Poznan University of Technology
Faculty of Computing and Telecommunications

Doctoral dissertation

Experimental analysis of the properties of models and decision support methods in the context of the use of holistic preferences

Michał Wójcik

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

Poznań, 2024

Abstract

Over the years, the development and growing popularity of the Multi-Criteria Decision Aiding field has brought many models and algorithms capable of solving various decision problems. Although the greater variety of methods has undoubtedly enriched the literature on the subject, the lack of a comprehensive comparison of different approaches has caused an increased difficulty in selecting an appropriate algorithm for the considered problem. This issue is important from the perspective of both decision makers and analysts, who strive to obtain high-quality recommendations, and the selection of an algorithm is one of the crucial steps in the decision aiding process. Addressing this problem was the main goal of this doctoral dissertation. As part of the research, several comparative analyses of methods and models were conducted, followed by the provision of observations and conclusions facilitating the selection of an appropriate approach based on the specificity of the considered problem. Moreover, new algorithms and models were proposed based on stochastic analysis and adapted to various assumptions regarding the decision-maker's preferences. Quality measures were also proposed, which allowed for the examination of the quality, robustness, and expressiveness of the approaches considered. Lastly, an approach to the preference learning problem using nature-inspired optimization algorithms was proposed.

List of publications

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

[P1] M. Kadziński, M. Wójcik, and K. Ciomek, "Review and experimental comparison of ranking and choice procedures for constructing a univocal recommendation in a preference disaggregation setting", *Omega*, vol. 113, p. 102715, 2022

DOI: 10.1016/j.omega.2022.102715.

Number of citations[1]:

- according to Web of Science: 6
- according to Google Scholar: 8

[P2] 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*, vol. 278, p. 110871, 2023

DOI: 10.1016/j.knosys.2023.110871.

Number of citations[1]:

- according to Web of Science: 2
- according to Google Scholar: 5

[P3] M. Wójcik and M. Kadziński, "Nature-inspired Preference Learning Algorithms Using the Choquet Integral", in *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '24, (New York, NY, USA), p. 440–448, Association for Computing Machinery, 2024

DOI: 10.1145/3638529.3654054.

[P4] M. Kadziński, M. Wójcik, and M. Ghaderi, "From investigation of expressiveness and robustness to a comprehensive value-based framework for multiple criteria sorting problems", *Omega*, 2024. Status: accepted for publication

[1] as on September 23, 2024

Contents

Chapter 1

Introduction

Multiple Criteria Decision Analysis (MCDA) concentrates on research and systematization of knowledge on solving decision problems in which alternatives are evaluated using more than one criterion. Research in this area aims to provide best practices that constitute the decision-making process. The most commonly used approach is to synergize information on the evaluation of alternatives based on individual criteria and preferences obtained from the Decision Maker (DM) and then to provide valuable conclusions and recommendations. To achieve this ultimate goal, over the years, many techniques, models, methods, and algorithms have been developed, improving the decision support process and thus enriching the literature on the subject.

The decision-making process consists of many steps, among which it is essential to determine the type of problem, to elicit the DM's preferences, and to select an appropriate decision support model or method. The development and popularization of the field have led to many practical applications of MCDA, solving real-world problems such as software evaluation [5], environmental management [6], energy policy [7], health care [8], conservation prioritization and planning [9], sports players' evaluation [10], energy systems analysis [11], and sustainability of insulating materials [12].

According to [13], in MCDA, four basic types of problems are distinguished, which are: choice (α), sorting (β), ranking (γ), and description (δ) problems; each of them aims at a different structure of recommendations about the considered alternatives. In the choice problem, one or a few best decision variants are selected; the idea of the sorting problem is to assign alternatives to preferentially-ordered classes; the ranking problem orders the alternatives from most to least preferred; and the description problem provides information about the consequences of choosing a particular action.

In the context of preference elicitation, the multitude of problems and methods for solving them goes hand in hand with the variety of possibilities for expressing preferences by the DM. [14] distinguished input and output-

oriented preference information. Input-oriented preferential statements can refer to the intra-criterion preferences, which assess the significance of differences in the ratings obtained by different alternatives concering a given criterion, and to the inter-criterion preferences, which define weights and trade-offs between the importance of individual criteria. On the other hand, expressed preferences may also refer to desired output, showing holistically the DM's attitude towards the alternatives considered. The most common ways of expressing preferences in this way are comparisons between pairs of actions [15], indicating which one is preferred by the DM, and assigning example alternatives to predefined decision classes [16]. It is commonly believed that this type of preference representation is less cognitively demanding and more intuitive for DMs. Moreover, it does not require detailed domain knowledge and familiarity with specific decision support methods. Another advantage is that the holistic preferences expressed in this way are compatible with many different approaches, which makes them universal.

The development of the MCDA field and the introduction of new methods and algorithms have made it possible to obtain more robust and qualitative recommendations and to solve more sophisticated problems. On the other hand, it has also made the work of decision analysts more demanding, because selecting a method tailored to the considered problem has become more difficult. The lack of analyses examining the characteristics and comparing different approaches constitutes a gap in the literature, and filling it was the main motivation for this study. In turn, the research hypothesis of this dissertation assumes that based on the results of the experimental analysis of the properties of decision support models and methods, it is possible to formulate guidelines for decision analysts that will facilitate the selection of adequate and qualitative approaches to the considered problems.

This dissertation focuses on presenting the analyzed decision support models and methods, the diversity in their ability to solve specific problems, provide different quality of recommendations, and consider different assumptions regarding recommendation expectations. These approaches were analyzed in terms of different features that are desirable in the field of MCDA, such as predictive abilities, robustness, expressiveness, and solving preference learning problems. A number of quality measures were proposed, which were then used to conduct an experimental comparative analysis. The scheme of the experiments that were conducted was also described in detail, and the results that were obtained were discussed. Finally, based on the results, guidelines for decision analysts were developed to facilitate their work.

The remainder of this doctoral thesis is organized in the following way. Chapter 2 presents the necessary theoretical background; in particular, it describes the considered decision problems and the methods used to solve them. Chapter 3 describes the obtained results of the conducted research work. Chapter 4 contains a summary of the research.

Chapter 2

Theoretical background

This chapter describes the basic theories and issues in the field of MCDA, which are essential for a better understanding of this dissertation and for defining its scope. The information included here concerns the decision problems considered, as well as the algorithms and models analyzed. It also describes the different perspectives on the quality of models, which were the focus of the conducted analyses.

2.1 Considered problems

In MCDA, decision problems are usually defined by providing a set of decision alternatives together with their evaluation on various criteria. At the stage of defining the problem, it is also necessary to determine its type, and thus the expected structure of the received results. To solve this problem and obtain satisfactory results, it is also essential to provide knowledge about the DM's preferences so that the resulting recommendations can accurately reflect them.

The set of alternatives $A = \{a_1, \ldots, a_n\}$ represents all the possible actions, alternatives, and options available to the DM when considering the problem being solved. It can be provided as a complete set of already existing actions considered in a specific decision-making context. Another approach is to represent the set of alternatives indirectly by specifying a range of values, constraints, and trade-offs between ratings on particular criteria. This approach is typically used for design problems in which arbitrary alternatives can be generated with specified characteristics that satisfy assumptions and constraints specific to the domain of the problem under consideration. Regardless of how the set of actions is defined, it also determines the space of all possible solutions, i.e., all possible assignments of alternatives to classes, all possible rankings, or all subsets of selected alternatives, depending on the nature of the problem.

The set of criteria $G = \{g_1, \ldots, g_m\}$ contains all the attributes that are important for the specific decision-making process. To consider a multi-criteria problem, it should contain $m \geq 2$ criteria, which describe various features of individual actions. These attributes should not be redundant, i.e., they should not refer to the same aspects of the alternatives. In addition, they should be easily interpretable since they often constitute the basis for expressing preferences. Moreover, the DM should be able to indicate the nature of individual criteria, i.e., determine their type:

- *gain* – the higher the performance of an alternative, the more it is preferred;
- *cost* – the lower the performance of an alternative, the more it is preferred;
- *non-monotonic* – there is no monotonic relationship between the intensity of preferences and the attribute values obtained by the alternatives.

In most cases considered in this paper, criteria are treated independently, but it is also possible to define more sophisticated ways of representing DM preferences, e.g., by taking into account inter-criteria interactions. The value of the alternative for the j-th criterion is typically denoted as $g_j(a)$.

Information about DM's preferences is crucial to improve the quality of the recommendations received. The better and more accurately they reflect perceptions of alternatives, the more accurate recommendations can be expected. DMs can express their preferences in various ways, referring to specific elements of the problem as well as to specific aspects of the approach used. Indications regarding the algorithm used (e.g., "the weight of criterion g_3 in model M should take the value of 0.4", "the veto threshold v_1 for criterion g_1 should be 10") are usually easy to apply directly to a given approach, but they require advanced knowledge and full understanding of the specific decision-making procedure and also prevent the use of this information in other approaches. On the other hand, information directly related to the alternatives ("a_3 is preferred over a_7", "a_5 should be assigned to at least class C_2", "a_4 should be among the top 10 alternatives in the ranking") and/or evaluation criteria ("criterion g_5 is more important than g_3", "a_2 and a_4 are indifference according to criterion g_1") are easy to interpret, do not require much effort and domain knowledge from the DM, and are universal for various decision support algorithms.

In the research described in this dissertation, problems with a pre-defined set of alternatives were considered, containing from a dozen to several hundred (in case of preference learning problems) alternatives, and evaluated using two to nine attributes. Within the problems considered, the types of criteria, the presence or absence of preference monotonicity, and possible

inter-criteria interactions were indicated. The preference information was tailored to the expected recommendation structure, i.e., pairwise comparisons of alternatives were provided for the ranking and choice problems, as well as example class assignments for the sorting problems.

2.2 Models and algorithms

This section provides an introduction to the models and algorithms whose study and comparison were the subject of this dissertation. The most important assumptions concerning the discussed approaches, their applications, and variants are presented. It is also shown why their comparison is an important research issue.

The UTA-like methods

Basic assumptions

The concept of the UTA (UTilités Additives) method, proposed by [15], based on the Multiple Attribute Value Theory (MAVT), uses the preference *disaggregation-aggregation* paradigm, as described in [17]. The concept of preference disaggregation assumes the use of example DM's decisions to develop a decision model consistent with the DM's preferences. On the other hand, the aggregation paradigm assumes that the utility of a given alternative is directly indicated by the marginal utilities on the individual criteria. In the case of UTA, the global utility score of the alternatives is determined by Additive Value Function (AVF), which aggregates marginal scores by adding them up. Global utility values are usually denoted as $U(a)$, and MVF as a function of the value of the alternative on a specific criterion: $u_j(g_j(a))$. Thus, the AVF value is determined as the sum of the MVF values: $U(a) = \sum_{j=1}^{m} u_j(g_j(a))$.

In the standard approach, two main assumptions are also used: normalization and monotonicity of MVFs. Normalization assumes that all functions $u_j(a)$ have the lowest value equal to 0 and that the sum of their highest values is 1. These assumptions ensure that the values of the AVF will be in the range $[0, 1]$, where 0 denotes the least and 1 is the most preferred, potentially existing alternative. Monotonicity, in turn, assumes that the function $u_j(a)$ is consistent with the type of the j-th criterion, i.e., it is non-increasing in the case of a *cost* criterion and non-decreasing in the case of a *gain* criterion.

The basic version of the model assumes that the marginal functions are defined for the interval containing all possible values for a given criterion. In addition to satisfying the properties resulting from the above assumptions, the functions are piecewise linear, so they consist of $\gamma_j \geq 1$ intervals of equal length, separated by *characteristic points* for which the function values are explicitly determined. On the other hand, the evaluation of the function

within the intervals is determined based on linear interpolation of the function values for two characteristic points that are the boundaries of a specific interval. Such a formulation of the function results in the fact that, on the one hand, it maintains great flexibility in shaping the DM's preferences with respect to a specific criterion. At the same time, it allows for the representation of monotonicity and normalization constraints in the form of linear constraints.

It is worth noting that in addition to reducing the model assumptions to linear constraints, this formulation of the model also allows the DM's preferences to be expressed in the same way. In the case of a pairwise comparison of alternatives, if the DM indicates that a is preferred over b , denoted as: $a \succ^{DM} b$, an additional linear constraint can be introduced: $U(a) \geq U(b)+\varepsilon$, where ε is a small, constant positive value. This constraint ensures that the desired preference relation for this pair of alternatives will be preserved in the resulting model. Similarly, if the DM indicates an indifference between a pair of alternatives, then a constraint $U(a) = U(b)$ is added, which ensures that both alternatives obtain the same global utility value.

In the case of sorting problems, the conducted research used the UTADIS model proposed in [18], which introduces a *threshold-based* approach by enriching the model with additional *threshold* values $(t_0, \ldots, t_p)$, symbolizing the boundaries between decision classes. Due to the assumed preferential order of classes, which is a feature of multi-criteria sorting problems, the threshold values must also satisfy the monotonicity constraints, i.e., $t_l - t_{l-1} \geq \varepsilon$. They should also split the interval of all possible global utility scores into smaller sub-intervals that unambiguously assign the utility value to a specific class. Threshold values defined in this way allow for an effective solution of sorting problems. Given the preferential information provided by the DM, if the reference assignment indicates that alternative a should be assigned to class C_l, then it is necessary to introduce constraints that ensure that the value of $U(a)$ is between t_{l-1} and t_l. This approach allows preferential information to be included directly in the assumptions of the decision model.

Both the linear constraints resulting from the model assumptions and from the DM's preferences define the space of all feasible solutions, denoted as $\mathcal{U}^R$. In the case of inconsistency within the DM's preferences or between the preferences and the model assumptions, it may happen that there will be no solution satisfying all of the constraints – then $\mathcal{U}^R = \emptyset$. To deal with such situations, the original proposal by [15] introduced $\sigma(a)$ variables, symbolizing "potential error relative to the utility", which allowed for the acceptance of solutions that were not fully consistent with the DM's preferences, but rather were a compromise solution in which the sum of errors was minimized. Another approach to this problem might be to point out the source of the inconsistency to the DM and then elicit preferences leading to a feasible solution. On the other hand, it also happens that the set contains an infinite

number of feasible solutions, which provides an opportunity to analyze the properties and relationships occurring in the models included in the set. The most popular tools for analyzing the mentioned aspects of the set of all solutions consistent with DM's preferences are Robust Ordinal Regression (ROR) and Stochastic Ordinal Regression (SOR).

Robust Ordinal Regression

According to [19], robust ordinal regression considers all solutions compatible with the DM's preferences and provides information about the *necessary* (compatible with all models) and *possible* (compatible with at least one model) "consequences of applying all compatible preference models to the considered set of alternatives". Using robustness analysis, inferences are drawn about possible and necessary relations in a set of solutions, e.g., about preferences between pairs of alternatives or assignments to decision classes. To verify a hypothesis concerning a possible property of all feasible solutions, it is enough to find at least one solution that confirms it. On the other hand, to check the validity of the necessary relation, one must prove by contradiction that there is no model consistent with the DM's preferences that does not satisfy the given properties.

Regardless of the issue or relationship under consideration, ROR enables the verification of hypotheses and obtaining qualitative information about all solutions, which allows for checking the possibility or necessity of a given phenomenon, resulting from both the structure and assumptions of the model and the DM's preferences. This information is useful for better understanding the possible consequences of the decisions made and can support the interactive recommendation building process in cooperation with the DM. The decision maker can react to the feedback received on the ROR results, e.g., by indicating desirable and unacceptable relations in the expected outcome of the procedure. Based on the DM's indications, the preferential information can be updated, which should lead to more satisfactory recommendations.

Furthermore, ROR is also used in some approaches to select a representative model to provide univocal recommendations. The most popular approach is to use observed relationships through procedures that try to emphasize their significance and highlight them in the obtained outcome. Moreover, the information provided by ROR can also be used to formulate conclusions about the robustness of recommendations and expressiveness of models.

Stochastic Ordinal Regression

SOR, similarly to ROR, also attempts to describe phenomena occurring in the set of all solutions. However, unlike ROR, it does not focus on the qualitative analysis of the necessity or possibility of relations but tries to capture them in a quantitative aspect. Stochastic analysis examines how often

a given relation occurs in the set of all compatible models. Thanks to this, it is possible to indicate the most popular relations occurring between pairs of alternatives or alternatives and decision classes.

Accurately assessing the frequency of occurrence of phenomena in a set of infinite elements using analytical methods is difficult. For this reason, algorithms from the class of Monte Carlo (MC) methods are used, which are one of the most popular approaches to this problem. They use random sampling of the solution space, and based on the obtained subset, approximate values of specific measures are determined. One of the approaches to generating uniformly distributed points over bounded regions called hit-and-run (HAR) algorithm, was first proposed by [20]. The algorithm consists of iteratively generating a sequence of points in N-dimensional space. Each subsequent point is obtained by randomly selecting a line passing through the previous point, and then randomly selecting a point from the selected segment, bounded by a feasible space of points. This algorithm allows for efficiently obtaining a set of uniformly distributed samples from any convex polytope, and this is exactly the shape of the solution space defined based on the assumptions of the described model. For this reason, it can be successfully used to perform stochastic analysis of the solution space of multi-criteria decision support problems solved by methods from the UTA family.

The SOR results provide important information about the set of feasible solutions and are used in a similar way to the ROR outcomes. Quantitative aspects of the phenomena occurring in the solution space are useful from the perspective of procedures exploiting this information and are used to define adequate quality measures that allow assessing how well-selected univocal recommendations represent the entire solution space.

Construction of univocal recommendations

The literature on this topic provides many strategies for creating unambiguous recommendations based on the UTA-like approach, which is used to solve the selection, ranking and sorting problems. They can be classified according to the way of building recommendations as follows:

- **Representative utility function.** This is a basic approach to solving the above-mentioned problem. This concept assumes building recommendations based on one, arbitrarily selected AVF, found in the set of all feasible solutions. This is usually done by defining an appropriate objective function, which, in combination with linear constraints imposed by the model assumptions, constitutes the Linear Programming (LP) problem. Various functions are known that try to capture the most discriminant, parsimonious, benevolent or aggressive AVF. Due to the convexity of the solution space, it is also possible to obtain

a unique model by selecting many feasible models arbitrarily or randomly and then averaging them. It is also possible to determine the *central* solution, for example, by introducing additional constraints and variables into the model. Finally, it is also possible to exploit outcomes from ROR and SOR to emphasize relationships and class assignments that occur most frequently in the set of all solutions, making univocal recommendations *representative* in the context of the entire set of solutions. However, what all these approaches have in common is that the recommendations are derived directly from one of the feasible models.

- **Decision rules.** These rules, unlike representative feature selection, do not use a single AVF to directly obtain recommendations, but take into account knowledge about several solutions or statistics derived from the analysis of the entire solution space. These approaches create decision rules that shape recommendations based on various aspects of the evaluations of alternatives, e.g., extreme utility values of alternatives, highest and lowest positions in the obtained rankings, or extreme assignments to classes in the entire feasible solution set. They can also apply information from the ROR and SOR results to the generated recommendations, e.g., by assigning alternatives to the most probable classes or positions in the rankings. It is worth noting that there may be no feasible model that provides recommendations identical to those obtained by applying a given decision rule. For this reason, one can perceive these approaches as a certain extension of the space of potential solutions.

- **Scoring procedures.** These approaches, like decision rules, also attempt to take into account a more complex perspective on the DM's preferences than using a single solution that satisfies all the model constraints. They mainly use the results of stochastic and robustness analysis to determine the *score* for each alternative and then, based on the obtained evaluations, generate recommendations. Scoring procedures use the results of stochastic and robustness analysis to determine *score* for each alternative and then provide recommendations based on them. They are often based on indices estimating the frequency of preference relations for pairs of alternatives or taking into account the ratio of models for which a specific assignment to a decision class occurred.

- **Most robust solution.** This concept is also based on the use of LP, but in a different way than to find acceptable or optimal AVF parameters. This approach consists of exploiting the results of Stochastic Ordinal Regression. It gives information about the percentage of models in which a preference or indifference relation holds for a pair of alternatives or how often an alternative is assigned to a specific class. Depending on the problem considered, the task of LP is to find a ranking or as-

signment of alternatives to classes that maximizes the aggregate values of statistics derived from stochastic analysis. Modeling is usually done by introducing binary variables that determine the position of an alternative in a ranking or assign it to a specific class. These variables are included in constraints indicating the uniqueness of the assignment to the class or ranking position and the compatibility of the relation with DM's preferences. On the other hand, in the objective function, they are assigned weights adequate to the statistics derived from SOR. This method allows obtaining a compromise ranking, choice, or sorting solution that is maximally acceptable among all feasible models. Hence, it can be considered as the most robust solution.

Model modifications

With the development of the MCDA field, a multitude of different approaches have been proposed, based on the UTA concept. As mentioned above, various procedures extending the basic approach allow obtaining different univocal solutions. In addition to them, extensions of the model representing preferences have also been proposed, which can be observed, among others, in MCDA-MSS - software supporting the selection of an adequate approach to the considered problem, proposed by [21], which provides a collection of over 200 decision support methods, 27 of which contain "UTA" in their name, indicating a connection with the approach described above.

One of them is UTA$^{\text{GMS}}$-INT, proposed by [22], which addresses one of the issues of the basic approach, i.e. the inability to represent interactions between criteria occurring in DM preferences. The authors proposed to extend the AVF with additional functions representing positive and negative interactions occurring for pairs of criteria. In this way, it allows modeling more demanding scenarios concerning DM's preferences. The method works in two stages: in the first stage, it identifies pairs of interacting criteria, based on the provided preference information by solving the Mixed-Integer Linear Programming (MILP) problem. In the second phase, it provides an extended additive value function, which can then be used to evaluate alternatives and provide valuable recommendations and conclusions.

Other models based on UTA focus on a different issue, which is the inability to reflect non-monotonic preferences for individual criteria, resulting directly from the constraints imposed on the model. One of them is the approach described in [23], which, at the stage of solving the linear programming problem, abandons the constraints related to the monotonicity of marginal functions and the normalization of the achievable global utility. Instead, it introduces bound constraints, which limit the range of MVF values, but allow them to take any shape, preserving their piecewise-linear nature. This procedure allows for the representation of non-monotonic prefer-

ences for individual criteria, and the resulting model can be easily normalized *a posteriori* to be consistent with the normalization assumptions used in the basic UTA method.

Another approach to modeling nonmonotonicity was proposed in [24]. In this case, the non-monotonic nature of preferences was modeled as a composition of two value functions with opposite preference directions – non-increasing for cost type marginal function component and non-decreasing for its opposite counterpart – gain type. As a result of the composition of these functions, it was possible to provide non-monotonic MVFs, which then also had to be normalized in accordance with the assumptions of UTA. Nevertheless, the remaining elements of the method are fully compatible with the assumptions of the above-mentioned model.

Choquet Integral

A slightly different approach, which also provides recommendations based on the estimation of the utility of alternatives, is the Choquet Integral model, which was first published in [25] and named after its inventor. This function aggregates the evaluations of alternatives for individual criteria while allowing for the representation of complex interactions using non-additive measures. In order to obtain a model, it is necessary to determine the values of the model parameters, called *capacities*, which represent the importance of the evaluations of different subsets of the set of all considered criteria – starting from single attributes to combinations of all available criteria.

The basic assumption of the method is that the strength of the criteria coalition is indicated by the minimum value that the alternative obtained for these criteria. The comprehensive value for an alternative is therefore determined as the product of subsequent capacities (μ_M) assigned to the subsets of criteria and the lengths of the intervals between the values obtained by the alternative: $Ch_M(a) = \sum_{j=1}^{m}[g_{(j)}(a) - g_{(j-1)}(a)] \cdot \mu_M(G_{(j)})$, where $(\cdot)$ is a permutation of criteria indices that sorts the values obtained by the alternative on the criteria from smallest to largest, and $G_{(j)}$ contains the subset of criteria: $\{g_{(j)}(a), \ldots, g_{(m)}(a)\}$. In this way, each successive term of the sum is the result of multiplying the interval between the ratings on the criteria with increasingly higher ratings and the capacity for increasingly smaller subsets of criteria. The last term is equal to the difference between the two highest evaluations of an alternative among all criteria and the capacity for the subset containing the criterion with the highest rating.

