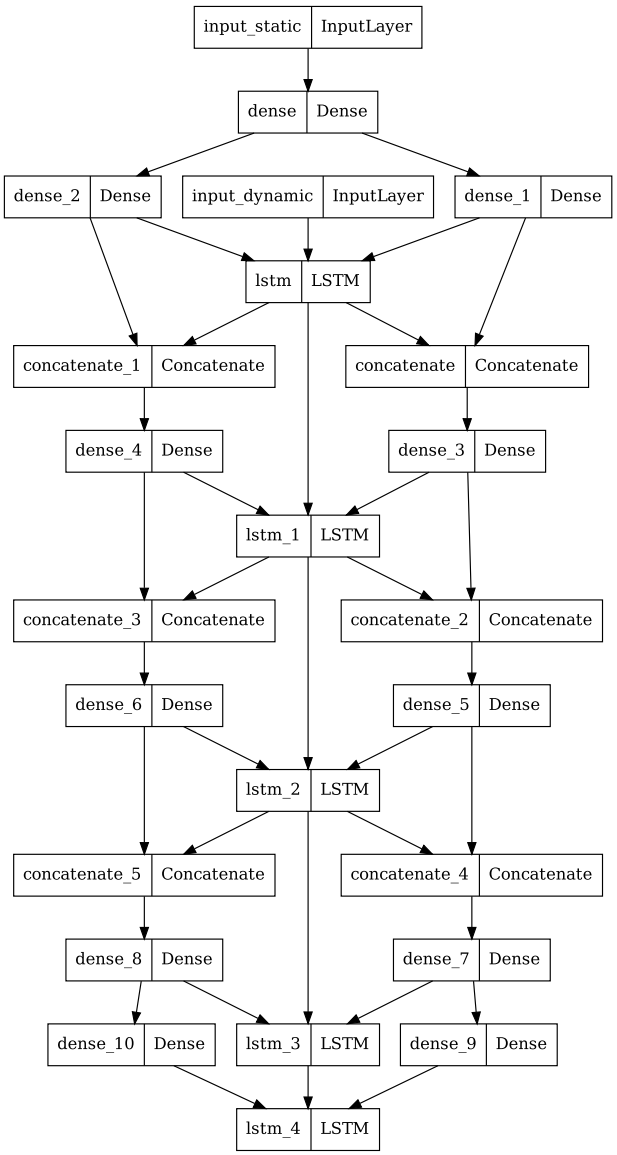


Fig. 3. The architecture of a recurrent neural network enriched with static metadata.

opens a gateway for a variety of new applications. One might consider time series of a different nature or sequences that do not involve the time domain at all.

3. Experiments

In this section, we present and discuss the results achieved with the RGQN approach on simulated pipeline transporting liquid medium where the process variables are pressure time series. We chose to simulate the process data to explore the full potential of the proposed model by testing different scenarios and obtaining pressure values for any point in the pipeline. Such a flexible playground is impossible to achieve with productive data for many reasons, e.g. due to fixed pipeline topology, set of hydraulic events, etc. However, in the following, we will also consider the existing pipeline topology, which can be considered a more realistic use case.

Our focus here is on hydraulic events, i.e. pressure phenomena that cause certain disturbances in the pipeline. The modeling of hydraulic events is of key importance in LDS, because pipeline ruptures cause such events, and efficient algorithms that capture them are critical from the leak detection perspective. Hydraulic events are typically reflected in pressure profiles as certain patterns. These are usually simple increases or decreases, oscillations, or other more complex shapes depending on the nature of the disturbance. Other factors that affect pressure include geodetic elevation and distance from the location where the event occurred. In the following study, we consider the following 4 scenarios:

(1) Measurements are available for all points in a pipeline (see Sec. 3.2).
(2) Measurements are available for a fixed number of equidistant points (see Sec. 3.3).
(3) Breakdown of a subset of sensors, so that measurements are available for a variable number of irregularly distributed points determined by the pipeline's topology (see Sec. 3.4).
(4) Leak occurrence, indicated by growing prediction errors (see Sec. 3.5).

All calculations were performed in the Python environment under the Ubuntu 18.04 LTS operating system. The neural model has been created with Keras library,[58] the backend was Tensorflow 2.8.[59] The entire list of languages and libraries with version numbers is presented below:

- Python, 3.9.10,
- Keras, 2.8.0,[58]
- TensorFlow, 2.8.0,[59]
- Matplotlib, 3.5.1,[60]
- pandas, 1.4.0,[61]
- scikit-learn, 1.0.2.[62]

The development was carried out in Jupyter notebooks. The computationally demanding calculations, in particular the training of the neural models, were carried out with a GPU support. Here, we used

Nvidia GeForce 2080Ti graphics cards with CUDA 11.6 libraries.

3.1. *Data generator*

In order to reflect the most important pipeline conditions, we proposed a parameterized data generator to simulate the fluid behavior in the pipeline. It adopts the following parameters:

- Set of points describing geodetic altitude at a given position with assumed linear interpolation between them (values from 0.75 to 1.25).
- Base pressure – the mean value of the pressure before the event occurs (value from the uniform distribution $U(-0.5, 0.5)$).
- Pattern – a curve reflecting a particular hydraulic event.
- Speed – the speed of the pattern propagation ($U(0.5, 1.5)$).
- Noise – the scale of the Gaussian noise added to the time series.

These parameters determine the generated pressure time series after an event for a given location and time in a specific pipeline. For simplicity, it is assumed that events are always generated at the first station and propagate in one direction towards the end of the pipeline. Each pattern is random:

(1) Generate a pair of starting and ending points from ($U(-0.5, 0.5)$).
(2) Generate a third of window size points from equidistant linear interpolation of the pair above.
(3) Generate random power from $1, 2, 3$ set and raise all points to that power.
(4) Divide all points by a maximum of absolute value.
(5) Generate a divisor from ($U(1.2, 2.2)$) and divide all points by it.

The following procedure is used to generate the pressure time series:

(1) Fill the vector with a *base pressure*.
(2) Multiply the vector with the scaling factor for an *altitude* of a given position.
(3) Divide the *pattern* by the logarithm of the distance from the event occurrence.
(4) Add the *pattern* to a vector at a proper index calculated using *speed* and *position*.
(5) Add *noise* from the normal distribution multiplied by a given factor ($U(-0.002, 0.002)$).

For each pipeline defined by a set of points describing geodetic altitude, 768,000 cases were generated and used in the training phase while validation and test sets consist of 12,800 cases each. The start of any event is taken as the base, where time is zero as the event wave head starts to propagate outwards from the source location. For measurements other than at the source location, time is relative and describes the amount of time before or after time zero. The training was carried out with Adam optimizer[63] available in Tensorflow library.[59] The training stop condition was determined based on the detection of a plateau on the loss function curve calculated with independent validation data.

3.2. *All points available*

In this experiment we assumed unlimited measurements anywhere in the pipeline. Such a scenario is unrealistic; however, it demonstrates the model's learning capabilities. In both the training and validation phases, pressure values from any location were supplied to the model, thus it was able to generate pressure time series for any location in the pipeline.

The observed error metrics, presented in Table 1, are comparable to the order of magnitude of the noise (4.0×10^{-6} for mean squared error and 1.6×10^{-3} for mean absolute error), the mean squared error is less than 10^{-4} and the mean absolute error is less than 10^{-2}; thus, we can assume that the model was capable of generating pressure time series based on similar measurements at other locations. The model can handle a variety of patterns, as shown in Fig. 4. To examine the performance of the RGQN model, we have developed an auxiliary GQN architecture that is

Table 1. Comparison of mean squared error and mean absolute error for an experiment with all points available during the training phase for 3 pipeline topologies applying a novel RGQN approach or an adjusted classical GQN.

	RGQN		GQN	
Pipeline ID	*MSE*	*MAE*	*MSE*	*MAE*
P1	**7.54e-5**	**4.72e-3**	4.35e-4	1.37e-2
P2	**8.06e-5**	**5.48e-3**	4.77e-4	1.83e-2
P3	**4.72e-5**	**4.77e-3**	2.54e-4	1.22e-2

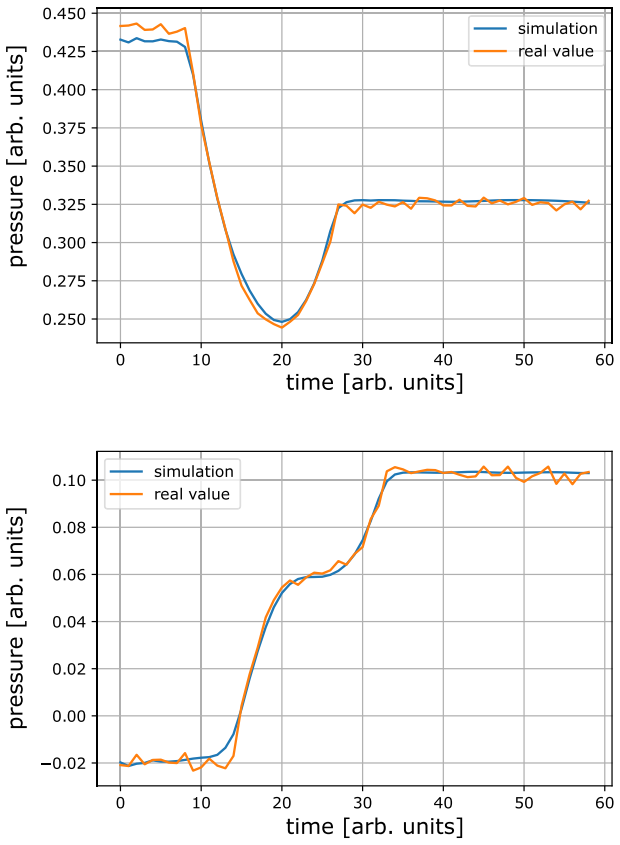


Fig. 4. Comparison of model output – simulation, with the ground truth – real value.

capable of accepting time series. The only difference, compared to RGQN, is the replacement of recurrent units by 1D convolutional units and slightly different handling of static meta-information. The novel approach RGQN performs significantly better than the auxiliary GQN model. The prediction errors are one order of magnitude smaller in favor of the RGQN approach.

3.3. *Limited points available*

In the following experiment, only 10 uniformly distributed measurement locations were selected. This is a more realistic scenario in which, for example, a buried pipeline cannot be fully monitored, and only pre-installed sensors provide data on its condition. The tests were conducted in two different ways. First, time series generation was tested at points from a predefined space of sensor locations that participated in the training phase. Second, to test the generalization capability, the model was required to generate pressures for locations that were not part of the training dataset, that is, the model was never exposed to data from these locations. The test points were chosen such that the height between them was always a linear interpolation. This ensures that the results for all points are comparable because the model is equipped with all the information necessary to generate a time series for a given location.

Table 2. Mean squared error and mean absolute error for experiment with selected points available during the training phase and tested on the same set of points and at intermediate positions for 3 different topologies of the pipeline.

	Selected points		All points	
Pipeline ID	*MSE*	*MAE*	*MSE*	*MAE*
P1	4.76e-5	4.75e-3	7.06e-4	1.31e-2
P2	3.01e-5	4.18e-3	1.09e-3	1.10e-2
P3	6.28e-5	5.54e-3	6.18e-4	1.07e-2

The performance obtained on the predefined sensor positions available during the training phase is comparable to that of the experiment in which all measurement points were available during training and testing, see Table 2. As can be expected, the performance for any location on the pipeline is worse than for the actual measurement points, but the accuracy obtained is still at a satisfactory level. The quality of the results for both approaches is stable when pipelines with different topologies are used.

During the training phase the performance in both cases, i.e. all points and arbitrary points, stagnates early (see Figs. 5 and 6).

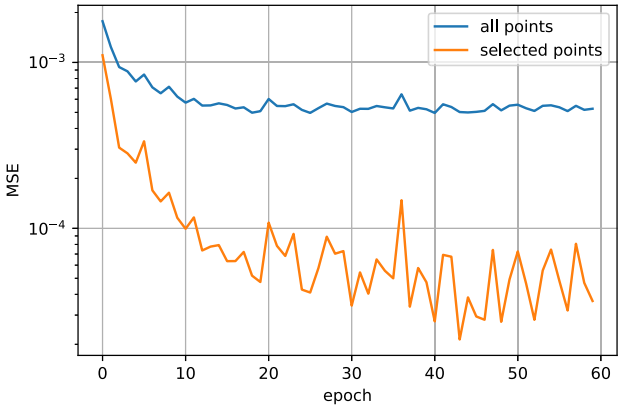


Fig. 5. Performance on validation sets through learning phase expressed as mean squared error.

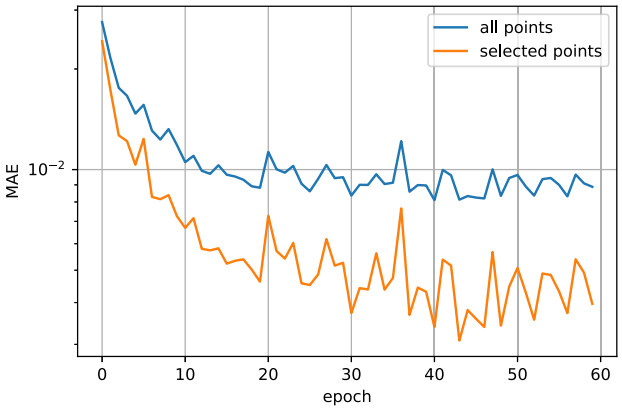


Fig. 6. Performance on validation sets through the learning phase expressed as mean absolute error.

3.4. *Real-life scenario: Breakdown of sensors*

For the purpose of presenting the following experiments in this and in the next section, the measurement stations were organized according to the topology of the actual pipeline. The pipeline segment under consideration consists of 23 stations equipped with pressure sensors, distributed irregularly along the pipeline. For reasons of anonymity, the positions of individual stations, which are also distances from the upstream entrance of the pipeline, were expressed in an arbitrary unit that roughly corresponds to the distance traveled by the pressure wave in a few tens of milliseconds. The distances between the adjacent stations range from a few hundred meters to several dozen of kilometers, and the length of the pipeline segment can be several hundred kilometers.

The experiment presented in this section is a generalization of the scenario described in Sec. 3.3. In a productive application, the measurements of the 23 stations provide data to a hydraulic model that can analyze the condition of the entire pipeline. However, it may happen that some of the measurement points are inaccessible due to failures, connection problems, maintenance, etc. The purpose of the model is to generate pressure time series for the missing sensors, based on the remaining available measurements.

In order to examine how model performance changes as the number of active sensors increases, we conducted the following experiment. For a given number of active sensors, at each trial, we randomly excluded the corresponding subset of sensors and

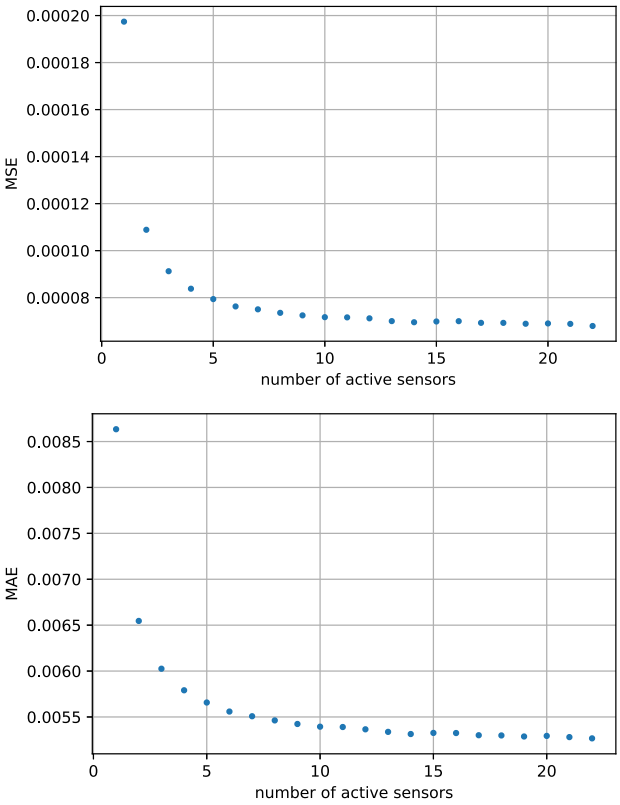


Fig. 7. Performance of the model depends on the number of active sensors expressed as mean squared error and mean absolute error, respectively.

predicted only for a randomly selected inactive sensor. Figure 7 shows the MSE and MAE for all trials (200,000) for each number of active sensors. The results obtained are consistent with intuition; the largest error occurs when only one sensor is available. In this case, the model does not have information on the wavefront velocity responsible for signal propagation with time along the pipeline, so it cannot correctly reconstruct the pressure time series. The experiment also shows that the availability of at least 2 sensors is sufficient to reconstruct the signal sonic velocity. However, the higher the number of active sensors, the smaller the errors that occur in the reconstruction phase.

It is also interesting to look at the model's behavior for a more extreme situation, i.e. one in which the last sensor is reconstructed from a given number of preceding sensors (starting with the beginning of the pipeline). As in routine operations, technological events appear first at these points; the reconstruction is very important to avoid false alarms for leak

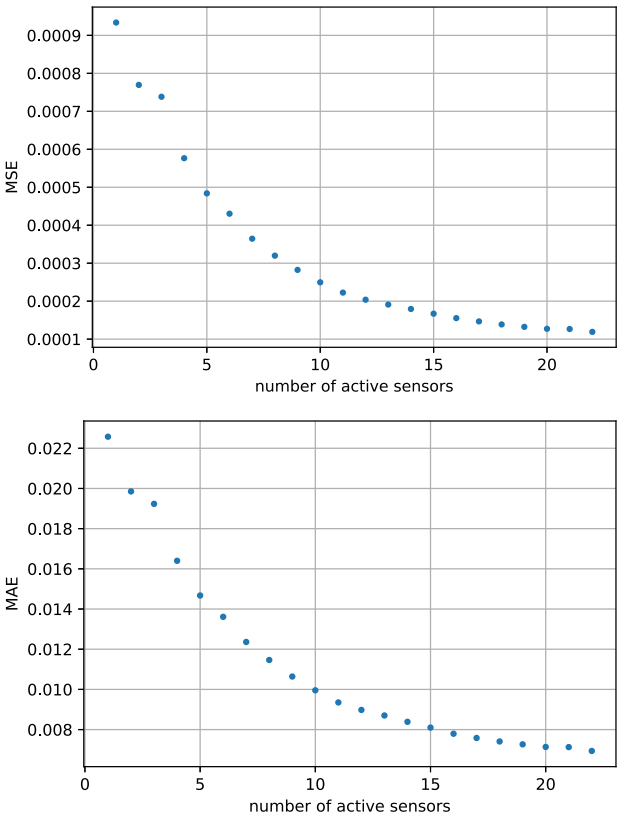


Fig. 8. Performance of the model depending on the number of active sensors expressed as mean squared error, where active sensors are taken from the beginning or the end of the pipeline, respectively.

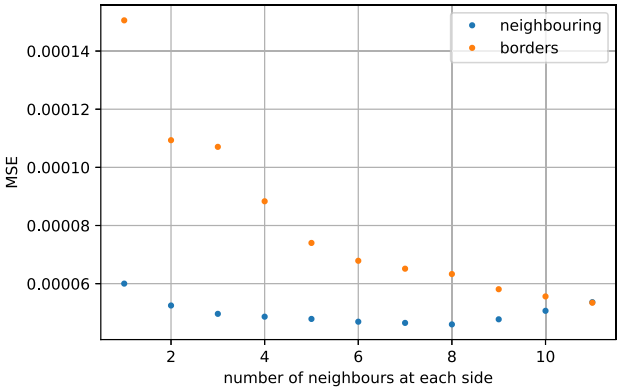


Fig. 9. Performance of the model depends on the number of active sensors expressed as mean squared error for the neighboring sensors and border sensors, respectively.

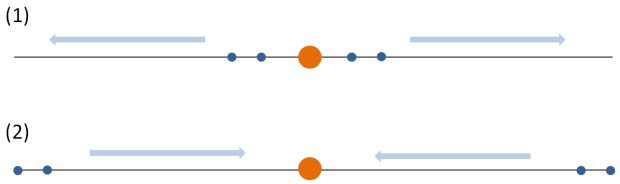


Fig. 10. Experiment cases: (1) predictions for the middle sensor are based on the neighboring sensors, the error of the prediction depends on the number of sensors at each side and (2) predictions for the middle sensor are based on the edge sensors, the error of the prediction depends on the number of sensors at each side.

locations near borders. In this case, Fig. 8 shows the dependency of the prediction residuals (expressed as MSE and MAE) as a function of the number of sensors supplied sequentially to the model. The largest error is observed, as intuitively expected, for the case when the scene is formed from just a single sensor. The model is unable to adequately reproduce the velocity of the event, resulting in a poor reconstruction of the time series. Successively adding more sensors improves the reconstruction and ultimately removes the residuals almost completely when all available sensors are used to create the scene.

Moreover, we observed that predictions based on neighboring sensors have less error than predictions based on distant measurements, see Fig. 9. As an experiment, the predictions for a sensor located in the center of the pipeline were based on a varying number of neighboring sensors on each side. In the second case, the predictions were based on sensors placed at the periphery of the pipeline, the number of which was the same on each side of the sensor under consideration. The method of sensor selection in both cases is explained in Fig. 10. The orange circle reflects the sensor which is being reconstructed, the little blue circles are the sensors used for the scene creation. The light blue arrows show the direction of subsequent pairwise incorporation of the sensors, either from the center to the edge (1) or from the edge to the center (2).

3.5. *Real-life scenario: Leak detection*

The aim of the last experiment is to examine the behavior of the model in the presence of a leak, which manifests itself as a pressure drop of decreasing amplitude, propagating in both directions along the pipeline. As expected in such a case, the prediction errors will generally be larger than in the situation without leakage, since no data loaded with leakage effects were provided in the training phase. To rephrase, during the training phase, we only provide data containing standard events reflected by

pre-defined patterns. Therefore, the model is not capable of predicting the additional effect associated with leakage, which ultimately leads to an increase in the prediction error. The prediction error should be inversely proportional to the distance from the leakage site because the leakage effect is more pronounced at sites close to the leak location. This is exactly the result of our experiments.

In order to make our model capable of detecting leakage, we re-trained the previously applied model with data in which events originate from both the beginning and the end of the pipeline. This means that the data were expanded so that events in both the left and right directions were made available to the model. As mentioned above, only ordinary events are provided to the model during training, reflecting the prospective business application. The model is exposed to standard events that are associated with regular pipeline operation, i.e. no-leak scenarios. In the case of a leak event, the model receives qualitatively different data that includes additional hydraulic effects associated with the leak and propagating from the location of leak origin. This happens exclusively in the inference phase when the previously trained model is simply exposed to a stream of currently acquired process data.

The prediction error comparisons (MSE) between the leaky and no-leak cases, shown in Figs. 11–13, fully support the assumptions made above. First, the cases without leaks (orange dots) are almost perfectly predicted by the model. This is exactly the expected behavior — the model has been trained

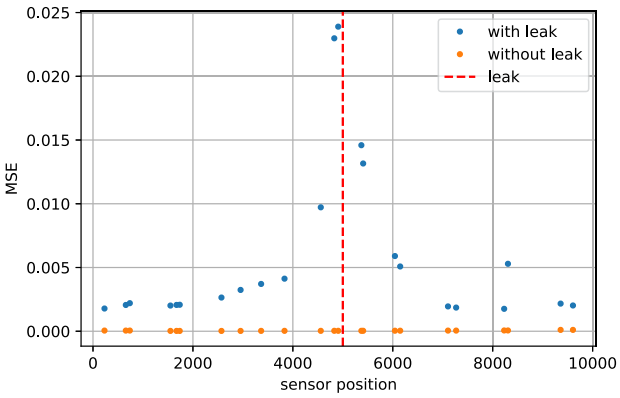


Fig. 11. Comparison of the model's performance for leak occurring in the middle of the pipeline's length and nonleak case expressed as a mean squared error.