Moreover, capacities satisfy monotonicity and normalization constraints: $\mu_M(\emptyset) = 0, \mu_M(G) = 1, \mu_M(G_1) \leq \mu_M(G_2)$, for each $G_1 \subseteq G_2 \subseteq G$. Furthermore, the model requires that the evaluation of individual criteria be represented by values from the same scale. In particular, if they take values in the range $[0, 1]$, then the comprehensive value of alternatives also takes val-

ues in this range, which is identical to the range of obtainable comprehensive values in methods from the UTA family. In addition, it is also possible to enrich the model with additional parameters constituting boundary values for decision classes, allowing the adaptation of the approach to solving multi-criteria sorting problems in a threshold-based manner. Finally, it is also worth mentioning that preference information can be modeled analogously to the UTA method, i.e., by comparing the utility of alternatives in the case of indicating a preference for one alternative over another or assigning an alternative to the desired class by limiting its comprehensive value to the range between the thresholds of the relevant class.

It is worth noting that several applications use the Choquet integral in decision support problems, such as the evaluation of research institutions described in [26] and customer segmentation proposed in [27]. Nevertheless, finding optimal values of capacities is challenging, especially in the case of problems with a large number of criteria, as the number of model parameters grows exponentially with the number of attributes on which the alternatives are evaluated. In such cases, a simplified 2-additive version of the model may be helpful, in which capacities are explicitly determined only for the subsets containing one and two criteria.

2.3 Analyzed issues of algorithms and solutions

This section presents the issues that were analyzed in the publications that contribute to this dissertation. Issues such as accuracy, robustness, and expressiveness will be discussed. In addition, the issue of Preference Learning (PL) will be presented.

Accuracy of a solution

One of the most important determinants of the quality of recommendations suggested by the decision support approach is their accuracy. However, it should be clearly noted that it should not be associated with and interpreted in the same way as the concept of accuracy known from other fields, such as Machine Learning (ML). In these domains, accuracy is usually measured by aggregating the results of comparisons between expected outputs and predictions obtained using a given method. However, this requires knowledge of the "ground truth" about the analyzed data, which is often unavailable or even impossible in the case of decision-making problems. In the problems considered, the DM's preferences are not explicitly expressed, but rather inferred based on incomplete knowledge about them. The role of discovering "ground truth" is on the side of decision support models, hence the difficulty in assessing the quality of the recommendations they provide.

Nevertheless, there are different approaches that attempt to determine the accuracy of specific models and procedures. Naturally, this would be possible if the expected recommendations were completely known a priori – in which case it would be possible to compare the recommendations with them. Another approach is to perform a retrospective analysis on already solved decision problems, in which the DM was fully satisfied with the recommendations obtained. In such a case, it is possible to provide the same input data to different models and procedures and then compare the recommendations to those that were acceptable to the DM.

In the context of accuracy, experimental analysis offers greater possibilities than in the case of analyzing this aspect for real decision problems. In this case, it is possible to create an artificial DM that is able to provide complete and consistent preferential information which can be easily transformed into a form containing the expected recommendations. The preferences of the artificial DM can also be provided in different ways. On the one hand, it is possible to use any decision model for this purpose, whose recommendations are then transformed into the DM's preferential information. On the other hand, it is possible to generate completely random sample decisions, which are then fed to the methods. Finally, it is possible to provide random, but also guided or biased information that may simultaneously reflect the DM's consistent belief system but also introduce some inaccuracies, thus simulating real-world use cases of these methods where preferential information is often inconsistent.

Turning to the quality measures used, they are oriented to the type of problem being considered. For sorting problems, traditional measures are used to compute the fraction of alternatives for which the expected and predicted decision classes are the same. For the choice problem, a commonly used quality measure is the Hit Ratio [28], which returns 1 in case of a correct prediction of the expected best alternative and 0 otherwise. In the case of experimental analysis and repeated experiments, the average value of this measure can indicate the probability of the method making a correct choice. In the case of ranking problems, measures based on similarity of rankings are used, such as rank correlation coefficients proposed by [29] and [30]. In addition, there are measures oriented to the differences in the comparison of two rankings: expected and predicted. These measures examine how much the rankings differ in terms of the assignments of alternatives to ranking positions and vice versa.

Robustness of recommendation

As stated in [31], the term *robustness* can be perceived in various ways and relate to different phenomena and features of decision support models. Bernard Roy points out that "depending on the situation, this notion can be related

to, or integrated into, the notions of flexibility, stability, sensitivity, and even equity". In general, robustness refers to all the features of the decision-making process related to various uncertainties. These uncertainties may concern input data, preferential information, a specific decision support procedure, and the usefulness of proposed recommendations in the future. The proposed decision-making process can be described as *robust* if it copes with these issues and allows for obtaining valuable conclusions.

In the context of this work, which is focused on methods and procedures providing univocal recommendations, the main emphasis is placed on obtaining robust solutions, and the study of this aspect mainly refers to the recommendations provided by the methods. This dissertation does not consider the issue of robustness in the context of the provided input data or the uncertainty associated with the DM's preferences. Therefore, this work offers a dual perspective on the robustness of various decision support approaches and the results they obtain.

First, the research includes the robustness analysis by verifying the consistency and representativeness of the recommendations provided. The analyses conducted attempt to see to what extent the results obtained by the individual procedures aimed at selecting univocal recommendations are confirmed by all other feasible solutions for the assumed preference representation model. This makes it possible to draw conclusions about the robustness of these recommendations by determining the level of credibility they obtained in comparison to other solutions.

On the other hand, the self-consistency of all feasible solutions is also examined. Assuming no uncertainty regarding the input data and the DM's preferences, checking such properties allows us to determine the quality of the model and increases confidence in the recommendations it provides. In both cases, SOR and ROR are indispensable, providing the features and statistics obtained by the models when they are applied to the considered problem.

The quality measures considered verify, on the one hand, to what extent the provided recommendations, such as positions and relations between alternatives in a ranking or their assignments to classes in a sorting, are confirmed among all other possible solutions. On the other hand, statistics on the frequency of occurrence of specific phenomena can be successfully used to estimate the robustness of individual models, e.g., by evaluating entropy-based measures for all possible pairwise comparisons or assignments to classes.

Expressiveness of a preference model

The concept of expressiveness refers to the model's ability to reproduce and represent DM's preferences, expressed in an indirect way. The analysis of this issue allows us to determine the degree of universality of a given model and its applicability to various problems. Models with high expres-

siveness are able to represent richer DM's preferences, which are also exposed to a greater risk of containing inconsistencies.

For this reason, it can be said that this issue is closely related to the robustness described earlier. The fundamental difference between them is that expressiveness concentrates on the ability to represent DM's beliefs and ignores the aspect of the quality of the recommendations provided, the study of which is the domain of robustness. Moreover, it can be said that these aspects of the models are opposed because the more flexible the model with respect to the inaccuracy of the information provided, the more it is exposed to problems related to the robustness of recommendations.

The assessment of the ability to express DM's preferences and solve various decision problems is problematic from a theoretical point of view. Instead, an empirical approach is used, which verifies for what kind of problems a specific model is able to reproduce DM's beliefs. This proves that in this context, conducting an experimental analysis of the properties of models is crucial to investigate and understand this phenomenon.

Ability to solve preference learning problems

According to [32], PL is a subfield of machine learning focused on predicting or inferring preferences. As stated in [33], PL problems are challenging due to the need to deal with large amounts of incomplete and inconsistent preference information and to provide interpretable outcomes. These models, similarly to machine learning, should generalize well to knowledge about complex preferences. They should be resistant to biased data and enable the delivery of valuable and robust recommendations, even in the case of strong inconsistencies in both the input data and the model itself. Moreover, among entities formulating decision-making problems, the requirement for full explainability of recommendations generated by intelligent decision support systems has recently become popular [34].

In general, the majority of MCDA methods and procedures are designed to provide high-quality and interpretable recommendations and solve problems involving a relatively small number of alternatives. On the other hand, approaches used in ML are often oriented towards training and prediction on large datasets, neglecting the aspect of understandability of the delivered judgments. For this reason, researchers are interested in the possibility of combining and synergizing the features of these approaches to obtain procedures that have the capacity for both explainability and scalability of recommendations.

The research literature in the field of MCDA contains several studies and proposals of methods that are able to effectively address PL problems. In [35] the authors introduced a statistical framework for PL classification with monotonicity constraints, [36] proposed an optimization approach us-

ing an additive value function model enriched to handle inter-criteria interactions, and [37] presented several approaches to optimize the parameters of a sorting model with class profiles.

In the case of PL problems, the traditional approach to determining the values of model parameters by solving the LP problem is difficult to implement due to the large number of constraints resulting from the rich preferential information. Moreover, inconsistencies and contradictions of preferences with model assumptions make obtaining a model fully consistent with DM's beliefs in this way time-consuming and often impossible. For this reason, different approaches are used to solve this problem. Instead of obtaining fully compatible solutions, the aim is to obtain the most satisfactory recommendations possible, reflecting as closely as possible the DM's information.

One way to provide models capable of solving PL problems is to optimize the parameters of models used in popular methods for solving multi-criteria sorting problems, which can be achieved using e.g. logistic regression [38] or artificial neural networks [39]. However, development and research considering these aspects are limited and there are many unexplored ways to optimize models. In particular, it is possible to use nature-inspired metaheuristics and other optimization approaches, which have been proven to give excellent results in various applications.

To assess the consistency of the recommendation with the DM's beliefs, one should also use a quality measure related to the accuracy of the solution, thus measuring the ability to infer preferences. It should be remembered that for real-world decision problems, knowledge about the desired ranks of alternatives or class assignments is limited and often unavailable. Nevertheless, experimental analysis, using artificially generated decision scenarios, allows for verification of the quality of the delivered solutions based on both the reference and validation subsets of alternatives. Ultimately, the higher the ability of a method to correctly assess non-reference alternatives, the higher the ability to generalize knowledge about the DM's preferences and to solve challenging PL problems.

Chapter 3

Results of comparative analyses

This chapter contains a description of the research work conducted, based on the experimental analysis of the properties of models and decision support procedures, along with the obtained results.

3.1 Predictive performance and robustness concerns of value-based preference disaggregation ranking and choice methods

As previously mentioned, the MCDA field's development has brought many models and procedures, allowing the solving of multi-criteria decision problems. In particular, many approaches have emerged that are oriented towards solving the ranking and selection problem by providing univocal recommendations. Various models and procedures, based on different assumptions about the way and richness of preference representation, have positively impacted the field and contributed to further development. Among them, one can distinguish approaches based on complete preferential information, which requires high cognitive effort and domain knowledge from the DM. An alternative concept is to provide incomplete holistic judgments, which is an intuitive and relatively simple way to express the DM's beliefs.

Unfortunately, incomplete preferential information often leads to multiple or infinitely many solutions and recommendations compatible with indirect statements which reflect DM's preferences. The multiplicity of solutions results in ambiguous recommendations, which, due to difficulty and low interpretability, do not provide valuable answers to the questions posed by the decision maker. This issue can be addressed in various ways. One of them is to use the preference elicitation leading to enrichment of the provided information, but this leads to additional costs and is not always possible. For

this reason, another effective way to address this issue is to construct or derive univocal recommendations that provide a precise answer to the DM's dilemma.

Despite the availability of many concepts and proposals presenting approaches to obtaining univocal recommendations, the scientific literature has so far demonstrated the lack of a comprehensive approach allowing for their comparison and indicating in what circumstances the use of a given procedure may be beneficial. This became the motivation for the research described in the publication **P1**, which proposes a comprehensive review of the approaches described in the literature and their comparison by conducting a series of computational experiments and then analyzing the results to determine the properties of the examined decision-making procedures and the recommendations obtained.

The research included thirty-five different approaches extending the UTA method and capable of solving ranking and choice problems. The compared decision algorithms represent four groups of methods. The first group aimed at selecting a representative value function, which was achieved by modifying the goal function and introducing additional parameters to the LP model. The group included methods with various interpretation of *representativeness*, e.g. as the selection of most discriminant, central, average, aggressive, benevolent, parsimonious or robust AVF. The second group used decision rule-based approaches. These methods based their rules on characteristics of individual alternatives, such as the best or worst comprehensive score of an alternative among all feasible solutions or the frequency of obtaining a specific position in the resulting ranking. Another group used scoring functions to evaluate alternatives and then create a ranking based on them. Scoring functions exploited the relationships between pairs of alternatives – in particular, extreme differences in utilities between alternatives and differences in the frequency of outranking relationships. The last group includes methods oriented towards constructing a robust ranking. They use information obtained from the stochastic analysis to maximize the support of the provided recommendations among all compatible solutions, using LP techniques. The publication contains a detailed description and mathematical notation of existing approaches in the literature, along with an explanation of their motivation.

The experimental analysis scheme assumed examining the properties and quality of solutions obtained by individual methods, depending on the characteristics of the problems considered. The problems differed in the number of decision alternatives ($M \in \{6, 8, 10, 12, 14\}$), the number of criteria ($E \in \{3, 4, 5\}$), the richness of preferential information (the number of pairwise comparisons – $C \in \{4, 6, 8, 10\}$) and the ability to adapt the shape of marginal functions (the number of characteristic points – $P \in \{2, 3, 4\}$). The experimental setup assumed solving problems created based on each

combination of the above parameters. This allowed to provide results taking into account diverse problems of different difficulty levels. For each combination, 1000 problem instances were considered in order to increase the reliability of the obtained results. Overall, the conducted experiments included solving 180,000 decision problems using all considered procedures.

The performance of each of the considered methods was quantified using seven quality measures, four of which were oriented towards the quality of representing the DM's actual preferences and three towards the robustness of the proposed recommendations. It was assumed that each of these subgroups should be distinguished by one measure oriented towards solving the choice problem, while the remaining ones concerned the ranking problem. The measures assessing the quality of preference mapping were based on the correlation of rankings (*Kendall's* τ), the similarity in individual ranking positions (*Normalized Hit Ratio – NHR*, *Rank Agreement Measure – RAM*) or the differences in positions for individual alternatives (*Rank Difference Measure – RDM*). On the other hand, the measures oriented toward assessing the robustness of recommendations exploited the information provided by the SOR, indicating the dominant phenomena in all feasible solutions. In particular, on the indices measuring the frequency of alternatives' assignments to individual ranking positions (*First Rank Acceptability Index – FRAI*, *Mean Rank Acceptability Index – MRAI*) and checking how often specific relations occur for pairs of alternatives (*Mean Pariwise Relation Acceptability Index – MPRI*).

Having completed the experiments, their results were analyzed in detail to identify the best approaches, distinguishing between the type of problem (ranking or choice), the aspect analyzed (correct representation of DM's preferences or all feasible solutions), and the parameters of the considered problem (number of alternatives, number of criteria, etc.). For each quality measure, a separate analysis was performed, and the individual procedures were compared. To determine the statistical significance of the observed relationships, a statistical test was used, specifically the Wilcoxon signed-rank test [40] with a p-value equal to 0.05.

Both quality measures oriented towards the choice problem (NHR, FRAI) confirmed that the best procedure in their context was one of the approaches from the decision rule group – **BESTRAI**. This procedure was based on a rule that selected the best action from among the entire collection of alternatives based on the lexicographic objective and the results of stochastic analysis. The primary aim sorted alternatives according to the best possible position in any of the feasible rankings, and the secondary aim broke ties by favoring actions with a higher *Rank Acceptability Index* ($RAI : A \times N \to \mathbb{R} \in [0, 1]$) value for a given position, which is the approximate frequency of a given alternative at a specific position in the ranking, among all rankings compatible with the DM's preferences.

In the case of ranking problems, the remaining quality measures consistently indicated that the best results were provided by approaches from the group providing recommendations by constructing robust rankings. These methods can be classified into two subgroups according to the information provided by the SOR, which they exploit to build recommendations. The first subgroup of methods is based on the above-mentioned RAI values and shares a common prefix in their name (RANK-), as these values focus on the relations between alternatives and ranking positions. The second subgroup (REL-) uses *Pairwise Winning Index* ($PWI : A \times A \to \mathbb{R} \in [0, 1]$) values, which represent the share of all compatible AVFs for which a preference relation holds between one alternative and another.

The common feature of all the approaches discussed is the method of obtaining recommendations by solving the MILP problem, which assigns alternatives to specific ranking positions using binary variables. It is worth noting that for each of the considered approaches and accompanying MILP problems, the assumptions of the UTA method are neglected. Instead, they rely exclusively on the results of stochastic analysis, incorporating, depending on the procedure, the values of RAI or PWI into the objective function as weights of the binary variables representing the assignment of the alternative to the rank position. These methods also differ in formulating the objective function, presenting different approaches to aggregating the values of stochastic indices for individual alternatives to obtain a uniform evaluation of the entire ranking. In detail, three types of aggregation in the objective function are distinguished and introduced as suffixes to the names of the procedures – summation (**-SUM**), product (**-PROD**), and maximin (**-MM**) problems. It is also worth noting that these methods also had their counterparts marked with an additional suffix (**-IND**), which allowed for the reflection of indifference relations by assigning more than one alternative to a given ranking position. Nevertheless, these method variants achieved comparable but statistically significantly worse results.

Going deeper into the details of the analysis, from the perspective of quality measures focused on pairwise relations (Kendall's τ, MPRI) and minimization of rank differences (RDM), the best average results were obtained by methods from the **REL-** group. On the other hand, quality measures emphasizing the correct positioning of alternatives in the ranking positions (RAM, MRAI) confirmed the dominance of procedures from the **RANK-** group.

Among the classical approaches based on selecting a representative AVF, taking into account all quality measures, the best results were obtained by the **REPROC** method, which also enriched the UTA model with information provided by SOR. This was done by solving the maximin problem, focused on maximizing the differences in comprehensive values for such pairs of alternatives for which the preference relation occurred more often in the sample

set. Among the decision rules, apart from the aforementioned BESTRAI, which dominated the quality measures related to the choice problem, the **EXPRANK** method, which consists of evaluating and sorting the alternatives with respect to the expected position calculated on the basis of all feasible rankings, kept the best results for all five measures related to the ranking problem. For scoring methods, **MINPOI** performed best in choice problems, and **SUMPOI** in ranking problems. Both methods assign a score to an alternative based on the obtained differences in the values of *Preference Outranking Index* ($POI : A \times A \rightarrow \mathbb{R} \in [0, 1]$) compared to all other alternatives, with the former aggregating these values by minimization and the latter by summation.

It is also worth noting that the conducted analyses, apart from the global perspective aggregating quality measures among all considered problems, also confirmed the priority of the above-mentioned methods, regardless of the specific parameters of the problem, such as the problem's size or the amount of preferential information. Thus, the clear indication of the best procedures, taking into account the considered aspects of the quality of the recommendations provided, confirmed the hypothesis put forward in this thesis.

3.2 Predictive performance and robustness concerns of value-based preference disaggregation sorting methods

Similarly to the ranking and choice problems, decision support methods oriented towards solving the multi-criteria sorting problems face similar challenges. It should be emphasized again that there are many MCDA procedures for solving this type of problem, which can be categorized according to the type of preferential information required. Low cognitive cost and high interpretability make procedures adopting the preference disaggregation paradigm, based on the DM-provided example decisions, popular in this respect. In the context of multi-criteria sorting, the DM's decisions are usually expressed as assignments of reference alternatives to decision classes. One of the most popular sorting methods operating in this paradigm and accepting exemplary class assignments is UTADIS, which provides recommendations based on AVF and threshold values separating ranges of comprehensive values that unambiguously assign alternatives to specific classes.

The incompleteness of preference information implies a potentially infinite number of solutions which, although fully consistent with DM's expressed beliefs, may lead to different recommendations. As previously mentioned, conclusions derived from the robustness analysis, which provides a holistic view of the relationships and features of the complete set of all feasible solutions, are difficult for a DM to interpret without domain knowledge. To deal with

this issue, some procedures propose preference elicitation, which makes the decision support process more demanding and is not always possible; therefore, another popular tool to facilitate the understanding of the problem solution is to provide univocal recommendations that are easy to interpret and address the DM's doubts about the choices made.

The popularity of this approach has led to the emergence of many competing procedures offering different perspectives on the selection of a representative sorting model. However, the scientific literature has not provided a clear answer to the question about the usefulness of individual approaches and how to select an adequate procedure for the problem under consideration. This question is crucial for decision analysts whose task is to identify the appropriate tools to solve a given problem. In this context, it is worth noting the important contribution of [41], which compares four different procedures providing univocal recommendations. Nevertheless, this work does not cover a large number of other relevant approaches in this context; hence, the conclusions drawn from this publication are limited. To address this literature gap, a comparative experimental analysis was conducted and described in the publication **P2**.

In addition to presenting the formal definition of the model and the issues related to ROR and SOR, the research work included a detailed description and mathematical formulation behind fourteen different approaches to obtaining univocal recommendations based on the UTADIS model. The analysis performed included procedures providing recommendations based on a selected representative value function. Depending on the method, most discriminant [42], parsimonious [43], average [44, 45], central [42, 41], and robust [42] models were selected.

In addition, four robust approaches based on stochastic outcomes were also described, including three novel approaches presented in the mentioned publication. These approaches were based on stochastically derived *Class Acceptability Indices* ($CAI : A \times C \rightarrow \mathbb{R} \in [0, 1]$), which provide information about the fraction of all feasible models in which an alternative was assigned to a particular class, and *Assignment-based Pairwise Outranking Indices* ($APOI : A \times A \rightarrow \mathbb{R} \in [0, 1]$), which specify the fraction of compatible models in which an outranking relation holds for a pair of alternatives. These procedures were designed to find an AVF that best reflected the relations satisfied in the largest possible number of all feasible solutions, constituting their representative form providing univocal recommendations. Moreover, for all fourteen considered approaches, an illustrative study was also provided, presenting their performance on the practical multi-criteria sorting problem.

The experimental setting was designed to capture three different aspects of the solutions that were obtained. One of them was classification accuracy, defining how often the examined methods correctly recommended class

assignments for non-reference alternatives that were not included in the provided preferential information. To measure accuracy, it was necessary to define complete reference information containing the assignment of all alternatives to specific classes. For this purpose, for each of the considered problems, a value function and threshold values were randomly selected, representing the DM's internal belief system. The second quality measure was oriented towards assessing the robustness of the solution by checking the compatibility of non-reference class assignments with other feasible solutions, using CAI values derived from stochastic analysis. The third aspect considered was the similarity between the reference model and the model obtained by a given approach, which was assessed by comparing the corresponding marginal functions, comprehensive values, and thresholds.

The considered problems differed in the number of classes ($p \in \{2, 3, 4, 5\}$), the number of criteria ($m \in \{3, 5, 7, 9\}$), the number of characteristic points in MVFs ($\gamma_j \in \{2, 4, 6\}$) and the number of reference assignments per class ($R \in \{3, 5, 7, 10\}$). The setting formulated in this way made it possible to examine the methods in the context of solving problems of varying complexity. For each combination of parameters, 100 different problem instances were solved to provide more reliable outcomes and conclusions, which in total resulted in solving 19200 problem instances.

The highest accuracy in reproducing desired class assignments was obtained by the central **ACUTADIS** approach, determining the analytical center of the polytope that represents the space of all feasible solutions. The statistically significant advantage over the other methods was confirmed by the Wilcoxon signed-rank test with p-value $= 0.05$, comparing it with the other methods. Given the comparison, the next method was **CENTROID**, which determines the average solution from all samples used to perform SOR. In the next positions, we can observe novel approaches (**CAI**, **APOI**, **COMB**) and another approach to finding the central solution – **CHEBYSHEV** that seeks the central solution, i.e., the center of the largest Euclidean ball contained in the polytope. Moreover, the performed multivariate analysis confirmed the dominance of **ACUTADIS** over the others while emphasizing greater differences in the obtained mean values of the quality measure for problems of greater complexity, i.e., with a larger number of criteria, alternatives, and characteristics points and a lower richness of preference information.

The group of the above-mentioned six best solutions remained the same for the *Mean Class Acceptability Index* (MCAI) analysis, averaging the CAI values corresponding to the assignments of the alternatives to the classes provided by a particular method. In the case of this measure, all three novel approaches performed best, led by **CAI**, which attempted to obtain a model with the greatest possible support among all feasible solutions in terms of assigning alternatives to specific classes. Among the methods based on searching for a representative value of a function, the best performing one was

CENTROID approach, followed by **UTACHEB** and **ACUTADIS**. In this case, the statistically significant advantage of **CAI** over the other methods was also confirmed in the multivariate analysis taking into account subsets of results with specific problem characteristics.

Taking into account the similarity of the reference model and the solution used to create the recommendations, the best results were obtained by the three approaches mentioned above oriented on the central and average representative models. It is worth mentioning **REPDIS**, which was also one of the leading ones in this context. This method emphasizes the advantage of alternatives, which SOR revealed as those assigned more often to a higher priority class than others. It should be noted that from the DM's perspective, the quality and robustness of the recommendations provided are much more important aspects. Hence, the similarity of the models provides limited indications of the benefits of using a given procedure.