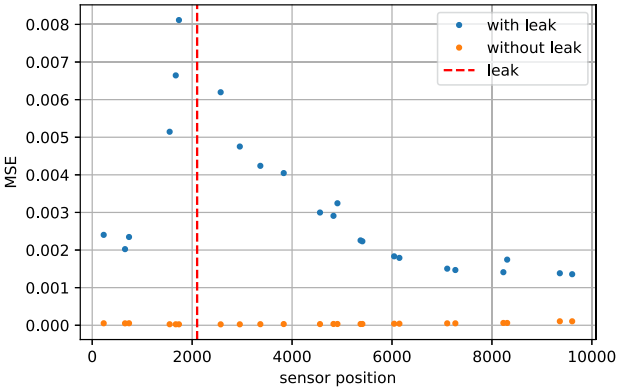


Fig. 12. Comparison of the model's performance for leak occurring in the first quarter of the pipeline's length and nonleak case expressed as a mean squared error.

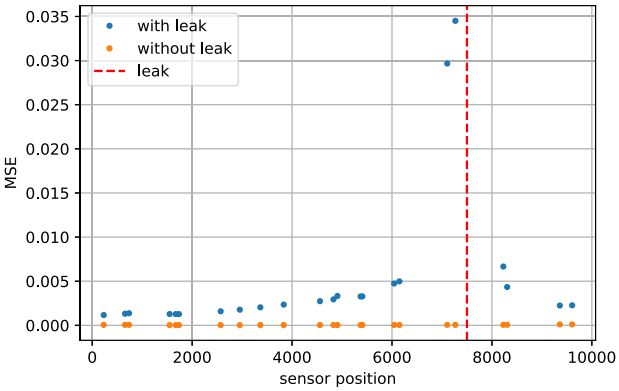


Fig. 13. Comparison of the model's performance for leak occurring in the third quarter of the pipeline's length and nonleak case expressed as a mean squared error.

on such data and is therefore able to predict pressure values of standard events almost perfectly. On the other hand, the blue dots reflect the predictions residual for cases where the simulated leakage effect was included in the data during the inference phase. Here we observe a general deterioration of the prediction capability, but more importantly, the residual values increase exponentially with decreasing distance from the leak site (leak site is denoted by a vertical dashed red line). This can be easily explained by the fact that the hydraulic effect of leakage is simply stronger at locations close to the leak, meaning that the data are more misleading to the model, what is directly reflected in higher residual values. Furthermore, the relative residual values can be used to estimate the location of the leak.

In this experiment, each sensor signal has been reconstructed from a subset of all other sensors that

were used to create the scene. Three cases, differing in the location of the leakage, were investigated, namely:

- the leak in the middle of the pipeline (between 12th and 13th sensor),
- the leak around the first quarter of the pipeline's length (between 6th and 7th sensor),
- the around a third quarter of the pipeline's length (between 19th and 20th sensor).

Similar results can also be obtained for arbitrary leak locations. As mentioned earlier, the largest errors occur exactly in the leak location area. This can be explained by the fact that the model only knows the propagation of the events occurring at the beginning and at the end of the pipeline, so a sudden increase in the amplitude at the intermediate location of the pipeline is an unexpected issue for the model.

It is clearly visible from Figs. 11–13 that the approximate leak location can be determined by analyzing the prediction residuals. They form an exponential curve, in Fig. 14, we present fitted curves to the functional form $y = A \times \exp(B \times x) + C$, where A, B, C are fitting parameters. The intersections point can be considered as an approximated leak location, which in this case was ca. 5006. This compares well with the assumed value of 5000.

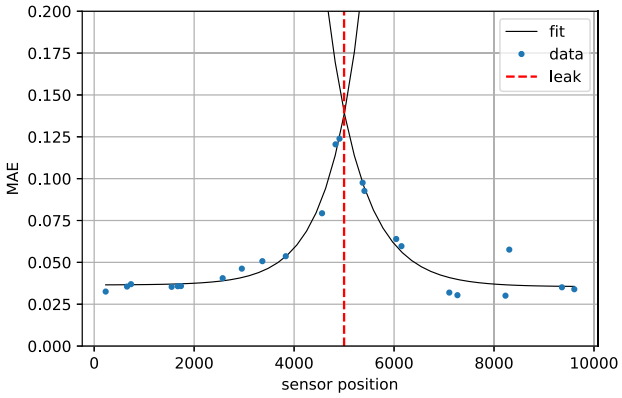


Fig. 14. The simple methodology for estimating the leak location. The points at the left and right sites with respect to the leak position were independently fitted to the curve of the form: $y = A \times \exp(B \times x) + C$. The leak location can be estimated as an intersection point between these two curves. The values of the coefficients (A, B, C) are: $(2.06e-07, 2.37e-03, 2.25e-03)$ and $(5.46e+02, -1.98e-03, 2.45e-03)$, respectively, for the left and right curves.

In summary, we would like to emphasize that the methodology presented here can be considered a reliable basic component of the LDS. The model can be routinely exposed to incoming data that are usually devoid of leak effect. This assumption implies that the predictive model incorporates the standard hydraulic profile picture of the pipeline during training. Subsequently, during the inference phase, the model is able to quantify the abnormal effects associated with leaks. This directly opens up the possibilities for LDS applications.

4. Conclusions and Future Efforts

We have proposed a novel approach based on the GQN framework suitable for processing the time series. The model was proven to be capable of predicting pressure values at a given time and position in a pipeline based on the other measurements. This input information is transformed into a latent representation reflecting the current state of a pipeline, which is then used to make predictions for a given query. This approach was tested in four different scenarios. Within the first one, the most idealistic case, the model was exposed to data from every possible point in the pipeline in the training phase. In the second scenario, a more realistic one, the model was exposed to data from a limited number of evenly distributed measurement points. The model has proven to have generative capabilities for predefined points as well as for any arbitrary location in the pipeline. In the third example, utilizing a real-life pipeline's topology, measurement points were distributed irregularly along the pipeline and the prediction phase was applied only to these points, in order to reflect a business case of sensor's breakdown. In the last experiment, we have shown the capability of the RGQN framework to be the key component of the LDS. By a simple analysis of the prediction error obtained from the data comprising the leak effect, one can detect the presence of the leak, as well as calculate the approximate leak location.

It is noteworthy that the approach presented here can be applied not only to pressure prediction but to all other use cases where time series are accompanied by some static descriptor, e.g. road traffic, energy consumption, weather forecast, etc. As a future effort, we plan to apply this methodology to realistic data obtained from the pipeline transporting

real media. Another direction for future work will be the application of the framework presented here to the time series reflecting energy consumption or generation, especially in the context of prosumers and renewables.

Acknowledgments

This study has been carried out in the frame of the Polish National Center for Research and Development (NCBR) project: "The leak detection and location system for multiphase fluids pipelines" under the Smart Growth Operational Programme 2014-2020, Measure 1.1 Research and development projects of enterprises, Sub-measure 1.1.1 Industrial research and development works implemented by enterprises, project identifier: POIR.01.01.01-00-0300/19. The data and the anonymized pipeline topology can be provided upon request.

References

1. A. Abadi, T. Rajabioun and P. A. Ioannou, Traffic flow prediction for road transportation networks with limited traffic data, *IEEE Trans. Intell. Transp. Syst.* **16**(2) (2014) 653–662.
2. N. F. M. Radzuan, Z. Othman and A. A. Bakar, Uncertain time series in weather prediction, *Proc. Technol.* **11** (2013) 557–564.
3. A. Shamshad, M. Bawadi, W. Wanhussin, T. Majid and S. Sanusi, First and second order Markov chain models for synthetic generation of wind speed time series, *Energy* **30** (2005) 693–708.
4. L. Li, Z. Yang, Z. Dang, C. Meng, J. Huang, H. Meng, D. Wang, G. Chen, J. Zhang, H. Peng and Y. Shao, Propagation analysis and prediction of the COVID-19, *Infect. Dis. Model.* **5** (2020) 282–292.
5. G. Miebs, M. Mochol-Grzelak, A. Karaszewski and R. A. Bachorz, Efficient strategies of static features incorporation into the recurrent neural network, *Neural Process. Lett.* **51** (2020) 2301–2316.
6. S.-E. Park, N. G. Laxpati, C.-A. Gutekunst, M. J. Connolly, J. Tung, K. Berglund, B. Mahmoudi and R. E. Gross, A machine learning approach to characterize the modulation of the hippocampal rhythms via optogenetic stimulation of the medial septum, *Int. J. Neural Syst.* **29** (2019) 1950020.
7. H. Balzter, Markov chain models for vegetation dynamics, *Ecol. Model.* **126** (2000) 139–154.
8. Y.-Z. Li and J.-C. Niu, Forecast of power generation for grid-connected photovoltaic system based on markov chain, in 2009 *Asia-Pacific Power and Energy Engineering Conf.* (IEEE, Wuhan, 2009), pp. 652–655.
9. J. P. Amezquita-Sanchez, N. Mammone, F. C. Morabito and H. Adeli, A new dispersion entropy and fuzzy logic system methodology for automated classification of dementia stages using electroencephalograms, *Clin. Neurol. Neurosurg.* **201** (2021) 106446.
10. A. Burns, H. Adeli and J. A. Buford, Upper limb movement classification via electromyographic signals and an enhanced probabilistic network, *J. Med. Syst.* **44** (2020) 176.
11. C. A. Perez-Ramirez, J. P. Amezquita-Sanchez, M. Valtierra-Rodriguez, H. Adeli, A. Dominguez-Gonzalez and R. J. Romero-Troncoso, Recurrent neural network model with bayesian training and mutual information for response prediction of large buildings, *Eng. Struct.* **178** (2019) 603–615.
12. B. K. Oh, K. J. Kim, Y. Kim, H. S. Park and H. Adeli, Evolutionary learning based sustainable strain sensing model for structural health monitoring of high-rise buildings, *Appl. Soft Comput.* **58** (2017) 576–585.
13. Z. Li and H. Adeli, Control methodologies for vibration control of smart civil and mechanical structures, *Expert Syst.* **35** (2018) e12354.
14. M. Gutierrez Soto and H. Adeli, Semi-active vibration control of smart isolated highway bridge structures using replicator dynamics, *Eng. Struct.* **186** (2019) 536–552.
15. S. Zhang, L. Zhou, X. M. Chen, L. Zhang, L. Li and M. Li, Network-wide traffic speed forecasting: 3d convolutional neural network with ensemble empirical mode decomposition, *Comput. Aided Civil Infrastruct. Eng.* **35** (2020) 1132–1147.
16. A. Karim and H. Adeli, Incident detection algorithm using wavelet energy representation of traffic patterns, *J. Transp. Eng.* **128** (2002) 232–242.
17. A. Karim and H. Adeli, Fast automatic incident detection on urban and rural freeways using wavelet energy algorithm, *J. Transp. Eng.* **129** (2003) 57–68.
18. J. Bright, C. Smith, P. Taylor and R. Crook, Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data, *Sol. Energy* **115** (2015) 229–242.
19. S. Li, H. Huang and W. Lu, A neural networks based method for multivariate time-series forecasting, *IEEE Access* **9** (2021) 63915–63924.
20. M. Khashei and M. Bijari, An artificial neural network (p, d, q) model for timeseries forecasting, *Expert Syst. Appl.* **37** (2010) 479–489.
21. D. Ruta and B. Gabrys, Neural network ensembles for time series prediction, in *2007 Int. Joint Conf. Neural Networks* (IEEE, Orlando, FL, 2007), pp. 1204–1209.
22. Y. Xue, Q. Zhang and F. Neri, Self-adaptive particle swarm optimization-based echo state network for time series prediction, *Int. J. Neural Syst.* **31** (2021) 2150057.

23. E. Bayro-Corrochano, S. Solis-Gamboa, G. Altamirano-Escobedo, L. Lechuga-Gutierres and J. Lisarraga-Rodriguez, Quaternion spiking and quaternion quantum neural networks: Theory and applications, *Int. J. Neural Syst.* **31** (2021) 2050059.
24. D. S. Jodas, T. Yojo, S. Brazolin, G. D. N. Velasco and J. P. Papa, Detection of trees on street-view images using a convolutional neural network, *Int. J. Neural Syst.* **32** (2022) 2150042.
25. P. Kovács, G. Bognár, C. Huber and M. Huemer, Vpnet: Variable projection networks, *Int. J. Neural Syst.* **32** (2022) 2150054.
26. A. Olamat, P. Ozel and A. Akan, Synchronization analysis in epileptic EEG signals via state transfer networks based on visibility graph technique, *Int. J. Neural Syst.* **32** (2022) 2150041.
27. G. Zhang, H. Rong, P. Paul, Y. He, F. Neri and M. J. Pérez-Jiménez, A complete arithmetic calculator constructed from spiking neural P systems and its application to information fusion, *Int. J. Neural Syst.* **31** (2021) 2050055.
28. M. Zhu, Q. Yang, J. Dong, G. Zhang, X. Gou, H. Rong, P. Paul and F. Neri, An adaptive optimization spiking neural P system for binary problems, *Int. J. Neural Syst.* **31** (2021) 2050054.
29. P. Lara-Benítez, M. Carranza-García and J. C. Riquelme, An experimental review on deep learning architectures for time series forecasting, *Int. J. Neural Syst.* **31** (2021) 2130001.
30. R. J. Frank, N. Davey and S. P. Hunt, Time series prediction and neural networks, *J. Intell. Robot. Syst.* **31** (2001) 91–103.
31. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, Generative adversarial nets, in *Proc. 27th Int. Conf. Neural Information Processing Systems — Volume* 2, *NIPS'14* (MIT Press, Cambridge, MA, 2014), pp. 2672–2680.
32. M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger and H. Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, *Neurocomputing* **321** (2018) 321–331.
33. C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri and H. Nakayama, GAN-based synthetic brain MR image generation, in 2018 *IEEE* 15*th Int. Symp. Biomedical Imaging* (*ISBI* 2018) (IEEE, Washington, DC, 2018), pp. 734–738.
34. T. Karras, T. Aila, S. Laine and J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, preprint (2017), arXiv:1710.10196.
35. M. Mirza and S. Osindero, Conditional generative adversarial nets, preprint (2014), arXiv:1411.1784.
36. M. Arjovsky, S. Chintala and L. Bottou, Wasserstein generative adversarial networks, in *Proc 34th Int. Conf. Machine Learning*, Vol. 70, eds. D. Precup and Y. W. Teh (PMLR, 2017), pp. 214–223.
37. P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, Image-to-image translation with conditional adversarial networks, in *2017 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*) (IEEE, Honolulu, HI, 2017), pp. 5967–5976.
38. E. Brophy, Z. Wang, Q. She and T. Ward, Generative adversarial networks in time series: A survey and taxonomy, preprint (2021), arXiv:2107.11098.
39. J. Yoon, D. Jarrett and M. van der Schaar, Time-series generative adversarial networks, in *Advances in Neural Information Processing Systems*, Vol. 32, eds. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett (Curran Associates, 2019), pp. 5508–5518.
40. S. M. A. Eslami, D. J. Rezende, F. Besse, F. Viola, A. S. Morcos, M. Garnelo, A. Ruderman, A. A. Rusu, I. Danihelka, K. Gregor, D. P. Reichert, L. Buesing, T. Weber, O. Vinyals, D. Rosenbaum, N. Rabinowitz, H. King, C. Hillier, M. Botvinick, D. Wierstra, K. Kavukcuoglu and D. Hassabis, Neural scene representation and rendering, *Science* **360** (2018) 1204–1210.
41. M. A. Adegboye, W.-K. Fung and A. Karnik, Recent advances in pipeline monitoring and oil leakage detection technologies: Principles and approaches, *Sensors* **19**(11) (2019) 2548.
42. N. Turner, Hardware and software techniques for pipeline integrity and leak detection monitoring, in *Offshore Europe* (Society of Petroleum Engineers, Aberdeen, 1991), pp. 139–148.
43. D. Hovey and E. Farmer, Pipeline leak probability and where to target the next maintenance dollar, *Pipes Pipelines Int.* **44** (1999) 41–43.
44. M. R. Delgado and O. B. Mendoza, A comparison between leak location methods based on the negative pressure wave, in 2017 14*th Int. Conf. Electrical Engineering, Computing Science and Automatic Control* (*CCE*) (IEEE, 2017), pp. 1–6.
45. Q. Chen, G. Shen, J. Jiang, X. Diao, Z. Wang, L. Ni and Z. Dou, Effect of rubber washers on leak location for assembled pressurized liquid pipeline based on negative pressure wave method, *Process Saf. Environ. Prot.* **119** (2018) 181–190.
46. H. Li, D.-Y. Xiao and X. Zhao, Morphological filtering assisted field-pipeline small leakage detection, in 2009 *IEEE Int. Conf. Systems, Man and Cybernetics* (IEEE, 2009), pp. 3769–3774.
47. P. Ostapkowicz, Leak detection in liquid transmission pipelines using simplified pressure analysis techniques employing a minimum of standard and non-standard measuring devices, *Eng. Struct.* **113** (2016) 194–205.

48. J. Wan, Y. Yu, Y. Wu, R. Feng and N. Yu, Hierarchical leak detection and localization method in natural gas pipeline monitoring sensor networks, *Sensors* **12**(1) (2011) 189–214.
49. G. Geiger, State of the art in leak detection and localisation, *Proc. Pipeline Technology 2006 Conf.* 25 April 2006, Hannover, pp. 1–25.
50. P.-S. Murvay and I. Silea, A survey on gas leak detection and localization techniques, *J. Loss Prev. Process Ind.* **25** (2012) 966–973.
51. O. Akinsete and A. Oshingbesan, Leak detection in natural gas pipelines using intelligent models, *SPE Nigeria Annual Int. Conf. and Exhibition*, 5–7 August 2019, Lagos, Nigeria, D023S009R001.
52. Z. Hu, B. Chen, W. Chen, D. Tan and D. Shen, Review of model-based and data-driven approaches for leak detection and location in water distribution systems, *Water Supply* **21** (2021) 3282–3306.
53. I. Amini, Y. Jing, T. Chen, A. Colin and G. Meyer, A two-stage deep-learning based detection method for pipeline leakage and transient conditions, in 2020 *IEEE Electric Power and Energy Conference* (*EPEC*) (IEEE, Edmonton, AB, 2020), pp. 1–5.
54. B. Arifin, Z. Li, S. L. Shah, G. A. Meyer and A. Colin, A novel data-driven leak detection and localization algorithm using the kantorovich distance, *Comput. Chem. Eng.* **108** (2018) 300–313.
55. K. Angelopoulos and G. O. Glentis, Performance assessment of correlation methods for the velocity estimation of vibro-acoustic signals propagating in fluid-filled pipelines, in 2021 10*th Int. Conf. Modern Circuits and Systems Technologies* (*MOCAST*) (IEEE, Thessaloniki, 2021), pp. 1–6.
56. J. Zheng, Y. Liang, N. Xu, B. Wang, T. Zheng, Z. Li, Q. Liao and H. Zhang, Deeppipe: A customized generative model for estimations of liquid pipeline leakage parameters, *Comput. Chem. Eng.* **149** (2021) 107290.
57. J. Li, Q. Zheng, Z. Qian and X. Yang, A novel location algorithm for pipeline leakage based on the attenuation of negative pressure wave, *Process Saf. Environ. Prot.* **123** (2019) 309–316.
58. F. Chollet *et al.*, Keras (2015), Software available from keras.io.
59. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems (2015), Software available from tensorflow.org.
60. J. D. Hunter, Matplotlib: A 2d graphics environment, *Comput. Sci. Eng.* **9**(3) (2007) 90–95.
61. J. Reback, Jbrockmendel, W. McKinney, J. Van Den Bossche, T. Augspurger, P. Cloud, S. Hawkins, M. Roeschke, Gfyoung, Sinhrks, A. Klein, P. Hoefler, T. Petersen, J. Tratner, C. She, W. Ayd, S. Naveh, M. Garcia, J. H. M. Darbyshire, J. Schendel, R. Shadrach, A. Hayden, D. Saxton, M. E. Gorelli, F. Li, M. Zeitlin, V. Jancauskas, A. McMaster, P. Battiston and S. Seabold, pandas-dev/pandas: Pandas 1.4.0 (2022).
62. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.* **12** (2011) 2825–2830.
63. D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, preprint (2014), arXiv:1412.6980.

Publication [P10]

Gehrlein, J., Miebs, G., Brunelli, M., and Kadziński, M. (2023). An active preference learning approach to aid the selection of validators in blockchain environments. In Omega (Vol. 118, p. 102869). Elsevier BV., DOI: 10.1016/j.omega.2023.102869.

Number of citations[8]:

- according to Google Scholar: 1

Contribution of co-authors (excluding the author and the supervisor of this dissertation):

- Jonas Gehrlein:
  - Co-authorship of the idea underlying the paper
  - Consultation of the context of case study and interpretation of its results
  - Co-authorship of the text of the publication
- Matteo Brunelli:
  - Co-authorship of the idea underlying the paper
  - Consultation of the problem structuring and methods for solving the study
  - Consultation of the experiment's design and interpretation of its results
  - Co-authorship of the text of the publication

[8] as on September 25, 2023

Omega 118 (2023) 102869

Contents lists available at ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

An active preference learning approach to aid the selection of validators in blockchain environments☆

Jonas Gehrlein[a], Grzegorz Miebs[b,*], Matteo Brunelli[c], Miłosz Kadziński[b]

[a] Web3 Foundation, Zug, Switzerland
[b] Institute of Computing Science, Poznan University of Technology, Piotrowo 2, Poznań 60-965, Poland
[c] Department of Industrial Engineering, University of Trento, Italy

ARTICLE INFO

Article history:
Received 17 October 2022
Accepted 7 March 2023
Available online 11 March 2023

Keywords:
Blockchain ecosystem
Validator selection
Multiple criteria decision analysis
Active learning
Preference learning
Multi-attribute value function

ABSTRACT

We consider a real-world problem faced in some blockchain ecosystems that select their active validators—the actors that maintain the blockchain—from a larger set of candidates through an election-based mechanism. Specifically, we focus on Polkadot, a protocol that aggregates preference lists from another set of actors, nominators, that contain a limited number of trusted validators and thereby influence the election's outcome. This process is financially incentivized but often overwhelms human decision makers due to the problem's complexity and the multitude of available alternatives. This paper presents a decision support system (DSS) to help the nominators choose the validators in an environment with frequently changing data. The system structures the relevant multiple attribute problem and incorporates a dedicated active learning algorithm. Its goal is to find a sufficiently small set of pairwise elicitation questions to infer nominators' preferences. We test the proposed solution in an experiment with 115 real nominators from the Polkadot ecosystem. The empirical results confirm that our approach outperforms the unaided process in terms of required interaction time, imposed cognitive effort, and offered efficacy. The developed DSS can be easily extended to other blockchain ecosystems.