The presented results allow for the formulation of clear recommendations regarding the selection of an adequate procedure for building univocal recommendations. In this context, the conducted research provides sufficient evidence to confirm the research hypothesis. Moreover, the three novel approaches introduced allow for the obtaining of representative and robust solutions, which is an additional contribution to the literature on the subject.

3.3 Recommendation robustness and model expressiveness in view of value-based sorting methods

In addition to the procedures addressing the selection of univocal recommendations, an important aspect is the structure of the model used and its ability to represent the DM's preferences. One of the most popular models for solving multi-criteria sorting problems is the UTADIS, operating in the preference disaggregation paradigm and providing recommendations using a threshold-based procedure for assigning alternatives to preferentially ordered classes. It is widely used due to its intuitive form of representing DM's beliefs, expressed through exemplary assignments of alternatives to classes, and providing explainable and easy-to-interpret recommendations.

When dealing with value-based sorting models operating in the preference disaggregation paradigm, one may encounter two important issues to address. On the one hand, indirect preferences are not always consistent with the assumed model, which may result in the lack of feasible solutions. On the other hand, in the case of achieving full compatibility between the method and preference information, many feasible models may provide ambiguous recommendations. These issues can be linked to the concepts of the expressiveness of models and the robustness of recommendations, respec-

tively. In this context, robustness concerns the credibility of the model and the recommendations it provides, whereas expressiveness reflects the model's ability to reproduce the DM's preferences and, thus, its flexibility towards inaccurate or inconsistent statements about the DM's beliefs. Both of these aspects can be seen as competing, since the more expressive models have richer possibilities of representing preferences, which may result in reduced robustness. Therefore, it is important to maintain a balance between these two phenomena in order to provide correct and as robust as possible recommendations. The UTADIS model's assumptions mean that it provides recommendations for problems assuming monotonic preferences for individual criteria and treats all attribute values of the alternatives independently. It is worth noting that there are problems for which such a representation model is insufficient, which has led researchers to propose several modifications to the basic model described in the scientific literature.

The UTA$^{\text{GMS}}$-INT approach presented in [22] proposed a modification of the basic UTA method commonly used to solve ranking problems, enriching the model based on AVF with additional functions representing positive and negative values related to interactions occurring for pairs of criteria. In this way, the authors changed the approach to the evaluation of alternatives, allowing for the reflection of the dependence of DM's preferences on multiple criteria simultaneously. The adaptation of LP problem formulation was handled in two ways. In order to maintain the greatest possible interpretability of the model, during the first phase, the smallest possible subset of criterion pairs was identified, for which it was necessary to enrich the representation of preferences with the mentioned *synergy* and *redundancy* functions. In turn, the second phase allowed for obtaining recommendations using robustness analysis. Moreover, by introducing additional parameters, the model allowed determining the maximum value of the influence of the introduced interactions on evaluating the utility of alternatives and limiting the maximum number of interactions in which a single criterion can be involved. In the publication **P4** discussing issues related to models with diverse assumptions about the preference structure, it was proposed to adopt the UTA$^{\text{GMS}}$-INT approach to multi-criteria sorting problems by introducing threshold values that allow to determine the assignment of alternatives to classes.

The second trend of modifications to UTADIS concerned the introduction of the possibility of reflecting the non-monotonic nature of preferences with respect to individual criteria for evaluating alternatives. One of the first approaches to solving this problem was UTA-NM introduced in [46], which modified the UTA-Star variant of the model proposed in [47] while maintaining all the assumptions concerning the solution except for the aforementioned monotonicity of preferences with respect to the criteria. However, as the author noted, this approach was highly inefficient due to the long time

of optimization of the LP model, even for trivial decision problems. Another approach was proposed by [23], which, in the phase of formulating LP problem, avoided assumptions regarding both monotonicity and normalization of marginal functions, replacing them with boundary constraints. This enabled a consistent interpretation of the obtained comprehensive scores and relations between alternatives with the results obtained by other methods from the UTA family. Another model requiring post-normalization of the obtained model was the approach proposed in [24]. This method modeled the DM's non-monotonic preferences by doubling the marginal functions for each criterion, treating each of them bidirectionally - as a gain and a cost simultaneously, by providing a non-decreasing and non-increasing MVF, respectively. Such a formulation of the problem did not guarantee that the best and the worst possible decision alternatives would obtain a comprehensive value equal to 1 and 0, respectively; hence, for this procedure, it was also necessary to normalize the model and the recommendations provided.

In total, the described study included a detailed presentation of the six models considered. In addition to the basic **UTADIS** concept, the properties of two approaches adapting the UTA$^{\text{GMS}}$-INT approach to the sorting problem were investigated, one of which limited the number of possible interactions to a maximum of one per criterion (**UTADIS-INT-1**) and the other without any limit in this respect (**UTADIS-INT-∞**). In addition, three approaches accepting non-monotonic preferences were proposed, two of which were adapted from the model presented in [23]. The first proposed procedure (**UTADIS-NM-1**) enriched the proposed model with threshold values in order to adapt the procedure to the sorting problem. This model, in its original form, tried to capture the concept of searching for the most discriminant value function on the one hand, and minimize slope changes in piecewise linear MVF on the other, which was supposed to discourage the model from radical, non-monotonic changes in preferences. The second proposal (**UTADIS-NM-2**) ignored this aspect, focusing only on maximizing the discriminative ability of the solution. The last approach (**UTADIS-NM-3**) introduced a marginal modification to the original model described in [24], which proposed a solution to problems with multiple decision attributes.

The publication also describes tools explaining preference models that were used for subsequent analysis of the results, such as ROR and SOR. Robustness analysis focused on verifying the set of classes to which, based on the used model and DM's preferences, it was possible to assign an alternative, which enabled determining the set of all *Possible Class Assignments* (PCA) for a given alternative. On the other hand, stochastic analysis based on statistical analysis of a sampled subset of all feasible solutions, provided *Class Acceptability Indices* (CAI), indicating how often a given alternative is assigned to a particular class. The determined values formed the basis for defining the model quality measures described below.

The aim of the comparative experimental analysis was to examine the characteristics of individual models regarding robustness and expressiveness as well as the trade-off between these two important issues. The simulation design included a solution for which the sample assignments of non-dominated alternatives to classes provided by the DM were generated completely randomly to avoid biased results and lower the quality of the conclusions drawn from the study. The considered decision problems differed in the number of classes (from 2 to 5), criteria (from 2 to 5), the number of characteristic points in MVFs (from 2 to 5), and the number of reference assignments provided as input (from 1 to 5 for each decision class). The non-dominated set of alternatives constituting the input was determined in two ways: either it was generated using an iterative algorithm drawing attribute values from a uniform distribution while maintaining the lack of dominance relations in the set or by drawing a point from the unit m-sphere and then changing the point coordinates to the attribute values. The multivariate analysis assumed the solution of 64,000 problems of varying structure and difficulty by all six models considered. Then, the models were evaluated using seven quality measures, two of which focused on the expressiveness of the model and five on the robustness of the model and the recommendations provided.

The measure addressing the question of the model's expressiveness was *Preference recoverability*, which assigned 1 if it succeeded in recovering the DM's preferences and 0 otherwise. Averaging the values of this measure allowed us to answer the question of how well the model can represent inconsistent information about DM's beliefs. The second measure focused on the size of the slack variable (δ^*), which determined the minimal distance between the comprehensive value of an alternative and the boundary value of the class to which the model was forced to assign it by appropriate constraints. The formulation of all six models providing univocal recommendations assumed maximization of this value in different ways. Hence, their comparison provided an additional premise proving the expressiveness of the model.

When considering the robustness of the models, the quality measures focused on, on the one hand, assessing the level of agreement between all feasible solutions and, on the other hand, on how well they were represented by the returned univocal recommendations. The first two quality measures (*Average possible class assignments*, *Certain assignment ratio*) captured the qualitative results of the ROR analysis, checking the stability of the recommendation, measured by the number of classes to which the model could assign individual non-reference alternatives. In this context, the most robust model with the lowest possible uncertainty level would provide the possible and necessary assignment of an alternative to a single decision class. The next quality measure (*Entropy class acceptability index*) used the quantitative results from SOR, validating the entropy measure based on CAI of the

alternatives. The last two measures aimed at evaluating univocal recommendations, on the one hand, checked the mean acceptability of the decisions made (*Mean class acceptability index*), and on the other hand, the stability of class assignments with respect to potential modifications of threshold values (*Certain class assignments*).

The results of the analyses confirmed the highest expressiveness in the INT-∞ model, followed by NM approaches and INT-1, while the UTADIS model, as the most constrained one, gave the least chance of reproducing problems. On the other hand, considering the problems that UTADIS was able to solve, it provided the most robust recommendations, the significance of which was confirmed by all five quality measures. For the remaining problems solved by both INT approaches and all NM models, the most robust solutions turned out to be those provided by INT-1, which was also the least expressive of those mentioned. Further, when comparing INT-∞ with all non-monotonic approaches, it obtained better robustness results than the others on average, but when making a more detailed analysis, it turned out that this statement is not always true. As concluded from the results, the advantage of INT-∞ was noticeable only for problems for which the provided recommendations involved at most two active pairs of interacting criteria. In the case of three or more active synergy functions, more robust results are expected from the NM group of methods. Moreover, within the group of models addressing non-monotonicity of preferences, the NM-2 model turned out to obtain the most robust results, although its advantage was not as evident as in the case of the other comparisons discussed earlier. Similar to the previously discussed publications, the significance of all observed relationships was confirmed by the Wilcoxon signed-rank test with a p-value of 0.05.

The analysis confirmed that more expressive models generally provide less robust recommendations. In order to support the work of analysts, this publication proposes a solution to the problem of selecting an adequate model by using the proposed framework. First, the most robust approach (UTADIS) should be used, and if it is impossible to obtain a compatible solution, more sophisticated models should be considered, starting from INT-1 and INT-∞, with the proviso that they should only be used if the obtained recommendations involve at most two pairs of interacting criteria. If they are also unable to provide a feasible solution, then the NM-2 model should be used, and if that fails too, preference elicitation in cooperation with the DM is necessary. Nevertheless, it should be clearly stated that the proposed framework should be used only when information about the nature of preferences for individual criteria is unavailable and there is no way to determine it. In the case of an explicit DM's indications about the non-monotonicity of preferences or the need to express interactions, the decision analyst should select an adequate decision model that meets the DM's expectations.

The contribution of this publication is threefold. First, it has adapted models that modify the basic assumptions of the UTADIS approach and proposed an experimental setting that examines the properties of model expressiveness and recommendation robustness along with seven quality measures. Second, the study that was conducted experimentally provided evidence of the opposite nature of expressiveness and robustness. Third, it has provided a framework that proclaims guidelines for decision analysts and facilitates the selection of an adequate model for the considered problem, which is also the ultimate goal of this dissertation and confirms the research hypothesis.

3.4 Nature-inspired preference learning Choquistic approaches

One of the most critical issues discussed in the contemporary scientific literature on decision support approaches is Preference Learning problems, which, in their essence, combine issues known from MCDA and ML [32]. On the one hand, similarly to MCDA problems, they are based on the provided, indirect, and incomplete preference information about decision alternatives and retain a solution structure similar to MCDA problems, such as ranking or sorting of alternatives. Decision-making procedures are also typically intuitive and easy to interpret, and they provide the tools and evidence to support the recommendations and conclusions provided. Unfortunately, they often lose the quality of the derived solutions with the increase in the complexity of the problem and the data provided. On the other hand, ML-based approaches are capable of providing high-quality predictions even for large datasets. However, due to sophisticated, non-linear methods, the explainability of the judgments made is limited. The optimal Preference Learning solution would, therefore, be to use an approach that combines the best features of both concepts by providing a solution that is correct, accurate, robust, and easy to interpret, even for a large dataset with incomplete and often inconsistent preferential information.

One approach that may prove helpful in this context is the Choquet integral model, which has already been used to solve Preference Learning problems in [38]. This model calculates comprehensive scores of alternatives based on a non-additive, fuzzy measure that provides importance weights for all subsets of criteria in the considered problem, which allows for the capture of advanced preferences related to inter-criteria interactions. The obtained comprehensive values of alternatives, together with thresholds, allow for obtaining an unambiguous classification of individual alternatives, which, combined with the complexity and large possibilities of representing preferences by the model, allows for solving challenging multi-criteria sorting problems. Nevertheless, as noted in the aforementioned publication, the

problem of selecting appropriate model parameters (*capacities*) in the case of a large data set and rich preferential information, is too complex for the standard model optimization approach based on the LP problem formulation. For this purpose, a technique for solving such complex problems called *Cutting-plane method*, was used, which in the optimization phase initially ignores and then gradually introduces subsequent constraints resulting from the provided preferential information, in this case, pairwise comparisons of alternatives. The proposed approach allows for solutions that are not fully compatible with all provided statements, which is a common practice when solving ML and PL problems, as obtaining a feasible solution using a given model is often very difficult or impossible, and could also indicate its overfitting and reduced predictive capabilities. However, there exist many other optimization techniques that can effectively address the problem of determining the parameters' values of the Choquet integral model, the investigation of which was the subject of the research work described in publication **P3**.

The publication, together with supplementary materials, contains a detailed definition of the problem, including a description of symbols and concepts, and, most importantly, a formal definition of the Choquet integral model, which is the basis for the threshold-based sorting procedure, together with an explanation and a practical demonstration of how the procedure evaluates and assesses alternatives. The publication also proposes eight different approaches to searching for an accurate model, inspired by well-known and commonly used optimization techniques. Two implement the classical approach to optimizing model parameters, frequently used in MCDA, by formulating and solving the Mathematical Programming problem. To reduce the complexity of the problem and the time needed to obtain a solution, the bagging-inspired approach [48] was used so that instead of solving the entire mathematical optimization problem by minimizing the number of errors (**MNR**) or the maximum error (**MMR**) across all preference-reflecting recommendations, these approaches solved many simple problems, which involved only a subset of the preference statements as input. In addition, each of the three methods has a given *patience* parameter, which indicates how long the algorithm should try to stick to the current solution in the absence of improvement over several iterations. If the number of iterations without improvement in the quality of the solution exceeds the parameter value, then the search starts with a newly drawn model.

The next group of methods included three procedures inspired by a local search in the solution space. The operation of all methods was based on the neighborhood relation, which was true for a pair of solutions for which the Euclidean distance between vectors containing all model parameters was smaller than the assumed radius. For these approaches to work, it was also necessary to introduce the *loss* and *regret* functions. Regret symbolizes the difference between the comprehensive value of the alternative and the range of values

assigned to the desired class. If the alternative is correctly assigned by the model, then regret is equal to 0, and otherwise, it reaches positive values. In turn, the model's loss function aggregates the regret values for all reference alternatives, by averaging them, hence models with the lowest possible loss function value are preferred. These concepts are important for understanding the above-mentioned methods, which differ from each other in their approach to accepting a newly selected, neighboring solution. The first one (**GLS**) generates one new solution in each iteration and accepts it if its loss function is lower than the previously chosen one. The second approach (**SLS**) generates multiple neighboring solutions simultaneously, selects the best candidate, and accepts it based on the same rule. The last approach (**SAN**), implementing the *Simulated Annealing* approach [49], generates one neighbor and then accepts it unconditionally in case of improvement of the solution quality or in case of deterioration, with a certain probability, depending on the difference of the loss function between the compared models.

The last group of methods are nature-inspired metaheuristics, which allow for optimization by considering and evolving multiple solutions simultaneously. The first one is the Genetic Algorithm (**GEN**) [50], based on the concept of natural selection, which, by mutating, crossing, and selecting the population of solutions, creates subsequent generations of models. By using evolutionary pressure, this approach can lead to gradual improvement and find better solutions. The second approach is Fish School Search (**FSS**) described in [51], inspired by the movement of schools of fish searching for food, the specificity of which is used to search for better solutions. The last concept is Particle Swarm Optimization (**PSO**), introduced in [52], inspired by the dynamics of movements in large swarms of birds. It assumes that each particle iteratively moves through the solution space according to a vector representing its velocity, which is attracted by the best solution found so far by the given particle and by the whole swarm.

Two post-optimization techniques have also been proposed. One of them selects such threshold values that maximize classification accuracy for reference alternatives. The other uses an approach inspired by the backpropagation algorithm, which is commonly used for training neural networks [53]. This idea consists of updating the model parameters in such a way as to reduce the regret for each of the considered reference alternatives. A preliminary analysis was also conducted to confirm this technique's usefulness and determine the optimal proportion of time spent on its operation in relation to the time spent on optimization using one of the eight algorithms. Additionally, to protect the optimization procedures and backpropagation algorithm from obtaining solutions that are inconsistent with the monotonicity and normalization assumptions of Choquet integral capacities, they are prevented by shortening the solution shift vector so that the resulting model is compatible with the Choquet integral assumptions.

The experimental analysis comparing the proposed approaches included the evaluation of the algorithms on five benchmark binary classification problems, containing from 4 to 7 criteria and from over 200 to 1000 alternatives, with three different ratios (20-80, 50-50, 80-20) of assigning alternatives to reference and non-reference subsets, respectively. The analysis measured two indicators of model quality: *acc* representing the classification accuracy, and *auc* indicating the ratio of correctly reflected pairwise comparisons for non-reference alternatives that were originally assigned to different classes.

It should be emphasized that each of the considered algorithms provided the possibility of influencing a diverse number of hyperparameters. For this reason, the study was divided into two phases. In the first phase, for each combination including the algorithm, the considered benchmark problem, and the proportion of the division of alternatives, the optimization of the algorithm hyperparameters was carried out, using 10-fold Monte Carlo Cross-Validation and setting the algorithm stop condition and post-optimization techniques to ten seconds. This phase aimed to select the best set of hyperparameters, for which the number of combinations varied from a dozen for mathematical programming approaches to over two hundred for the genetic algorithm. The second phase, aimed at comparing the proposed algorithms, performed 100-fold Monte Carlo Cross-Validation for algorithms with selected hyperparameter values, with a 30-second execution timeout.

The analysis of the average values for both quality measures consistently confirmed that regardless of the considered split ratio, the best results were obtained by nature-inspired approaches, consistently outperforming the other approaches according to the average position in the algorithm rankings created for each benchmark dataset. In particular, for problems with the lowest preferential information richness, containing 20% of alternatives with reference assignment, the best results were obtained by PSO, followed by GEN, and FSS. For the 50-50 split, the best performing algorithm was GEN, followed by PSO, while FSS obtained results comparable to those of the SLS approach. The results for problems with 80% of the reference alternatives highlighted the even greater dominance of GEN over the other approaches and showed a comparable quality of solutions generated by PSO and FSS.

The conducted research provides a new perspective on the optimization of the Choquet Integral model parameters in the context of its application in preference learning problems, along with the description of eight dedicated algorithms. Moreover, a comparative experimental analysis framework was proposed, assuming two phases of experiments and examining two different aspects of the recommendations provided. The results clearly indicated the dominance of nature-inspired approaches over the other methods considered. Finally, the conclusions from the conducted analysis provide guidelines for selecting an adequate approach to the problem, depending on the availability of information about preferences, which confirms the dissertation hypothesis.

Chapter 4

Summary

The increasing interest in MCDA methods, which indirectly process DM's statements and capture preferences in a holistic manner, has led to the development of numerous approaches consistent with the disaggregation paradigm. The popularity of these approaches is a direct result of their great usefulness and simplicity, high interpretability and universality, and the low cognitive effort required by the DM. Undoubtedly, the recent developments in this field have made a significant contribution to the scientific literature on decision support. However, this also means that the role of decision analysts has become more crucial than ever. They are now faced with the increasingly important task of selecting the right tools to effectively solve real-world problems and formulate valuable recommendations.

Analyzing issues related to the accuracy, robustness, expressiveness, intuitiveness, and interpretability of models is an important research aspect because it reveals the strengths and weaknesses of individual approaches. Observations of these aspects in an experimental environment, considering many problems with different characteristics, allow for the determination of the credibility of specific methods and indicate the circumstances in which a given decision procedure should be used. This may constitute a premise for formulating guidelines that, based on the characteristics of the problem under consideration, can support the process of selecting an adequate decision-making procedure. The research work aimed to demonstrate that it is possible to formulate such guidelines based on the results of an experimental comparative analysis of models and methods exploiting holistic DM's preferences.

One of the challenges posed in this dissertation concerned providing highly interpretable and, at the same time, qualitative recommendations for multicriteria choice, ranking, and sorting problems. Considering this issue, the scope was set on the family of UTA methods capable of providing univocal recommendations. Their diversity, as well as the lack of comprehensive consideration and comparison of these approaches in the scientific literature,

made it essential to address this issue. The conducted experimental analyses revealed features indicating the accuracy and robustness of the provided recommendations.

For the ranking and choice problems, thirty-five procedures originating from four method streams were compared based on seven quality indicators. The analysis was focused on the compliance of the provided recommendations with the DM's preferences, the compatibility of the provided univocal solution with the set of all feasible models, and the internal consistency of the entire space of compatible models and the recommendations they provide. The multivariate evaluation allowed for obtaining conclusions that clearly identify procedures that give significantly better results than the others, taking into account the analyzed priority indicator of decisions made. More specifically, the results confirmed the high usefulness of the procedures based on the exploitation of the Stochastic Ordinal Regression results for both types of problems considered. The conclusions provided allowed the formulation of a set of guidelines that facilitate the process of selecting an appropriate procedure for a specific multi-criteria problem.

A similar study was proposed for procedures oriented toward solving the sorting problem. Fourteen procedures were compared, including three novel approaches, focused on finding solutions based on the results of stochastic analysis, providing observations on all feasible solutions. Five quality measures were proposed, covering such aspects as the consistency of recommendations with DM's preferences, the credibility of recommended decisions by assessing their representativeness, and the similarity of the derived solution to the reference model representing DM's beliefs. The results of the experimental analysis proved that the novel procedures provided the most robust recommendations. On the other hand, the methods dedicated to central and average solutions, with the approach of determining the analytical center of the polygon constituting all feasible solutions, most accurately reproduced the DM's preferences. Regardless of the structure and size of the considered sorting problem, the derived results of multivariate analysis clearly confirm the advantage of the approaches mentioned above, constituting a basis for formulating recommendations for decision analysts.

The next issue considered was the comparison of different models solving sorting problems in the disaggregation preference paradigm. The scope of the study included examining the properties of one of the most popular representatives of this paradigm – the UTADIS model along with five modifications of this model, introducing the possibility of representing preferences with respect to inter-criteria interactions and their non-monotonic nature. The analysis performed included the assessment of the model expressiveness and verification of the robustness of the results provided by the models. The study reveals the contradictory nature of these two issues, showing that the more expressive the model is, the less robust recommendations it offers.

In order to provide qualitative statements representing DM's beliefs, the research results suggest applying models, starting from the most robust, basic UTADIS model. In case of a lack of ability to fully reflect DM's preferences, the framework further indicates models expressing inter-criteria interactions, assuming their limited use to a maximum of two active synergies for pairs of criteria. Otherwise, a non-monotonic approach should be used, and in case of further inability to represent DM's preferences, preference elicitation should be performed. Moreover, the developed guidelines should only be applied when the DM's attitude toward interacting and non-monotonic preferences is unknown and difficult to determine.

The last issue considered in this dissertation is a comparative analysis of approaches to solving Preference Learning problems in the context of providing solutions to the binary classification problem. The desired features of the provided solutions are: high accuracy and interpretability, the ability to efficiently process large data sets, and robustness to inconsistencies contained in preferential information. The Choquet integral model, capable of representing interacting DM's preferences, was chosen to represent preferences in this context. The experimental analysis included the study of different approaches to solve the problem of establishing the model parameters' values. Their determination is an important aspect because standard methods based on Linear Programming formulation are time-consuming.

In the conducted research, eight different optimization approaches were proposed and compared, along with two post-optimization techniques. Their evaluation was based on experimental comparative analysis using five benchmark datasets and assessment on two quality measures, reflecting the accuracy of statements referring to unambiguous assignments to classes and the relations for pairs of alternatives from different decision classes. The results confirmed the superiority of procedures based on nature-inspired metaheuristics over classical approaches based on mathematical programming and implementing local search strategies in both aspects. In particular, the Particle Swarm Optimization methods are recommended for problems with poor DM preference representation, and the Genetic Algorithm should be used in the remaining cases. Such clear guidelines provide grounds for selecting an adequate method for model optimization under time-constrained conditions.