1. Introduction

A blockchain is a decentralized data structure for storing information in blocks linked via cryptography. Each block contains transactions that are instructions on how the underlying state of the database is changed. The key problem that blockchains solve is how distributed computer nodes agree on those changes and obtain a valid and collectively controlled data structure. Thanks to their ability to decentralize data storage and allow for adequate data consistency checks, blockchains have emerged as a disruptive technology over the last decade [42]. They are vital to ensuring data validity and security without needing a trusted third party. Although they are best known for their crucial role of recording transactions in cryptocurrency systems, they have been considered a key enabling technology for innovation in fintech [39], the Internet of Things [1], Industry 4.0 [2,7], big data [8], and supply chain management [26,38].

In our study, we solve a decision problem prominent in various *Proof-of-Stake* (PoS) blockchains that employ an approval-based election for selecting *validators*, the actors that maintain the blockchain. Those validators are relevant for the security of the network and often protect billions of dollars worth of value. Crucially, a wide-spread underlying assumption to guarantee security is that at least two-thirds of the actors are trustworthy and adhere to the protocol rules [37]. Hence, the systems' security depends on a mechanism to prevent malicious entities from controlling a critical mass of validators. However, the permissionless nature of those blockchains allows anyone to become a validator candidate. As there is no automated way to judge the trustworthiness and capabilities of a candidate in this setting, protocols often introduce a second set of actors, the *nominators* (also called delegators). Those are human Decision Makers (DMs) tasked to scan all candidates and select suitable validators they deem trustworthy. To do so, the nominators individually submit unordered preference lists of sometimes multiple candidates, which are then aggregated and used to determine the active set of validators according to the respective election algorithm. In most protocols, elections occur frequently, and preference lists can be updated continuously. If nominators make no changes, their previous list is reused in the next election.

☆ Area Decision Analysis and Preference-Driven Analytics This manuscript was processed by Associate Editor Alessio Ishizaka.

* Corresponding author.
E-mail addresses: gehrlein.jonas@protonmail.com (J. Gehrlein), grzegorz.miebs@cs.put.poznan.pl (G. Miebs), matteo.brunelli@unitn.it (M. Brunelli), milosz.kadzinski@put.poznan.pl (M. Kadziński).

https://doi.org/10.1016/j.omega.2023.102869
0305-0483/

To motivate nominators and compensate them for their effort, nominators' financial incentives are tied to the future behavior of their nominated validators. Both entities need to lock some *stake*, a security deposit in the form of the blockchain's native token, to participate in the process. As long as their selected validators act according to the protocol's rules, both entities are then rewarded with newly minted tokens. In contrast, if validators misbehave and are caught by honest members, they and all their nominators share the punishment by destroying some or all of their stakes. While this mechanism makes choices economically meaningful and incentivizes the nominators to find honest validators, the decision problem is nevertheless complex. In addition to the aforementioned concept of security, nominators are also interested in optimizing their staking performance by selecting validators who consistently gain higher rewards. Also, many members of the blockchain space want to contribute to the network's decentralization with their vote by avoiding nominating larger operators that already have many validators in the active set.

We apply our developed DSS to the Polkadot network, which satisfies the mechanisms described above and provides a suitable testing ground. There, around 40,000 DMs (nominators) create and maintain their preference lists from 1500 validator candidates containing between one and sixteen trusted alternatives. The complexity of the problem has already reached a critical level. With the number of candidates and nominators expected to grow substantially in the near future, DSSs are required for a properly functioning network. From the point of view of nominators, selecting validators is a frequent decision process, as the election outcome is computed daily, complex due to the large size of the pool of validators and the different aspects of validators that shall be considered in the selection process, and financially relevant. At present, around 500 nomination activities occur every day. Those are currently largely unaided, and the nominators select their validators from a table in which the rows represent validators and columns their salient features, as explained in further detail in Section 2.

Given the setting of Polkadot, this manuscript makes three significant contributions to the literature. First, we formalize a multiple attribute decision problem designed to support a nominator (i.e., the DM) in selecting a small subset of validators (i.e., decision alternatives). In particular, we define a family of attributes containing the viewpoints relevant to the analysis. They refer to metrics of economic security, popularity, degree of independence, and performance. For each attribute, we discuss whether the monotonicity of preferences is justified given the problem's aim. For example, the self stake—indicating the amount the validators use to nominate themselves—can be interpreted as a gain-type criterion. In turn, it is impossible to associate a pre-defined preference direction with the cluster size that reflects the number of validators operated by the same entity. The considered problem is original in the literature at the crossroads of blockchain and Multiple Criteria Decision Analysis (MCDA) [16] due to considering an application *within* a specific blockchain protocol. On the contrary, recent studies concerned problems *across* protocols. In particular, [13] dealt with selecting the consensus protocol based on the identified criteria, priorities, and requirements. In turn, [12] delivered a framework for selecting the blockchain platform based on its functionality, adaptability, and compatibility, given the context of software-producing organizations. In addition, both works provided rich overviews of similar problems approached by other authors.

Our second contribution is elaborating a decision support system (DSS) that aids and facilitates the nomination process. By reducing the cognitive complexity of the problem and offering a system that efficiently elicits the preferences, nominators will select better-suited and secure validators based on recent data. Ultimately, this increases the entire blockchain's reliability and security. The proposed DSS involves a customized *preference learning* procedure. We make few ex-ante assumptions about the respective attributes of a validator because that would induce subjectivity, and a biased tool would receive little acceptance from the DMs.

The choice of the methodological basis underlying the DSS was dictated by the need to capture sufficiently complex preferences without adding excessive complexity or assumptions to the elicitation process. In particular, we formulated the following requirements [5]. First, when it comes to problem typology, we required a complete order of alternatives while considering a flat family of criteria and deterministic performances specified per alternative. Second, as far as the preference model is concerned, we faced quantitative performance scales, admitted full compensation between criteria, and demanded that the performances are compared with respect to the non-graded intensity of preference. At the same time, we did not intend to consider pairwise comparison thresholds, interactions between attributes, or criteria profiles. This led us to an additive value-based model that explains each alternative's position given the contributions derived from its performance on the individual attributes [11]. Third, we opted for using indirect preference information in the form of pairwise comparisons of reference alternatives. This was motivated by (a) a broad spectrum of DSS users not trained in using methods that require direct specification of model parameters, (b) the experience of nominators in making the holistic choices from among hundreds of validators, and (c) a low number of attributes making pairwise comparisons appropriate and feasible for DMs to answer accurately. The latter was confirmed in the initial tests. Incorporating pairwise comparisons allowed us to decompose the main problem into more minor and tractable problems, following a divide-and-conquer logic and keeping cognitive effort reasonably low. Finally, regarding the exploitation of the preference model, we required a univocal recommendation. These conditions led us to preference disaggregation methods, particularly the UTA family [35]. However, the existing approaches did not satisfy some other requirements, calling for further methodological developments and adjustments.

In particular, the traditional UTA method leaves the selection of pairs of alternatives to be compared to the DMs. With hundreds of alternatives to choose from in the considered problem, this would not be effective. Hence we propose a dedicated active learning procedure supporting the nominators. A comprehensive overview of procedures for selecting the most appropriate pairwise preference questions in the context of MCDA was presented by Ciomek et al. [6]. Our proposal is an inherent part of a stream aiming at a fast reduction of uncertainty regarding the parameters of a value function. However, the peculiar features of the considered problem, related to its great size and frequency, required a dedicated solution to speed up the computations. Specifically, we estimate the potential information gain offered by each pair of alternatives by computing the worst-case similarity between rankings induced by value functions compatible with either answer. The questions leading to a more significant rank correlation in the sub-spaces induced by the possible user's indications are deemed more informative for the preference elicitation process. The existing active learning approaches that quantify the information gain based on the exact outcomes of robustness analysis (e.g., the necessary preference relation or extreme ranks) were excluded due to being too demanding given their computational effort. In turn, the approaches built on the size of a polyhedron of feasible models or quantifying the entropy of possible results proved less favorable in the preliminary tests. The latter also confirmed the usefulness of accounting for the worst-case scenario when estimating the information. Techniques based on the mean or expected measures in view of unknown answers proved less appropriate in reducing the uncertainty when dealing with many users, as they all need to be guaranteed satisfactory results after a limited number of interactions. The basic variant of the proposed active learning algorithm

is similar to the state-of-the-art methods in using existing alternatives (validators) for the comparisons. However, we also propose an original method for constructing fictitious options through Bayesian optimization using the Gaussian process. Using fictitious alternatives aims to amplify the information gain compared to restricting the process to existing alternatives.

Further adaptations coupled with the proposed active learning procedures and imposed by the missed features of the existing methods are two-fold. On the one hand, some attributes in the considered problem were not associated with the pre-defined preference directions. Therefore, we had to adjust the proposed preference learning to handle potentially non-monotonic attributes. A review of existing value-based methods in this stream is provided by [21]. We used marginal value functions without imposing constraints on the monotonicity and controlling their shape. This allowed for their better adjustment to the nominators' preferences. On the other hand, we aimed to map the indirectly elicited preferences of nominators to a suitable, well-grounded recommendation. This required accounting for the variability of outcomes that can be attained given the incompleteness of the supplied pairwise comparisons. Therefore, we proposed to derive a mean model based on the representative and sufficiently large sample compatible with the nominator's indirect preferences [22]. In this way, we avoided extreme solutions in the feasible polyhedron and negligence of any attribute.

Our third contribution is validating the proposed approach in a real-world experiment involving 115 DMs familiar with the unaided nomination process. For this purpose, we tested the two proposed variants with the manual selection, not supported by any algorithmic solution, and a random selection, as control, in which the recommended alternatives are not linked to the elicited pairwise comparisons. The methods are assessed in terms of three dimensions: (i) the capacity of the algorithmic procedure to correctly capture and represent the preference of nominators, (ii) the possible reduction of the cognitive burden, and (iii) the degree of acceptance by the nominators. We discuss the advantages of our proposal from these viewpoints.

The remainder of the paper is organized as follows. Section 2 presents further details on the Polkadot network and the considered decision problem. In Section 3 we recall the fundamentals of Multi-attribute value theory (MAVT) and preference learning. In Section 4, we introduce the background theory and describe the proposed active preference learning algorithms. Section 5 outlines the social experiment, and Section 6 presents the results. The last section concludes the paper.

2. Application background

In this section, we provide a brief overview of the Polkadot decentralized network [3], where we apply the proposed DSS to aid tens of thousands of DMs. Polkadot is a fully sharded, permissionless blockchain that offers the infrastructure for a network of closely-connected blockchains that can communicate in a trustless and secure manner. The *relay chain* at the heart of the network is a central blockchain that interconnects all participating blockchains, referred to as *parachains* or *shards*.

Sharding is a database architecture design technique that horizontally partitions the data into various silos. Thereby the data can be distributed, and computation can be made on each shard separately and in parallel. Consequently, nodes allocated to one shard are not required to do computations on data belonging to other shards, which frees up valuable resources. This opposes non-sharded designs, such as Bitcoin [31], where all nodes in the network must process every transaction. As a result, it is possible to *scale* the throughput, reduce the costs per interaction, and enable wide adoption.

This architecture allows for *shared security*, a mechanism to scale security by lifting the need for every parachain to provide its own economic security. This crucial feature of Polkadot is achieved by sharing *validators* of the relay chain with all connected parachains. Validators are actors who validate and execute the instructions on how to change the database through transactions. They run software on their dedicated machines and are responsible for securing billions of dollars worth of value. As shareholders of the network [32], they are guardians of a robust and properly functioning blockchain. Polkadot operates securely when 2/3 of the active validators are honest and adhere to the protocol's rules [37], a concept called Byzantine Fault Tolerance [33]. This means it is fundamental for Polkadot to build and maintain an honest and reliable set of validators, as they secure not only the relay chain, but also all value contained in their interconnected shards.

Polkadot's permissionless nature allows anyone to participate and interact with it and, in particular, to become a validator candidate. This poses risks to the system's security because anonymous actors have substantial financial incentives to exploit the network. This fundamental challenge boils down to coordinating mutually distrusting entities to make agreed-upon changes to the underlying distributed ledger that are coherent with the rules. In particular, invalid actions such as double-spending or maliciously generating new tokens must be prevented. In short, protocols must ensure that all participants are in *consensus* about the current state and proposed changes. To achieve that, PoS systems, such as Polkadot, require their validators to lock up financial resources (in the blockchain native token) as a *stake* and expose those resources to the risk of being seized or burned. When other participating validators detect misconduct, the stakes of the offenders are *slashed*, and they lose (part of) their stake. The degree of the slashing depends on the offense and often amounts to 100%. Thus, the staking system models the economic incentives such that misconduct is very costly.

Because there is no ex-ante concept of a trustworthy validator, selecting new validators relies on human intuition about various trade-offs concerning quantitative properties, their communication, and reputation. To utilize the human capabilities of screening validators, Polkadot applies the *Nominated-Proof-of-Stake* (NPoS) protocol [4] that refines the election rules of how the active set of validators is determined. In particular, NPoS allows candidates to compete in the election with their own stake and those of another set of actors, the *nominators*. While the validators run the protocol's software on dedicated hardware and thereby add blocks to the blockchain, nominators do not need to run any software. Their task is to submit unordered preference lists containing a (small) subset of validators from the large set of candidates that they trust and are willing to support. The protocol's staking system then uses the combined stake of elected validators and their nominators as collateral. The incentive mechanism then rewards both parties alike with newly minted tokens (DOT) when things go well. When validators are caught offending, however, their stakes and the stakes of all their nominators are slashed by the network. Thereby, nominators have a vested interest in finding and selecting validators whom they deem trustworthy and do not expect to misbehave. Thus, by making rational and selfish choices, nominators contribute to a higher quality validator set on Polkadot and are, thereby, a fundamental part of the system's security.

This, however, requires nominators, who are human DMs, to process all the available information and accurately derive optimal choices based on their preferences. Furthermore, it assumes that nominators make an effort to update their selection based on new data because candidates frequently change, and a previously optimal choice might not be optimal anymore. With growing data, this mental processing has become a significant challenge.

Table 1
Summary of the main characteristics of the considered attributes.

j	Attribute name	Measurement unit	Performance range	Preference direction
1	Commissions	%	[0,10]	possibly monotonic with a maximum
2	Self stake	DOTs	[1,624586]	monotonic increasing
3	Total stake	DOTs	[1971785.4, 3430973.7]	monotonic decreasing
4	Era points	Number	[18383.33, 25765.71]	monotonic increasing
5	Voters	Number	$\{48, \ldots, 4142\}$	possibly non-monotonic
6	Cluster size	Number	$\{1, \ldots, 15\}$	possibly non-monotonic

In addition to honest validators, nominators generally try to optimize other metrics when making their selection. For example, they often want to increase their staking rewards by selecting validators that have more consistent performance and thus offer a higher expected payout. This is influenced by node operators deploying their validating software on superior hardware with better infrastructure. While this naturally leads to higher costs for validators, this leads to dependable block production, which positively reflects in more rewards. Incorporating other concerns about the degree of independence (i.e., decentralization) and popularity of a validator makes this a complex task that is solved differently with respect to a nominator's individual preferences, opinions, and beliefs.

Each validator has a set of quantifiable metrics that make it possible for each nominator to judge the trade-offs mentioned above. We chose six attributes that are prominent among nominators and hold valuable information. Specifically, we consider the following attributes:

- *Commission*: Represents a validator's cut for providing their services. A lower commission, all other performance being equal, means more rewards for the nominator. However, monotonicity is not enforced because many nominators are suspicious of validators who offer their services for free (i.e., zero commission) or hold fairness concerns and want to split their rewards. Therefore, many DMs might prefer a low but non-zero value.
- *Self stake*: The amount of DOT the validators use to nominate themselves. Those tokens are within the ownership of a validator and thereby measure the operator's economic mass at stake. A higher amount, ceteris paribus, means more skin in the game. As nominators prefer higher self stake values, we define this attribute as monotonic increasing.
- *Total stake*: The total amount of DOT the validator uses to compete in the election. To avoid a concentration of power in the hands of few validators, the rewards are distributed such that a lower total stake means that any nominated DOTs have, ceteris paribus, a larger share of the total stake and thereby increase the payoff. By doing so, excessively large total stakes are discouraged. Because nominators prefer smaller total stakes, this attribute is monotonic decreasing.
- *Era points*: A metric that describes the performance of a validator, e.g., how many blocks they produced or other actions that are beneficial for the network. The total reward to all validators is distributed based on the relative share of the era points gathered between each election of one validator compared to other validators. Therefore, this attribute is monotonic increasing.[1]
- *Voters*: The number of nominators voting for a validator. This indicates the popularity of a validator among other nominators. This attribute is non-monotonic because, at lower values, the increase in voters signals trust from other nominators. In turn, at higher values, it could indicate centralization, a property many nominators actively try to minimize.
- *Cluster size*: The number of (known) validators operated by the same entity. A higher number might indicate higher technical proficiency but could also point out more centralization and a higher risk of being slashed (because a mistake is likely to affect the whole infrastructure, that is, several validators, which is more severely punished by the network). Therefore, nominators might find their optimum between the minimum and the maximum, and the attribute is defined as non-monotonic.

The main characteristics of the above attributes are summarized in Table 1.

These six attributes comply with the requirements discussed by Keeney and Gregory [23] as they are unambiguous, direct, operational, understandable, and comprehensive. Let us note that a single dimension of the decision problem in the form of the validators' identities (with regard to name, homepage, etc.) is missing in the proposed family of attributes. However, a crucial aspect of identity is the number of nodes a validator controls, which is represented by the attribute *Cluster size*. This should mitigate the shortcoming of not being able to incorporate identity in the decision-making process and make the set of attributes sufficiently comprehensive.

3. Multi-attribute value theory and preference learning

The problem considered in this paper consists of aiding the nominators in Polkadot to rank a set of validators given the multiple attributes relevant to their assessment. In terms of MCDA, each nominator is treated as a DM; validators are interpreted as a non-empty finite set of m decision alternatives $A = \{a_1, \ldots, a_m\}$, and their characteristics form a set of n attributes (also called criteria) $G = \{g_1, \ldots, g_n\}$. The performance of validator $a_i \in A$ on attribute g_j is denoted by $g_j(a_i)$.

We apply MAVT to rank the validators from the best to the worst. It is an axiomatically grounded theory underlying the evaluation of alternatives according to their relevant attributes [24]. Its major assumption consists of assigning a marginal value $u_j(g_j(a_i))$ to each performance $g_j(a_i)$, reflecting the quality of alternative a_i with respect to attribute g_j as perceived by the DMs. For the sake of simplicity, we assume that marginal value functions u_j, $j = 1, \ldots, n$ are piecewise linear interpolations of γ_j ($\gamma_j \geq 2\ \forall j$) characteristic points. That is, each function u_j is the linear interpolation of a set of points with $g_j^1 < g_j^2 < \cdots < g_j^{\gamma_j}$ values in abscissa.

The marginal values $u_j(g_j(a_i))$ on all attributes g_j, $j = 1, \ldots, n$, are aggregated into a comprehensive measure of desirability, called a comprehensive value $U(a_i)$. In particular, under mild conditions, that is, preference independence [36], such aggregation can be performed using an additive value model:

$$U(a_i) = \sum_{j=1}^{n} u_j(g_j(a_i)) = \sum_{j=1}^{n} u_j(a_i).$$

It represents the preference such that:

$$U(a_i) \geq U(a_k) \quad \Leftrightarrow \quad a_i \succsim a_k \quad \forall a_i, a_k \in A$$

[1] As a technical side note: to account for statistical noise introduced through some correlated random process, we use the average over a more extended period.

where $a_i \gtrsim a_k$ means that a_i is at least as good or weakly preferred to a_k. Hence a value function U can be used to order the alternatives from the best to the worst. We will denote a ranking induced by function U on the set of alternatives A by $R_U(A)$. It points out a subset of r most preferred options corresponding to the r top-ranked alternatives in $R_U(A)$. The rank attained by alternative $a_i \in A$ will be denoted by $R_U(a_i, A)$.

As discussed by Dias et al. [9], various analytic methods exist to elicit marginal value functions. However, in some cases, the direct elicitation of parameters related to the functions' shapes and their maximal shares in the comprehensive value may be excessively lengthy and cognitively demanding. Then, it is more sensible to ask the DMs for holistic preferences on a reference set of alternatives. Under such a scenario, the method infers u_j, $j = 1, \ldots, n$, that are compatible with the provided judgments. This process is called *preference learning* [15] or *preference disaggregation* [18]. In MCDA, it has been first implemented in the UTA method [17,22], finding use in such various fields as e-governance [34], pharmacy [29], or marketing [27].

According to this approach, the DMs express their preferences on pairs of alternatives (real or fictitious). We assume that only preference statements in the form $a_i \succ a_k$ can be provided. We denote by S the set of pairs satisfying this strict preference relation, that is, if we consider A to be the set of real or fictitious alternatives, then:

$$S = \{(a_i, a_k) \in A \times A \mid a_i \succ a_k\}.$$

The set of value functions compatible with the preferences collected in set S is called $\mathcal{U}(S)$. The following three types of constraints define it. First, all functions in $\mathcal{U}(S)$ need to reproduce the desired relations between reference alternatives:

$$U(a_i) > U(a_k), \ \forall (a_i, a_k) \in S.$$

Second, for each attribute, we account for constraints related to the desired shapes of marginal value functions as implied by the respective preference directions. In particular, for a gain-type criterion, a monotonic non-decreasing function is expected by demanding $0 = u_j(g_j^1) \leq \cdots \leq u_j(g_j^{\gamma_j})$. Similarly, the monotonic non-increasing value function is modeled for a cost-type criterion, i.e., $u_j(g_j^1) \geq \cdots \geq u_j(g_j^{\gamma_j}) = 0$. In turn, we do not impose monotonicity constraints if attributes cannot be associated with a pre-defined preference direction.

Third, comprehensive values for all alternatives $a_i \in A$ are constrained to the interval $[0, 1]$, i.e., $0 \leq U(a_i) \leq 1$. After sampling from the polyhedron of compatible functions, to increase the interpretability of results, we ensure that for each function, the anti-ideal (ideal) alternative receives a comprehensive value of zero (one). For this purpose, for each marginal function associated with a possibly non-monotonic attribute, we subtract the minimal observed value for any performance. This way, the worst performance on each criterion is assigned a value of zero. Then, we scale all marginal functions by dividing them by the sum of the maximal observed marginal values for all attributes. This ensures that the values associated with the best performances sum up to one. These operations do not influence the ranking of alternatives.

When using preference information in the form of holistic judgments, typically, $\mathcal{U}(S)$ is composed of infinitely many compatible value functions. The more preference judgments are provided, the smaller the set of compatible value functions; that is,

$$S \subseteq S' \implies \mathcal{U}(S') \subseteq \mathcal{U}(S).$$

Even if all functions in $\mathcal{U}(S)$ reproduce the provided pairwise comparisons, the rankings implied by these functions may differ. To deliver an unambiguous order of alternatives, it is possible to elicit a sufficiently informative set of holistic judgments. However, if the numbers of alternatives and attributes are high, the number of such statements is prohibitively large. Then, it is more reasonable to select a single representative value function. In this paper, we follow the approach proposed by Kadziński et al. [20] by sampling a sufficiently large subset of compatible value functions from $\mathcal{U}(S)$ and averaging them to find a central model in the feasible space.

4. Active preference learning for ranking a set of validators

Finding a sufficiently small and informative set of compatible value functions $\mathcal{U}$ is crucial. At the same time, the cognitive effort of the DMs should be reduced by limiting the number of questions they are expected to answer. In the traditional preference disaggregation methods [17,18,30], the DMs need to select the pairs for comparison on their own or set S is considered given. More recent contributions focus on the optimal selection of the elements of S. The general idea consists of asking the pairwise elicitation questions that are the most informative given the type of expected recommendation [6].

Following [6], we consider the process of acquiring a DM's preferences as an iterative question-answer procedure. In each step, (i) a question maximizing the information gain is generated by heuristic active learning approach H, (ii) the DM is asked to answer this question by performing a pairwise comparison between two alternatives, and (iii) the preference is added to set S. The choice of each question depends on previous answers, and the set of all possible sequences of questions forms a binary tree whose depth depends on the number of questions to be asked.

4.1. Active preference learning with comparisons of real alternatives

Let us first focus on the algorithms considering real alternatives corresponding to validators in set A. When it comes to the set of candidate questions $Q(\mathcal{U}(S))$ at each stage of interaction, we consider only pairs $\{a_i, a_k\}$ leading to non-empty sets of compatible value functions irrespective of the provided answer, that is, $\mathcal{U}(S \cup (a_i, a_k)) \neq \emptyset$ and $\mathcal{U}(S \cup (a_k, a_i)) \neq \emptyset$. Such pairs cannot be compared in the same way by all functions in $\mathcal{U}(S)$.

The crucial part of the question selection strategy regards the principles according to which the information gain for each candidate pair is estimated. We follow the general scheme proposed by Ciomek et al. [6] and quantify the uncertainty in the set of value functions compatible with the DM's answer by referring to the variability of rankings induced by these functions. For this purpose, we compare the rankings $R_{U'}$ and $R_{U''}$ induced by value functions U' and U'' using the Spearman rank correlation coefficient [25]:

$$\rho(R_{U'}, R_{U''}, A) = 1 - \frac{6 \cdot \sum_{a_i \in A} [R_{U'}(a_i, A) - R_{U''}(a_i, A)]^2}{m(m^2 - 1)}.$$

Intuitively, $\rho(R_{U'}, R_{U''}, A)$ is high, that is, close to 1, when the differences between ranks of alternatives are small or non-existing. In turn, the coefficient is low, that is, close to -1, when the ranks attained by all alternatives in $R_{U'}$ and $R_{U''}$ differ greatly. Such comparison is performed for all pairs of rankings induced by the compatible value functions, and the minimum value is adopted as the measure of uncertainty related to the recommended ranking:

$$f_\rho(\mathcal{U}(S)) = \min_{U', U'' \in \mathcal{U}(S)} \rho(R_{U'}, R_{U''}, A).$$

In practice, because $\mathcal{U}(S)$ contains infinitely many compatible value functions, the computations consider $M = 1000$ uniformly distributed functions $U \in \mathcal{U}(S)$ sampled with the hit-and-run algorithm [40]. If $f_\rho(\mathcal{U}(S))$ equals one, all functions impose the same ranking on the set of alternatives A.

The measure assessing the quality of each candidate question needs to aggregate the information gain given the two possible answers. Preliminary results lead us to consider a pessimistic approach in which the score of question $\{a_i, a_k\} \in Q$ is represented

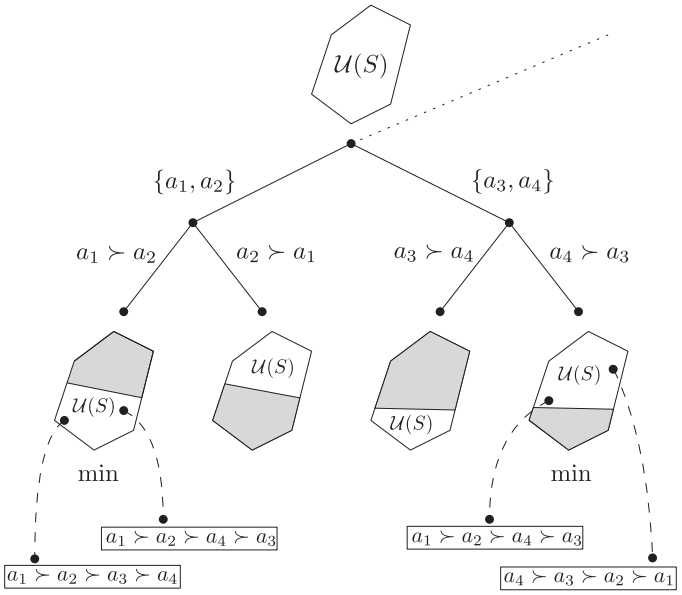


Fig. 1. A snapshot of the next question selection procedure: At a given step of the questioning process, a sample of possible pairs (here, for illustrative purposes, only two: $\{a_1, a_2\}$ and $\{a_3, a_4\}$) is considered for the next question. The spaces of compatible value functions derived from the potential questions, split with respect to both their respective answers, are probed, and a sufficiently large set of compatible rankings is sampled from each (in this figure, the variability of rankings in set $\mathcal{U}(S)$ is assumed to be related to the "size" of $\mathcal{U}(S)$). Then the selected question will be the one that maximizes the information gain for the answer associated with the least gain. In this highly simplified case, $\{a_1, a_2\}$ is representative of the selected question, as in the worst case marked with min, the ambiguity of $\mathcal{U}(S)$ is the smallest.

by the *worst* case similarity between rankings compatible with any answer. The question for which such similarity is the highest is selected; that is,

$$H(\mathcal{U}(S), A) = \arg\max_{\{a_i,a_k\}\in Q(\mathcal{U}(S))} f_\rho(a_i, a_k, \mathcal{U}(S)) = \arg\max_{\{a_i,a_k\}\in Q(\mathcal{U}(S))} \min\{f_\rho(\mathcal{U}(S \cup (a_i, a_k)), f_\rho(\mathcal{U}(S \cup (a_k, a_i)))\}.$$

Figure 1 sketches a simplified case of the next question selection procedure.

4.2. Active preference learning with comparisons of fictitious alternatives

Due to the availability of a finite number of real alternatives, it may be hard to find a question that would offer satisfactorily high information gain. For this reason, one can resort to fictitious alternatives, as sometimes advocated in the literature [41]. Then, the problem of finding the next best question becomes continuous. Similarly to $H(\mathcal{U}(S), A)$, the goal is to find a pair of fictitious alternatives (a_i, a_k) maximizing the worst-case similarity between rankings compatible with either answer.

We optimize $f_\rho(a_i, a_k, \mathcal{U}(S))$ by modifying $2n$ performances corresponding to performances of a_i and a_k on g_j, $j = 1, \ldots, n$. Let us denote the variables corresponding to these performances by $X = [x_1, x_2, \ldots, x_{2n}]$. To estimate their values, we apply the Bayesian optimization using Gaussian processes [14]. This technique is suitable for optimizing a black box function with evaluation taking a long time. It is based on multiple iterations of function estimation and evaluation at selected points (see Algorithm 1). In the first stage, the estimation is inaccurate due to a lack of data. Hence the value of function f_ρ is computed for e_0 random points. In the second stage, estimation based on all previous evaluations is used for selecting the next evaluation point. The estimation model provides the expected value $\mu(X)$ and standard deviation $\sigma(X)$ of f_ρ at any point X. Based on these values, acquisition function a, called the upper confidence bound, is used to determine the next point. In this approach, X maximizing $\mu(X) + \beta\sigma(X)$ is selected. This process is repeated over E iterations. Finally, the vector of variables corresponding to the performances of two fictitious alternatives leading to the greatest value of function f_ρ is returned as the algorithm's output. We remark that all the fictitious alternatives are reasonable and were perceived by participants as feasible. They were generated within the space induced by the minimum and maximum values of each attribute.

Algorithm 1 Basic pseudo-code for Bayesian optimization of function f_ρ.

Input:
function to be optimized f_ρ
number of initial evaluations e_0
number of evaluations E
acquisition function a

$i \leftarrow 0$
$y^{best} = -\infty$
while $i < e_0$ **do**
  $X^i \leftarrow random$
  $y^i = f_\rho(X^i)$
  if $y^i > y^{best}$ **then**
    $y^{best} \leftarrow y^i$
    $X^{best} \leftarrow X^i$
  end if
  $i \leftarrow i + 1$
end while
while $i < E$ **do**
  Update estimation of f_ρ using all i points. Estimation can be used to obtain the expected value μ and standard deviation σ of any $f_\rho(X)$.
  $X^i \leftarrow arg\,min_X a(\mu(X), \sigma(X), y_{best})$
  $y^i = f_\rho(X^i)$
  if $y^i > y^{best}$ **then**
    $y^{best} \leftarrow y^i$
    $X^{best} \leftarrow X^i$
  end if
  $i \leftarrow i + 1$
end while
Return (y^{best}, X^{best})

4.3. An example illustrating the interaction with the nominator

In this section, we illustrate the use of the proposed active learning algorithm and report the interaction results with an arbitrarily selected real-world nominator. The problem involves 85 validators to be ordered from the best to the worst. The performance for a subset of ten validators on the six attributes is provided in Table 2.

We employed a variant of the algorithm that generates questions involving fictitious alternatives. Table 3 shows how the process was carried out by indicating six questions (each composed of a pair of alternatives), the justification for selecting the specific pairs (i.e., their scores indicating the worst-case similarity between rankings compatible with each answer), and the nominator's response corresponding to the first alternative in each pair.

Alternatives a_1^* and a_2^* considered in the first question differ only in commission, total stake, and cluster size. The selection of a_1^* over a_2^* is motivated by its lower commission that outweighs the higher total stake and cluster size. The value of measure f_ρ leading to the question selection equals $\min\{-0.14, -0.26\} = -0.26$, and indication of a_1^* as a preferred alternative leads to a subset of com-

Table 2
Performances, marginal and comprehensive values of the top- and bottom-ranked validators considered in the illustrative example (the marginal values are provided in round brackets).

Commission	Self Stake	Total Stake	Era Points	Cluster Size	Voters	Rank	U
1.0 (0.180)	500.0 (0.070)	2182629.7 (0.072)	21047.45 (0.030)	6 (0.153)	4142 (0.184)	1	0.689
1.0 (0.180)	624586.0 (0.263)	2182629.8 (0.072)	21241.96 (0.031)	1 (0.055)	2330 (0.018)	2	0.619
0.0 (0.222)	3.0 (0.002)	2182549.5 (0.072)	25765.71 (0.067)	9 (0.167)	240 (0.025)	3	0.554
0.5 (0.186)	11039.0 (0.125)	2182623.3 (0.072)	24398.82 (0.056)	12 (0.111)	1284 (0.003)	4	0.553
1.0 (0.180)	500.0 (0.070)	2182629.8 (0.072)	22674.90 (0.042)	6 (0.153)	2262 (0.017)	5	0.533
0.0 (0.222)	3.0 (0.002)	2182645.6 (0.072)	22690.59 (0.042)	9 (0.167)	2113 (0.014)	6	0.518
1.0 (0.180)	500.0 (0.070)	2182627.9 (0.072)	21799.22 (0.035)	6 (0.153)	1728 (0.006)	7	0.516
			...				
10.0 (0.0)	1.0 (0.000)	2182605.9 (0.072)	21325.88 (0.032)	1 (0.055)	80 (0.028)	295	0.187
10.0 (0.0)	1.0 (0.000)	2182670.4 (0.072)	22048.44 (0.037)	1 (0.055)	333 (0.023)	296	0.187
10.0 (0.0)	1.0 (0.000)	2182606.4 (0.072)	21241.96 (0.031)	1 (0.055)	534 (0.019)	297	0.177

Table 3
Questions involving fictitious alternatives asked of the nominator in the illustrative example. Marginal and comprehensive values were assigned to the compared alternatives by the selected representative value function (the marginal values are provided in the round brackets).

Q.	Alt.	f_ρ	Commiss.	Self Stake	Total Stake	Era Points	Clust. Size	Voters	U
1	a_1^*	-0.14	0.0 (0.222)	624586.0 (0.263)	3430973.8 (0.000)	18,383 (0.000)	15 (0.000)	4142 (0.184)	0.669
	a_2^*	-0.26	10.0 (0.000)	624586.0 (0.263)	1971785.4 (0.092)	18,383 (0.000)	1 (0.055)	4142 (0.184)	0.594
2	a_3^*	-0.04	7.4 (0.095)	430345.7 (0.219)	3380420.5 (0.002)	23,413 (0.047)	6 (0.153)	4092 (0.178)	0.696
	a_4^*	0.06	4.6 (0.139)	261829.4 (0.181)	3430973.8 (0.000)	23,693 (0.050)	11 (0.148)	3816 (0.146)	0.665
3	a_5^*	0.11	7.7 (0.084)	412866.9 (0.215)	3255236.4 (0.008)	20,190 (0.021)	2 (0.076)	68 (0.029)	0.433
	a_6^*	0.18	9.1 (0.032)	280371.1 (0.186)	1997078.0 (0.090)	18,383 (0.000)	15 (0.000)	3244 (0.080)	0.387
4	a_7^*	0.28	6.8 (0.117)	23657.7 (0.128)	2047989.5 (0.085)	18,648 (0.003)	11 (0.148)	2336 (0.018)	0.499
	a_8^*	0.15	7.4 (0.097)	350714.4 (0.201)	3052037.2 (0.018)	24,576 (0.057)	2 (0.076)	2076 (0.013)	0.461
5	a_9^*	0.50	9.6 (0.016)	614132.0 (0.260)	3051378.2 (0.018)	21,305 (0.032)	9 (0.167)	3244 (0.080)	0.572
	a_{10}^*	0.37	5.9 (0.128)	624586.0 (0.263)	2373425.2 (0.054)	23,548 (0.048)	15 (0.000)	1433 (0.000)	0.494
6	a_{11}^*	0.68	5.1 (0.135)	597929.4 (0.257)	3430973.7 (0.000)	18,535 (0.002)	10 (0.171)	4142 (0.184)	0.749
	a_{12}^*	0.59	0.0 (0.222)	248409.8 (0.178)	2401013.3 (0.052)	25,765 (0.067)	6 (0.153)	68 (0.029)	0.701

patible value functions that offer slightly more-similar rankings. In the second iteration, the performances of compared alternatives a_3^* and a_4^* differ substantially. Indication of the former as more preferred is motivated by its higher self stake and the number of voters, even if its commission is slightly worse. In turn, the selection of a_5^* over a_6^* in the third iteration is justified mainly by its lower commission and cluster size and higher self stake even if the total stake associated with a_5^* is much higher. The interaction with the nominator continues over six iterations, exhibiting various trade-offs between the considered attributes. With each following question, the similarity between rankings induced by the compatible value functions becomes greater. Specifically, the values of measure f_ρ increase from -0.26 through -0.04, 0.11, 0.15, and 0.37 to 0.59. Once the final answer is provided, the minimal Spearman rank correlation between feasible rankings equals 0.68, confirming the high agreement.

When the set of compatible value functions is significantly reduced, we choose a single representative model to deliver a univocal ranking. This is attained by averaging a large sample of feasible functions. The obtained marginal value functions are presented in Fig. 2. As confirmed by Table 3, this function reproduces the six pairwise comparisons. For example, a_1^* is preferred to a_2^* because $U(a_1^*) = 0.669 > U(a_2^*) = 0.594$. Also, the marginal functions respect the pre-defined model assumptions. For example, the function for era points is increasing, whereas the one for the total stake is decreasing. The marginal functions for the monotonic attributes exhibit various shapes. In particular, the function for self stake is convex, whereas the function for era points is concave. Nevertheless, the deviations from neutrality represented by a linear function are not substantial. For the potentially non-monotonic attributes, the model confirms a high appreciation of zero commission, the $\wedge$-shape for cluster size with the maximum just above 10, and the $\vee$-shape for the number of voters with the minimum around 1400. The greatest share in the comprehensive value is associated with self stake (0.263) and commission (0.222). Hence, these attributes

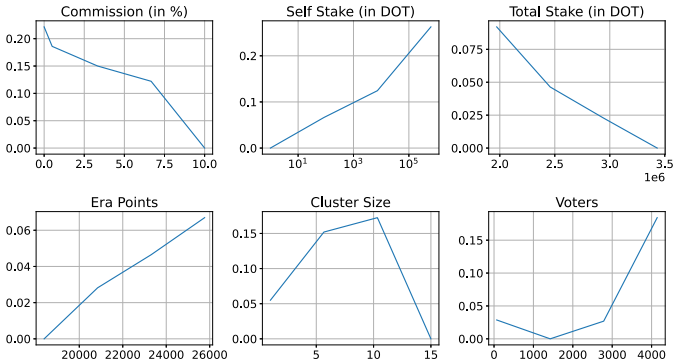


Fig. 2. Marginal value functions selected in the illustrative example.

can be deemed as most important for the nominator. In contrast, the least impact on the overall assessment of validators is related to total stake and era points.

Ultimately the selected model can be used to order all alternatives. In fact, the alternatives reported in Table 2 correspond to the top seven and bottom three options. Analyzing their performance in the context of the obtained value function justifies the constructed ranking. The most preferred validators derive their strengths mainly from low commissions, relatively high self stakes, and intermediate cluster sizes. In contrast, the worst validators are characterized by very high commissions, low self stakes, cluster sizes, and numbers of voters. Interestingly, the total stakes do not differentiate the validators ranked at the two extremes.

5. Methodology and social experiment

We conducted a social experiment with community members from the Polkadot ecosystem to validate the algorithm's effectiveness in the application-specific context. These members are famil-

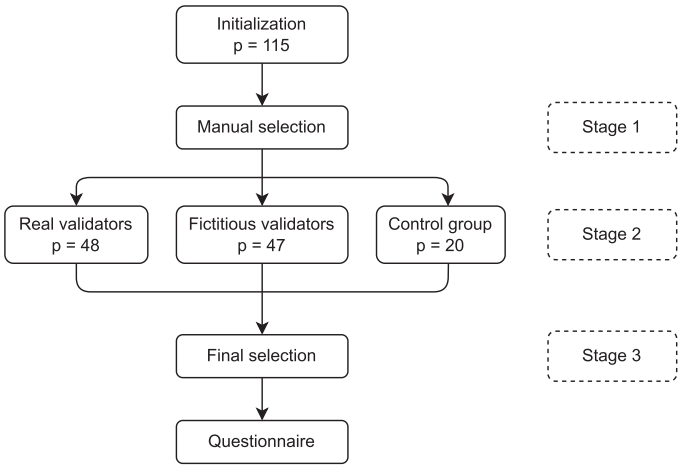


Fig. 3. The flow of the experiment: p is the number of participants in each part. All nominator participants first completed the manual selection process. Afterward, they were randomly assigned to one of the three treatments fictitious, real, or control). Finally, all participants made their final selection and completed a questionnaire.

iar with the validator selection process, where the predominant method is to select validators manually from an extensive list of alternatives with some limited decision aids like filters and column ordering. In this experiment, we tested two algorithm variations, fictitious and real, alongside a control group. Those different conditions, called *treatments*, allow us to analyze the performance of our algorithm under real-life conditions with the ability to make causal inferences on the difference between the two versions and compare them to the manual selection process and the control group. Additionally, our findings have high external validity, because they are generated by DMs familiar with the real-world problem.

In treatment *fictitious*, the algorithm constructs pairwise comparisons of validator alternatives that are not part of the underlying data set. In *real*, the algorithm was restricted to present alternatives in the data set and thereby are real validators on the network. In *control*, we provided random recommendations not based on our proposed DSS solution. The experiment was conducted online between 12 April 2022 and 29 April 2022 with 115 unique participants. It was designed to gather quantitative and qualitative measures to test three key dimensions of our decision aid. First, we tested how well our DSS captures the nominator's preferences. Second, we collected data to approximate the cognitive effort required in each decision step. Third, we elicited the nominator's acceptance of our DSS. Figure 3 illustrates the flow of the experiment. For later reference, the main points for DMs are labeled stages.

To analyze whether the algorithm benefits the DMs compared to the unaided process, we asked the DMs to manually select seven validators from the complete list of all validators at the beginning of the experiment. In the second step, all participants were exposed to the respective algorithm of their treatment and received a recommendation about another seven validators. In the third step, we combined the list of manually selected validators with those of the recommendation, and we asked the DMs to make a final selection of seven alternatives. Each participant only partook in one treatment and was unaware of the other conditions. This setup allows us to compare the individual recommendations with their baseline choices and infer the algorithm's quality, as well as compare the different variations of algorithms with each other. The data set with the list of validators and their attributes was obtained from the blockchain around the time the experiment was conducted. The data were held constant across the experiment for all participants, which means we could pre-compute the interaction tree for all users.

Stage 1: Manual selection

In the first stage of the experiment, participants were asked to manually select validators from a list of 297 validators, each with six attributes as described in Table 1. The manual selection was designed to mimic the selection process that nominators are accustomed to. In addition to familiar features of the interface, we offered the possibility to order by attributes. In this stage, we asked users to take their time to select their seven most-preferred validators. We ensured that they knew that we were interested in their personal preferences and that there were no right or wrong decisions. In this stage, we obtain a baseline of each subject's preferences to later compare with the recommendation of our algorithm and create quantitative metrics about the efficacy of the preference elicitation.

Stage 2a: Active learning process and recommendation

In the second stage of the experiment, we employed our algorithm to generate a recommendation using the preference learning technique presented in Section 4. Simulations before the experiment concluded that asking for six pairwise comparisons gives reasonable accuracy while minimizing the time users must spend in the process. Here, we run the two different variations of our algorithm, one that was restricted to constructing the pairwise comparisons from real alternatives (see Section 4.1) and the other that created fictitious alternatives (see Section 4.2) to optimize the information gain. After participants finished all pairwise comparisons and thus reached the end node of the binary tree, they were presented with the respective set of the most preferred validators. Then, we asked participants to mindfully examine the recommendations.

Stage 2b: Control group

In the control group, we created a recommendation by *randomly* sampling seven real validators from the underlying data set. The goal of the control group is to account for behavioral effects that might confound the results when subjects assess the quality of our algorithm. One prominent factor in behavioral research is the *experimenter demand effect* that describes a change of behavior of humans based on their subjective perception of appropriate behavior in that situation (see, for example, [43]). In our case, we asked participants to rate a novel algorithm we developed to deliver a recommendation. Although we applied objective language in all instructions, subjects might have inferred that we created the algorithm with the prospect of improving the selection process. Therefore, participants might feel obligated to act benevolently toward the algorithm recommendations. By adding the control treatment that provides a recommendation independent of the DM's preferences, we can both infer the relative quality of our DSS as well as obtain insights into whether our participants would express distaste for any of our approaches.

Stage 3: Final selection

The final stage of the decision-making process for participants was to present them with a randomized list of the combined alternatives from their manual selection (stage 1) and the algorithm recommendation (stage 2). We indicated the origin of each alternative. If an alternative was chosen both by the participant and the algorithm, we indicated this. Users were asked to select their seven most preferred alternatives with whom they wanted to stake their participation reward. This ensured that their choices were economically meaningful. More information on this is provided in Section 5.1. This step allows us to elicit how nominators evaluate their selection compared with the recommendation.

Questionnaire

After the final selection of validators, we concluded the experiment by gathering additional information from participants. We

used Likert scales, simple yes/no questions, and open-ended questions where participants could give detailed comments. This complements our analysis of quantitative data gathered from the subjects and contributes to a more robust evaluation of the three dimensions of our analysis.