The common feature of all the research conducted was their experimental nature, providing empirical evidence of differences in the broadly understood *quality* of solutions offered by specific procedures. Additionally, to increase the comprehensibility of the presented issues, each of the discussed studies was enriched with a detailed description of the considered procedures, models, and decision support methods, along with a practical visualization of their application based on an illustrative study. The study identified the most important aspects to consider when selecting an adequate decision aiding procedure.

Moreover, the high usability and effectiveness of the proposed comparative experimental analyses have been proven. This approach provided many observations and evidence regarding qualitative aspects of approaches oriented towards providing representative, univocal preferences, the contrasting nature of expressiveness of models, and the robustness of recommendations dependent on the assumptions of a specific preference disaggregation model and the accuracy of statements provided by various Choquet integral model optimization procedures addressing preference learning problems.

In addition to the analyses performed, newly proposed procedures and algorithms, as well as adaptations of existing approaches to the problems being solved, have contributed significantly to the state-of-the-art knowledge in the field of MCDA. Furthermore, several quality measures related to various features of methods and recommendations have been proposed, which facilitate the consideration and empirical capture of different aspects desired in decision support approaches.

Noticing and measuring the similarities and differences between the various procedures, models, and algorithms allowed for the formulation of comprehensive guidelines indicating the best approaches in the domains and types of problems. It should be strongly emphasized that the presented conclusions and the guidelines derived from them are applicable only under the condition of maintaining specific features of the decision-making process, the considered problem, and, in some cases, the characteristics of the solutions obtained. Overall, the conducted research fulfilled the research objectives. The provided conclusions, results, and the above-mentioned formulated guidelines confirm the research hypothesis.

Future research directions may include further development of evidence-based guidelines for decision analysts to support their work. It would also be desirable to create universal principles for evaluating newly proposed approaches, allowing them to be presented along with indications of their advantages and evidence of their high utility in a given context. These principles could include a comprehensive set of diverge measures focused on particular quality aspects of the models, along with a specific collection of benchmark datasets, in order to achieve repeatability of the analysis and increase the credibility of the presented results. Lastly, it would be beneficial to create a universal decision support meta-procedure, capable of recommending an appropriate method based on limited information about the considered problem, without detailed indications regarding the characteristics of the alternatives, criteria, and the DM's preferences. Such a method would fit into the above-mentioned paradigm of preference disaggregation, deriving recommendations based on indirect statements and reducing the cognitive effort needed to apply such an approach.

Bibliography

[1] M. Kadziński, M. Wójcik, and K. Ciomek, "Review and experimental comparison of ranking and choice procedures for constructing a univocal recommendation in a preference disaggregation setting", *Omega*, vol. 113, p. 102715, 2022.

[2] 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*, vol. 278, p. 110871, 2023.

[3] M. Wójcik and M. Kadziński, "Nature-inspired Preference Learning Algorithms Using the Choquet Integral", in *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '24, (New York, NY, USA), p. 440–448, Association for Computing Machinery, 2024.

[4] M. Kadziński, M. Wójcik, and M. Ghaderi, "From investigation of expressiveness and robustness to a comprehensive value-based framework for multiple criteria sorting problems", *Omega*, 2024. Status: accepted for publication.

[5] E. Paschetta and A. Tsoukiàs, "A real-world mcda application: evaluating software", *Journal of Multi-Criteria Decision Analysis*, vol. 9, no. 5, p. 205–226, 2000.

[6] J. Siskos and N. Assimakopoulos, "Multicriteria highway planning: A case study", *Mathematical and Computer Modelling*, vol. 12, no. 10–11, p. 1401–1410, 1989.

[7] A. Nikas, H. Doukas, E. Siskos, and J. Psarras, *International Cooperation for Clean Electricity: A UTASTAR Application in Energy Policy*, p. 163–186. Springer International Publishing, 2018.

[8] I. Khan, L. Pintelon, and H. Martin, "The application of multicriteria decision analysis methods in health care: A literature review", *Medical Decision Making*, vol. 42, no. 2, p. 262–274, 2021.

[9] B. Adem Esmail and D. Geneletti, "Multi-criteria decision analysis for nature conservation: A review of 20 years of applications", *Methods in Ecology and Evolution*, vol. 9, no. 1, p. 42–53, 2018.

[10] B. Kizielewicz and L. Dobryakova, "Mcda based approach to sports players' evaluation under incomplete knowledge", *Procedia Computer Science*, vol. 176, p. 3524–3535, 2020.

[11] M. Cinelli, P. Burgherr, M. Kadziński, and R. Słowiński, "Proper and improper uses of mcda methods in energy systems analysis", *Decision Support Systems*, vol. 163, p. 113848, 2022.

[12] 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*, vol. 173, p. 281–290, 2018.

[13] B. Roy, "The optimisation problem formulation: Criticism and overstepping", *Journal of the Operational Research Society*, vol. 32, no. 6, p. 427–436, 1981.

[14] V. Mousseau, "A general framework for constructive learning preference elicitation in multiple criteria decision aid". working paper or preprint, 2005.

[15] E. Jacquet-Lagrèze and Y. Siskos, "Assessing a set of additive utility functions for multicriteria decision making: the UTA method", *European Journal of Operational Research*, vol. 10, pp. 151–164, 1982.

[16] M. Doumpos and C. Zopounidis, *Disaggregation Approaches for Multicriteria Classification: An Overview*, p. 77–94. Springer International Publishing, 2018.

[17] Y. Siskos, E. Grigoroudis, and N. F. Matsatsinis, *UTA Methods*, p. 297–334. Springer-Verlag, 2005.

[18] J. Devaud, G. Groussaud, and E. Jacquet-Lagreze, "Utadis: Une méthode de construction de fonctions d'utilité additives rendant compte de jugements globaux", *European Working Group on Multicriteria Decision Aid, Bochum*, vol. 94, pp. 285–298, 1980.

[19] S. Greco, R. Słowiński, J. R. Figueira, and V. Mousseau, *Robust Ordinal Regression*, p. 241–283. Springer US, 2010.

[20] R. L. Smith, "Efficient monte carlo procedures for generating points uniformly distributed over bounded regions", *Operations Research*, vol. 32, no. 6, p. 1296–1308, 1984.

[21] 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", 2021.

[22] S. Greco, V. Mousseau, and R. Słowiński, "Robust ordinal regression for value functions handling interacting criteria", *European Journal of Operational Research*, vol. 239, no. 3, p. 711–730, 2014.

[23] M. Ghaderi, F. Ruiz, and N. Agell, "A linear programming approach for learning non-monotonic additive value functions in multiple criteria decision aiding", *European Journal of Operational Research*, vol. 259, no. 3, p. 1073–1084, 2017.

[24] M. Kadziński, K. Martyn, M. Cinelli, R. Słowiński, S. Corrente, and S. Greco, "Preference disaggregation method for value-based multi-decision sorting problems with a real-world application in nanotechnology", *Knowledge-Based Systems*, vol. 218, p. 106879, 2021.

[25] G. Choquet, "Theory of capacities", *Annales de l'institut Fourier*, vol. 5, p. 131–295, 1954.

[26] R. Pelissari, A. Abackerli, and L. T. Duarte, "Choquet capacity identification for multiple criteria sorting problems: A novel proposal based on Stochastic Acceptability Multicriteria Analysis", *Applied Soft Computing*, vol. 120, p. 108727, 2022.

[27] H. Q. Vu, G. Beliakov, and G. Li, "A Choquet Integral Toolbox and Its Application in Customer Preference Analysis", in *Data Mining Applications with R* (Y. Zhao and Y. Cen, eds.), pp. 247–272, Boston: Academic Press, 2014.

[28] F. H. Barron and B. E. Barrett, "Decision quality using ranked attribute weights", *Management Science*, vol. 42, no. 11, p. 1515–1523, 1996.

[29] M. Kendall, *Rank Correlation Methods*. C. Griffin, 1948.

[30] C. Spearman, "The proof and measurement of association between two things", *The American Journal of Psychology*, vol. 15, no. 1, p. 72, 1904.

[31] B. Roy, "Robustness in operational research and decision aiding: A multi-faceted issue", *European Journal of Operational Research*, vol. 200, no. 3, p. 629–638, 2010.

[32] J. Fürnkranz and E. Hüllermeier, *Preference Learning*. Springer, 2011.

[33] E. Hüllermeier and R. Słowiński, "Preference learning and multiple criteria decision aiding: differences, commonalities, and synergies–part i", *4OR*, 2024.

[34] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision making and a "right to explanation"", *AI Magazine*, vol. 38, no. 3, p. 50–57, 2017.

[35] W. Kotlowski and R. Slowinski, "On nonparametric ordinal classification with monotonicity constraints", *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 11, p. 2576–2589, 2013.

[36] J. Liu, M. Kadziński, X. Liao, and X. Mao, "Data-driven preference learning methods for value-driven multiple criteria sorting with interacting criteria", *INFORMS Journal on Computing*, 2020.

[37] M. Kadziński and A. Szczepański, "Learning the parameters of an outranking-based sorting model with characteristic class profiles from large sets of assignment examples", *Applied Soft Computing*, vol. 116, p. 108312, 2022.

[38] A. F. Tehrani, W. Cheng, and E. Hullermeier, "Preference learning using the choquet integral: The case of multipartite ranking", *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 6, pp. 1102–1113, 2012.

[39] K. Martyn and M. Kadziński, "Deep preference learning for multiple criteria decision analysis", *European Journal of Operational Research*, vol. 305, no. 2, p. 781–805, 2023.

[40] F. Wilcoxon, "Individual comparisons by ranking methods", *Biometrics Bulletin*, vol. 1, no. 6, p. 80, 1945.

[41] M. Doumpos, C. Zopounidis, and E. Galariotis, "Inferring robust decision models in multicriteria classification problems: An experimental analysis", *European Journal of Operational Research*, vol. 236, no. 2, pp. 601–611, 2014.

[42] S. Greco, M. Kadziński, and R. SŁowiński, "Selection of a representative value function in robust multiple criteria sorting", *Computers & Operations Research*, vol. 38, no. 11, p. 1620–1637, 2011.

[43] J. Branke, S. Greco, R. Slowinski, and P. Zielniewicz, "Learning value functions in interactive evolutionary multiobjective optimization", *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 1, p. 88–102, 2015.

[44] 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*, vol. 10, no. 2, pp. 151–164, 1982.

[45] M. Kadziński, M. Cinelli, K. Ciomek, S. R. Coles, M. N. Nadagouda, R. S. Varma, and K. Kirwan, "Co-constructive development of a green chemistry-based model for the assessment of nanoparticles synthesis", *European Journal of Operational Research*, vol. 264, no. 2, p. 472–490, 2018.

[46] T. Kliegr, "Uta-nm: Explaining stated preferences with additive non-monotonic utility functions", *Preference Learning*, vol. 56, 2009.

[47] Y. Siskos and D. Yannacopoulos, "Utastar: An ordinal regression method for building additive value functions", *Investigaçao Operacional*, vol. 5, no. 1, pp. 39–53, 1985.

[48] L. Breiman, "Bagging predictors", *Machine Learning*, vol. 24, no. 2, p. 123–140, 1996.

[49] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing", *Science*, vol. 220, no. 4598, p. 671–680, 1983.

[50] A. Lambora, K. Gupta, and K. Chopra, "Genetic algorithm- a literature review", in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, IEEE, 2019.

[51] C. J. A. Bastos Filho, F. B. de Lima Neto, A. J. C. C. Lins, A. I. S. Nascimento, and M. P. Lima, "A novel search algorithm based on fish school behavior", in *2008 IEEE International Conference on Systems, Man and Cybernetics*, IEEE, 2008.

[52] J. Kennedy and R. Eberhart, "Particle swarm optimization", in *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4 of *ICNN-95*, p. 1942–1948, IEEE, 1995.

[53] Hecht-Nielsen, "Theory of the backpropagation neural network", in *International Joint Conference on Neural Networks*, IEEE, 1989.

Publication reprints

Publication [P1]

M. Kadziński, M. Wójcik, and K. Ciomek, "Review and experimental comparison of ranking and choice procedures for constructing a univocal recommendation in a preference disaggregation setting", *Omega*, vol. 113, p. 102715, 2022

DOI: 10.1016/j.omega.2022.102715.

Number of citations[1]:

- according to Web of Science: 6
- according to Google Scholar: 8

[1] as of September 23, 2024

Omega 113 (2022) 102715

Contents lists available at ScienceDirect

Omega

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

Review and experimental comparison of ranking and choice procedures for constructing a univocal recommendation in a preference disaggregation setting☆

Miłosz Kadziński*, Michał Wójcik, Krzysztof Ciomek

Institute of Computing Science, Poznan University of Technology, Piotrowo 2, Poznań 60-965, Poland

ARTICLE INFO

Article history:
Received 23 September 2021
Accepted 11 June 2022
Available online 14 June 2022

Keywords:
Multiple criteria decision analysis
Preference disaggregation
Ranking
Univocal recommendation
Robustness analysis

ABSTRACT

We account for the preference disaggregation setting given multiple criteria ranking and choice problems. An assumed preference model is a set of additive value functions compatible with the Decision Maker's pairwise comparisons of reference alternatives. The incompleteness of such indirect preferences implies the multiplicity of feasible functions and the ambiguity in indicating the most preferred alternative or ordering alternatives from the best to the worst. We review approaches that construct a univocal recommendation under such scenarios. They represent four groups of methods: procedures selecting a representative value function, decision rules, scoring methods, and mathematical models for constructing a robust ranking. The use of all thirty-five approaches is illustrated on a simple decision problem. Then, they are compared in an extensive computational study in terms of their abilities to reconstruct the DMs' true preferences and robustness of delivered recommendations given the support they are given in the set of all compatible models. The results are quantified in terms of seven performance measures. Their analysis indicates that in the context of choice, it is beneficial to consider the rank acceptabilities for the best ranks. For ranking problems, the most advantageous outcomes are attained by procedures that emphasize the most frequent relations or positions in the feasible polyhedron. Apart from the average results, we discuss how the performance of all approaches changes for different parameterizations of the decision problem and preference model.

1. Introduction

Multiple criteria ranking and choice are among the most frequent real-world decision problems [18]. The former aims at ordering the set of alternatives from the best to the worst, whereas the latter is oriented towards selecting a small subset of the most preferred options. Both types of problems are solved using a relative comparison approach that combines two sorts of information: a dominance relation and Decision Maker's (DM's) preferences [30]. The preference information enriches the dominance, making the alternatives more comparable given the conflicting nature of the criteria.

The preferences about the problem and model parameters used in the Multiple Criteria Decision Aiding (MCDA) methods may be complete or incomplete [11,13,48]. The complete preferences guarantee precision and direct impact of the DMs on how their value systems are represented within the method. Their use is advised when the DMs have a thorough understanding of the employed MCDA approach and the respective model parameters and feel confident about providing precise inputs. On the contrary, the incomplete preferences take the form of imprecise statements that are translated into constraints on admissible parameter values of an assumed preference model [48]. Alternatively, they may emerge as incomplete holistic judgments concerning a small subset of reference alternatives [26]. In this way, the DMs are not forced to provide exact estimates of the parameter values. Moreover, the use of incomplete preferences requires lesser cognitive effort on the part of DMs, allowing them to exercise their decisions [11]. However, this is at the cost of trusting the mechanism of disaggregating holistic statements into the compatible parameter values, the ambiguity of representing the DM's preferences by the assumed model, and, typically, accepting some level of equivocality in the suggested recommendation.

In this paper, we focus on the most popular preference disaggregation method, called UTA [25]. It incorporates pairwise com-

☆ Area: Decision Analysis and Preference-Driven Analytics. This manuscript was processed by Associate Editor Banu Lokman.

* Corresponding author at: Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland.

E-mail addresses: milosz.kadzinski@cs.put.poznan.pl (M. Kadziński), 96michal.wojcik@gmail.com (M. Wójcik), k.ciomek@gmail.com (K. Ciomek).

https://doi.org/10.1016/j.omega.2022.102715
0305-0483/

parisons of reference alternatives to infer a set of compatible additive value functions. For this purpose, it employs dedicated linear programming techniques [51]. Usually, among many functions consistent with the DM's preferences, a single one is selected to impose a complete order on the set of alternatives or identify the most preferred option. The UTA method has been appreciated in the MCDA community for using a highly interpretable preference model that differentiates between inter- and intra-criteria attractiveness and exhibiting a direct link between the input preferences and output recommendation. Such engaging characteristics motivated its use in real-life decision problems concerning, e.g., marketing and development of new products [42], environmental management [50], energy policy [44], project portfolio selection [59], e-government benchmarking [49], or pharmaceutical strategy determination [38].

The basic variant of UTA has been extended in numerous ways. When it comes to the accepted preference information, the enriched approaches accept other types of holistic judgments, including preference intensities [15], rank-related requirements [32], and uncertain pairwise comparisons [10]. Moreover, some active learning strategies have been proposed to maximize the information gain from the provided indirect preference information and minimize the number of iterations needed to arrive at a sufficiently decisive recommendation [8,9]. As far as the employed preference model is concerned, the revised variants of UTA accept general [21], recursive exponential [2], polynomial, or splined [53] marginal functions instead of piecewise linear ones. Other model-oriented developments admit data-driven selection of characteristic points [29], non-monotonicity of per-criterion preferences [17,45], and interactions between criteria [23,24]. The preference disaggregation procedures were also extended to define various errors quantifying the consistency between the supplied and obtained comparisons or rankings [51]. Moreover, some consistency restoration procedures were devised to suggest which DM's statements should be modified or withdrawn [21].

Further methodological advancements have been devoted to robustness analysis, providing dedicated explanations, and addressing various structures and types of decision problems. The robust methods exploit the multiplicity of compatible value functions to quantify the necessary, possible [21], extreme [30], and probabilistic [34] consequences of their application on the set of alternatives. Other robustness indices for quantifying the stability of both an additive value model (e.g., the average range of the preferential parameters and average stability index) and recommended results have been proposed in [40]. Furthermore, techniques for generating dedicated explanations of such outcomes were introduced in [28], whereas measures for quantifying the stability of results were proposed in [8,54]. Also, [12] adopted UTA to handle criteria organized in hierarchical structures, whereas [51] adjusted it to problems under uncertainty. Finally, group decision methods were elaborated for arriving at a consensus recommendation in a preference disaggregation setting [19,41]. A plethora of real-world applications and methodological extensions confirm the usefulness and importance of UTA. However, the major problem related to the practical use of incomplete preferences in UTA derives from multiple or even infinitely many instances of the preference model compatible with the DM's indirect statements. It is so because their application on the set of alternatives potentially leads to ambiguous recommendations [21]. In general, preference information can be completed by eliciting additional preference judgments. However, in many scenarios, the possibility of continuing such an elicitation process is limited [9,48]. As a result, the robustness analysis methods mentioned above often leave the problem far from being solved, failing to provide a complete ranking or indicate the most preferred alternative. Moreover, the analysis of multiple preference model instances is too abstract for many users who, in turn, are used to analyzing a synthetic, precise solution to the problem at hand [31].

This paper deals with methods that derive a univocal recommendation in the preference disaggregation setting. We review approaches that support the DMs in concluding which alternative is the most preferred or ordering alternatives from the most to the least preferred, even if their preferences are incomplete. These techniques can be divided into four groups. First, we can select a single value function representing the feasible polyhedron [3,31]. The procedures serving this purpose are based on different principles, identifying the most discriminant [3], average [25], central [4], benevolent [5], parsimonious [22], or robust [31,34] model. However, they all deliver a function that can be displayed to the DM and derive a precise recommendation. The second group comprises decision rules that implement arbitrary criteria for developing a univocal outcome [33,48]. Examples include maximax and minimax rules, maximization of expected value or likelihood, and minimization of regret or unlikelihood. The third subset includes scoring procedures that exploit the outcomes of robustness analysis for deriving a comprehensive measure of desirability. In this case, the intermediate results concern pairwise value differences between alternatives [39] or the shares of value functions confirming the advantage of some options over others [33]. Finally, the last stream proposes mathematical programming models for constructing a complete ranking based on the stochastic results [58]. They maximize the support given to the elementary pairwise relations or assignments of alternatives to specific ranks by all compatible value functions. Overall, we describe 35 procedures that find use in the context of multiple criteria ranking and/or choice.

The scientific literature offers limited guidance as to which methods for deriving a univocal recommendation should be used [48]. The arguments that can be taken into account when conducting such a selection are diverse. The first builds on whether the recommendation is associated with singling out a compatible preference model instance [31,58]. Then, we can refer to the ordinal [58] or cardinal [3,39] character of the scale leading to the ranking or choice. We may also consider if the robustness concern is incorporated in the process by referring to the outcomes obtained with all feasible models [31,39,58]. The computational cost may be accounted for because the execution of some procedures is time-consuming [31,34], and others are hardly applicable for large sets of alternatives due to the low efficiency of contemporary solvers [58]. Furthermore, some researchers point out that the suitability of methods may differ depending on the problem context. For example, when modest stakes are involved, one may consider some central estimates; however, more precautionary procedures should prevail with high stakes [48].

The above aspects are evident from the description of each method. Their consideration may lead to a subjective selection of the most suitable procedure for a given problem. In this paper, we add a pair of more objective features that may be accounted for when deciding on which approach for constructing a univocal recommendation should be used. On the one hand, we refer to the ability to reconstruct the entire ranking or indicate the most preferred alternatives based on incomplete preference information. On the other hand, we verify the robustness of provided recommendation in terms of the support all compatible value functions give it. These general ideas are materialized with seven measures, making the comparison of all procedures meaningful. The results are derived from an extensive computational study involving problems with different numbers of alternatives, criteria, characteristic points of marginal value functions, and pairwise comparisons of reference alternatives. In the experiment, we focus on additively rational DMs whose pairwise comparisons are consistent with an assumed model and relatively small MCDA problems for which both the choice and ranking recommendation may be of interest

to the DMs. We discuss the average results attained for all scenarios and the performance trends observable with increasing problem's complexity, preference model's flexibility, and availability of holistic judgments.

The paper's remainder is organized in the following way. Section 2 reminds UTA and its robust extensions. In Section 3, we discuss various procedures for constructing a univocal ranking and choice recommendations in a preference disaggregation setting. Their use is illustrated on a didactic example in the e-Appendix (supplementary material available online). Section 4 presents the results of an extensive experimental study. The last section concludes the paper.

2. Reminder on UTA and robustness analysis

The following notation is used in the paper:

- $A = \{a_1, a_2, \ldots, a_i, \ldots, a_n\}$ – a finite set of n alternatives; each evaluated in terms of m criteria;
- $A^R = \{a_1^*, a_2^*, \ldots a_r^*\}$ – a finite set of r reference alternatives; $A^R \subseteq A$;
- $G = \{g_1, g_2, \ldots, g_j, \ldots, g_m\}$ – a finite set of m evaluation criteria, $g_j : A \to \mathbb{R}$ for all $j \in J = \{1, \ldots, m\}$; without loss of generality, we assume that all of criteria in G are of gain type;
- $X_j = \{g_j(a_i), a_i \in A\}$ – a finite set of performances of all alternatives in A on criterion g_j;
- $x_j^1, x_j^2, \ldots, x_j^{n_j(A)}$ – the ordered values of X_j, $x_j^{k-1} < x_j^k, k = 2, \ldots, n_j(A)$, where $n_j(A) = |X_j|$ and $n_j(A) \leq n$;
- $rank : A \to N \in \{1, \ldots, n\}$ – a function indicating the alternative's rank.