5.1. Recruitment and participation compensation

The experiment was announced through social media in the Polkadot ecosystem and was targeted toward experienced nominators. This ensures that they understand the selection problem and are familiar with the presented attributes. To prevent users from signing up multiple times, we took a snapshot of the member list of the key social media channel before announcing the experiment. Only those users who were part of the channel at the time of the snapshot were allowed to participate in the experiment. Additionally, we did not emphasize the financial rewards that could be earned by participating. Examining the data confirms that we successfully avoided this issue, that our observations are from unique DMs, and we could prevent this common issue of online experiments. We collected sign-ups for around a week before sending out the first wave of invitations to estimate the size of the accessible subject pool. Given the initial traction, we expected to achieve 120 sign-ups within a reasonable time. We allocated around 20% to the control group and 40% to each algorithm. Allocation to the treatments was random under the previously mentioned targeted ratio. All participants were asked not to share any details with others or discuss the study on social media channels. This ensures that new sign-ups would be unbiased and that later observations remain comparable to earlier ones. We recruited $n = 115$ participants who were split into $n_{control} = 20$, $n_{fictitious} = 48$, and $n_{real} = 47$.

It is crucial to compensate subjects for their participation and make their choices incentive compatible. This means that their final payoff must, at least partly, be driven by their choices. General compensation for participation was achieved by promising a base fee of 2 DOT (at the time around $36) for around 25 min of work. This is an above-average hourly rate and ensures that participants are sufficiently motivated. Furthermore, we withheld their base fee for around 38 days and simulated staking rewards with their final selection of validators. That means they would be exposed to the consequences, both positive and negative, of their selected validators from stage 3 and thereby motivate them to optimize their selection based on their preferences.

6. Results

This section summarizes the effectiveness of the proposed algorithms based on the experimental results. We consider the following three viewpoints: (i) the degree to which the recommendations of the algorithms fit with the nominators' manual choices, (ii) the comparative performance in terms of time and cognitive effort, and (iii) the willingness of the nominators to accept the results and use the algorithm in the future. We statistically compared the performance differences between various treatments. The reported p-values of continuous metrics are computed using a t-test, while ordinal data are tested using a two-sided Mann–Whitney U test. In addition, we discuss the staking reward for different treatments. Finally, we present the most frequent interaction patterns when considering the preferences of nominators using the proposed active preference learning methods.

6.1. Effectiveness of preference elicitation

The fit of the algorithm with the real preferences of nominators was studied both quantitatively and qualitatively. Table 4 summarizes the relevant metrics per treatment.

Table 4
Key metrics to assess how well the algorithm elicits the preferences of the nominators (standard deviations in parenthesis).

	Overlap	Recommended
Control	0.30 (0.47)	1.45 (1.76)
Treatment: Real	2.68 (1.60)	4.77 (1.15)
Treatment: Fictitious	1.77 (1.48)	4.50 (1.50)

A strong indication of efficient preference elicitation is when the algorithm recommends a validator who was previously selected in the manual step. The variable *Overlap*, as indicated in Table 4, states this metric separated by treatment. This value naturally goes from 0 to 7. For example, in A_{real}, the algorithm recommended, on average, 2.68 alternatives already present in the nominator's manual selection. This number, given 297 alternatives, is quite encouraging. Naturally, there was little overlap in $A_{control}$, as validators were selected randomly. We can also observe that the recommendation from $A_{fictitious}$ was inferior to those of A_{real}, and the difference is statistically significant ($p < 0.01$).

The second metric, *Recommended*, measures how many alternatives were taken from the algorithm recommendation in stage 3, the final selection of a nominator. We can see that both algorithms outperformed the control ($p < 0.001$). Additionally, testing against a 50:50 allocation (i.e., a value of $7/2 = 3.5$), nominators chose more than half the validators from the recommendations in both algorithms ($p < 0.001$) and many fewer from the control algorithm ($p < 0.001$).

In the questionnaire, at the end of the experiment, we showed participants the seven recommended validators in a table and asked how well they felt their preferences were represented. Results are presented in Fig. 4. While there was no significant difference between the perceived quality of both algorithms, both scored significantly higher in the quality rating than the control group (both $p < 0.001$). The presented metrics consistently indicate that our DSS, in both variations, efficiently elicits the preferences of the DMs.

6.2. Assessment of cognitive effort

We used one qualitative and one quantitative metric to approximate the cognitive effort of using the algorithm. Fig. 5 presents the seconds (on a log-scale) required to complete stage 1 (manual selection) compared to the two variants of pairwise comparisons in stage 2. The time required to solve the manual selection task in stage 1 (404 s on average) is a multitude higher than the time required to complete the pairwise comparisons (120 s in *Real* and 133 s in *Fictitious*; both differences are statistically significant at $p < 0.001$).

Notably, we asked nominators to select seven validators. Under field conditions, nominators want to fill all sixteen available slots of their preference list, making the required time to do it manually even longer. Crucially, the time required for pairwise comparisons for the algorithm depends on the number of questions, not the number of selected validators. Hence, in most real-world scenarios, this completion time difference between the manual and algorithmic selection processes will grow even larger. There is no statistically significant difference between the time required to solve the pairwise comparison between the algorithms ($p = 0.57$).

To complement the quantitative approximations of cognitive effort, we explicitly asked participants how difficult they perceived the tasks of manual selection and comparing the pairwise comparisons. Pooling the difficulty assessment (see Fig. 6) of participants in both algorithms, we can confirm that they assess the process to be easier than the manual selection ($p < 0.001$). Notably, the

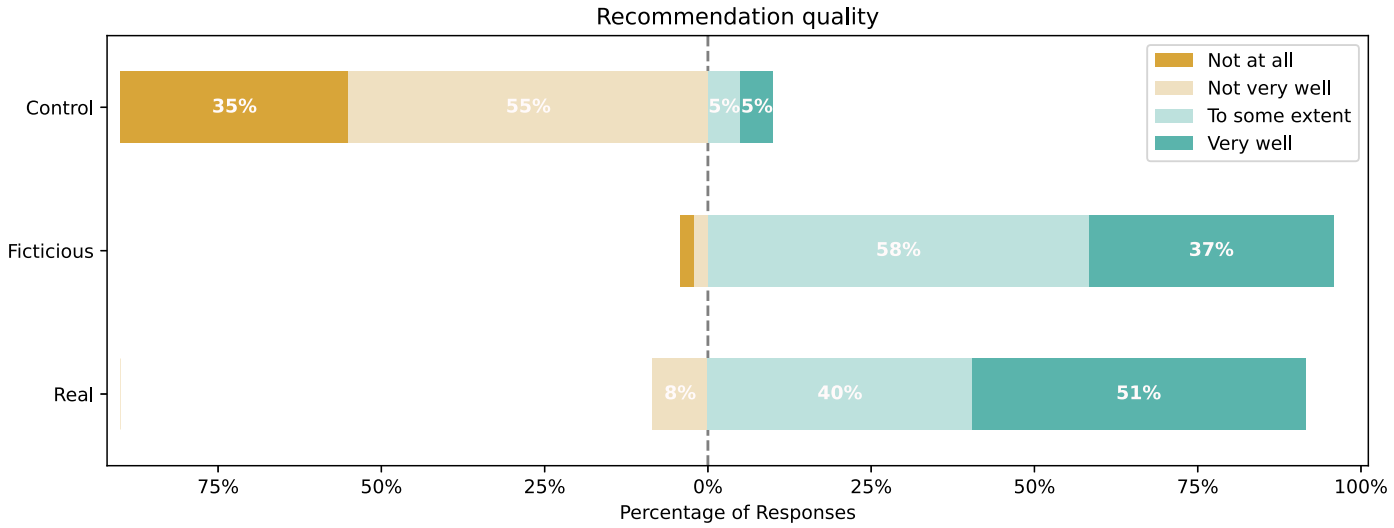


Fig. 4. Answers to the question: "How well does the recommendation suit your preferences?".

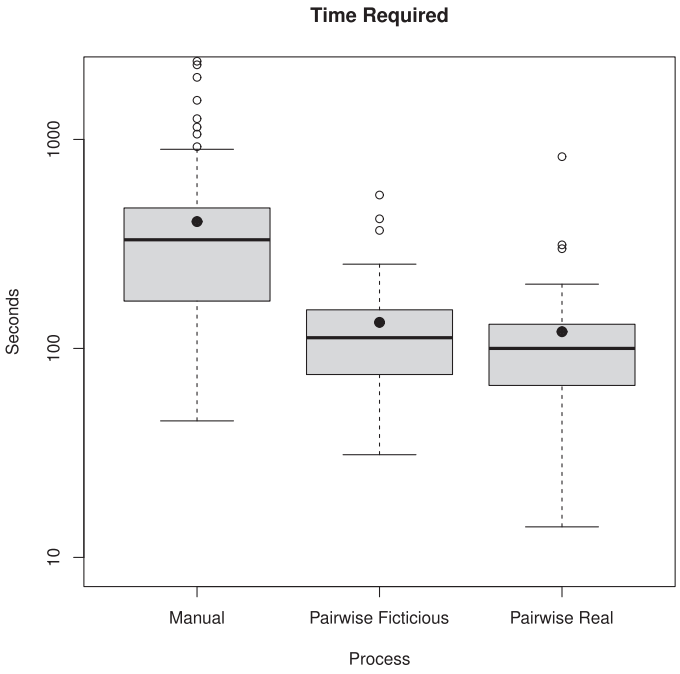


Fig. 5. The time required to complete stage 1 for all participants ("Manual") compared with the time required in stage 2 for the two variations of the algorithm. Values on a logarithmic scale.

pairwise choices in *Real* were perceived as easier than those in *Fictitious* ($p < 0.01$). We can then conclude that the algorithm is less cognitively demanding in terms of the time required and perceived difficulty.

When it comes to the two variants using real and fictitious alternatives, there is a trade-off between cognitive effort and effectiveness. Specifically, although the questions involving fictitious alternatives are more cognitively demanding (see Fig. 6), they are also more effective. Remember that we considered the worst Spearman correlation with respect to the pairs of rankings compatible with $\mathcal{U}(S)$ as a proxy measure of its intrinsic uncertainty. Figure 7 shows the average improvement of this indicator with respect to the number of questions asked for the two proposed variants of the questioning strategy.

The Spearman coefficients are increasing for both real and fictitious alternatives. This confirms that when more questions are asked, the set $\mathcal{U}(S)$ becomes less ambiguous. However, after the third question, fictitious alternatives become more valuable as they offer significantly more precise questions. The fact that the strategy with fictitious alternatives initially performs worse may not be representative because, in the absence of any information, the space of possible questions is relatively large, and most questions are of poor quality. This makes it hard for an optimization algorithm to find a reasonable question within a limited number of evaluations. In turn, solutions made from real alternatives, despite not being optimal, still have better quality than most possible solutions.

6.3. Acceptance of the algorithm

The literature suggests that many human DMs are reluctant to take advice from algorithms, a bias called *algorithm aversion*, despite the clear benefits of such advice [28]. Therefore, it is important to evaluate the acceptance of the DSS. While the participants' responses to how well the algorithm covers their preferences, as presented in Fig. 4, provide evidence that DMs are satisfied with the DSS, we explicitly asked participants which method they would prefer. The results are presented in Fig. 8.

The data illustrate a broad acceptance of the algorithm as a DSS. The results also show that DMs still value manual selection, which is a crucial part of the process because the algorithm lacks certain important aspects of the selection process, namely the reputation and identity of individual operators. We analyzed written responses indicating why they gave the respective answers. While comments were diverse, the most common pattern seemed not to consider the algorithm as opposed to manual selection. In fact, many participants saw the algorithm as an aid to the selection process, which could easily co-exist with (and boost) the manual selection. In this sense, the algorithm achieves what it was envisioned for, as a learning tool for the least expert nominators.

6.4. Staking performance

Table 5 shows the estimated staking rewards (in one-hundredth of a DOT, or cDOT) that the seven validators relevant at each step would generate on 2 DOTs, based on their attributes during the remaining staking period (38 days). The data was updated daily. The average reward from the manual selection was 2.70 cDOT across

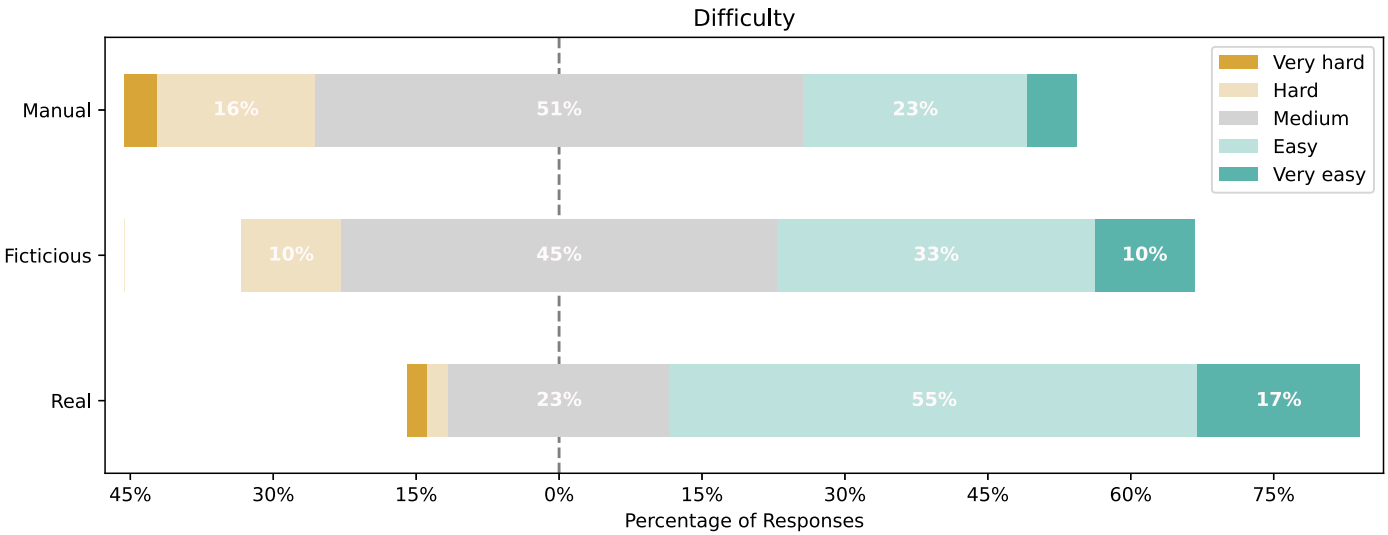


Fig. 6. Answers to the question: "How would you rate the effort it took for [manually selecting validators / the pairwise comparisons]?".

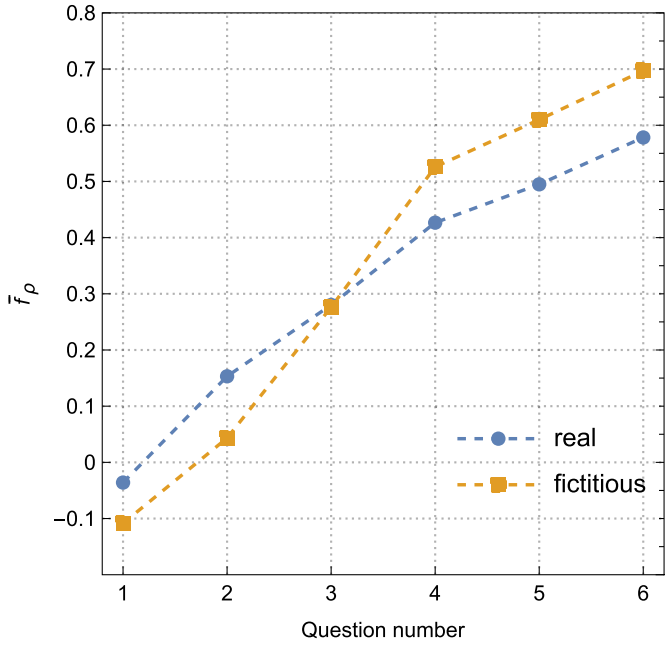


Fig. 7. The average values of the worst Spearman correlation coefficients over the six interaction stages given the treatments involving real and fictitious alternatives (the value $\bar{f}_\rho$ represents the lowest detected Spearman correlation index calculated on 1000 randomly sampled value functions in $\mathcal{U}(S)$).

Table 5
Average estimated staking rewards (in cDOT) at the end of the staking period of the experiment averaged over the seven validators relevant at each step.

	Manual Selection	Recommendation	Final Selection
Control	3.00	0.25	2.72
Real	2.69	3.07	2.80
Fictitious	2.58	3.12	3.01

all treatments. The algorithmic selection proved more profitable by 0.37 and 0.42 cDOT in the case of strategies incorporating real and fictitious alternatives ($p < 0.001$), respectively. The results of the final selection, where nominators chose among the validators derived from the manual screening and the DSS, were slightly better than those of the manual-only selection but slightly worse than those obtained with only the algorithms. The worst outcomes (0.25 cDOT) were obtained for the control group that was provided with random recommendations.

6.5. Preference of nominators

The question-answer interaction tree for either real or fictitious treatment comprises 64 (2^6) paths. An inspection of the answers provided by the nominators reveals that these paths were not chosen equally. Figures 9 and 10 reveal the paths supported by at least five nominators in the fictitious and real treatment, respectively. The validators involved in the pairwise comparisons depicted in these trees are represented as performance vectors $[g_1(a_i), g_2(a_i), \ldots, g_6(a_i)]$. The similarity of preferences for significant shares of nominators participating in the experiment can hardly be explained by randomness. For example, in the root of the tree for the fictitious treatment, 29 out of 48 nominators preferred an alternative with a lower total stake, a smaller cluster size, and a higher commission (see Fig. 9). However, the answer provided by the majority of nominators (18 out of 29) on the most frequent path already opted for a lower commission, given different trade-offs between the remaining criteria than in the initial stage of interaction. Interestingly, even if some paths in the fictitious treatment were supported by a significantly greater number of nominators than others, there was no single complete path composed of six question-answer iterations that would be supported by at least five experiment participants.

The interaction tree of the real treatment is less balanced than that of the fictitious one (see Fig. 10). In particular, 25 of 47 (53%) nominators made the same choices over the first five stages. This observation is even more interesting, given that this path has no obvious preference pattern. For example, even if the nominators opted for a higher self stake in all five questions, they preferred a lower commission when answering three questions, and they favored a higher number of era points in only two cases. The sixth question on this most frequent path splits the group into 12 and 13 nominators. The former sub-group opts for a validator with 1971 voters and a cluster of size 9, whereas the latter prefers a validator with 1673 voters and a cluster of 1.

The answers provided by the nominators are reflected in the corresponding value functions. Figure 11 presents five functions supported by the most significant shares of nominators. The three functions (U_1^R, U_2^R, and U_3^R) representing the preferences of the

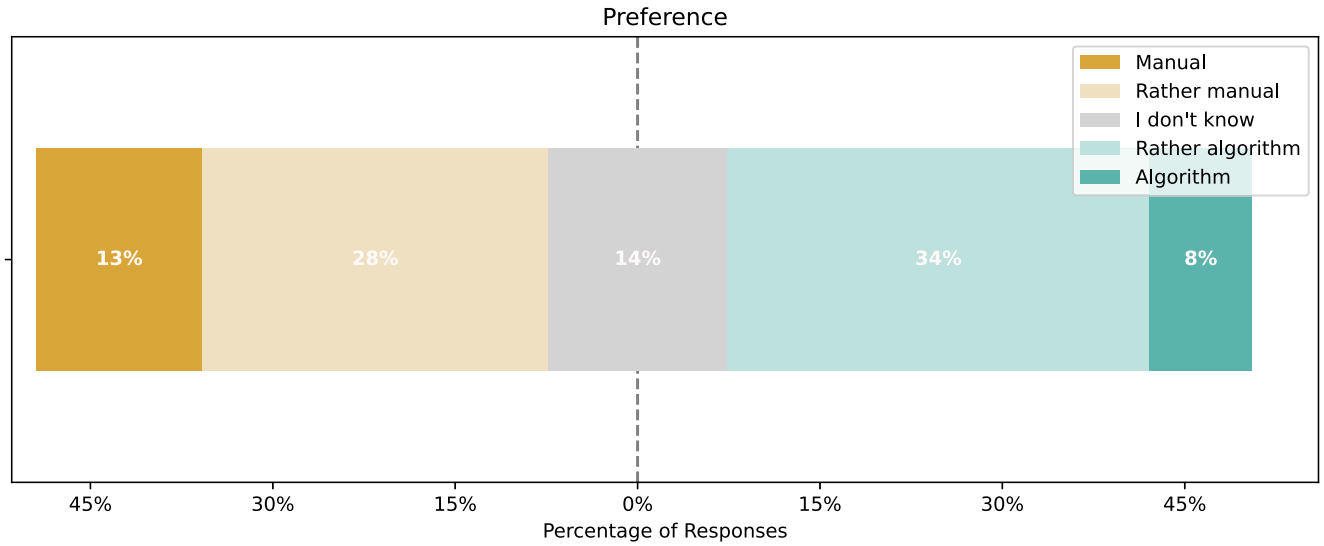


Fig. 8. Answers to the question: "Given a choice, which method would you prefer?".

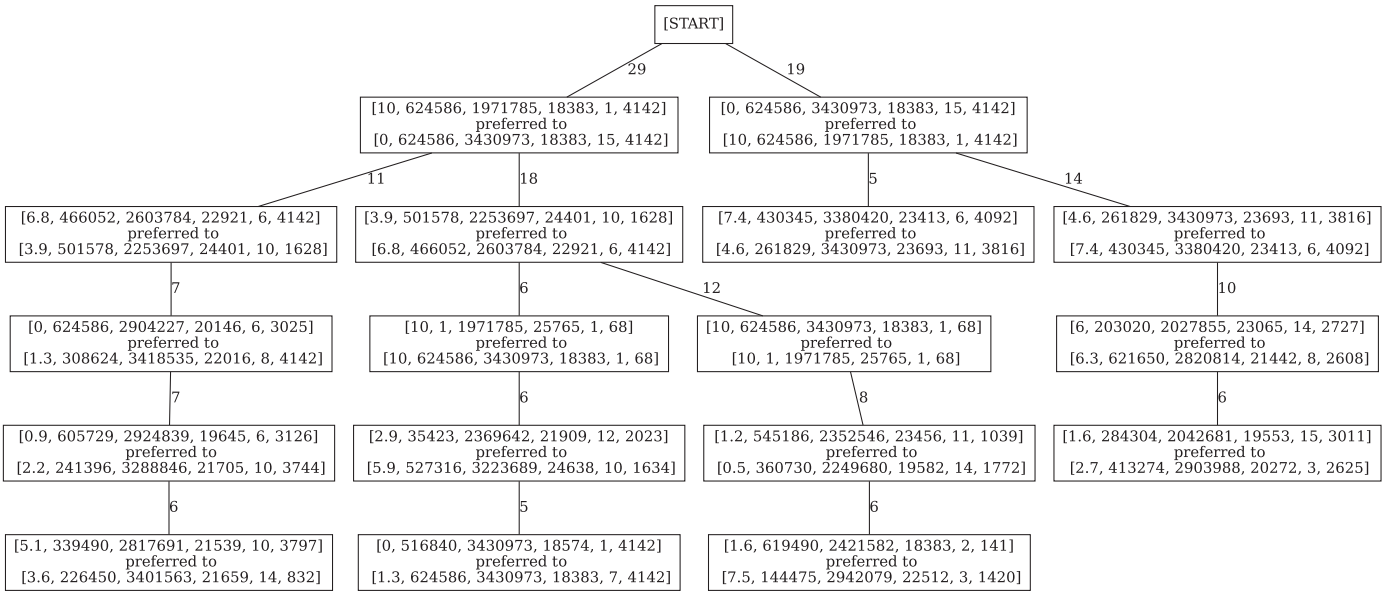


Fig. 9. An interaction tree presenting the most-often-selected question-answer paths with the support of at least five nominators in the fictitious treatment.

greatest subsets of nominators are distinguished with symbols introduced in Fig. 10. Notably, commission, self and total stakes, and era points have a significantly greater impact on the comprehensive value than the remaining two attributes. For the total stake, the differences between maximal shares for the five functions are minor (0.20–0.23), whereas, for the remaining attributes, they are more significant (e.g., 0.17–0.27 for commission and 0.18–0.27 for total stake). The nominators' answers justify these differences. For example, U_3^R is characterized by a more significant impact of era points because, as opposed to U_1^R and U_2^R, the nominators whose preferences are represented by U_3^R emphasized the importance of a greater number of era points in their answers in five of six iterations. The marginal value functions for the cluster size and the number of voters differ greatly among the five most-frequent models. According to the assumptions made, they are all non-monotonic but depending on the nominators' preferences, the non-monotonicity is either of the $\wedge$- or the $\vee$-shape. The reasons underlying these differences can be best explained for U_1^R and U_2^R. The functions are consistent with the same answers provided in the five initial iterations but differ in the response given in the last stage. In particular, the marginal function for cluster size for U_1^R reaches its minimum at 1. In contrast, a marginal value assigned to this performance in U_2^R is the highest because this aspect proved essential for the nominators in the last question-answer interaction.