To compute a comprehensive score of alternative $a \in A$, UTA [25] considers an Additive Value Function (AVF) [35]:

$$U(a) = \sum_{j=1}^{m} u_j(g_j(a)), \forall a \in A \tag{1}$$

where u_j, $j = 1, \ldots, m$, are Marginal Value Function (MVF) being piecewise linear monotonic and defined by a pre-defined number γ_j of equally distributed characteristic points $\beta_j^1, \beta_j^2, \ldots, \beta^{\gamma_j}$, such that:

$$\beta_j^s = x_j^1 + (x_j^{n_j(A)} - x_j^1)\frac{s-1}{\gamma_j - 1}, \ j = 1, \ldots, m, s = 1, \ldots, \gamma_j. \tag{2}$$

A comprehensive value is normalized in the $[0, 1]$ range by assuming that $u_j(\beta_j^1) = 0$, for $j = 1, \ldots, m$, and $\sum_{j=1,\ldots,m} u_j(\beta_j^{\gamma_j}) = 1$. To enable control over the difference between marginal values assigned to the subsequent characteristic points, we consider the ρ variable defined as follows:

$$u_j(\beta_j^s) - u_j(\beta_j^{s-1}) \geq \rho, \ j = 1, \ldots, m, s = 2, \ldots, \gamma_j. \tag{3}$$

In the basic setting, ρ is set to zero. The marginal value for performance $x_j^k \in [\beta_j^s, \beta_j^{s+1}]$ can be computed using a linear interpolation:

$$u_j(x_j^k) = u_j(\beta_j^s) + (u_j(\beta_j^{s+1}) - u_j(\beta_j^s))\frac{x_j^k - \beta_j^s}{\beta_j^{s+1} - \beta_j^s}, \ j = 1, \ldots, m, \ k = 1, \ldots, n_j(A). \tag{4}$$

UTA infers the parameters of AVF from the DM's pairwise comparisons of reference alternatives $a^*, b^* \in A^R$, indicating either indifference ($a^* \sim b^*$) or preference ($a^* \succ b^*$) relation. Such holistic judgments are reproduced via preference disaggregation as follows:

$$\forall_{a^*, b^* \in A^R} \ a^* \sim b^* \Rightarrow U(a^*) - U(b^*) = 0, \tag{5}$$

$$\forall_{a^*, b^* \in A^R} \ a^* \succ b^* \Rightarrow U(a^*) - U(b^*) \geq \delta, \tag{6}$$

where δ is an arbitrarily small positive value, implying $U(a^*) > U(b^*)$. In the original UTA method [25], the comprehensive value of alternative $a \in A$ is expressed as $U'(a) = U(a) + \sigma(a)$, where $\sigma(a) \geq 0$ is a potential error relative to $U'(a)$. Alternatively, in the UTASTAR method [52] – an improved variant of UTA – it is expressed as $U'(a) = U(a) - \sigma^+(a) + \sigma^-(a)$, where $\sigma^+(a), \sigma^-(a) \geq 0$ are the over- and underestimation errors. The following linear program minimizing the sum of deviations is solved to estimate a value function:

$$\text{Minimize } F = \sum_{a^* \in A^R} \sigma(a^*) \text{ for UTA} \quad \text{or} \quad F = \sum_{a^* \in A^R} \sigma^+(a^*) + \sigma^-(a^*) \text{ for UTASTAR}, \tag{7}$$

subject to:

$$\left.\begin{array}{l} \left.\begin{array}{l} u_j(\beta_j^1) = 0, j = 1, \ldots, m, \\ \sum_{j=1}^m u_j(\beta_j^{\gamma_j}) = 1, \end{array}\right\}(E^N) \\ \left.\begin{array}{l} u_j(\beta_j^s) - u_j(\beta_j^{s-1}) \geq \rho, j = 1, \ldots, m, s = 2, \ldots, \gamma_j, \end{array}\right\}(E^M) \\ U'(a^*) - U'(b^*) = 0, \text{ for } a^*, b^* \in A^R : a^* \sim b^*, \\ U'(a^*) - U'(b^*) \geq \delta, \text{ for } a^*, b^* \in A^R : a^* \succ b^*, \\ \sigma(a^*) \geq 0 \text{ for UTA or } \sigma^+(a^*), \sigma^-(a^*) \geq 0 \text{ for UTASTAR}. \end{array}\right\}(E^{A^R}_{UTA}) \tag{8}$$

Then, the stability analysis of the provided results is conducted. If the optimum F^* is equal to zero, then the polyhedron of compatible value functions is non-empty. The polyhedron of near-optimal solutions is defined by $E^{A^R}_{UTA} \cup F \leq F^* + k(F^*)$, where $k(F^*)$ is a threshold being a small proportion of F^*. As noted in [40], in most applications of UTA, one usually seeks value functions that are free of errors ($F^* = 0$), and no relaxation from the minimal error is allowed ($k(F^*) = 0$). In case the optimal solution is non-unique, the original postulate of UTA is to partially explore the polyhedron by first finding the functions that either maximize or minimize $u_j(\beta_j^{\gamma_j})$, for $j = 1, \ldots, m$, and then averaging thus obtained extreme functions into the final solution. For details and other possible exploitation algorithms, see [51].

To make the presentation of other UTA-like methods more straightforward, we will use a simplified notation, where a set $\mathcal{U}^R$ of compatible AVF [21] is defined by the following set of linear constraints:

$$\left.\begin{array}{l} E^N, E^M, \\ U(a^*) - U(b^*) = 0, \text{ for } a^*, b^* \in A^R : a^* \sim b^*, \\ U(a^*) - U(b^*) \geq \delta, \text{ for } a^*, b^* \in A^R : a^* \succ b^*. \end{array}\right\}(E^{A^R}) \tag{9}$$

When E^{A^R} is feasible, $\mathcal{U}^R$ consists of at least one compatible AVF. Typically, when the compatible AVF is non-unique, there are infinitely many such functions. This paper assumes that the DM's preference information is consistent with an assumed additive value model, and hence $\mathcal{U}^R$ is non-empty. For the exemplary algorithms dealing with the potential inconsistency, see [3,21,51].

In what follows, we discuss the approaches for robustness analysis, whose results will be exploited by some procedures constructing a univocal recommendation. Robust Ordinal Regression (ROR) exploits $\mathcal{U}^R$ to verify the stability of the recommendation. In particular, the necessary relation $\succsim^N$ holds if a weak preference relation $\succsim$ is unanimously confirmed by all compatible AVF, i.e. [21]:

$$\forall_{a,b \in A} \ a \succsim^N b \iff \forall U \in \mathcal{U}^R : U(a) \geq U(b). \tag{10}$$

Its truth is verified by solving the following Linear Programming (LP) problem:

$$\text{Minimize } U(a) - U(b), \text{ s.t. } E^{A^R}. \tag{11}$$

Let us denote its optimal solution by $D(a, b)$, indicating the minimal value difference between a and b. In case $D(a, b) \geq 0$, then

$a \succsim^N b$. Otherwise, there exists at least one compatible value functions such that $U(b) > U(a)$, and hence $\neg(a \succsim^N b)$.

In Stochastic Ordinal Regression (SOR), $\mathcal{U}^R$ is exploited with the Monte Carlo simulations to derive a large set $S \subseteq \mathcal{U}^R$ of uniformly distributed AVF that are representative for all feasible preference model instances [34]. The results obtained for these models are summarized in the form of stochastic acceptabilities [37], which are estimates of actual shares of compatible value functions confirming a specific outcome:

- Rank Acceptability Index ($RAI(a,k)$) is the share of all compatible AVF that rank alternative $a \in A$ at k-th position, i.e.:

$$\forall_{a\in A}\forall_{k\in\{1,\ldots,n\}} \quad RAI(a,k) = \frac{|\{U \in S : rank(a) = k\}|}{|S|}; \tag{12}$$

- Pairwise Winning Index ($PWI(a,b)$) is the share of all compatible AVF for which alternative a is strictly preferred to alternative b, i.e.:

$$\forall_{a,b\in A} \quad PWI(a,b) = \frac{|\{U \in S : U(a) > U(b)\}|}{|S|}; \tag{13}$$

- Pairwise Outranking Index ($POI(a,b)$) is the share of all compatible AVF for which alternative a is weakly preferred to alternative b, i.e.:

$$\forall_{a,b\in A} \quad POI(a,b) = \frac{|\{U \in S : U(a) \geq U(b)\}|}{|S|}; \tag{14}$$

- Pairwise Indifference Index ($PII(a,b)$) is the share of all compatible AVF for which alternative a is indifferent with alternative b, i.e.:

$$\forall_{a,b\in A} \quad PII(a,b) = \frac{|\{U \in S : U(a) = U(b)\}|}{|S|}. \tag{15}$$

The above relation can be revised to an approximate indifference, which represents a scenario when the comprehensive values of two alternatives differ by no more than a pre-defined threshold [58].

Note that $\forall_{a\in A} \quad \sum_{k=1}^{n} RAI(a,k) = 1$ and $\forall_{a,b\in A} PWI(a,b) = 1 - POI(b,a)$. In this paper, we sample from set $\mathcal{U}^R$ using the Hit-And-Run (HAR) algorithm [56] implemented in [7].

3. Methods for constructing a univocal recommendation

This section reviews thirty-five methods for constructing a univocal ranking and choice recommendation in a preference disaggregation setting. The underlying assumption is that the space of compatible AVF is non-empty.

3.1. Selection of a representative value function

In this section, we review different procedures for selecting a representative value function in the context of UTA. For this purpose, they optimize different objective functions subject to the constraint set E^{A^R} that defines a set of compatible value functions. The selected function can be displayed to the DMs, who can analyze the shapes of MVF, criteria weights, and alternatives' comprehensive scores, leading to a univocal recommendation.

Let us start with the max-min formulations that seek the most discriminant AVF. This idea was first implemented in **UTAMP1** [3], which postulates maximizing the minimal difference between comprehensive values of reference alternatives related by the preference relation, i.e.:

$$\text{Maximize } \delta, \text{ s.t. } E^{A^R}. \tag{16}$$

This approach highlights the DM's indirect preferences, reproducing them boldly and robustly. Another procedure, called **UTAMP2** [3], optimizes δ along with the difference between marginal values assigned to all pairs of consecutive characteristic points, i.e.:

$$\text{Maximize } \delta + \rho, \text{ s.t. } E^{A^R}, \tag{17}$$

where $\rho \geq 0$. The method favors strictly monotonic MVF with steeper linear components and greater slopes.

The procedure underlying the selection of a parsimonious AVF aims at MVF, which minimally deviate from the linearity [6,31]. The model corresponding to this idea is called a Minimal Slope Change Value Function (**UTAMSCVF**). It can be obtained by solving the following Linear Programming (LP) model, which is applicable when at least three characteristic points are considered:

$$\text{Minimize } \phi$$

$$s.t. \begin{left}\begin{array}{l} E^{A^R} \\ \dfrac{u_j(\beta_j^k) - u_j(\beta_j^{k-1})}{\beta_j^k - \beta_j^{k-1}} - \dfrac{u_j(\beta_j^{k-1}) - u_j(\beta_j^{k-2})}{\beta_j^{k-1} - \beta_j^{k-2}} \leq \phi \\ \dfrac{u_j(\beta_j^{k-1}) - u_j(\beta_j^{k-2})}{\beta_j^{k-1} - \beta_j^{k-2}} - \dfrac{u_j(\beta_j^k) - u_j(\beta_j^{k-1})}{\beta_j^k - \beta_j^{k-1}} \leq \phi \end{array}\right\} \text{for } j = 1,\ldots,m,\ k = 3,\ldots,\gamma^j \Bigg\} (E^{A^R}_{MSCVF}) \tag{18}$$

A different idea consists in finding a model that sheds a positive light on all alternatives considered jointly. Such a benevolent procedure, called Maximal Sum of the Scores Value Function (**UTAMSVF**) [6], maximizes a sum of comprehensive values for all reference alternatives:

$$\text{Maximize } \sum_{a^*\in A^R} U(a^*), \text{ s.t. } E^{A^R}. \tag{19}$$

Another group of procedures derive a representative subset of AVF and average them to approximate the centroid of the polyhedron of feasible models. In **UTAJLS**, $2 \cdot m$ AVF are generated by optimizing the maximal share of each MVF. We revise this idea by optimizing the sum of marginal values associated with all characteristic points on a given criterion, i.e., for $j = 1,\ldots,m$:

$$\text{Maximize / Minimize } \sum_{k=1}^{\gamma_j} u_j(\beta_j^k), \text{ s.t. } E^{A^R}. \tag{20}$$

In this way, we consider extreme models representing the maximal and minimal impacts that each criterion has on the comprehensive score. They can also be interpreted as the most concave (when maximizing) or the most convex (when minimizing) marginal functions.

An alternative procedure, called **UTAAVE**, averages a large sample $S = \{U^1, U^2, \ldots U^{|S|}\}$ of compatible AVF considered in SOR hence obtaining a more accurate approximation of the central solution:

$$u_j(\beta_j^s) = \frac{1}{|S|}\sum_{i=1}^{|S|} u_j^i(\beta_j^s), \ j = 1,\ldots,m, s = 1,\ldots,\gamma^j. \tag{21}$$

A central model can be looked for directly, without considering a sample of feasible AVF. **UTACHEB** is an adaptation of the procedure proposed in [14] for sorting problems. It seeks for the center defined as a mid-point of the largest Euclidean ball that fits in the polyhedron of feasible models:

Maximize r,

$$\left.\begin{array}{l} E^N, \\ u_j(\beta_j^s) - u_j(\beta_j^{s-1}) - r \geq 0, \text{ for } j = 1, \ldots, m, s = 2, \ldots, \gamma_j, \\ U(a^*) - U(b^*) = 0, \text{ for } a^*, b^* \in A^R : a^* \sim b^*, \\ U(a^*) - U(b^*) - c_i r \geq 0, \text{ for } a^*, b^* \in A^R : a^* \succ b^*, \end{array}\right\} (E^{A^R}_{CC}) \quad (22)$$

where c_i is the Euclidean norm of the decision variables' (except r) coefficients in the constraint in which they occur [14]. The centrality of such a model derives from being equally distant from all essential monotonicity and preference disaggregation constraints. In the same spirit, **ACUTA** selects an analytic center of the feasible polyhedron [4]. It is identified by maximizing the logarithmic barrier function of the slacks ($d_{a^*,b^*}, d_{j,s}$) involved in the essential constraints of E^{A^R}, using Newton's method:

$$\text{Maximize} \sum_{\forall_{a^*,b^* \in A^R : a^* \succ b^*}} \log d_{a^*,b^*} + \sum_{j=1}^{m} \sum_{s=2}^{\gamma_j} \log d_{j,s},$$

$$\left.\begin{array}{l} E^N, \\ u_j(\beta_j^s) - u_j(\beta_j^{s-1}) = d_{j,s}, \text{ for } j = 1, \ldots, m, s = 2, \ldots, \gamma_j \\ U(a^*) - U(b^*) = 0, \text{ for } a^*, b^* \in A^R : a^* \sim b^*, \\ U(a^*) - U(b^*) = d_{a^*,b^*}, \text{ for } a^*, b^* \in A^R : a^* \succ b^*. \end{array}\right\} (E^{A^R_{AC}}) \quad (23)$$

The solution of the above model is always unique. The last stream of procedures aims at selecting AVF that is representative in the sense of robustness preoccupation. **UTAROB** emphasizes the necessary consequences of applying all compatible AVF on the set of alternatives [31]. In the first stage, it maximizes the minimal value difference for pairs of alternatives related by $\succ^N$, which is evidence of a robust advantage of some alternatives over others:

Maximize ω,

$$\left.\begin{array}{ll} E^{A^R}, & \\ U(a) - U(b) \geq \omega & \forall_{a,b \in A} \; (a \succsim^N b) \wedge \neg(b \succsim^{\rightarrow,N} a). \end{array}\right\} (E^{A^R}_{ROB_I}) \quad (24)$$

Then, it minimizes a value difference for pairs that are incomparable in terms of $\succsim^N$, suggesting that their order in the ranking depends on the compatible AVF. This is conducted while respecting the optimization of the previous target (i.e., setting $\omega = \omega^*$):

Minimize λ,

$$\left.\begin{array}{l} E^{A^R}_{ROB_I}, \\ \omega = \omega^*, \\ U(c) - U(d) \leq \lambda, \; \forall_{c,d \in A} \neg(c \succsim^N d) \wedge \neg(d \succsim^N c), \\ U(d) - U(c) \leq \lambda, \; \forall_{c,d \in A} \neg(c \succsim^N d) \wedge \neg(d \succsim^N c). \end{array}\right\} (E^{A^R}_{ROB_{II}}) \quad (25)$$

In turn, **REPROC** exploits the results of SOR. It emphasizes the advantage of alternatives that are more preferred over others for a more significant share of compatible AVF [34]. This is attained by maximizing the minimal value difference for pairs $a, b \in A$ such that $PWI(a,b) > PWI(b,a)$, i.e.:

Maximize κ,

$$\left.\begin{array}{ll} E^{A^R}, & \\ U(a) - U(b) \geq \kappa(a,b), & \forall_{a,b \in A} \; PWI(a,b) > PWI(b,a), \\ \kappa(a,b) \geq \kappa, & \forall_{a,b \in A} \; PWI(a,b) > PWI(b,a). \end{array}\right\} (E^{A^R}_{PWI_I}) \quad (26)$$

In the second stage, we optimize the sum of elementary value differences $\kappa(a,b)$, while respecting the results of the first stage by setting $\kappa = \kappa^*$, i.e.:

$$\text{Maximize} \sum_{\forall_{a,b \in A} PWI(a,b) > PWI(b,a)} \kappa(a,b), \text{ s.t. } E^{A^R}_{PWI_I} \cup \kappa = \kappa^*. \quad (27)$$

3.2. Decision rules

Decision rules have been elaborated to impose a complete ranking or arbitrarily indicate the most preferred alternative in case there is no agreement concerning the recommended decision in set $\mathcal{U}^R$. Their name should not be confused with "if-then" rules that are used in MCDA as a preference model [20]. The recommendation is suggested without singling out a compatible preference model instance, referring, in turn, to the extreme or expected scores, value differences, or ranks. In what follows, we formulate all rules in a way that favors alternatives with greater scores.

The first group of decision rules derives the recommendation from the ranges of comprehensive values that alternatives attain in the feasible polyhedron. In **MAXIMAX**, the alternatives are ranked according to their highest possible values, i.e.:

$$\text{Maximize } U(a), \text{ s.t. } E^{A^R}. \quad (28)$$

Hence, each alternative is let to select a value function that is the most advantageous for it. As a result, there is no common basis for the comparison because these functions may differ from one alternative to another. Analogously, in **MAXIMIN**, the ranking is determined by the lowest possible values, i.e.:

$$\text{Minimize } U(a), \text{ s.t. } E^{A^R}. \quad (29)$$

Thus, this rule favors alternatives that are the best in the least advantageous scenario compatible with the DM's preferences. Please note that **UTAAVE** can be seen as a decision rule, which considers expected values rather than extreme ones while assuming a uniform distribution of compatible AVF.

Other decision rules consider how favorable is the performance of alternatives relative to others. The most prevailing procedure among them is called **MM-REGRET**. It first considers the maximal loss of value for each pair of alternatives, indicating how much the value of other alternatives can exceed that of the potentially selected option, i.e.:

$$\text{Maximize } U(b) - U(a), \text{ s.t. } E^{A^R}. \quad (30)$$

Let us call the optimal solution of the above model by $regret(a,b) = -D(a,b)$. Intuitively, the greater $regret(a,b)$, the more significant the loss when choosing a rather than b. The comprehensive score for alternative $a \in A$ is derived from its worst-case comparison against some other alternative $b \in A \setminus \{a\}$. Overall, to favor alternatives with the greatest maximal regrets, we order them from the best to the worst by considering the following scores:

$$Sc_{MM-REGRET}(a) = -\max_{b \in A \setminus \{a\}} regret(a,b). \quad (31)$$

The remaining three rules build on the outcomes of SOR concerning the stability of ranks attained by alternatives in $\mathcal{U}^R$. **EXPRANK** derives the comprehensive score of each alternative from its expected rank in the feasible polyhedron, i.e., $\sum_{k=1}^{n} -k \cdot RAI(a,k)$ [33]. While this rule represents an average performance given incomplete preference information, the remaining two procedures stand for the optimistic and pessimistic scenarios.

The **BESTRAI** rule is a generalization of the maximal likelihood principle to ranking problems. Specifically, the alternatives are ordered by their best possible ranks and, in case of a draw, by the

highest probability of attaining these most favorable ranks. Such scores can be synthetically represented as:

$$Sc_{BESTRAI}(a) = -BestRank(a) + RAI(a, BestRank(a)), \times \text{ where } BestRank(a) = \min_{k=1,\ldots,n} \{RAI(a,k) > 0\}. \quad (32)$$

Note that this rule ranks at the top potentially optimal alternatives and, in addition, favors those that attain the first ranks for the greatest share of compatible AVF.

Analogously, **WORSTRAI** generalizes the minimal unlikelihood principle by ordering alternatives according to their worst ranks, and breaking the ties, by favoring options with lower probabilities of attaining these unfavorable ranks. Such scores can be represented as:

$$Sc_{WORSTRAI}(a) = -WorstRank(a) - RAI(a, WorstRank(a)), \times \text{ where } WorstRank(a) = \max_{k=1,\ldots,n} \{RAI(a,k) > 0\}. \quad (33)$$

This rule ranks at the bottom the alternatives, which are ranked last for the most significant share of compatible AVF.

3.3. Scoring procedures

The role of scoring procedures is to exploit the outcomes of comparisons for all pairs of alternatives to derive a comprehensive measure of desirability for each alternative. Hence, similarly to the decision rules, a cardinal scale is driving the recommendation, but no single model is associated with it.

The first group of approaches prioritizes the alternatives by exploiting the minimal differences between their comprehensive scores. Note that such a score for pair $a, b \in A$ is denoted by $D(a, b)$ (see Section 2). Such differences are called intensities of dominance, and hence the respective approaches are considered as dominance measuring methods. In what follows, we discuss a pair of procedures, called **AP1** and **AP2**, proposed in [1]. **AP1** orders the alternatives according to a comprehensive dominance measure defined as a sum of dominance intensities of one alternative over the remaining ones, i.e.:

$$Sc_{AP1}(a) = \sum_{b \in A \setminus \{a\}} D(a, b). \quad (34)$$

In **AP2**, the dominating and dominated measures are considered jointly to combine the arguments in favor of each alternative's strength and weakness, i.e.:

$$Sc_{AP2}(a) = \sum_{b \in A \setminus \{a\}} D(a, b) - D(b, a). \quad (35)$$

Thus, the more an alternative dominates others, and the less the remaining ones dominate it, the higher its position in the comprehensive ranking.

Similar idea has been implemented in dominance measuring extensions, called **DME1** and **DME2** [39]. The motivation for their development derived from the observation that **AP1** involves a trade-off between positive and negative values of dominating measures, whereas **AP2** duplicates the dominated measures. To address these problems, in **DME1**, one considers the positive and negative dominating (α_a^+ and α_a^-) and dominated (β_a^+ and β_a^-) measures:

$$\alpha_a^+ = \sum_{b \in A \setminus \{a\} \wedge D(a,b) > 0} D(a, b) \text{ and } \alpha_a^- = \sum_{b \in A \setminus \{a\} \wedge D(a,b) < 0} D(a, b), \quad (36)$$

$$\beta_a^+ = \sum_{b \in A \setminus \{a\} \wedge D(b,a) > 0} D(b, a) \text{ and } \beta_a^- = \sum_{b \in A \setminus \{a\} \wedge D(b,a) < 0} D(b, a). \quad (37)$$

Note that a is necessarily strictly preferred to b if $D(a, b) > 0$, while being ranked lower for all $U \in \mathcal{U}^R$ when $D(b, a) > 0$. A total score of dominance intensity is computed as the difference between proportions representing both the strength of a given alternative dominating the remaining ones and its weakness derived from being dominated by others, i.e.:

$$Sc_{DME1}(a) = \frac{\alpha_a^+}{\alpha_a^+ - \alpha_a^-} - \frac{\beta_a^+}{\beta_a^+ - \beta_a^-}. \quad (38)$$

In this way, the stronger the intensity of preference of a over others and the weaker the preference intensity of others over a, the more preferred is a. The **DME2** procedure is similar in the sense of exploiting dominance measures, but they are transformed into preference intensities (*PI*; also called – dominance probabilities). Specifically, $PI(a, b) = 1$, if $D(a, b) \geq 0$ (indicating the evident advantage of a over b); $PI(a, b) = 0$, if $D(b, a) \geq 0$ (indicating a clear weakness of a compared to b), and $PI(a, b) = \frac{-D(b,a)}{-D(b,a)-D(a,b)}$, otherwise (i.e., when the results of a comparison between a and b are ambiguous). The alternatives are ordered in the non-increasing order according to the following dominance probability measure that captures the comprehensive strength of alternative's preference over all remaining options:

$$Sc_{DME2}(a) = \sum_{b \in A \setminus \{a\}} PI(a, b). \quad (39)$$

The other group of scoring procedures exploits the results of pairwise comparisons capturing the share of compatible value functions confirming the preference of some alternatives over the others [33]. Specifically, we refer to the difference of *POI*s for all pairs of alternatives while aggregating them using different operators:

- **MAXPOI** derives the maximal *POI* difference, capturing the most favorable pairwise comparison for alternative $a \in A$:

$$Sc_{MAXPOI}(a) = \max_{b \in A \setminus \{a\}} [POI(a, b) - POI(b, a)]; \quad (40)$$

- **MINPOI** computes the minimal *POI* difference, reflecting the least advantageous pairwise comparison for alternative $a \in A$:

$$Sc_{MINPOI}(a) = \min_{b \in A \setminus \{a\}} [POI(a, b) - POI(b, a)]; \quad (41)$$

- **SUMPOI** aggregates *POI*s supporting each alternative's strength and weakness, hence indicating its average performance against all remaining alternatives:

$$Sc_{SUMPOI}(a) = \sum_{b \in A \setminus \{a\}} [POI(a, b) - POI(b, a)]. \quad (42)$$

Intuitively, in the above three procedures, the alternatives derive their scores from the comparison with the best, the worst, or all alternatives. Moreover, the potential ties are broken by applying the same procedure limited in scope to a subset of alternatives attaining the same score.

3.4. Construction of a robust ranking

The methods for constructing a robust ranking exploit the probabilistic information provided by the stochastic acceptabilities. They do not infer a representative value function nor associate a score with any alternative. In turn, they aim at figuring the order supported by a large share of compatible AVF.

The first group of models constructs a ranking by solving an assignment problem of alternatives to ranks [58]. For this purpose, they consider binary variables $x_{ik} \in \{0, 1\}$ such that $x_{ik} = 1$ means that alternative a_i is assigned to k-th position. The most straightforward method, denoted by **RANK-SUM-IND**, maximizes the sum

of *RAI*s supporting the constructed ranking, while respecting that each alternative is assigned to one rank, i.e.:

$$\text{Maximize} \sum_{i=1}^{n}\sum_{k=1}^{n} RAI(a_i,k)\cdot x_{ik}, \tag{43}$$

$$\left.\begin{array}{l} s.t. \sum\limits_{k=1}^{n} x_{ik}=1, \quad \forall i=1,\ldots,n. \end{array}\right\}(E_{RAI}) \tag{44}$$

A revised model, called **RANK-SUM**, additionally assumes that exactly one alternative must be assigned to each rank, hence preventing the indifference relation. It leaves the same objective function as the former model, thus maximizing the average probability, but considers an enriched set of constraints:

$$\left.\begin{array}{l} E_{RAI} \cup \sum_{i=1}^{n} x_{ik}=1, \quad \forall k=1,\ldots,n. \end{array}\right\}(E_{RAI}^{IND}) \tag{45}$$

Another model, called **RANK-PROD**, considers the joint probability of the entire ranking by maximizing the product of *RAI*s associated with the assignments of alternatives to their respective ranks. The linear form of the considered model is as follows:

$$\text{Maximize} \sum_{i=1}^{n}\sum_{k=1}^{n} log(RAI(a_i,k))\cdot x_{ik}, \ \text{s.t. } E_{RAI}^{IND}. \tag{46}$$

The last model exploiting *RAI*s, called **RANK-MM**, maximizes the minimal support of any assignment:

$$\begin{array}{l} \text{Maximize } f_{MM}^{RAI}, \ \text{s.t. } E_{RAI}^{IND} \cup f_{MM}^{RAI} \le RAI(a_i,k)+(1-x_{ik}), \quad \forall i \\ =1,\ldots,n, \quad \forall k=1,\ldots,n. \end{array} \tag{47}$$

It is also possible to consider the variants of the last two models that admit an indifference relation.

The second group of models – also originally proposed in [58] – constructs a ranking by determining pairwise preference relations. The approaches that do not admit indifference consider binary variables $y_{ij}\in\{0,1\}$ such that $y_{ij}=1$ means that alternative a_i is strictly preferred to alternative a_j. The considered set of constraints ensure that the relation imposed on the set of alternatives is complete, asymmetric, irreflexive, and transitive:

$$\left.\begin{array}{l} y_{ij}+y_{ji}=1, \quad \forall i\neq j, \\ y_{ii}=0, \quad \forall i=1,...,n, \\ y_{ij}\ge y_{ik}+y_{kj}-1.5, \quad \forall k\neq i,j. \end{array}\right\}(E_{PWI}) \tag{48}$$

Then, three different objectives can be optimized to maximize the support given by *PWI*s to the ranking constructed from the relations assigned to pairs of alternatives:

- for **REL-SUM** – the sum of preference probabilities, i.e., Maximize $\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} PWI(a_i,a_j)\cdot y_{ij}$, s.t. E_{PWI};
- for **REL-PROD** – the joint probability, i.e., Maximize $\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} \log(PWI(a_i,a_j))\cdot y_{ij}$, s.t. E_{PWI};
- for **REL-MM** – the minimal probability of an established preference, i.e.:

$$\begin{array}{l} \text{Maximize } f_{MM}^{PWI}, \ \text{s.t. } E_{PWI}\cup f_{MM}^{PWI} \le PWI(a_i,a_j)+(1-y_{ij}), \\ \forall i=1,\ldots,n, \quad \forall j=1,\ldots,n. \end{array}$$

For the counterparts of the above models that tolerate indifference, one needs to consider binary variables $y_{ij}^{w}\in\{0,1\}$ such that $y_{ij}^{w}=1$ implies weak preference of a_i over a_j, and z_{ij} such that $z_{ij}=1$ when a_i and a_j are indifferent. The considered set of constraints ensure that the weak preference is complete and transitive. In contrast, indifference is instantiated when the weak preference holds for an ordered pair of alternatives and its inverse counterpart:

$$\left.\begin{array}{l} y_{ij}^{w}+y_{ji}^{w}\ge 1, \quad \forall i=1,...,n-1, \ j=i+1,...,n, \\ y_{ij}^{w}\ge y_{ik}^{w}+y_{kj}^{w}-1.5, \quad \forall k\neq i,j, \\ z_{ij}=y_{ij}^{w}+y_{ji}^{w}-1, \quad \forall i=1,...,n-1, \ j=i+1,...,n. \end{array}\right\}(E_{REL-IND}) \tag{49}$$

The three objectives maximizing the support given by *PWI*s and *PII*s to the constructed ranking are as follows:

- for **REL-SUM-IND** – an additive objective related to the sum of probabilities of established relations:

$$\begin{array}{l} \text{Maximize} \sum\limits_{i=1}^{n-1}\sum\limits_{j=i+1}^{n} PWI(a_i,a_j)\cdot(y_{ij}^{w}-z_{ij}) \\ +PWI(a_j,a_i)\cdot(y_{ji}^{w}-z_{ij})+PII(a_i,a_j)\cdot z_{ij}, \ \text{s.t. } E_{REL-IND}. \end{array} \tag{50}$$

- for **REL-PROD-IND** – a multiplicative objective related to the product of probabilities of established relations that can be translated into the following linear form:

$$\begin{array}{l} \text{Maximize} \sum\limits_{i=1}^{n-1}\sum\limits_{j=i+1}^{n} \log(PWI(a_i,a_j))\cdot(y_{ij}^{w}-z_{ij}) \\ +\log(PWI(a_j,a_i))\cdot(y_{ji}^{w}-z_{ij}) \\ +\log(PII(a_i,a_j))\cdot z_{ij}, \ \text{s.t. } E_{REL-IND}. \end{array} \tag{51}$$

  Technically, in the above objective function, the stochastic acceptabilities are increased by an arbitrarily small positive value ϵ to avoid an undefined value of $log(0)$.
- for **REL-MM-IND** – the max-min objective optimizing the least probability of an established relation:

$$\text{Maximize } f_{MM}^{REL} \tag{52}$$

$$\left.\begin{array}{l} E_{REL-IND}, \\ s.t. f_{MM}^{REL}\le PWI(a_i,a_j)\cdot(y_{ij}^{w}-z_{ij})+PWI(a_j,a_i)\cdot(y_{ji}^{w}-z_{ij}) \\ +PII(a_i,a_j)\cdot z_{ij}, \ i=1,...,n-1, \ j=1,...,n. \end{array}\right\}$$

The last group adopts the ranking techniques originally proposed as part of other MCDA methods to a new context. In this regard, we propose to use distillation procedures known from ELECTRE III [46,47] for exploiting a valued preference relation formed by *POI*s rather than an outranking relation constructed via concordance and discordance tests. We consider the downward and upward distillations, denoted by **DOWN-DIST** and **UP-DIST**. Their detailed formulations can be found in [47]. Let us note that **DOWN-DIST** constructs a ranking in a top-down fashion, retaining alternatives with the greatest quality first and iteratively applying the same procedure until all alternatives are added to the preorder. The quality is interpreted as the difference between strength and weakness. Roughly, for alternative $a\in A$, they are interpreted as the numbers of other alternatives $b\in A\setminus\{a\}$ such that $POI(a,b)$ is significantly large and substantially greater than $POI(b,a)$ or vice versa. In turn, **UP-DIST** is conducted analogously with the proviso that the preorder is constructed bottom-up, and the alternatives with the least quality are retained first.

Using all 35 methods for constructing a univocal recommendation is illustrated in a simple didactic example in the e-Appendix. This description emphasizes the specificity of all methods by referring to the results they exploit and the objectives they optimize. The reference to a particular, small problem and precise obtained results gives a better chance of comprehending the discussed procedures.

4. Computational experiments

This section is devoted to the computational experiments performed to verify the quality of procedures for constructing a univocal recommendation. First, we define the measures employed to compare the 35 considered approaches. Second, we specify an experimental setting. Finally, we discuss the average results across all considered problem instances and some trends when the value of a single problem- or model-related parameter is changed. The extreme (minimal and maximal) results for all measures are discussed in the e-Appendix. Note that whenever we claim that some procedures are the best or the worst, perform favorably or poorly, such conclusions are limited by the considered experimental setting, in particular, the decision problem characteristics and the way of simulating the DM's preferences.

4.1. Comparative measures

The performance of procedures for selecting a single sorting model will be quantified in terms of seven measures. On the one hand, they capture the similarity between the DM's simulated reference model and the recommendation derived with various procedures in the context of choice or ranking based on incomplete and indirect preference information. On the other hand, they refer to the robustness of provided recommendations in view of the support they are given in the set of all compatible AVF.

First, we refer to the measures quantifying the similarity between the true DM's ranking and the recommendation obtained with procedure P. They are recalled after [33], where detailed definitions and explanations can be found. We denote the highest ($r^*(M,a)$) and the lowest ($r_*(M,a)$) ranks attained by alternative $a \in A$ according to the DM ($M = DM$) or procedure P ($M = P$). If there are no shared ranks, then each $a \in A$ is assigned precisely to a single rank, and hence $r^*(M,a) = r_*(M,a)$. Otherwise, $r^*(M,a) < r_*(M,a)$ for at least two alternatives $a \in A$. For example, the ranking: $a \succ b \sim c \succ d$ translates into the following positions: $r^*(M,a) = r_*(M,a) = 1$, $r^*(M,b) = r^*(M,c) = 2$, $r_*(M,b) = r_*(M,c) = 3$, and $r^*(M,d) = r_*(M,d) = 4$. Such an interpretation corresponds to acting with prudence, i.e., avoiding arbitrary tie-breaking by assigning, e.g., the best or the worst admissible rank to all indifferent alternatives contained in the same equivalence class. A subset of alternatives that according to $M \in \{DM, P\}$ are ranked in the r-the position is denoted by:

$$M(r) = \{a \in A : r^*(M,a) \leq r \leq r_*(M,a)\}, \text{ for } r = 1, \ldots, n. \quad (53)$$

To compare alternatives $a, b \in A$ according to M, we use function $p(M,a,b)$. It is equal to 1 if a is preferred to b, 0.5 when a and b are indifferent, and 0 when a is worse than b.

The only measure that considers the similarity of recommendations in the context of the choice problem is *Normalized Hit Ratio* (*NHR*). It compares the subsets of alternatives that are ranked at the top according to DM and procedure P, similarly to the Jaccard's coefficient:

$$NHR(DM,P) = \frac{|DM(r=1) \cap P(r=1)|}{|DM(r=1) \cup P(r=1)|} \in [0,1]. \quad (54)$$

If the same subset of alternatives is ranked first by the DM and P, then *NHR* is equal to one. When there is no intersection between the two subsets, then *NHR* is zero. The remaining three similarity measures consider the entire rankings.

The *Kendall's* τ quantifies the similarity given the pairwise relations observed for all pairs of alternatives. It computes the comprehensive distance between these relations and normalizes it to the $[-1,1]$ interval as follows:

$$\tau(DM,P,n) = 1 - 2 \cdot \frac{\sum_{(a,b) \in A \times A} |p(DM,a,b) - p(P,a,b)|}{n \cdot (n-1)}. \quad (55)$$

Note that the distance between relations observed for $a, b \in A$ is the least when these relations are the same and the greatest when comparing preference with inverse preference. As a result, if the relations for all pairs of alternatives are the same, τ is equal to 1, whereas if one ranking is inverted with respect to the other, $\tau = -1$.

In turn, *Rank Difference Measure* (*RDM*) quantifies the similarity between the attained ranks. For alternative $a \in A$, the respective difference in the ranks assigned by the DM and procedure P can be computed as follows:

$$rdm(DM,P,a) = \frac{\sum_{r_1=r^*(DM,a)}^{r_*(DM,a)} \sum_{r_2=r^*(P,a)}^{r_*(P,a)} |r_1 - r_2|}{[r_*(DM,a) - r^*(DM,a) + 1] \cdot [r_*(P,a) - r^*(P,a) + 1]}. \quad (56)$$

For example, consider alternative $a \in A$ with $r^*(DM,a) = r_*(DM,a) = 2$, and two procedures $P1$ and $P2$ such that $r^*(P1,a) = r_*(P1,a) = 4$ and $r^*(P2,a) = 3 < r_*(P2,a) = 6$. Then, $rdm(DM,P1,a) = |2-4|/(1 \cdot 1) = 2$ and $rdm(DM,P2,a) = [|2-3| + |2-4| + |2-5| + |2-6|]/(1 \cdot 4) = 2.5$. Further, *RDM* aggregates these differences for all alternatives and normalizes them to the [0,1] interval:

$$RDM(DM,P,n) = 1 - \frac{\sum_{a \in A} rdm(DM,P,a)}{max_{diff}^{rank}(n)} \in [0,1], \quad (57)$$

where $max_{diff}^{rank}(n)$ is $(n/2) \cdot n$ when n is even or $(n/2) \cdot (n-1)$ if n is odd. Overall, *RDM* is equal to one, when all alternatives attain exactly the same rank(s), and it is equal to zero, when the differences between their positions are the greatest possible.

The last similarity measure is called *Rank Agreement Measure* (*RAM*). It generalizes *NHR* to all ranks, hence investigating if exactly the same subsets of alternatives attain the same positions $r = 1, \ldots, n$:

$$RAM(DM,P,n) = \frac{1}{n} \cdot \sum_{r=1}^{n} \frac{|DM(r) \cap P(r)|}{|DM(r) \cup P(r)|} \in [0,1]. \quad (58)$$

In the case of a perfect agreement, *RAM* is equal to 1. On the contrary, when each alternative $a \in A$ attains a different rank according to the DM and procedure P, *RAM* is 0.

To investigate the robustness of the recommendation delivered by procedure P, we define three measures. They capture the support that is offered to such a univocal recommendation by all compatible value functions $\mathcal{U}^R$. Such support is quantified by the stochastic acceptabilities derived from SOR. As far as the choice recommendation is concerned, we refer to the *First Rank Acceptability Index* (*FRAI*), which quantifies an average share of compatible AVF that assign the first position to the alternatives ranked at the very top by procedure P, i.e.:

$$FRAI(P) = \frac{1}{|P(r=1)|} \cdot \sum_{a \in P(r=1)} RAI(a,1) \in [0,1]. \quad (59)$$

When a single alternative is unanimously the most preferred according to P, *FRAI* reflects the probability that this alternative is ranked first in $\mathcal{U}^R$. *FRAI* is equal to 1 if all feasible models rank such an alternative at the top, whereas it is equal to 0 if none compatible AVF indicates it as the most preferred option.

Mean Rank Acceptability Index (*MRAI*) generalizes *FRAI* to all ranks by investigating an average *RAI*-based support that is given to the ranks $(r = r^*(P,a), \ldots, r_*(P,a))$ assigned to each alternative $a \in A$ by procedure P:

$$MRAI(P,n) = \frac{1}{n} \cdot \sum_{a \in A} \frac{\sum_{r=r^*(P,a)}^{r_*(P,a)} RAI(a,r)}{r_*(P,a) - r^*(P,a) + 1} \in [0,1]. \quad (60)$$

Consequently, *MRAI* is equal to 1 when all alternatives are ranked at the same position(s) by procedure P and all compatible AVF. On

the contrary, *MRAI* is 0, when there is no feasible model supporting the rank attained by any alternative according to *P*.

In turn, *Mean Pairwise Relation Acceptability Index* (*MPRI*) investigates the support that is given to the pairwise relations observed for all pairs of alternatives in the ranking determined with *P*. Let us denote such a support for pair $(a, b) \in A \times A$ by $PRI(P, a, b)$. When a is preferred to b according to P, PRI is equal to $PWI(a, b)$; if a and b are indifferent, we consider $PII(a, b)$, and in case b is ranked better than a, $PRI(P, a, b)$ is set to $PWI(b, a)$. Then, *MPRI* is an average *PRI*-based support that is given to all pairs of alternatives:

$$MPRI(P, n) = 2 \cdot \frac{\sum_{a,b \in A \wedge a \neq b} PRI(P, a, b)}{n \cdot (n-1)} \in [0, 1]. \quad (61)$$

Note that *MPRI* is equal to 1 if all compatible AVF compare all pairs of alternatives in the same way as procedure *P*. On the other extreme, *MPRI* is 0 when no feasible model supports the relation observed in *P* for any pair of alternatives.

4.2. Experimental setting

When generating instances of test problems, we considered various settings for the dimensionality of data:

- the number of alternatives – $M \in \{6, 8, 10, 12, 14\}$;
- the number of criteria – $E \in \{3, 4, 5\}$;
- the number of characteristic points for each criterion g_j – $P \in \{2, 3, 4\}$;
- the number of pairwise comparisons provided by the DM – $C \in \{4, 6, 8, 10\}$.

In this way, we focus on MCDA problems with a small size for which both the choice and ranking recommendation may be of interest to the DMs. For each combination of parameter values, we averaged the results over 1000 problem instances with randomly drawn performances [14]. Hence we considered $5 \cdot 3 \cdot 3 \cdot 4 \cdot 1000 = 180,000$ instances in total. In general, some considered parameter combinations represent less plausible scenarios (e.g., eliciting a limited number of pairwise comparisons for a problem with numerous alternatives, criteria, and characteristic points, or considering rich preference information when the values of other problem dimensions are small). However, we did not filter them so that the subsequent analysis of the impact of values of a single parameter on the attained results is more reliable and independent of the values assigned to other parameters.

For each instance, we randomly generated AVF serving as the DM's reference model. To ensure its consistency with an assumed model, the number of characteristic points γ_j for the respective MVFs was equal to *P* in the considered problem setting. This function was used to rank *M* alternatives evaluated in terms of *E* criteria. For this ranking, *C* pairs of alternatives were randomly selected, and the relations observed for them were supplied as a simulated DM's indirect and incomplete preference information. Then, for each instance, we performed robustness analysis in the spirit of ROR and SOR. Finally, 35 methods for constructing a univocal recommendation were run, and their respective recommendations were compared with the DM's true model. Note that UTAMSCFV was not run for instances with $P = 2$ characteristic points as its objective function makes sense only when the MVF are piecewise linear. Overall, we performed 6,24 million executions of all procedures.

4.3. Results

4.3.1. Similarity between the DM's simulated model and the derived recommendation

In this section, we discuss the similarities in recommendations provided by the reference model and the procedures exploiting incomplete preference information. The values of NHR, Kendall's τ, RDM, and RAM averaged over all considered problem instances are provided in Table 1.

Let us start by discussing the agreement for the most preferred alternatives. The difference between the best and worst-performing procedures in terms of NHR is large (over 0.25). This means the best procedures can correctly identify the most preferred alternative in 25% more scenarios than average worst performers. The most advantageous NHR is attained by BESTRAI (0.7671). This method is specifically oriented toward identifying the most preferred alternative because it derives the ranking from the analysis of stochastic acceptabilities for the best rank of each alternative. As a result, the alternative attaining the first position for the most significant share of compatible value functions is ranked at the very top. Its advantage over all remaining approaches is statistically significant. This is confirmed by the Hasse diagram presented in Fig. 1, indicating a partial order established based on the Wilcoxon test for paired samples with *p*-value equal to 0.05 [9].

Only slightly worse results are attained by methods that construct a robust ranking by exploiting the stochastic acceptabilities for pairwise relations. They include all six REL procedures that emphasize the most frequent relations in the set of feasible models (*NHR* equal to 0.7657 or 0.7654). As a result, the best-ranked alternative is most likely the one that is preferred to the remaining ones for the majority of compatible value functions. Though modeled slightly differently, similar objectives are considered by MINPOI (0.7656) and REPROC (0.7654). The computational process underlying MINPOI is based on a simple scoring function. Hence it is computationally less complex than the REL methods. Moreover, REPROC associates a value function with the provided recommendation. Advantageous results are also attained by UTAAVE (0.7650). By averaging a large sample of feasible models, it ranks an alternative with the greatest expected comprehensive value at the top. The differences between these nine approaches are not statistically significant (see Fig. 1).

The best performing procedures exploit the outcomes derived with SOR. This also holds for the next group attaining favorable results in terms of average NHR. Among them, the best outcomes are attained by RANK-SUM. Its advantage over the remaining RANK procedures is statistically significant: RANK-SUM $\succ$ RANK-PROD $\succ$ RANK-SUM-IND $\succ$ RANK-MM, with an average advantage of RANK-SUM over RANK-MM being more than 0.03. In general, these procedures aim at assigning an alternative that is most frequently ranked at the top. However, as confirmed by the slight differences, the form of an optimized objective function impacts the attained results. Other NHR values confirm that it is slightly more advantageous to consider expected values than ranks (UTAAVE vs. EXPRANK), the worst-case pairwise comparison rather than all pairwise comparisons at once (MINPOI vs. SUMPOI), and construct the entire ranking at once rather than in subsequent iterations (the REL methods vs. UP-DIST and DOWN-DIST).

The best results among the procedures that do not take robustness concern into account are attained by ACUTA (0.7492) and UTACHEB (0.7427). For example, their advantage in terms of average NHR over UTAMP2 and UTAMP1 is about 0.08 and 0.13, respectively. The latter two approaches are among the bottom ones. This confirms the validity of selecting a central rather than the most discriminant model. When it comes to the methods exploiting dominance intensities, AP2 attains better results than AP1, DME1, and DME2. The worst NHR-based results are attained by UTAMSVF (0.5291) and MAXIMAX (0.5028). Both exploit the greatest comprehensive values attained by alternatives, though UTAMSVF optimizes them for all alternatives jointly, whereas MAXIMAX finds the most advantageous value function for each alternative considered individually. Let us emphasize that the results reported for UTAM-

Table 1
Average values of measures quantifying similarity between the DM's simulated model and the derived recommendation over all problem instances (* – UTAMSCVF was run for instances with at least three characteristic points).