6.6. Feedback

Participants were asked to respond to an open-ended question with their comments about the study. A sentiment analysis technique [19] was used to transform comments into numbers on a scale from -1 to 1. Numbers below zero mean the text was negative, while numbers above zero correspond with the positive text. Neutral text is transformed into zero. The histogram of these values is presented in Fig. 12. Although many inputs were neutral, there is a strong dominance of positive feedback over negative. Samples of these comments with corresponding sentiment analysis values are listed in Table 6.

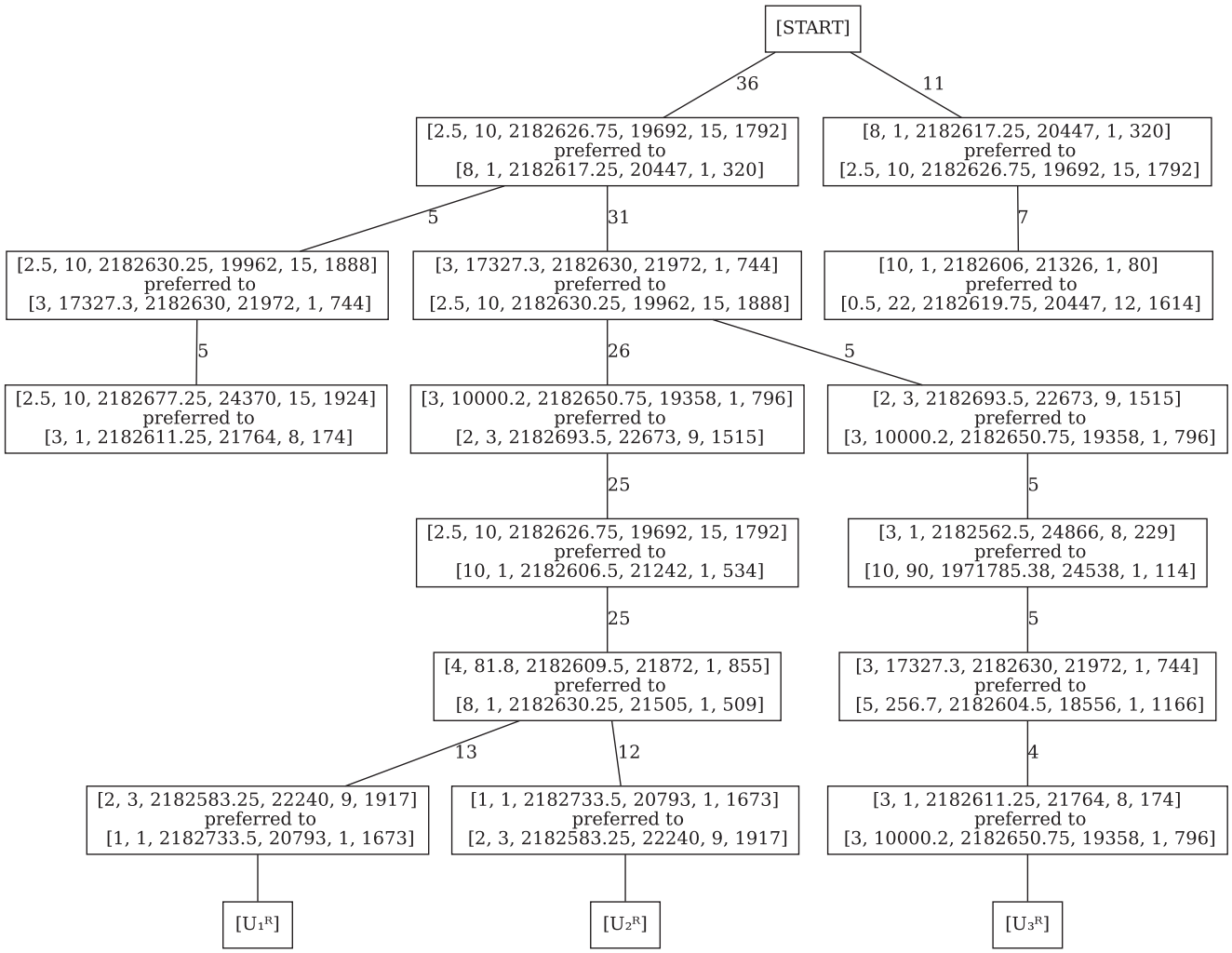


Fig. 10. An interaction tree presenting the most-often-selected question-answer paths with the support of at least five nominators in the real treatment. When representing the choice alternatives, the following order of the attributes is used [Commission, Self stake, Total stake, Era points, Cluster size, Voters].

Table 6
Examples of the participants' comments with the corresponding sentiment analysis value.

Comment	Sentiment
"The number of validators to choose from when manually selecting looks, initially at least, overwhelming. In particular, when you are new to Dotsama and to staking on the js browser. The pairwise route reduces the surface area of the decision space. Less choice is often a relief! Particularly for newbies."	0.57
"The algorithm was especially useful for confirming selections that I had made. Anything that scored AB, I was very confident, was a good selection. I felt like the algorithm was able to understand the similarities between the selections I made."	0.9
"Validators comparison was quite hard, and at some point, I had a feeling it's not going into the expected direction, especially hard choices on very close data, wasn't sure what's important then—like mentioned previously, e.g., voters vs. era points."	−0.24
"I am interested in self stake and the number of clusters, the algorithm did justice to that. Having to pick manually is a bit hectic and might have missed something; the only thing I have against the algorithm is the fact that it did not put the commission into consideration. Would have preferred a relatively higher one, which is why I chose the 'Rather algorithm' option."	0.64

One frequent complaint was the lack of control over the algorithm selection. This problem will not appear in the final implementation of the system because the algorithm will not make the selections but will sort alternatives according to its comprehensive values. To make the process even clearer, plots showing the value functions can be presented to a DM. Some participants also complained about algorithm recommending validators with a too-low commission. In this study, we assumed the commission is a cost-type attribute except for zero commission, which might be suspicious. This assumption seems true for most participants; however, some nominators set this threshold higher and treat a commission below three as suspicious. The best option to solve this issue would be treating commission as a fully non-monotonic attribute, resulting in higher flexibility. Hence, a higher number of questions might be required for the same recommendation quality.

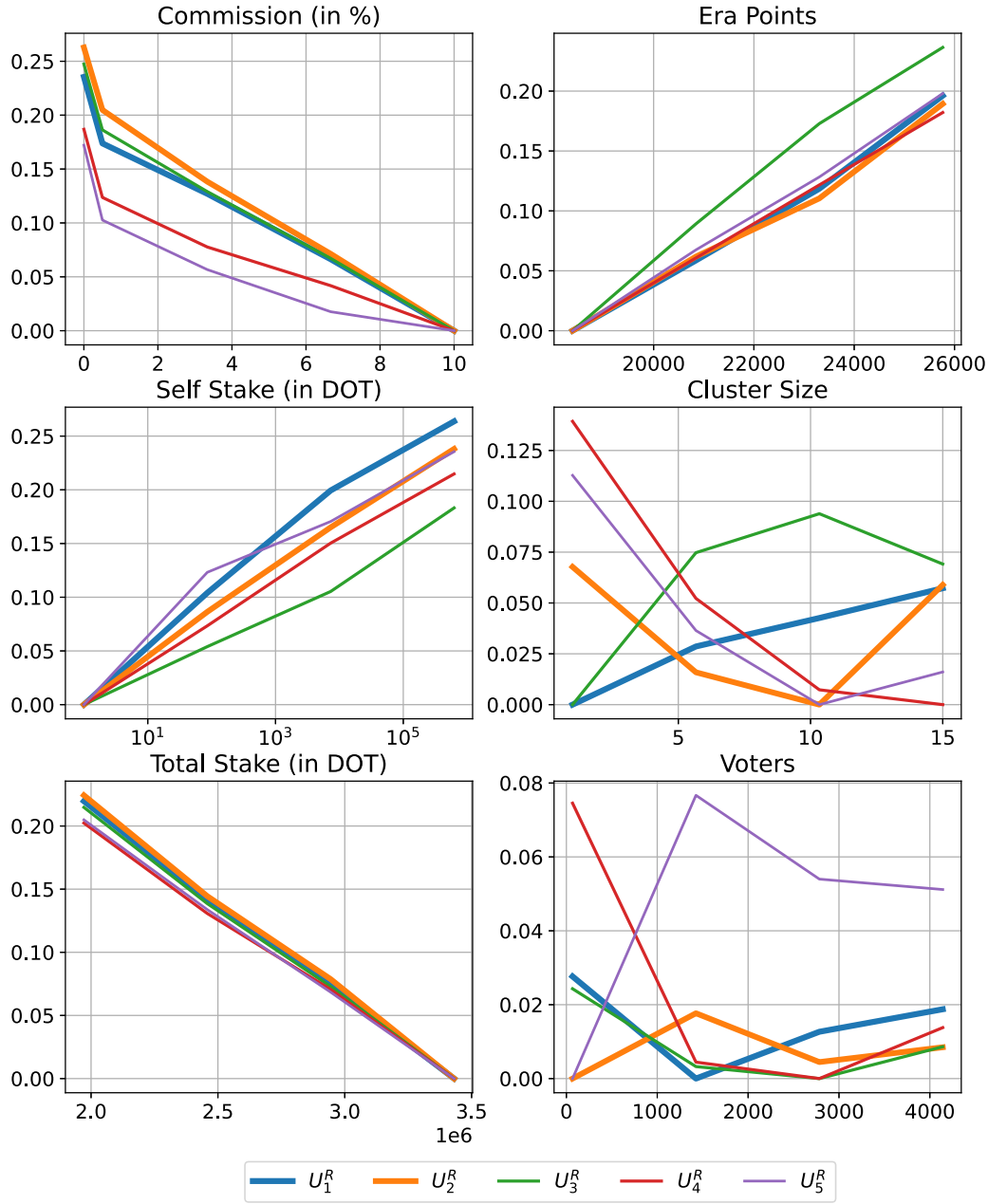


Fig. 11. The value functions representing the preferences of the most significant number of nominators in the real treatment (the thicker the line, the greater the support).

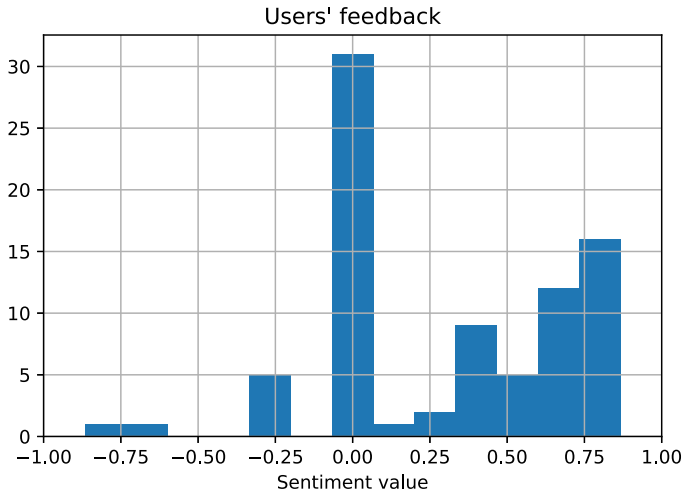


Fig. 12. Histogram of the sentiment analysis values from the participants' feedback.

6.7. Decentralization

A limited number of paths lead to a limited number of possible recommendations. This fact may raise concerns about decentralization because some validators might never appear in the recommendations, regardless of a DM's answers. Figure 13 presents the number of validators included in at least one recommendation, with changing n top-ranked validators. With the number of selected validators equal to seven, as used in this experiment, over 50% of validators have a chance to be recommended. When 16 validators are chosen, which is the limit of the preference list on Polkadot, this fraction exceeds 80%. Assuming the DMs will consider the first 25 validators, the share of potential validators is close to 90%.

Those numbers already propose a sufficiently decentralized recommendation, especially under field conditions. To further mitigate this concern, and in addition to an automatic selection of top n validators with the highest comprehensive value, it is trivial to show

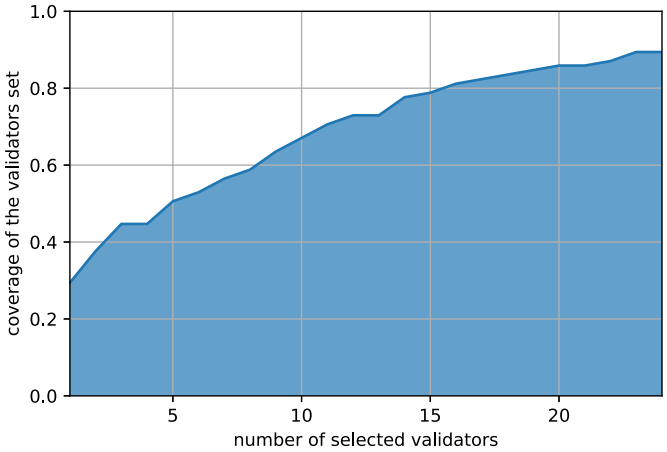


Fig. 13. Fraction of validators recommended at least once when selecting n top-ranked alternatives.

all available validators with their calculated scores. Then, the nominators can make their final selection manually and with the value score as an additional attribute, which does not prevent validators with lower ranks from being selected.

7. Conclusions

In this paper, we proposed a customized active preference learning algorithm for an application in the blockchain domain with frequent decisions and a large set of alternatives. The algorithm's goal is to decompose the problem into smaller problems and learn the preferences of nominators, here acting as DMs. The preferences are captured and represented fully intelligibly and interpretably using an additive value model defined over the family of six relevant attributes. Also, the pairs of alternatives they need to compare are selected or constructed to maximize the information gain interpreted in terms of the similarity between rankings induced by compatible value functions. This enables us to recommend suitable validators who, theoretically, optimally satisfy preferences elicited from the DMs.

To test our theory, we conducted an experiment involving 115 nominators from the Polkadot ecosystem. We compared two variants of the developed algorithm based on pairwise comparisons of either real or fictitious validators against the manual selection and a random recommendation. The results showed that the proposed methods provided more accurate recommendations. Also, they were parsimonious regarding required time and cognitive effort and led to more profit. Furthermore, the analysis of direct feedback provided by the experiment participants suggested that most of them were willing to use the tool on top of manual selection, allowing them to filter a subset of their favorite validators based on personal relationships and reputation.

This paves the way for our proposed DSS to be implemented in the field and used by thousands of DMs who seek aid in their nominations. By offering our tool to real nominators, we can reduce the pain of making qualified nomination decisions that are tailored to individual preferences. This ultimately improves the quality of the active validator set and thereby increases the reliability and security of Polkadot and all its parachains. Additionally, our proposed DSS can easily be deployed on other blockchains that follow similar validator elections that require votes from some other DMs. To do so, the criteria and/or the size of the recommendation set ought to be adjusted for the respective context. For example, the validators of the Kusama network have almost identical criteria as the one in Polkadot but allow nominators to vote for 24 (instead of 16) alternatives. Other Proof-of-Stake blockchains might only allow delegating stake to one validator but, nevertheless, require DMs to choose from a large set of alternatives with competing criteria.

Given the insights of this study, there are several interesting future developments, both theoretically and for practical implementation in the field. First, after making our DSS accessible for all nominators, we could *dynamically extend* the proposed system to incorporate natural changes in the underlying data set. For this, stored preferences from an initial elicitation step could be updated by adding complementary pairwise comparisons that are tailored to cover only the differences in the newest data. This rolling horizon approach guarantees continuously optimal nominations at low costs, which benefits nominators and the Polkadot network. Second, after observing the limited cognitive effort imposed on our DMs by asking for six pairwise comparisons, we could increase that number slightly to seven or eight. That would increase the number of distinct recommendations to 128 and 256, respectively. While the costs of doing so must be considered carefully, given that computational complexity increases exponentially, this further contributes to decentralization. Specifically, this would lead to 86% coverage with one and 94% with two more questions, given we recommend the top 16 ranked validators (the metric as described in Fig. 13). This also leads to a more nuanced recommendation that should fit preferences even better, with the average values of the worst Spearman correlation coefficient reaching 0.74 in the seventh and 0.8 in the eighth question. Third, the attributes used to describe a validator could be improved upon. While our set of attributes made sense for our study to guarantee that DMs are familiar with them, results suggest that certain metrics, such as a total stake or era points, contributed little to the overall utility scores. This was mainly due to the low variance of the metric over the set of validators. Instead of those metrics, we could add others or combine existing ones. For example, participants suggested in the questionnaire to indicate (by yes/no) whether a validator has a verified identity. This would further mitigate the issue of not including the identity of a validator in the algorithm while providing an additional metric important to users. Fourth, we observed some feedback from users who strive for more control over the algorithm. While the general response of our subjects was positive toward algorithmic decision support, we could further increase the acceptance by allowing for customizable features. As demonstrated in Dietvorst et al. [10], this has a further positive impact on the acceptance of the algorithm. Possible candidates for such individualization would be an individually defined set of attributes, filters that constrain the set of considered validators (e.g., remove those above or below a specific threshold for a given attribute), or a choice of how many pairwise comparisons are elicited.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Jonas Gehrlein: Conceptualization, Writing – original draft, Investigation, Funding acquisition, Data curation. **Grzegorz Miebs:** Methodology, Software, Investigation, Formal analysis, Writing – review & editing, Data curation. **Matteo Brunelli:** Conceptualization, Validation, Formal analysis, Writing – original draft. **Miłosz Kadziński:** Conceptualization, Methodology, Writing – original draft, Supervision.

Data availability

The data that has been used is confidential.

Acknowledgment

We thank the Area Editor and three anonymous reviewers for their comments. We gratefully acknowledge the support of the Polkadot Treasury for the financial resources (DOT) to compensate participants of the experiment as well as the Polkadot community members for making choices as our nominators. We further thank Roman Přikryl for designing and setting up the user interface of the experiment. Grzegorz Miebs acknowledges support from the Polish Ministry of Science and Higher Education under the Diamond Grant project (grant no. DI2018 004348). Miłosz Kadziński acknowledges support from the Polish National Science Center under the SONATA BIS project (grant no. DEC-2019/34/E/HS4/00045).

References

[1] Alladi T, Chamola V, Parizi RM, Choo K-KR. Blockchain applications for industry 4.0 and industrial IoT: a review. IEEE Access 2019;7:176935–51.
[2] Bodkhe U, Tanwar S, Parekh K, Khanpara P, Tyagi S, Kumar N, et al. Blockchain for industry 4.0: a comprehensive review. IEEE Access 2020;8:79764–800.
[3] Burdges J., Cevallos A., Czaban P., Habermeier R., Hosseini S., Lama F., Alper H.K., Luo X., Shirazi F., Stewart A., et al. Overview of Polkadot and its design considerations. arXiv preprint arXiv:2005134562020;.
[4] Cevallos A, Stewart A. A verifiably secure and proportional committee election rule. In: 3rd ACM conference on advances in financial technologies; 2021. p. 29–42.
[5] Cinelli M, Kadziński M, Miebs G, Gonzalez M, Słowiński R. Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system. Eur J Oper Res 2022;302(2):633–51.
[6] Ciomek K, Kadziński M, Tervonen T. Heuristics for prioritizing pair-wise elicitation questions with additive multi-attribute value models. Omega (Westport) 2017;71:27–45.
[7] da Rosa Righi R, Alberti AM, Singh M. Blockchain technology for industry 4.0. Springer; 2020.
[8] Deepa N, Pham Q-V, Nguyen DC, Bhattacharya S, Prabadevi B, Gadekallu TR, et al. A survey on blockchain for big data: approaches, opportunities, and future directions. Future Gen Comput Syst 2022;131:209–26.
[9] Dias LC, Morton A, Quigley J. Elicitation: the science and art of structuring judgement. International series in operations research & management science. Springer; 2018.
[10] Dietvorst BJ, Simmons JP, Massey C. Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. Manage Sci 2018;64(3):1155–70.
[11] Eisenführ F, Weber M, Langer T. Rational decision making. Springer; 2010.
[12] Farshidi S, Jansen S, España S, Verkleij J. Decision support for blockchain platform selection: three industry case studies. IEEE Trans Eng Manage 2020;67(4):1109–28.
[13] Filatovas E, Marcozzi M, Mostarda L, Paulavičius R. A MCDM-based framework for blockchain consensus protocol selection. Expert Syst Appl 2022:117609.
[14] Frazier PI. Bayesian optimization. In: Recent advances in optimization and modeling of contemporary problems. INFORMS; 2018. p. 255–78. doi:10.1287/educ.2018.0188.
[15] Ghaderi M, Ruiz F, Agell N. A linear programming approach for learning non–monotonic additive value functions in multiple criteria decision aiding. Eur J Oper Res 2017;259(3):1073–84.
[16] Greco S, Figueira J, Ehrgott M. Multiple criteria decision analysis, vol 37. Springer; 2016.
[17] Jacquet-Lagreze E, Siskos J. Assessing a set of additive utility functions for multicriteria decision-making, the UTA method. Eur J Oper Res 1982;10(2):151–64.
[18] Jacquet-Lagreze E, Siskos Y. Preference disaggregation: 20 years of MCDA experience. Eur J Oper Res 2001;130(2):233–45.
[19] Jongeling R, Datta S, Serebrenik A. Choosing your weapons: on sentiment analysis tools for software engineering research. In: 2015 IEEE international conference on software maintenance and evolution (ICSME). IEEE; 2015.
[20] Kadziński M, Cinelli M, Ciomek K, Coles S, Nadagouda M, Varma R, et al. Co–constructive development of a green chemistry-based model for the assessment of nanoparticles synthesis. Eur J Oper Res 2018;264(2):472–90.
[21] Kadziński M, Martyn K, Cinelli M, Słowiński R, Corrente S, Greco S. Preference disaggregation method for value-based multi-decision sorting problems with a real-world application in nanotechnology. Knowl Based Syst 2021;218:106879.
[22] Kadziński M, Wójcik M, Ciomek K. Review and experimental comparison of ranking and choice procedures for constructing a univocal recommendation in a preference disaggregation setting. Omega (Westport) 2022;113:102715.
[23] Keeney RL, Gregory RS. Selecting attributes to measure the achievement of objectives. Oper Res 2005;53(1):1–11.
[24] Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value trade-offs. Cambridge University Press; 1993.
[25] Kokoska S, Zwillinger D. CRC standard probability and statistics tables and formulae, student edition. CRC Press; 2000.
[26] Kumar A, Liu R, Shan Z. Is blockchain a silver bullet for supply chain management? Technical challenges and research opportunities. Decis Sci 2020;51(1):8–37.
[27] Liu J, Liao X, Huang W, Liao X. Market segmentation: a multiple criteria approach combining preference analysis and segmentation decision. Omega (Westport) 2019;83:1–13.
[28] Mahmud H, Islam AN, Ahmed SI, Smolander K. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. Technol Forecast Soc Change 2022;175:121390.
[29] Mastorakis K, Siskos E. Value focused pharmaceutical strategy determination with multicriteria decision analysis techniques. Omega (Westport) 2016;59:84–96.
[30] Matsatsinis NF, Grigoroudis E, Siskos E. Disaggregation approach to value elicitation. In: Elicitation. Springer; 2018. p. 313–48.
[31] Nakamoto S. Bitcoin: a peer-to-peer electronic cash system. Decentralized Bus Rev 2008:21260.
[32] Saleh F. Blockchain without waste: proof-of-stake. Rev Financ Stud 2021;34(3):1156–90.
[33] Shostak R, Pease M, Lamport L. The byzantine generals problem. ACM Trans Program Lang Syst 1982;4(3):382–401.
[34] Siskos E, Askounis D, Psarras J. Multicriteria decision support for global e-government evaluation. Omega (Westport) 2014;46:51–63.
[35] Siskos Y, Grigoroudis E, Matsatsinis N. UTA Methods. In: Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis: state of the art surveys. Boston: Springer; 2005. p. 297–344.
[36] Smith JE, Dyer JS. On (measurable) multiattribute value functions: an expository argument. Decis Anal 2021;18(4):247–56.
[37] Stewart A., Kokoris-Kogia E.. Grandpa: a byzantine finality gadget. arXiv preprint arXiv:2007015602020;.
[38] Sundarakani B, Ajaykumar A, Gunasekaran A. Big data driven supply chain design and applications for blockchain: an action research using case study approach. Omega (Westport) 2021;102:102452.
[39] Tapscott A, Tapscott D. How blockchain is changing finance. Harv Bus Rev 2017;1(9):2–5.
[40] Tervonen T, van Valkenhoef G, Basturk N, Postmus D. Hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis. Eur J Oper Res 2013;224(3):552–9.
[41] Vetschera R, Weitzl W, Wolfsteiner E. Implausible alternatives in eliciting multi-attribute value functions. Eur J Oper Res 2014;234(1):221–30.
[42] Zheng Z, Xie S, Dai H-N, Chen X, Wang H. Blockchain challenges and opportunities: a survey. Int J Web Grid Serv 2018;14(4):352–75.
[43] Zizzo DJ. Experimenter demand effects in economic experiments. Exp Econ 2010;13(1):75–98.

Extended abstract in Polish

Nowe komputerowe metody wspomagania decyzji odporne na niedoskonałości danych uczących

Wprowadzenie

Problemy decyzyjne, zwłaszcza w zastosowaniach praktycznych, są często wielokryterialne. Obiekty, zwane wariantami lub alternatywami, oceniane są za pomocą przynajmniej dwóch, zazwyczaj wzajemnie sprzecznych, kryteriów. Za najprostszy przykład można podać takie kryteria jak cena i jakość. W większości przypadków są one negatywnie skorelowane – im lepsza (niższa) cena, tym gorsza jakość. W takim przypadku nie jest możliwe obiektywne wskazanie najlepszego rozwiązania, gdyż alternatywy te są nieporównywalne. Za pomocą klasycznych narzędzi matematycznych wyznaczyć można jedynie zbiór Pareto-optymalny. Składa się on z rozwiązań, które nie są zdominowane przez żadną inną alternatywę, jednak względem siebie pozostają nieporównywalne. Aby możliwe było ich rozróżnienie, konieczne jest pozyskanie informacji preferencyjnej od decydenta i uwzględnienie jej w procesie decyzyjnym. Informacja ta odzwierciedla punkt widzenia użytkownika i jego system wartości. Może być ona wyrażona na przykład poprzez przypisanie wag do poszczególnych kryteriów. Metody wielokryterialnego wspomagania decyzji pozwalają na wypracowanie rekomendacji w wielokryterialnych problemach decyzyjnych z użyciem tej dodatkowej informacji od decydenta.

Rzeczywiste problemy decyzyjne często obarczone są różnego rodzaju niedoskonałościami. Niedoskonałość ta może dotyczyć opisu alternatyw, przekazanej informacji preferencyjnej czy też doboru procedury. W niektórych przypadkach możliwa jest próba wyeliminowania problemu u źródła. Jed-

nak często jest to niemożliwe i konieczne jest bazowanie na niedoskonałych danych. W tej sytuacji konieczne jest, aby algorytm był odporny na tego rodzaju niedoskonałości, gdyż w przeciwnym razie mogą mieć one istotny i jednocześnie niechciany wpływ na wypracowaną rekomendację.

Algorytmy wielokryterialnego wspomagania decyzji różnią się rodzajem problemu decyzyjnego, jaki rozwiązują, spodziewanym formatem danych wejściowych oraz wyjściowych czy sposobem interakcji z użytkownikiem. Ze względu na te różnice niemożliwe jest wypracowanie generycznego podejścia do obsługi niedoskonałości związanych z danymi uczącymi. Mechanizm ten musi uwzględniać zarówno rodzaj niedoskonałości, jak i specyfikę konkretnej metody.

W niniejszej pracy doktorskiej zaproponowano podejścia dedykowane konkretnym metodom wielokryterialnego wspomagania decyzji poprzez rozszerzenie lub modyfikację sposobu ich działania. Opracowane zostały też mechanizmy agnostyczne względem konkretnego algorytmu, a dedykowane wybranym rodzajom niedoskonałości. Nowe metody zostały zweryfikowane na rzeczywistych problemach decyzyjnych oraz użyte przez docelowych użytkowników.

1. Odporne metody wyznaczania wag kryteriów

Zaproponowane zostało rozszerzenie procedury Simos-Roy-Figueira (SRF) służącej do wyznaczenia wag kryteriów. W oryginalnym podejściu informacja preferencyjna w postaci rankingu kryteriów oraz dokładnie podanego stosunku ważności kryterium najistotniejszego do najmniej ważnego przekształcana jest w arbitralny sposób do precyzyjnego wektora wag kryteriów zgodnie z ustaloną procedurą. Wynik ten jest tylko jednym spośród potencjalnie nieskończenie wielu wektorów wag spójnych z preferencjami decydenta. Niedoskonałość wynika tu z porządkowej formy informacji preferencyjnej, która nie definiuje wag w sposób precyzyjny. Dodatkowo konieczność specyfikacji dokładnej wartości stosunku wag kryteriów może skutkować podaniem przez decydenta wartości nieodzwierciedlającej w pełni jego preferencji, gdyż jest to zadanie obarczone dużym wysiłkiem kognitywnym.

W opracowanym podejściu wyjściem nie jest jeden konkretny zestaw wag. Co więcej, decydent nie musi precyzyjnie definiować stosunku istotności między pierwszym a ostatnim kryterium. Informacja preferencyjna służy do wyznaczenia przestrzeni wag z nią spójnych. Przestrzeń ta jest następnie próbkowana z wykorzystaniem symulacji Monte Carlo. Symulacja służy wyznaczeniu indeksów akceptowalności, które odwzwierciedlają rozmiar przestrzeni parametrów potwierdzających określony wynik. Następnie przekładają się one na wypracowanie finalnego odpornego rozwiązania.

Podejście to zostało zaadoptowane do dwóch typów problemów. Pierwszy problem to wybór z wykorzystaniem metody ELECTRE I. W tym wariancie

indeksy akceptowalności dotyczyły porównania par wariantów i relacji zachodzących pomiędzy nimi. Posłużyły one do zbudowania grafu relacji przewyższania, z którego następnie wyekstrahowane zostało jądro grafu. Zarówno graf, jak i jego jądro uzyskane w taki sposób, cechują się znacznie większą odpornością w stosunku do podejścia opartego na pojedynczym wektorze wag. Motywacją do opracowania tego rozwiązania był problem wyboru dostawcy usług logistycznych.

Drugim problemem było przypisanie alternatyw do predefiniowanych uporządkowanych klas. Indeksy akceptowalności opisywały przypisanie do klas dla różnych wariantów. Rozważane były więc wszystkie modele spójne z preferencjami decydenta a nie tylko arbitralnie wybrany jeden z nich, potencjalnie prowadzący do skrajnych rozwiązań. Zaadoptowana procedura została wykorzystana z metodą ELECTRE TRI-rC. Algorytm ten został zastosowany do klasyfikacji materiałów izolacyjnych wykorzystywanych w jednym z obszarów wiejskich we Włoszech.

2. Wykorzystanie niepełnej informacji preferencyjnej w postaci przykładowych decyzji

Jedną z form przekazania preferencji jest podanie przykładowych decyzji. W zależności od rodzaju problemu mogą one przyjmować postać na przykład przypisania alternatyw do klas lub porównania parami wybranych alternatyw. Informacja taka z natury jest niedoskonała, gdyż dotyczy niewielkiego podzbioru wszystkich alternatyw. Cechują ją więc niekompletność. Co więcej, zazwyczaj istnieje wiele modeli matematycznych prowadzących do tej samej ewaluacji alternatyw ocenionych przez użytkownika, ale różnią się w ocenie pozostałych wariantów, tzw. niereferencyjnych.

Jednym z efektów pracy doktorskiej jest wypracowanie odpornej procedury indukcji reguł decyzyjnych na podstawie przykładowych przypisań alternatyw do klas. W zaproponowanym podejściu reguły są indukowane niezależnie dla każdego reduktu. Następnie istotność poszczególnych reguł ważona jest jakością klasyfikacji danego reduktu. Podejście to cechuje się większą stabilnością rozwiązania w porównaniu do klasycznego podejścia, gdzie powstaje tylko jeden zbiór równoważnych reguł. Wynikiem opracowanej metody są indeksy akceptowalności przypisań do poszczególnych klas. Dodatkowo algorytm działa poprawnie również w przypadku niepełnej informacji preferencyjnej dotyczącej kierunku preferencji dla poszczególnych kryteriów. Decydent nie musi określać typu kryteriów. Jeśli tego nie zrobi, kryterium jest traktowane jako niemonotoniczne lub interpretowane jest jako atrybut, na którym oceny są tylko różne lub takie same, bez rozważenia kierunku preferencji.

Modele budowane są na podstawie danych podanych przez użytkownika. W związku z tym im więcej przykładowych decyzji zostanie dostarczonych,

tym większa będzie precyzja stworzonego modelu, a tym samym wypracowana rekomendacja cechować będzie się wyższą jakością. Niestety, rozmiar podanych decyzji jest ściśle powiązany z czasem i wysiłkiem użytkownika. Wraz ze wzrostem rozmiaru danych zwiększa się zmęczenie procesem elicytacji, co przełożyć może się na mniejsze skupienie, a w efekcie podanie informacji niespójnej z rzeczywistymi preferencjami. Obserwacja ta była jedną z motywacji do opracowania kolejnej procedury będącej efektem tej rozprawy – algorytmu aktywnego uczenia dla metody UTA, gdzie informacją preferencyjną są dokonane przez decydenta porównania alternatyw parami.

Zaproponowany został algorytm generujący pytania w postaci pary alternatyw do porównania przez użytkownika. Pytanie to jest wynikiem działania algorytmu optymalizacji, którego celem jest maksymalizacja zysku informacyjnego uzyskanego po odpowiedzi na to pytanie. Dzięki zastosowaniu tej procedury wymagane jest mniej odpowiedzi od użytkownika do wytworzenia modelu o wysokiej jakości w porównaniu do klasycznego podejścia, w którym to decydent sam tworzy pytania i udziela na nie odpowiedzi. Co więcej, opracowano również metodę, w której pytania bazują nie na rzeczywistych alternatywach, lecz na sztucznych obiektach wygnerowanych tylko i wyłącznie w tym celu. Pozwala to na jeszcze dalej idącą redukcję koniecznego rozmiaru informacji preferencyjnej przekazanej przez użytkownika.

Oba warianty zostały przetestowane na zbiorze ponad 100 rzeczywistych decydentów, będących ekspertami w dziedzinie, której dotyczył dany problem. W tym przypadku był to wybór walidatorów kryptowalut. Osoby te nie miały jednak żadnej wiedzy związanej z algorytmi wielokryterialnego wspomagania decyzji. Mimo tego były one w stanie skutecznie posługiwać się opracowanym narzędziem, co dowodzi niskiemu poziomowi wysiłku kognitywnego z tym związanego. Istotność tych obserwacji zwiększa fakt, iż w przeprowadzonym eksperymencie użytkownicy dysponowali pieniędzmi, które były inwestowane przez określony okres czasu zgodnie z rekomendacjami wypracowanymi przez nowatorski algorytm. Motywowało to więc do zachowań racjonalnych i maksymalizacji zysku.

3. Dobór odpowiedniej metody wspomagania decyzji

Na przestrzeni lat opracowane i udokumentowane zostały setki metod wielokryterialnego wspomagania decyzji wraz z ich rozszerzeniami i wariantami. Każdy z nich dedykowany jest konkretnej klasie problemów decyzyjnych oraz wymaga danych wejściowych konkretnego rodzaju o ściśle określonym formacie i typie. Ze względu na mnogość algorytmów oraz ściśle określone warunku ich użycia, sam problem wyboru metody wspomagania decyzji można uznać za problem wielokryterialnego wyboru, w którym to użytkownik potrzebuje wsparcia. Niedoskonały charakter danych w tym problemie wynika z ograniczonej możliwości zdefiniowania wymagań przez użytkownika.

W celu rozwiązania tego problemu opracowany i zaimplementowany został system MCDA-MSS, który wspiera wybór odpowiedniej metody, bazując na opisie problemu oraz oczekiwań użytkownika. W systemie tym uwzględniono ponad 200 metod wielokryterialnego wspomagania decyzji, a każda z nich została opisana na ponad 150 atrybutach. Naturalny wydaje się fakt, że w większości przypadków użytkownik nie będzie w stanie precyzyjnie określić swoich wymagań na każdym z nich. Co więcej, wyraźnie widać, iż liczba metod nie jest w stanie pokryć wszystkich możliwych kombinacji wartości na atrybutach. Z tego powodu możliwa jest sytuacja, w której nie istnieje żadna metoda w pełni odpowiadająca preferencjom decydenta. Opracowany system adresuje obie opisane sytuacje.

W przypadku braku metody spełniającej wszystkie kryteria podane przez użytkownika następuje identyfikacja kluczowych ograniczeń, które muszą być spełnione, oraz tych, które mogą być pominięte, jeśli to konieczne. Procedura ta jest kontynuowana aż do momentu, w którym istnieje przynajmniej jedna metoda spełniająca wszystkie kluczowe ograniczenia. System prowadzi użytkownika w taki sposób, aby uniknąć sytuacji odwrotnej – zbyt szczegółowe ograniczenia, których nie spełnia żadna metoda. Co więcej, mechanizm ten pozwala na określenie obszarów, w których brakuje odpowiednich metod, co pozwala więc na odpowiednie ukierunkowania badań nad opracowaniem nowych algorytmów.

W sytuacji gdy opis wymagań jest niepełny, system jest w stanie zwrócić wszystkie metody spełniające zadane ograniczenia. Jednocześnie rekomendowane jest pytanie, które pozwoli na maksymalne zawężenie obszaru poszukiwań i zmniejszenie liczby pasujących algorytmów. Każde pytanie oraz atrybut zostały szczegółowo i w wyczerpujący sposób opisane, co ułatwia użytkownikom przekazanie preferencji bez posiadana szerokiej specjalistycznej wiedzy na temat poszczególnych metod. System posiada też inne mechanizmy interakcji z użytkownikiem ułatwiającej korzystanie. Na bieżąco pokazywana jest liczba odpowiadających metod, użytkownik jest też świadomy, jak liczba ta zmieni się po podaniu kolejnych ograniczeń.

System ten pozwala zmniejszyć liczbę sytuacji, w których wybrana została nieodpowiednia metoda lub wykorzystana została w niewłaściwy sposób. Może to przyczynić się do zwiększenia popularności całego obszaru wielokryterialnego wspomagania decyzji, gdyż decydent bez większej wiedzy w tej dyscyplinie będzie w stanie wybrać właściwą metodę do danego problemu i odpowiednio ją zastosować.

4. Wspomaganie grupowego podejmowania decyzji przy niespójnych preferencjach decydentów

W wielu rzeczywistych przypadkach w problemie decyzyjnym występuję więcej niż jeden decydent. Co więcej, mogą oni reprezentować różne punkty

widzenia, przez co ich preferencje są wzajemnie sprzeczne. Prowadzi to do niespójnej informacji preferencyjnej, która musi być odpowiednio przetworzona, a finalna rekomendacja jest rozwiązaniem kompromisowym.

W ramach obszaru wielokryterialnego wspomagania decyzji opracowane zostało wiele metod. Kilka z nich stało się bardzo popularne, doczekało się licznych rozszerzeń oraz zastosowań. Są one dobrze przetestowane i obdarzone zaufaniem przez wielu użytkowników. Niestety, w większości są one dedykowane problemom decyzyjnym w których występuje tylko jeden decydent, stąd nie można ich użyć wprost do rozwiązania problemów grupowego wspomagania decyzji.

Celem rozwiązania zaproponowanego w ramach tej pracy doktorskiej było stworzenie procedury, w ramach której metody odpowiednie dla problemów z jednym decydentem, będą mogły być użyte dla problemów z wieloma decydentami. Jest to dużo bardziej uniwersalne podejście niż tworzenie kolejnej metody dedykowanej jednemu specyficznemu typowi problemów. Pozwala to na wykorzystanie dobrze przetestowanych i sprawdzonych metod w szerszym kontekście. Dodatkowo decydenci nie muszą zapoznawać się z charakterystyką nowej metody, gdyż mogą nadal korzystać z tej, którą używali do tej pory.

W opracowanym rozwiązaniu skupiono się na problemie, którego wynikiem jest ranking. Rozważono zarówno ranking zupełny jak i częściowy, w którym dopuszczona jest nieporównywalność alternatyw. W podejściu tym dowolna procedura zwracająca ranking aplikowana jest niezależnie dla każdego decydenta, a faza agregacji odbywa się już na wynikowych rankingach, bez ingerencji w sam proces ich powstawania. Problem ten został zamodelowany jako zadanie optymalizacyjne, którego celem jest znalezienie rankingu najbliższemu rankingom uzyskanym dla wszystkich decydentów. Rozważone zostały dwie funkcje celu tego rodzaju: podejście utylitarne, w którym szukamy rozwiązania minimalizującego średnią odległość od wszystkich rankingów, oraz podejście egalitarne, w którym minimalizowana jest maksymalna odległość

Do rozwiązania tego zadania optymalizacyjnego użyte zostały odpowiednio zaadoptowane algorytmy metaheurystyczne takie jak symulowane wyżarzanie czy podejścia genetyczne, oraz metody opracowane specjalnie dla tego problemu. Podejścia te zostały przetestowane na szerokiej bazie rankingów o specyficznych cechach wygenerowanych specjalnie na problemy tego testu, a także na rzeczywistych danych pochodzących z realnego problemu decyzyjnego. Uzyskane wyniki cechowały się wysoką jakością, a także efektywnością, jeśli chodzi o czas obliczeń. Z tego względu możliwe jest także zastosowanie tej procedury nie tylko dla problemów grupowego podejmowania decyzji, gdzie łączna liczba rankingów z reguły nie przekracza kilkudziesięciu, ale też dla rankingów pochodzących z analizy odporności rozwiązań, które mogą być liczone w tysiącach.

Zaproponowane zostało również podejście do grupowego wspomagania decyzji dla problemu sortowania. Motywowane ono było konkretnym przypadkiem użycia w problemie z ponad 30 decydentami. Rozwiązanie to bazowało na agregacji poprzez uśrednianie indeksów akceptowalności uzyskanych dla każdego decydenta niezależnie. W ten sposób wypracowana została rekomendacja kompromisowa.

5. Sieci neuronowe odporne na niedoskonałe dane

Procesy wspomagania podejmowania decyzji mogą wykorzystywać także algorytmy uczenia maszynowego. Podejścia te bazują niemal wyłącznie na danych treningowych. Skutkuje to tym, iż jakość danych finalnie przekłada się na jakość modelu. Jest to zwłaszcza widoczne we współczesnych sieciach neuronowych, które, aby odblokować swój pełny potencjał, potrzebują dużych wolumenów danych treningowych o dobrej jakości. W przeciwnym wypadku modele te są w stanie nauczyć się niedoskonałości występujących w danych i traktować je jako rzeczywisty sygnał. Z racji powszechności tego modelu opracowane zostało wiele technik mających na celu minimalizację ryzyka wystąpienia takiej sytuacji. Niestety w wielu sytuacjach są one niewystarczające.

W ramach przedmiotowej pracy zaproponowane zostały dwie rekurencyjne architektury neuronowe, które cechują się większą odpornością od podejść klasycznych. W obu przypadkach było to motywowane rzeczywistym problemem decyzyjnym do rozwiązania. Pierwszy z nich polegał na predykcji konsumpcji energii elektrycznej dla przedsiębiorstw różnego typu. Z racji na dużą różnorodność standardowym podejściem byłby trening niezależnego modelu dla każdego z przedsiębiorstw. W podejściu tym w przypadku błędnych danych dla jednego podmiotu model nie miałby szans na poprawne działanie. W tym celu zaproponowana została specjalna architektura rekurencyjnej sieci neuronowej, w której jednocześnie wraz z szeregiem czasowym, którym były w tym wypadku dane z licznika zużycia prądy, przetwarzał cechy statyczne charakteryzujące ten szereg takie jak rozmiar i typ przedsiębiorstwa czy jego lokalizacja. Dzięki temu podejściu możliwe było wyeksponowanie modelu jednocześnie na wiele przedsiębiorstw, gdyż dodatkowe cechy statyczne pozwalały na rozróżnienie między nimi. Dodatkowo możliwa była interpolacja z użyciem tych wartości, więc nawet jeśli dla jednej firmy dane były błędne, to model był w stanie dokonać poprawnej predykcji, bazując na krzywych zużycia energii dla sąsiednich wartości. Warto zaznaczyć, iż mimo faktu, że podejście to opracowane było na potrzeby konkretnego problemu to można je uogólnić i stosować dla problemów z innej domeny o zbliżonej charakterystyce.