Method	NHR	τ	RDM	RAM	Method	NHR	τ	RDM	RAM
UTAMP1	0.6180	0.7218	0.7859	0.4116	DME1	0.6960	0.7536	0.8072	0.4448
UTAMP2	0.6658	0.7492	0.8047	0.4415	DME2	0.7197	0.7969	0.8382	0.4776
UTAMSVF	0.5291	0.6519	0.7356	0.3590	MAXPOI	0.7130	0.7844	0.8290	0.4704
UTAJLS	0.6881	0.7815	0.8278	0.4809	MINPOI	0.7656	0.7843	0.8289	0.4704
UTAAVE	0.7650	0.8266	0.8605	0.5330	SUMPOI	0.7612	0.8253	0.8592	0.5226
UTACHEB	0.7427	0.8083	0.8472	0.5108	RANK-SUM-IND	0.7546	0.8174	0.8548	0.5329
ACUTA	0.7492	0.8123	0.8501	0.5165	RANK-SUM	0.7650	0.8184	0.8569	0.5473
UTAROB	0.6515	0.7447	0.8017	0.4326	RANK-PROD	0.7639	0.8181	0.8566	0.5472
REPROC	0.7654	0.8269	0.8607	0.5334	RANK-MM	0.7329	0.8104	0.8506	0.5299
MAXIMAX	0.5028	0.7040	0.7724	0.3941	REL-SUM	0.7657	0.8271	0.8608	0.5337
MAXIMIN	0.6735	0.7048	0.7730	0.3938	REL-PROD	0.7657	0.8271	0.8608	0.5336
MM-REGRET	0.6815	0.7094	0.7760	0.4011	REL-MM	0.7657	0.8271	0.8608	0.5336
EXPRANK	0.7612	0.8253	0.8592	0.5226	REL-SUM-IND	0.7654	0.8269	0.8606	0.5330
BESTRAI	0.7671	0.7749	0.8214	0.4521	REL-PROD-IND	0.7654	0.8269	0.8606	0.5331
WORSTRAI	0.6972	0.7750	0.8215	0.4517	REL-MM-IND	0.7654	0.8268	0.8606	0.5329
AP1	0.7093	0.7691	0.8178	0.4551	DOWN-DIST	0.7542	0.8192	0.8547	0.5151
AP2	0.7233	0.7967	0.8384	0.4893	UP-DIST	0.7574	0.8191	0.8546	0.5150
UTAMSCVF (*)	0.6014	0.6836	0.7569	0.3864					

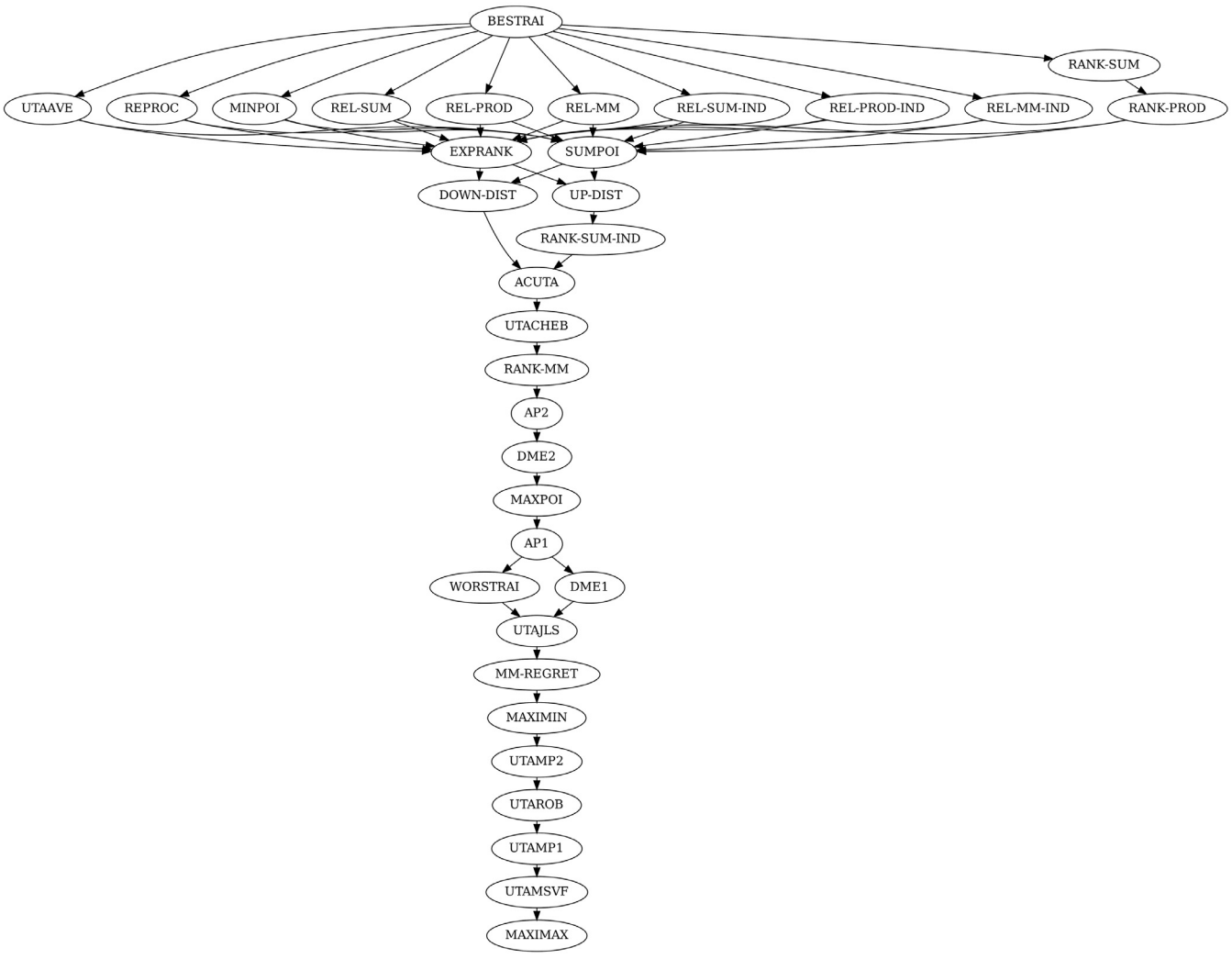


Fig. 1. The Hasse diagram indicating the statistically significant differences in terms of NHR based on the Wilcoxon test with p-value equal to 0.05.

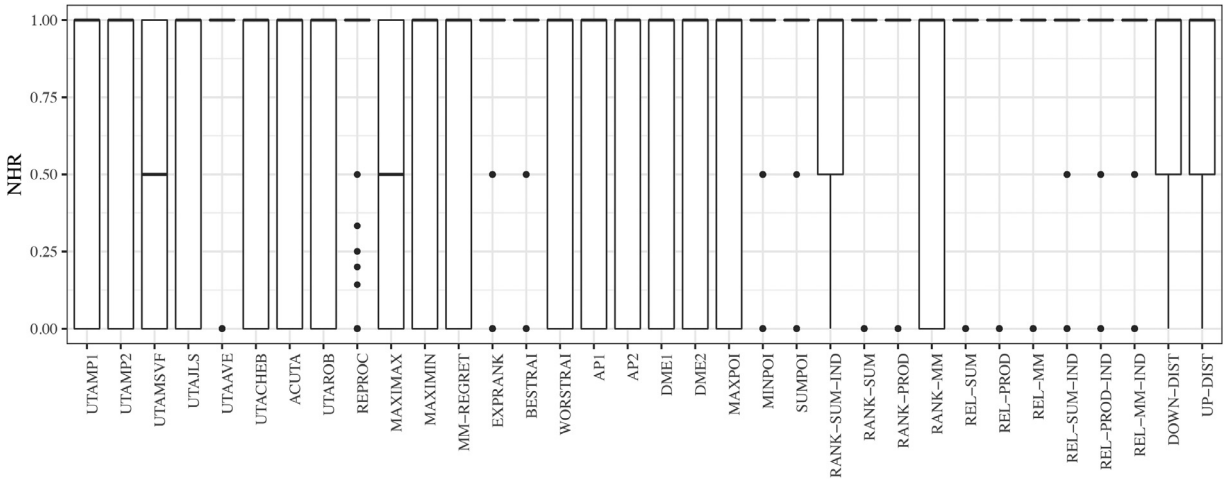


Fig. 2. Boxplot for Normalized Hit Ratio.

SCVF are derived from the analysis of problem instances with at least three characteristic points. However, even when considering only these instances, this procedure attains the NHR values better than UTAMSVF and MAXIMAX.

The respective boxplots for NHR are presented in Fig. 2. They confirm the stability of results attained by the best performing procedures. In fact, the box's length for many approaches is zero, which derives from Q1 and Q3 being equal to the median that, in turn, is equal to one. This means that for at least 75% problem settings, these methods attained the maximal NHR score. Some individual observations indicate fractional values (e.g., 1/2, 1/3, or 1/6), which correspond to the instances for which these procedures rank a few alternatives at the top, but only one of them is the DM's true most preferred alternative. On the contrary, the results for the worst performers, such as UTAMSVF or MAXIMAX, indicate greater variability of results. For these methods, the median is equal to 0.5, but Q3 is already equal to zero.

The ranking-oriented measures quantifying the similarity between the reference and resulting models will be discussed jointly. The analysis of Kendall's τ, RDM, and RAM (see Table 1) lead to similar conclusions. In the main paper, we focus on the average results. The respective boxplots and Hasse diagrams with statistically significant differences for the three measures are presented in the e-Appendix.

The best performing procedures in terms of rank similarity measures include procedures for constructing a robust ranking by solving dedicated mathematical programming models. They include:

- the REL procedures – 0.8268–0.8271 in terms of Kendall's τ, 0.8606–0.8608 for RDM, and 0.5329–0.5337 for RAM; in particular, REL-SUM, REL-PROD, and REL-MM attain statistically significant better results given Kendall's τ and RDM than all other methods; these results mean that they reproduce correctly over 91% pairwise relations and reveal over 86% consistency in terms of the differences between ranks attained by all alternatives; note that even though the differences between the REL methods are marginal, the approaches that do not admit indifference attain statistically better results than their IND counterparts; the favorable performance of the REL procedures in terms of Kendall's τ is consistent with the findings discussed in [58];
- the RANK-SUM and RANK-PROD procedures – 0.8181–0.8184 in terms of Kendall's τ, 0.8566–0.8569 for RDM, and 0.5472–0.5473 for RAM; in particular, they attain statistically significant better results in terms of RAM than all other approaches; the high values of RAM confirm that the similarity in terms of the share of alternatives attaining the same positions in the reference and resulting rankings is over 53% of the maximal possible consistency.

The differences in the top-ranked approaches for the above measures are understandable given their context and operational procedures. In fact, the REL procedures focus on emphasizing the most supported pairwise relations, which aligns with the pairwise perspective considered by Kendall's τ and RDM. In turn, the RANK methods emphasize the most frequent rank assignments, which is consistent with the focus on ranks implemented by RAM.

Favorable performances in terms of all three measures are attained by REPROC and UTAAVE. They are very close to the best-performing methods in Kendall's τ and RDM and just slightly worse given RAM. Nonetheless, these differences, even if marginal, are significant. In general, such statistically sound differences in rank similarity performances are observed for a greater number of method pairs in the upper half of the ranking than in the case of NHR.

An interesting observation concerns relatively good results attained by methods that focus on the stability of outcomes for all pairs of alternatives or all ranks. This is confirmed by the positions between ninth and thirteenth attained by SUMPOI and EXPRANK. They quantify the average strength from all pairwise relation acceptabilities or all ranks. Moreover, they are vastly better than their counterparts focusing only on the extreme outcomes, i.e., MAXPOI and MINPOI or BESTRAI and WORSTRAI. The latter methods derive rankings from the analysis of the most or the least favorable pairwise comparisons or ranks in the set of all compatible value functions.

Similar to NHR, the best results among approaches that do not incorporate robustness analysis are attained by UTACHEB and ACUTA. For example, ACUTA scores 0.8123 for Kendall's τ, 0.8501 for RDM, and 0.5165 for RAM. Both methods prove to be significantly better in terms of all three measures than procedures selecting a representative value function averaging only the extreme model (UTAJLS) or emphasizing the differences between comprehensive values of alternatives compared by the DM (UTAMP1 and UTAMP2) or related by the necessary preference (UTAROB). These methods are also outperformed by the approaches exploiting dominance intensities. Among them, the more advanced variants, called AP2 and DME2, attain better results than their simplified counterparts, AP1 and DME1.

In general, the lower halves of the rankings indicating statistically significant differences are the same for the three measures. In particular, they all agree that MM-REGRET is preferred to MAX-

Table 2
Average values of measures quantifying the robustness of recommendation derived by different methods over all problem instances (* – UTAMSCVF was run for instances with at least three characteristic points).

Method	MRAI	MPRI	FRAI	Method	MRAI	MPRI	FRAI
UTAMP1	0.4115	0.8535	0.6181	DME1	0.4453	0.8724	0.6965
UTAMP2	0.4408	0.8744	0.6647	DME2	0.4778	0.8984	0.7201
UTAMSVF	0.3590	0.8152	0.5298	MAXPOI	0.4701	0.8921	0.7127
UTAJLS	0.4807	0.8907	0.6884	MINPOI	0.4702	0.8921	0.7664
UTAAVE	0.5327	0.9132	0.7658	SUMPOI	0.5229	0.9127	0.7618
UTACHEB	0.5106	0.9040	0.7437	RANK-SUM-IND	0.5334	0.8961	0.7556
ACUTA	0.5160	0.9060	0.7485	RANK-SUM	0.5481	0.9093	0.7659
UTAROB	0.4319	0.8690	0.6512	RANK-PROD	0.5478	0.9091	0.7647
REPROC	0.5334	0.9134	0.7660	RANK-MM	0.5307	0.9053	0.7345
MAXIMAX	0.3936	0.8292	0.5027	REL-SUM	0.5336	0.9135	0.7664
MAXIMIN	0.3938	0.8373	0.6739	REL-PROD	0.5336	0.9135	0.7664
MM-REGRET	0.4010	0.8377	0.6822	REL-MM	0.5336	0.9135	0.7663
EXPRANK	0.5229	0.9126	0.7617	REL-SUM-IND	0.5331	0.9132	0.7660
BESTRAI	0.4517	0.8867	0.7681	REL-PROD-IND	0.5331	0.9132	0.7660
WORSTRAI	0.4515	0.8867	0.6967	REL-MM-IND	0.5330	0.9132	0.7660
AP1	0.4552	0.8845	0.7100	DOWN-DIST	0.5152	0.8935	0.7546
AP2	0.4893	0.8984	0.7238	UP-DIST	0.5152	0.8936	0.7579
UTAMSCVF (*)	0.3863	0.8418	0.6011				

IMIN and MINIMAX, which are, in turn, better than UTAMSVF. Nonetheless, these approaches are among the four worst performers given Kendall's τ, RAM, and RDM. This proves the limited usefulness of approaches exploiting the ranges of comprehensive values attained by the alternatives in the feasible polyhedron in reconstructing the entire ranking. Specifically, UTAMSVF is worse than the best performing procedures by over 0.17 for Kendall's τ, 0.12 for RDM, and almost 0.19 for RAM. Such great differences confirm the importance of selecting an appropriate method when constructing a ranking based on incomplete preference information.

When considering problem instances with at least three characteristic points, UTAMSCVF is better than MAXIMIN, MINIMAX, and UTAMSVF for all measures. Moreover, it proves to be better than MM-REGRET for RAM and RDM and better than UTAMP1 for RAM. The same relative comparisons are confirmed concerning the robustness of provided recommendations. This suggests that finding a parsimonious model that minimally deviates from the linearity is insufficient for reconstructing the DM's preferences generated using a potentially highly non-linear model.

4.3.2. Robustness of provided recommendations

In this section, we discuss the robustness of recommendations provided by the considered procedures understood in terms of the support all compatible value functions give them. The values of MRAI, MPRI, and FRAI averaged over all problem instances are provided in Table 2. In the main paper, we present the boxplot (see Fig. 3) and the Hasse diagram emphasizing statistically significant differences (see Fig. 4) only for MRAI. For the other two measures, the respective figures can be found in the e-Appendix.

The best average results in terms of MRAI are achieved by RANK-SUM (0.5481) and RANK-PROD (0.5478). This means that the position attained by each alternative in the ranking determined with these procedures is supported by almost 55% compatible value functions. Their advantage over the remaining approaches is statistically significant. In particular, it is over 0.014 greater when compared to REL-SUM, REL-PROD, and REL-MM and their IND counterparts. Such a beneficial performance of the two RANK procedures derives from optimizing the support given to the rank assignments in the objective functions, which is consistent with the robustness measure captured by MRAI. However, a slightly worse performance of RANK-MM and RANK-SUM-IND confirms that the form of both an objective function and constraints influences the attained results (see Fig. 4).

When it comes to the best performers given MPRI, they include REL-SUM, REL-PROD, REL-MM (0.9135), and their IND counterparts (0.9132). Such high values indicate that the relations established by these procedures for each pair of alternatives are supported, on average, by over 91% compatible value functions. When compared to the MRAI values, this means that the robustness of constructed rankings can be very high when considering pairwise comparisons, but the alternatives are not necessarily ranked in the same positions as in the feasible polyhedron. Slightly worse results attained by the best RANK procedures (0.9093 for RANK-SUM and 0.9091 for RANK-PROD) confirm that it is more beneficial to emphasize the robust results concerning the same perspective as captured by a specific measure. In the case of MPRI, these are relations holding for all pairs of alternatives, as done by the REL methods. Again, the performance of RANK-MM and RANK-SUM-IND is significantly worse than for the remaining methods constructing a robust ranking.

Among the procedures selecting a representative value function, the best results are attained by REPROC, which emphasizes the advantages derived from the analysis of *PWI*s. This procedure is ranked sixth in terms of MRAI and fourth given MPRI, being only marginally worse than the best performers. However, it outperforms many procedures that exploit the same stochastic acceptabilities, including the REL methods admitting indifference, SUMPOI, and both distillations. UTAAVE and ACUTA also attain highly favorable outcomes. The former exploits expected comprehensive values, outperforming the methods that build on the expected ranks (EXPRANK) or averages from the extreme models (UTAJLS). Furthermore, ACUTA proves to be better than UTACHEB, which suggests that the recommendation associated with the analytic center is more robust than the ranking corresponding to the Chebyshev center.

The differences between the best and the worst performing procedures are great. For MRAI – it is close to 0.2, and for MPRI – it is almost 0.1. Also, the rankings indicating the statistically significant differences are conclusive, leaving only a few pairs of methods incomparable (see, e.g., Fig. 4). Most of these incomparabilities concern approaches that exploit the same results, though in a slightly different way. In the lower half of the rankings for both measures, these pairs include (DOWN-DIST, UP-DIST), (MINPOI, MAX-POI), (BESTRAI, WORSTRAI), and (MAXIMAX, MAXIMIN). This suggests that the robustness of the entire ranking is similar irrespective if it is constructed up-down or bottom-up, based on the most or the least advantageous pairwise comparisons, when

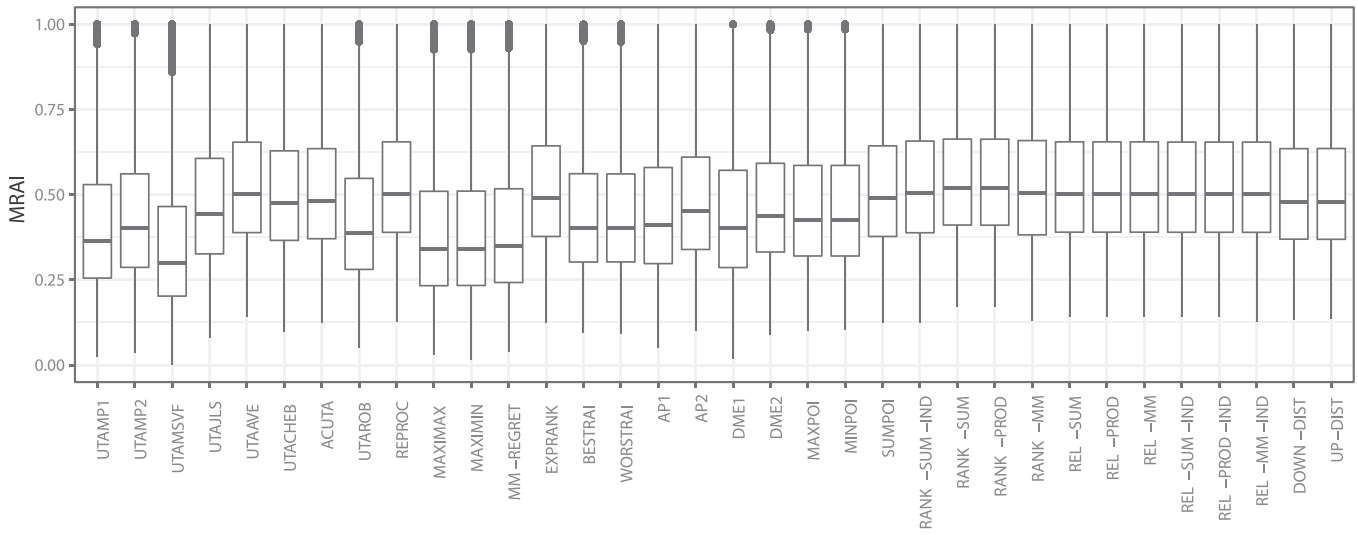


Fig. 3. Boxplot for Mean Rank Acceptability Index.

accounting for the best or the worst ranks, or when considering the greatest or the least comprehensive values.

Let us remind that BESTRAI and MINPOI were among the leaders when it comes to indicating the DM's true most preferred alternative. However, the robustness of the rankings determined with these procedures is poor. For example, the average support given to all ranks or pairwise relations for BESTRAI is lower by almost 0.1 and 0.03 than for the best-performing methods in MRAI and MPRI, respectively. Hence focusing on the extreme attainments in the set of feasible models may be suitable for identifying the best alternative. However, it is not sufficient for reconstructing the entire ranking nor producing a highly robust ranking.

The group of methods attaining the worst results in terms of MRAI and MPRI is the same. These include techniques selecting the most discriminant value function (UTAMP1, UTAMP2, and UTAROB), decision rules based on extreme outcomes (MM-REGRET, MAXIMIN, and MAXIMAX), and the benevolent procedure maximizing the sum of comprehensive values for all alternatives (UTAMSVF). The results attained by UTAMSVF are significantly worse than for all remaining methods.

The boxplot for MRAI is presented in Fig. 3. It indicates that all procedures attain the maximal consistency for at least one problem instance ($MRAI = 1$). When it comes to the least results, they range between 0.17 for RANK-SUM and RANK-PROD to zero for UTAMSVF. In general, the robustness of results depends on the considered problem instance. Relatively high variability is confirmed by the differences between $Q3$ and $Q1$, ranging between 0.25 and 0.3 for all methods. The average best performers attain the most stable outcomes. Still, a similar level of results' stability across all methods implies that the conclusions derived from the analysis of median, first and third quartiles are the same as for the mean.

The best average result in terms of FRAI is attained by BESTRAI (0.7671). This means that the alternative ranked at the top by this procedure is also indicated as the most preferred by almost 77% compatible value functions. It is followed by the REL procedures, MINPOI, and REPROC. These methods exploit the results of stochastic analysis for all pairs of alternatives, aiming to indicate as the most favorable option this alternative which is more preferred to all remaining ones for the majority of feasible models. Note that these approaches also proved to be the most advantageous in terms of reproducing the true pairwise relations in the entire ranking (see Kendall's τ and MPRI). When it comes to FRAI, they attain significantly better results than the RANK procedures. One could expect the opposite because FRAI captures the support given to the alternative ranked in the first place by all compatible value functions, and the RANK methods proved the best given MRAI. However, the first rank is only one out of many positions, specific in the sense that the conditions for attaining it can be easily defined with respect to pairwise comparisons. This gives a chance to the REL methods to outperform their RANK counterparts. Still, RANK-SUM and RANK-PROD need to be seen among the overall good performers because they scored the best in terms of RAM and MRAI and are only marginally worse than the best methods given NHR and FRAI.

As far as other procedures are concerned, high FRAI values are attained by UTAAVE, EXPRANK, and SUMPOI. In these cases, the average support given to the most preferred alternatives among all feasible models ranges between 76.17% and 76.58%. These procedures rank at the top an option that attains the greatest average comprehensive value, the highest expected rank, or the best POI-based support in all pairwise comparisons, respectively. When comparing with the robustness of the entire ranking, slightly better relative performance in terms of FRAI is attained by both distillation procedures. In particular, UP-DIST and DOWN-DIST were worse given MPRI than ACUTA, UTACHEB, DME2, and AP2, whereas the ranking is inverse when considering FRAI.

MAXIMAX (0.5027) and MSVF (0.5298) performed the worst given FRAI. In this case, the order is reversed compared to the measures quantifying the robustness of the entire ranking. However, both procedures are vastly outperformed even by the third worse method, i.e., UTAMP1 (0.6181). The results for FRAI confirm the benefits of constructing the recommendation based on stochastic acceptabilities. Among eight bottom-ranked procedures, none incorporates the shares of feasible models in its operational steps.