Dobrze świadczy o tym fakt, że architektura ta została wykorzystana jako składowa kolejnego modelu opracowanego w ramach tej pracy doktorskiej,

której celem była detekcja i lokalizacja wycieków w rurociągach transportujących paliwa ciekłe. Większość metod do tej problematyki jest wrażliwa na różnego rodzaju niedoskonałości danych. W praktyce możliwe są problemy z czujnikami rozmieszczonymi wzdłuż rurociągu, lub z komunikacją. Opracowana została architektura sieci neuronowej, w której jako wejście podawany jest zbiór obserwacji wraz z ich opisem, w tym wypadku była to lokalizacja czujnika na rurociągu oraz czas pomiaru. Na tej podstawie model buduje w tak zwanej zmiennej ukrytej obraz aktualnej sytuacji na rurociągu. W drugiej fazie obraz ten może zostać wykorzystany i po zadaniu zapytania, również w postaci lokalizacji i czasu, model generuje odpowiadający mu szereg czasowy. Ze względu na fakt, iż jako wejście podana może być dowolna liczba obserwacji, model ten jest odporny na problemy z jednym lub nawet kilkoma sensorami. Wystarczy w tym celu użyć niezależnie kilku kombinacji czujników. Również ten model może z powodzeniem być stosowany w problemach z innej domeny. Jako przykład może posłużyć jego zastosowanie do generowania cząsteczek chemicznych o zadanych właściwościach.

6. Wielokryterialna optymalizacja przy braku informacji preferencyjnej

Kolejnym rozważonym problemem była optymalizacja wielokryterialna przy braku informacji preferencyjnej. W tym celu zaadoptowany został algorytm genetycznej optymalizacji wielokryterialnej NSGA-II, tak aby znaleźć front Pareto, zawierający szeroki przekrój rozwiązań. Następnie na tym froncie zastosowano wariant metody ELECTRE dostosowany do tego typu problemu oraz formy przekazania informacji preferencyjnej, która odpowiadała decydentowi.

Podejście to zostało zastosowane do optymalizacji pracy kierowców i motorniczych w poznańskiej komunikacji miejskiej. Algorytm ten służy do wyznaczania grafików pracy (dni wolne, przydział zmian – rano, popołudnie, noc) oraz przypisania pracowników do konkretnych służb (linii tramwajowych lub autobusowych), optymalizując jednocześnie kilka kryteriów ekonomiczno-społecznych oraz spełniając liczne ograniczenia wynikające z kodeksu pracy czy wewnętrznych przepisów spółki.

Podsumowanie

Celem niniejszej rozprawy doktorskiej było stworzenie metod komputerowego wspomagania decyzji odpornych na różnego rodzaju niedoskonałości danych uczących. Metody te motywowane były charakterystyką rzeczywistych problemów decyzyjnych, skupiając się na konkretnym typie niedoskonałości danych.

Opracowane zostały metody wspierające odporne metody wyznaczania wag kryteriów. Zamiast jednego wektora wag rozważana była cała przestrzeń wag spójnych z preferencjami decydenta, co pozwoliło na uniknięcie arbitralnego wyboru, który mógł prowadzić do skrajnych rekomendacji. Dzięki temu rozwiązano problem istnienia niedoskonałości związanych z istnieniem wielu wektorów wag, z których każdy mógłby zostać wykorzystany do wypracowania ostatecznego wyniku.

Powstały również algorytmy bazujące na jednej z najprostszych form informacji preferencyjnej, którą są przykładowe decyzje. Dane te z natury są niekompletne; stąd wynika ich niedokładność. Jednak zaproponowane mechanizmy pozwalają wykorzystać je w zoptymalizowany i odporny sposób. Wykorzystanie technik aktywnego uczenia pozwala na wygenerowanie pytań maksymalizujących zysk informacyjny, a tym samym zmniejszenie liczby koniecznych przykładów decyzji. Zaproponowano także wariant indukcji reguł dla zbiorów przybliżonych, w których rozważane są podzbiory kryteriów, co pozwala na wygenerowanie kilku zestawów reguł. Mogą być one następnie wykorzystane do odpornego sortowania.

Kolejnym rozważonym tematem było wsparcie wyboru odpowiedniej procedury dla danego problemu decyzyjnego. Ze względu na mnogość algorytmów i ich wysoką specyfikację co do obszarów zastosowań dobór odpowiedniej procedury nie jest trywialny i wymaga wsparcia. Dzięki bazie danych zawierającej ponad 200 metod opisanych przy użyciu ponad 150 cech oraz interaktywnemu systemowi prowadzącemu przez proces wyboru nawet osoba bez wiedzy i doświadczenia w zakresie komputerowego wspomagania decyzji jest w stanie dobrać odpowiedni algorytm do swojego problemu.

Grupowe wspomaganie decyzji to następny temat poruszony w tej pracy doktorskiej. Zaproponowane zostało podejście, pozwalające na użycie dowolnych metod wspomagania decyzji rozwiązujących problem rankingu w ujęciu grupowym. Przy wykorzystaniu algorytmów optymalizacyjnych poszukiwany jest ranking kompromisowy minimalizujący odległość od rankingów uzyskanych dla poszczególnych decydentów. Rozważone zostało podejście utylitarne z minimalizacją średniej odległości oraz egalitarne z minimalizacją maksymalnej odległości. Opracowano także metodę odpornej analizy akceptowalności rozwiązań dla problemu sortowania. W pierwszym kroku analiza ta przeprowadzana jest niezależnie dla każdego decydenta, następnie na uzyskanych podczas niej indeksach akceptowalności następuje agregacja do grupowej rekomendacji.

Wreszcie rozważono również obszary poza głównym nurtem komputerowego wspomagania decyzji. Jednym z nich są sztuczne sieci neuronowe. Zaprojektowano dwie architektury, pozwalające na użycie szerszego zakresu danych treningowych oraz cechujące się większą odpornością na potencjalne niedoskonałości od podejść klasycznych. Opracowane zostało też podejście do zagadnienia optymalizacji przy braku preferencji. W tym celu wykorzystano

adaptacje istniejących algorytmów do rozwiązania postawionego problemu.

Większość badań przeprowadzonych w trakcie prac na rozprawą doktorską motywowanych było rzeczywistymi problemami decyzyjnymi i związanymi z nimi niedoskonałościami danych uczących. Co więcej, wiele z wypracowanych rozwiązań jest obecnie stosowanych w praktyce. Podkreśla to wysoki potencjał aplikacyjny stworzonych algrytmów.

Declarations

August 16, 2023

Marco Cinelli
Faculty Governance and Global Affairs
Leiden University College
The Hague, Netherlands

Declaration

I hereby declare the following contribution as an author of the following paper:

Cinelli, M., Kadziński, M., Miebs, G., Gonzalez, M., and Słowiński, R. (2022). Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system. In European Journal of Operational Research (Vol. 302, Issue 2, pp. 633–651)

- Co-authorship of the idea underlying the paper
- Co-authorship of the database with the MCDA methods' descriptions
- Co-authorship of the text of the publication

Marco Cinelli

Marco Cinelli

August 16, 2023

Rafał A. Bachorz
Advanced Analytics Team
PSI Poland, Towarowa 35
61-896, Poznań, Poland

Declaration

I hereby declare the following contribution as an author of the following papers:

Miebs, G., Mochol-Grzelak, M., Karaszewski, A., and Bachorz, R. A. (2020). Efficient Strategies of Static Features Incorporation into the Recurrent Neural Network. In Neural Processing Letters (Vol. 51, Issue 3, pp. 2301–2316). Springer Science and Business Media LLC.

- Co-authorship of the idea underlying the paper
- Consultation of the algorithms incorporating static metadata into a recurrent neural network
- Consultation of the context of case study and interpretation of its results
- Co-authorship of the text of the publication

Jurczak, M., Miebs, G., and Bachorz, R. A. (2022). Multi-criteria human resources planning optimisation using genetic algorithms enhanced with MCDA. In Operations Research and Decisions (Vol. 32, Issue 4). Politechnika Wroclawska Oficyna Wydawnicza.

- Consultation of the algorithm and results analysis
- Corrections to the first draft of the manuscript

Miebs, G., Wójcik, M., Karaszewski, A., Mochol-Grzelak, M., Wawdysz, P., and Bachorz, R. A. (2022). Predicting a Time-Dependent Quantity Using Recursive Generative Query Network. In International Journal of Neural Systems (Vol. 32, Issue 11). World Scientific Pub Co Pte Ltd

- Consultation of the neural network architecture and interpretation of its results
- Consultation of the context of case study and interpretation of its results
- Co-authorship of the text of the publication

Rafał A. Bachorz

August 23, 2023

Michael Gonzalez Ph.D.
U.S. Environmental Protection Agency
Office of Research and Development
26 West Martin Luther King Dr.
Cincinnati, 45268, OH, USA

Declaration

I hereby declare the following contribution as an author of the following paper:

Cinelli, M., Kadziński, M., Miebs, G., Gonzalez, M., and Słowiński, R. (2022). Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system. In European Journal of Operational Research (Vol. 302, Issue 2, pp. 633–651)

- Consultation of the system and methods' descriptions
- Editing of the manuscript

MICHAEL GONZALEZ

Digitally signed by MICHAEL GONZALEZ
Date: 2023.08.23 12:22:18 -04'00'

Michael Gonzalez

August 16, 2023

Luisa Paolotti
Department of Agricultural, Food and Environmental Sciences
University of Perugia
Borgo XX Giungno, 74, 06121
Perugia, Italy

Declaration

I hereby declare the following contribution as an author of the following papers:

Kadziński, M., Rocchi, L., Miebs, G., Grohmann, D., Menconi, M. E., and Paolotti, L. (2017). Multiple Criteria Assessment of Insulating Materials with a Group Decision Framework Incorporating Outranking Preference Model and Characteristic Class Profiles. In Group Decision and Negotiation (Vol. 27, Issue 1, pp. 33–59). Springer Science and Business Media LLC.

- Consultation of the context of case study and interpretation of its results
- Corrections to the first draft of the manuscript

Rocchi, L., Kadziński, M., Menconi, M. E., Grohmann, D., Miebs, G., Paolotti, L., and Boggia, A. (2018). Sustainability evaluation of retrofitting solutions for rural buildings through life cycle approach and multi-criteria analysis. In Energy and Buildings (Vol. 173, pp. 281–290).

- Consultation of the context of case study and interpretation of its results
- Editing of the manuscript

Luisa Paolotti

August 16, 2023

Antonio Boggia
Department of Agricultural, Food and Environmental Sciences
University of Perugia
Borgo XX Giungno, 74, 06121
Perugia, Italy

Declaration

I hereby declare the following contribution as an author of the following paper:

Rocchi, L., Kadziński, M., Menconi, M. E., Grohmann, D., Miebs, G., Paolotti, L., and Boggia, A. (2018). Sustainability evaluation of retrofitting solutions for rural buildings through life cycle approach and multi-criteria analysis. In Energy and Buildings (Vol. 173, pp. 281–290).

- Consultation of the context of case study and interpretation of its results
- Editing of the manuscript

Antonio Boggia

August 16, 2023

David Grohmann
Department of Agricultural, Food and Environmental Sciences
University of Perugia
Borgo XX Giungno, 74, 06121
Perugia, Italy

Declaration

I hereby declare the following contribution as an author of the following papers:

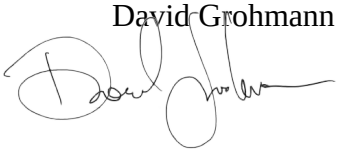
Kadziński, M., Rocchi, L., Miebs, G., Grohmann, D., Menconi, M. E., and Paolotti, L. (2017). Multiple Criteria Assessment of Insulating Materials with a Group Decision Framework Incorporating Outranking Preference Model and Characteristic Class Profiles. In Group Decision and Negotiation (Vol. 27, Issue 1, pp. 33–59). Springer Science and Business Media LLC.

- Consultation of the context of case study and interpretation of its results
- Corrections to the first draft of the manuscript

Rocchi, L., Kadziński, M., Menconi, M. E., Grohmann, D., Miebs, G., Paolotti, L., and Boggia, A. (2018). Sustainability evaluation of retrofitting solutions for rural buildings through life cycle approach and multi-criteria analysis. In Energy and Buildings (Vol. 173, pp. 281–290).

- Consultation of the context of case study and interpretation of its results
- Editing of the manuscript

David Grohmann



August 16, 2023

Lucia Rocchi
Department of Agricultural, Food and Environmental Sciences
University of Perugia
Borgo XX Giungno, 74, 06121
Perugia, Italy

Declaration

I hereby declare the following contribution as an author of the following papers:

Kadziński, M., Rocchi, L., Miebs, G., Grohmann, D., Menconi, M. E., and Paolotti, L. (2017). Multiple Criteria Assessment of Insulating Materials with a Group Decision Framework Incorporating Outranking Preference Model and Characteristic Class Profiles. In Group Decision and Negotiation (Vol. 27, Issue 1, pp. 33–59). Springer Science and Business Media LLC.

- Co-authorship of the concept of sorting algorithms and of the considered case study (Section 3)
- Co-authorship of the text of the publication

Rocchi, L., Kadziński, M., Menconi, M. E., Grohmann, D., Miebs, G., Paolotti, L., and Boggia, A. (2018). Sustainability evaluation of retrofitting solutions for rural buildings through life cycle approach and multi-criteria analysis. In Energy and Buildings (Vol. 173, pp. 281–290).

- Co-authorship of the concept of to the considered case study
- Data collection and analysis of results
- Co-authorship of the text of the publication

Lucia Rocchi

August 16, 2023

Maria Elena Menconi
Department of Agricultural, Food and Environmental Sciences
University of Perugia
Borgo XX Giungno, 74, 06121
Perugia, Italy

Declaration

I hereby declare the following contribution as an author of the following papers:

Kadziński, M., Rocchi, L., Miebs, G., Grohmann, D., Menconi, M. E., and Paolotti, L. (2017). Multiple Criteria Assessment of Insulating Materials with a Group Decision Framework Incorporating Outranking Preference Model and Characteristic Class Profiles. In Group Decision and Negotiation (Vol. 27, Issue 1, pp. 33–59). Springer Science and Business Media LLC.

- Consultation of the context of case study and interpretation of its results
- Corrections to the first draft of the manuscript

Rocchi, L., Kadziński, M., Menconi, M. E., Grohmann, D., Miebs, G., Paolotti, L., and Boggia, A. (2018). Sustainability evaluation of retrofitting solutions for rural buildings through life cycle approach and multi-criteria analysis. In Energy and Buildings (Vol. 173, pp. 281–290).

- Consultation of the context of case study and interpretation of its results
- Editing of the manuscript

Maria Elena Menconi

August 16, 2023

Francesca Torrieri
Department of Industrial Engineering
University of Naples Federico II
Naples, Italy

Declaration

I hereby declare the following contribution as an author of the following papers:

Oppio, A., Dell'Ovo, M., Torrieri, F., Miebs, G., and Kadziński, M. (2020). Understanding the drivers of Urban Development Agreements with the rough set approach and robust decision rules. In Land Use Policy (Vol. 96, p. 104678).

- Data and preference collection
- Corrections to the first draft of the manuscript

Francesca Torrieri
Francesca Torrieri

August 16, 2023

Adam Karaszewski
Advanced Analytics Team
PSI Poland, Towarowa 35
61-896, Poznań, Poland

Declaration

I hereby declare the following contribution as an author of the following papers:

Miebs, G., Mochol-Grzelak, M., Karaszewski, A., and Bachorz, R. A. (2020). Efficient Strategies of Static Features Incorporation into the Recurrent Neural Network. In Neural Processing Letters (Vol. 51, Issue 3, pp. 2301–2316). Springer Science and Business Media LLC.

- Consultation of the context of case study and interpretation of its results
- Corrections to the first draft of the manuscript

Miebs, G., Wójcik, M., Karaszewski, A., Mochol-Grzelak, M., Wawdysz, P., and Bachorz, R. A. (2022). Predicting a Time-Dependent Quantity Using Recursive Generative Query Network. In International Journal of Neural Systems (Vol. 32, Issue 11). World Scientific Pub Co Pte Ltd

- Consultation of the context of case study and interpretation of its results
- Corrections to the first draft of the manuscript

Adam Karaszewski

August 16, 2023

Paulina Wawdysz
Advanced Analytics Team
PSI Poland, Towarowa 35
61-896, Poznań, Poland

Declaration

I hereby declare the following contribution as an author of the following paper:

Miebs, G., Wójcik, M., Karaszewski, A., Mochol-Grzelak, M., Wawdysz, P., and Bachorz, R. A. (2022). Predicting a Time-Dependent Quantity Using Recursive Generative Query Network. In International Journal of Neural Systems (Vol. 32, Issue 11). World Scientific Pub Co Pte Ltd

- Consultation of the neural network architecture and interpretation of its results
- Corrections to the first draft of the manuscript

Paulina Wawdysz

August 16, 2023

Małgorzata Mochol-Grzelak
Advanced Analytics Team
PSI Poland, Towarowa 35
61-896, Poznań, Poland

Declaration

I hereby declare the following contribution as an author of the following papers:

Miebs, G., Mochol-Grzelak, M., Karaszewski, A., and Bachorz, R. A. (2020). Efficient Strategies of Static Features Incorporation into the Recurrent Neural Network. In Neural Processing Letters (Vol. 51, Issue 3, pp. 2301–2316). Springer Science and Business Media LLC.

- Co-authorship of the idea underlying the paper
- Consultation of the algorithms incorporating static metadata into a recurrent neural network
- Corrections to the first draft of the manuscript

Miebs, G., Wójcik, M., Karaszewski, A., Mochol-Grzelak, M., Wawdysz, P., and Bachorz, R. A. (2022). Predicting a Time-Dependent Quantity Using Recursive Generative Query Network. In International Journal of Neural Systems (Vol. 32, Issue 11). World Scientific Pub Co Pte Ltd

- Consultation of the neural network architecture and interpretation of its results
- Co-authorship of the text of the publication

Małgorzata Mochol-Grzelak

Małgorzata Mochol-Grzelak

August 16, 2023

Michał Wójcik
Advanced Analytics Team
PSI Poland, Towarowa 35
61-896, Poznań, Poland

Declaration

I hereby declare the following contribution as an author of the following paper:

Miebs, G., Wójcik, M., Karaszewski, A., Mochol-Grzelak, M., Wawdysz, P., and Bachorz, R. A. (2022). Predicting a Time-Dependent Quantity Using Recursive Generative Query Network. In International Journal of Neural Systems (Vol. 32, Issue 11). World Scientific Pub Co Pte Ltd

- Consultation of the neural network architecture and interpretation of its results
- Corrections to the first draft of the manuscript

Michał Wójcik
Michał Wójcik

September 5, 2023

Marta Dell'Ovo
Department of Architecture and Urban Studies
Politecnico di Milano
Milan, Italy

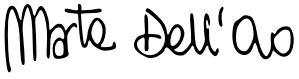
Declaration

I hereby declare the following contribution as an author of the following papers:

Oppio, A., Dell'Ovo, M., Torrieri, F., Miebs, G., and Kadziński, M. (2020). Understanding the drivers of Urban Development Agreements with the rough set approach and robust decision rules. In Land Use Policy (Vol. 96, p. 104678).

- Co-authorship of the idea underlying the paper
- Data and preference collection; interpretation of results
- Co-authorship of the text of the publication

Marta Dell'Ovo



August 16, 2023

To whom it may concern

Declaration

I hereby declare the following contribution as an author of the following paper:

Gehrlein, J., Miebs, G., Brunelli, M., and Kadzinski, M. (2023). An active preference learning approach to aid the selection of validators in blockchain environments. In Omega (Vol. 118, p. 102869).

- Co-authorship of the idea underlying the paper
- Consultation of the problem structuring and methods for solving the study
- Consultation of the experiment's design and interpretation of its results
- Co-authorship of the text of the publication

Matteo Brunelli
Associate Professor
University of Trento, Italy

August 16, 2023

Jonas Gehrlein
Web3 Foundation
Zug, Switzerland

Declaration

I hereby declare the following contribution as an author of the following paper:

Gehrlein, J., Miebs, G., Brunelli, M., and Kadzinski, M. (2023). An active preference learning approach to aid the selection of validators in blockchain environments. In Omega (Vol. 118, p. 102869).

- Co-authorship of the idea underlying the paper
- Consultation of the context of case study and interpretation of its results
- Co-authorship of the text of the publication

Jonas Gehrlein

September 25, 2023

Miłosz Kadziński
Institute of Computing Science
Poznan University of Technology
Poznan, Poland

Declaration

I hereby declare the following contribution of the author

Alessandra Oppio
Department of Architecture and Urban Studies
Politecnico di Milano
Milan, Italy

of the following paper:

Oppio, A., Dell'Ovo, M., Torrieri, F., Miebs, G., and Kadziński, M. (2020). Understanding the drivers of Urban Development Agreements with the rough set approach and robust decision rules. In Land Use Policy (Vol. 96, p. 104678).

- Co-authorship of the idea underlying the paper
- Data and preference collection; interpretation of results
- Co-authorship of the text of the publication

Miłosz Kadziński

Miłosz Kadziński

August 16, 2023

Roman Słowiński
Institute of Computing Science
Poznan University of Technology
Piotrowo 2, 60-965 Poznań, Poland

Declaration

I hereby declare the following contribution as an author of the following paper:

Cinelli, M., Kadziński, M., Miebs, G., Gonzalez, M., and Słowiński, R. (2022). Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system. In European Journal of Operational Research (Vol. 302, Issue 2, pp. 633–651)

- Co-authorship of the idea underlying the paper
- Consultation of the system's and methods' descriptions, including the characteristics and questions
- Editing of the manuscript

Roman Słowiński

Roman Słowiński

September 25, 2023

Miłosz Kadziński
Institute of Computing Science
Poznan University of Technology
Poznan, Poland

Declaration

I hereby declare the following contribution of the author

Kannan Govindan
Center for Sustainable Supply Chain Engineering
Department of Technology and Innovation
University of Southern Denmark, Campusvej 55,
Odense, DK-5230, Denmark

of the following paper:

Govindan, K., Kadziński, M., Ehling, R., and Miebs, G. (2019). Selection of a sustainable third-party reverse logistics provider based on the robustness analysis of an outranking graph kernel conducted with ELECTRE I and SMAA. In Omega (Vol. 85, pp. 1–15).

- Co-authorship of the idea underlying the paper (Section 1)
- Data and preference collection for the case study and interpretation of its results
- Co-authorship of the text of the publication

Miłosz Kadziński

Miłosz Kadziński

September 25, 2023

Miłosz Kadziński
Institute of Computing Science
Poznan University of Technology
Poznan, Poland

Declaration

I hereby declare the following contribution of the author

Ronja Ehling
Center for Sustainable Supply Chain Engineering
Department of Technology and Innovation
University of Southern Denmark, Campusvej 55,
Odense, DK-5230, Denmark

of the following paper:

Govindan, K., Kadziński, M., Ehling, R., and Miebs, G. (2019). Selection of a sustainable third-party reverse logistics provider based on the robustness analysis of an outranking graph kernel conducted with ELECTRE I and SMAA. In Omega (Vol. 85, pp. 1–15).

- Data and preference collection for the case study and interpretation of its results
- Corrections to the first draft of the manuscript

Miłosz Kadziński

Miłosz Kadziński

Poznan University of Technology
Faculty of Computing and Telecommunications
Institute of Computing Science
Typeset using LaTeX in Computer Modern.

BibTeX:

```
@phdthesis{ Miebs2023,
    author = "Grzegorz Miebs",
    title = "{New computer based decision support methods robust to imperfections of
learning data}",
    school = "Poznan University of Technology",
    address = "Pozna{\'n}, Poland",
    year = "2023",
}
```