4.3.3. Performance trends

In this section, we consider the impact that different parameterizations of the problem and preference model have on the performance of the considered methods. For this purpose, we report the values of performance measures for various numbers of alternatives (M), criteria (E), characteristic points (P), and pairwise comparisons provided by the DM (C). In the main paper, we discuss the observed trends for NHR and MRAI. In this way, we consider measures that are representative for choice (NHR) and ranking (MRAI) as well as for the consistency with the DM's true model (NHR) and robustness of results (MRAI). The detailed outcomes given the

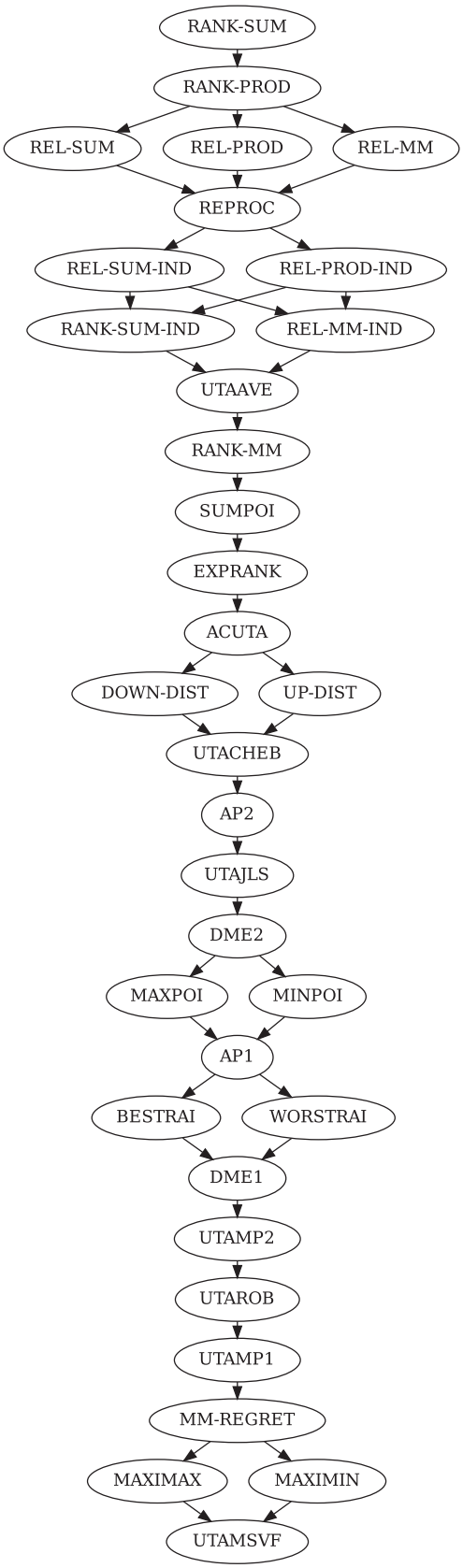


Fig. 4. The Hasse diagram indicating the statistically significant differences in terms of MRAI based on the Wilcoxon test with p-value equal to 0.05.

remaining measures are provided in the e-Appendix. Since the changes observed for them are analogous, we do not elaborate on them in detail.

In Table 3, we present the values of NHR attained by 35 methods for different problem settings. When it comes to the impact of the number of alternatives, with greater M, the results become worse. For example, BESTRAI correctly identifies the most preferred alternative in 86.54% scenarios for $M = 6$, whereas for $M = 14$ – such a consistency drops to 70.08%. A similar trend is observed for all procedures. This is understandable given an increased complexity of indicating the DM's true most preferred alternatives based on incomplete preferences when more alternatives are considered.

When considering the outcomes for different neighboring values of M, the greatest deterioration of NHR can be observed when passing from $M = 6$ to 8 (on average, 0.0866). On the contrary, the smallest decrease of 0.0265 can be observed between $M = 12$ and 14. Hence, when the set of alternatives is very small, adding additional ones increases the problem complexity of identifying the best option more than in the scenarios where the share of new alternatives is smaller.

The rankings of different procedures are consistent irrespective of the number of alternatives. In particular, BESTRAI attains the best results for all considered values of M, and its performance deterioration is the least with the increase of M. In turn, MAXIMAX and UTAMSVF are ranked at the bottom for all considered numbers of alternatives. Moreover, when moving from $M = 6$ to 14, their performance drop is twice as big as for the best methods (e.g., compare the difference of 0.3449 for MAXIMAX and 0.1646 for BESTRAI).

Also, the increase in the number of criteria has a negative impact on NHR attained by all procedures. For example, for BESTRAI, the consistency in terms of indicating the most preferred alternative drops from 79.51% for $E = 3$ through 76.47% for $E = 4$ to 74.14% for $E = 5$. With more criteria, the preference model becomes more flexible, and the set of compatible value functions becomes larger. As a result, the variety of choice recommendations delivered by the set of feasible models is greater, and it becomes more challenging to identify the most preferred option correctly. More significant differences are observed for fewer criteria. The performance decrease when moving from $E = 3$ to 4 is the least for the REL methods ($0.0292 - 0.0293$), MINPOI (0.0293), and REPROC (0.0294). The change from 4 to 5 criteria had the least significant impact on RANK-SUM and RANK-PROD (0.0204). In general, the impact of E on the attained results is smaller than in the case of M. The greatest difference between the extreme considered numbers of criteria is observed for UTAMSCVF (0.1221).

The greater number of characteristic points has, in general, a negative impact on the methods' ability to correctly identify the DM's true most preferred alternative. However, this trend is evident for all procedures only when moving from $P = 2$ to 3 characteristic points. For example, for BESTRAI, the average NHR value drops from 79.00% to 75.44%. Nevertheless, for some approaches, the change in P from 3 to 4 leads to better outcomes. For example, for DOWN-DIST and EXPRANK, the mean NHR increases by 0.005 and 0.0035, respectively. Overall, the correct indication of the top-ranked alternative is easier when marginal value functions are linear. However, when considering various piecewise linear functions, the trend depends on the specific procedure. The increase in the preference model's flexibility is much greater when adding a single breakpoint in the mid-range compared to the scenario where additional characteristic points are included allowing the modeling of even more complex curvatures of marginal functions.

The problem characteristic that positively impacts the efficiency of indicating the most preferred alternative is the number of pairs of reference alternatives compared by the DM. For example, for BESTRAI, the mean NHR is equal to 71.37% for $C = 4$ and 81.54% for

Table 3
Average Normalized Hit Ratios (NHRs) for different numbers of alternatives, criteria, characteristic points, and pairwise comparisons.

		Alternatives					Criteria			Characteristic points			Pairwise comparisons			
METHOD	AVG	6	8	10	12	14	3	4	5	2	3	4	4	6	8	10
UTAMP1	0.6180	0.7692	0.6545	0.5934	0.5489	0.5242	0.6612	0.6114	0.5815	0.6951	0.5957	0.5633	0.5099	0.5937	0.6559	0.7126
UTAMP2	0.6658	0.7982	0.6970	0.6430	0.6069	0.5841	0.7204	0.6592	0.6179	0.7319	0.6481	0.6176	0.5784	0.6451	0.6955	0.7443
UTAMSCVF	0.6014	0.7639	0.6465	0.5665	0.5280	0.5020	0.6666	0.5930	0.5445		0.5982	0.6045	0.5024	0.5794	0.6347	0.6890
UTAMSVF	0.5291	0.7209	0.5743	0.4962	0.4468	0.4075	0.5782	0.5216	0.4877	0.5820	0.5153	0.4902	0.4178	0.5031	0.5676	0.6280
UTAJLS	0.6881	0.8144	0.7196	0.6641	0.6348	0.6076	0.7370	0.6850	0.6424	0.7642	0.6760	0.6241	0.6092	0.6695	0.7156	0.7582
UTAAVE	0.7650	0.8643	0.7921	0.7475	0.7239	0.6970	0.7930	0.7623	0.7396	0.7868	0.7526	0.7555	0.7105	0.7517	0.7835	0.8141
UTACHEB	0.7427	0.8500	0.7706	0.7263	0.6948	0.6719	0.7755	0.7404	0.7123	0.7669	0.7275	0.7337	0.6894	0.7294	0.7601	0.7919
ACUTA	0.7492	0.8552	0.7761	0.7309	0.7026	0.6813	0.7775	0.7468	0.7234	0.7694	0.7358	0.7426	0.6965	0.7370	0.7660	0.7974
UTAROB	0.6515	0.7810	0.6776	0.6249	0.5945	0.5797	0.7050	0.6456	0.6040	0.7148	0.6295	0.6103	0.5681	0.6311	0.6800	0.7269
REPROC	0.7654	0.8652	0.7928	0.7486	0.7236	0.6966	0.7926	0.7632	0.7403	0.7878	0.7530	0.7553	0.7104	0.7532	0.7829	0.8150
MAXIMAX	0.5028	0.7153	0.5556	0.4656	0.4069	0.3704	0.5517	0.4992	0.4575	0.6312	0.4676	0.4096	0.3820	0.4714	0.5457	0.6120
MAXIMIN	0.6735	0.8086	0.7113	0.6523	0.6129	0.5826	0.7201	0.6677	0.6329	0.7112	0.6711	0.6383	0.6008	0.6560	0.6985	0.7388
MM-REGRET	0.6815	0.8167	0.7198	0.6599	0.6216	0.5895	0.7307	0.6753	0.6385	0.7396	0.6666	0.6382	0.6051	0.6633	0.7076	0.7500
EXPRANK	0.7612	0.8629	0.7888	0.7455	0.7183	0.6905	0.7890	0.7584	0.7362	0.7814	0.7494	0.7528	0.7061	0.7479	0.7790	0.8118
BESTRAI	0.7671	0.8654	0.7936	0.7507	0.7249	0.7008	0.7951	0.7647	0.7414	0.7900	0.7544	0.7568	0.7137	0.7551	0.7841	0.8154
WORSTRAI	0.6972	0.8213	0.7275	0.6750	0.6430	0.6191	0.7355	0.6921	0.6639	0.7218	0.6838	0.6860	0.6323	0.6800	0.7196	0.7569
AP1	0.7093	0.8263	0.7388	0.6879	0.6591	0.6343	0.7527	0.7044	0.6707	0.7469	0.6979	0.6830	0.6477	0.6911	0.7311	0.7671
AP2	0.7233	0.8284	0.7498	0.7061	0.6781	0.6540	0.7633	0.7200	0.6865	0.7744	0.7070	0.6885	0.6638	0.7084	0.7429	0.7780
DME1	0.6960	0.8154	0.7266	0.6733	0.6441	0.6205	0.7482	0.6921	0.6477	0.7472	0.6842	0.6566	0.6200	0.6766	0.7241	0.7632
DME2	0.7197	0.8275	0.7464	0.6997	0.6741	0.6508	0.7636	0.7162	0.6793	0.7626	0.7069	0.6897	0.6600	0.7023	0.7408	0.7757
MAXPOI	0.7130	0.8331	0.7406	0.6932	0.6619	0.6362	0.7544	0.7092	0.6754	0.7462	0.6984	0.6945	0.6499	0.6964	0.7352	0.7705
MINPOI	0.7656	0.8653	0.7930	0.7487	0.7238	0.6974	0.7929	0.7636	0.7405	0.7879	0.7534	0.7556	0.7108	0.7534	0.7829	0.8154
SUMPOI	0.7612	0.8630	0.7888	0.7455	0.7183	0.6905	0.7890	0.7584	0.7363	0.7815	0.7494	0.7528	0.7061	0.7479	0.7790	0.8118
RANK-SUM-IND	0.7546	0.8615	0.7866	0.7369	0.7084	0.6797	0.7847	0.7516	0.7275	0.7765	0.7406	0.7468	0.6964	0.7411	0.7737	0.8073
RANK-SUM	0.7650	0.8643	0.7919	0.7478	0.7231	0.6979	0.7934	0.7610	0.7406	0.7866	0.7524	0.7560	0.7108	0.7529	0.7826	0.8136
RANK-PROD	0.7639	0.8643	0.7909	0.7462	0.7218	0.6961	0.7925	0.7598	0.7393	0.7858	0.7508	0.7550	0.7101	0.7514	0.7812	0.8127
RANK-MM	0.7330	0.8566	0.7767	0.7169	0.6737	0.6408	0.7696	0.7287	0.7005	0.7611	0.7171	0.7207	0.6697	0.7190	0.7520	0.7910
REL-SUM	0.7657	0.8653	0.7930	0.7488	0.7239	0.6974	0.7929	0.7637	0.7405	0.7880	0.7534	0.7556	0.7109	0.7535	0.7829	0.8154
REL-PROD	0.7657	0.8653	0.7929	0.7488	0.7239	0.6975	0.7929	0.7637	0.7405	0.7880	0.7535	0.7556	0.7108	0.7535	0.7829	0.8154
REL-MM	0.7657	0.8653	0.7930	0.7488	0.7239	0.6976	0.7929	0.7637	0.7405	0.7880	0.7535	0.7557	0.7110	0.7535	0.7829	0.8154
REL-SUM-IND	0.7654	0.8651	0.7929	0.7483	0.7233	0.6972	0.7924	0.7631	0.7406	0.7875	0.7532	0.7555	0.7103	0.7531	0.7830	0.8151
REL-PROD-IND	0.7654	0.8651	0.7929	0.7483	0.7233	0.6973	0.7924	0.7631	0.7406	0.7875	0.7532	0.7555	0.7103	0.7531	0.7830	0.8151
REL-MM-IND	0.7654	0.8651	0.7929	0.7481	0.7236	0.6972	0.7926	0.7633	0.7402	0.7876	0.7529	0.7556	0.7107	0.7526	0.7831	0.8151
DOWN-DIST	0.7542	0.8570	0.7835	0.7361	0.7109	0.6832	0.7821	0.7512	0.7292	0.7735	0.7420	0.7470	0.6978	0.7413	0.7718	0.8057
UP-DIST	0.7574	0.8586	0.7855	0.7409	0.7136	0.6885	0.7862	0.7543	0.7317	0.7794	0.7445	0.7484	0.7017	0.7443	0.7757	0.8079

$C = 10$. When moving from 4 to 6 pairwise comparisons, the average increase for all procedures is 0.0506; between $C = 6$ and 8 – it is 0.0381, and when considering 10 rather than 8 reference pairs, the improvement is by 0.0379. On the one hand, the general increasing trend is understandable given the additional information gain offered by each pairwise comparison. The space of feasible models becomes more constrained, and the variability of alternatives that could be judged as the DM's most preferred alternatives becomes lesser. On the other hand, the increase in NHR becomes smaller when accounting for richer preference information. Hence the benefit offered by each additional pairwise comparison in terms of correctly identifying the most preferred alternative becomes lesser when numerous comparisons have been already elicited. This is consistent with the findings presented in [8,9] where the increase in the robustness of results offered by initially provided pairwise comparisons is significantly greater than by the preference statements supplied when the set of compatible value functions is already significantly constrained. This effect is additionally strengthened in the study by the random selection of reference pairs. There already exist algorithms for selecting such pairs to maximize the potential information gain (see, e.g., [8,9]). A more general conclusion is that care should be taken to elicit the diverse comparisons reliably reflecting the DM's policy.

The average MRAI results obtained for different problem settings are presented in Table 4. The increase in the number of alternatives implies significant deterioration of MRAI for all procedures. For most of them, when moving from $M = 6$ to 14 alternatives, the average MRAI becomes over twice lesser. For example, for EXPRANK, the results attained for the extreme considered M values are 74.79% and 37.12%. This can be easily explained because larger M translates into a greater number of possible rankings. As a result, with the same amount of preference information, the support given to the alternatives' positions in the set of compatible value functions becomes lesser for problems involving more alternatives.

The rankings of different approaches are very alike irrespective of M. Specifically, RANK-SUM achieves the highest average MRAI for all considered numbers of alternatives, followed closely by RANK-PROD. Their advantage over the remaining methods increases for greater M. For example, when 6 alternatives are considered, the difference between RANK-SUM and REL-SUM is 0.0045, whereas in the case of 14 alternatives – it is already 0.0212. This suggests that the competitive advantage of the procedures directly optimizing the support given to different ranking positions becomes more evident when the number of alternatives is greater.

Similarly, increasing the number of criteria leads to lower MRAI values for all methods. The performance deterioration is not as rapid as for different numbers of alternatives. For example, for RANK-SUM, average support given to the ranks to which the alternatives are assigned in the set of all compatible value functions drops from 59.28% for $E = 3$ through 54.14% for $E = 4$ to 51.00% for $E = 5$. The least absolute and relative deterioration in the robustness of assigned ranks is observed for DOWN-DIST and UP-DIST. It implies that both distillation procedures performed better for problems with 5 criteria than ACUTA and UTACHEB, even though the relation is inverse when 3 criteria are considered. A similar effect can be observed for DME2, which recorded the highest relative decrease in the average MRAI. As a result, the robustness of ranks delivered by DME2 is worse than for MIN-POI and MAX-POI for problems involving 5 criteria, even if it was clearly higher when accounting for 3 criteria. This observation emphasizes the importance of directly exploiting the stochastic acceptabilities for more complex problems than relying on the dominance intensities that capture only the extreme value differences in the set of compatible value functions.

As far as the number of characteristic points is concerned, its impact on MRAI is negative for all methods when moving from linear to piecewise linear marginal value functions. However, when comparing the results attained for $P = 3$ and $P = 4$, the trend depends on the specific approach. For most procedures, it is still decreasing. However, the absolute difference is smaller than when moving from $P = 2$ to 3. For other methods, including, e.g., the REL procedures, UTACHEB, and EXPRANK, there is even a marginal increase in MRAI.

The greatest relative deteriorations are observed for methods that identify the most discriminant models (e.g., UTAMP1) or derive their recommendations from analyzing extreme comprehensive values (MAXIMIN and MAXIMAX). This is understandable because with a greater number of characteristic points, the marginal value functions become more flexible, and the space of all feasible models becomes larger. As a result, such extreme results may become more distant from the regularities – captured by the stochastic acceptability indices – observed in the entire set of all compatible values functions. In turn, the smallest relative decreases of MRAI are noted for procedures that exploit *PWI*s or *RAI*s when constructing a recommendation. In particular, for UP-DIST, DOWN-DIST, SUM-POI, and EXPRANK, the number of characteristic points has no direct impact on the way these procedures work. It indirectly influences their operational steps by increasing the variability of results and decreasing stochastic acceptabilities' stability when more characteristic points are allowed.

The number of pairwise comparisons is again the only parameter that, when increasing, positively affects the performance of all procedures. This is particularly visible for procedures such as UTAMP1, UTAMSVF, MAXIMAX, MAXIMIN, MM-REGRET, and UTAMSCVF that cope badly when the DM compares only a few pairs. For example, UTAMP1 attained the highest relative increase of MRAI from 0.3180 for $C = 4$ comparisons to 0.4994 for $C = 10$. When additional preference information is available, the compatible rankings receive greater support in the set of all feasible models. This is related to the shrink of the feasible polyhedron when indirect preferences become more complete. The least relative increase in MRAI is noted for procedures such as RANK-SUM and RANK-PROD. These methods can produce robust rankings even if the number of pairwise comparisons is low and the space for improvement is much lesser than for the underperforming methods. Still, the robustness of their recommendations increases significantly when preference information becomes richer. For example, the average support given to the ranks produced by RANK-SUM increases from 0.4788 for $C = 4$ through 0.5279 and 0.5716 for the intermediate numbers of pairwise comparisons to 0.6139 for $C = 10$.

5. Summary and future research

We considered the preference disaggregation setting in the context of multiple criteria ranking and choice. We assume the Decision Maker specifies pairwise comparisons of reference alternatives, translated into parameters of an additive value function. The indirectness and incompleteness of such preference information imply the multiplicity of feasible models. In this case, the necessary preference relation is unlikely to help determine a complete ranking or the most preferred alternative. In such scenarios, it is necessary to seek other ways to arrive at a sufficiently decisive and conclusive recommendation.

We reviewed thirty-five methods for constructing a univocal recommendation given an indetermination of the preference model. They are divided into four sub-groups with the proviso that some procedures may be assigned into more than one category. These include (i) methods for selecting a representative value function, (ii) decision rules, (iii) scoring procedures, and (iv) approaches for constructing a robust ranking. Only procedures from the first group associate a feasible model with the provided recommenda-

Table 4
Average Mean Rank Acceptability Indices (MRAIs) for different numbers of alternatives, criteria, characteristic points, and pairwise comparisons.

		Alternatives					Criteria			Characteristic points			Pairwise comparisons			
METHOD	AVG	6	8	10	12	14	3	4	5	2	3	4	4	6	8	10
UTAMP1	0.4115	0.6360	0.4632	0.3714	0.3134	0.2734	0.4595	0.4037	0.3713	0.4864	0.3878	0.3602	0.3180	0.3852	0.4433	0.4994
UTAMP2	0.4408	0.6657	0.4955	0.4024	0.3407	0.2995	0.4963	0.4322	0.3937	0.5055	0.4248	0.3919	0.3554	0.4155	0.4699	0.5221
UTAMSCVF	0.3863	0.6340	0.4404	0.3396	0.2783	0.2390	0.4412	0.3769	0.3407		0.3844	0.3881	0.3015	0.3605	0.4147	0.4683
UTAMSVF	0.3590	0.6066	0.4082	0.3116	0.2531	0.2156	0.4012	0.3513	0.3245	0.4033	0.3450	0.3288	0.2787	0.3324	0.3854	0.4395
UTAJLS	0.4807	0.7083	0.5405	0.4435	0.3786	0.3328	0.5322	0.4746	0.4355	0.5435	0.4657	0.4330	0.4007	0.4570	0.5081	0.5571
UTAAVE	0.5327	0.7538	0.5970	0.4997	0.4308	0.3821	0.5755	0.5261	0.4964	0.5662	0.5151	0.5167	0.4599	0.5116	0.5574	0.6019
UTACHEB	0.5106	0.7333	0.5720	0.4757	0.4093	0.3626	0.5573	0.5037	0.4708	0.5494	0.4900	0.4923	0.4415	0.4897	0.5337	0.5774
ACUTA	0.5160	0.7393	0.5781	0.4815	0.4142	0.3671	0.5598	0.5091	0.4793	0.5518	0.4976	0.4987	0.4471	0.4953	0.5391	0.5826
UTAROB	0.4319	0.6482	0.4815	0.3935	0.3370	0.2991	0.4876	0.4240	0.3840	0.4990	0.4089	0.3877	0.3507	0.4071	0.4593	0.5104
REPROC	0.5334	0.7544	0.5979	0.5007	0.4315	0.3824	0.5762	0.5268	0.4971	0.5671	0.5158	0.5173	0.4606	0.5124	0.5581	0.6025
MAXIMAX	0.3936	0.6386	0.4491	0.3486	0.2866	0.2453	0.4408	0.3862	0.3538	0.4560	0.3796	0.3453	0.3051	0.3666	0.4241	0.4788
MAXIMIN	0.3938	0.6383	0.4487	0.3491	0.2877	0.2453	0.4408	0.3864	0.3541	0.4563	0.3807	0.3443	0.3049	0.3673	0.4243	0.4787
MM-REGRET	0.4010	0.6445	0.4570	0.3567	0.2949	0.2515	0.4459	0.3935	0.3635	0.4575	0.3839	0.3615	0.3130	0.3756	0.4310	0.4843
EXPRANK	0.5229	0.7479	0.5866	0.4887	0.4200	0.3712	0.5640	0.5167	0.4880	0.5540	0.5059	0.5087	0.4495	0.5014	0.5477	0.5929
BESTRAI	0.4517	0.6966	0.5083	0.4091	0.3443	0.3000	0.4930	0.4452	0.4167	0.4835	0.4354	0.4361	0.3802	0.4280	0.4750	0.5234
WORSTRAI	0.4515	0.6962	0.5082	0.4084	0.3444	0.3003	0.4928	0.4444	0.4173	0.4836	0.4349	0.4360	0.3800	0.4278	0.4751	0.5231
AP1	0.4552	0.6918	0.5152	0.4147	0.3500	0.3044	0.5076	0.4480	0.4101	0.5134	0.4360	0.4162	0.3765	0.4310	0.4814	0.5319
AP2	0.4893	0.7099	0.5475	0.4533	0.3902	0.3455	0.5433	0.4824	0.4421	0.5526	0.4690	0.4462	0.4136	0.4665	0.5148	0.5622
DME1	0.4453	0.6761	0.4999	0.4052	0.3440	0.3015	0.5091	0.4372	0.3897	0.5139	0.4287	0.3934	0.3582	0.4189	0.4756	0.5286
DME2	0.4778	0.6987	0.5332	0.4413	0.3796	0.3362	0.5317	0.4711	0.4306	0.5321	0.4603	0.4410	0.4024	0.4550	0.5031	0.5506
MAXPOI	0.4701	0.7093	0.5276	0.4288	0.3644	0.3205	0.5160	0.4633	0.4312	0.5108	0.4508	0.4488	0.3986	0.4470	0.4940	0.5409
MINPOI	0.4702	0.7093	0.5281	0.4291	0.3644	0.3200	0.5161	0.4636	0.4309	0.5103	0.4510	0.4491	0.3989	0.4472	0.4938	0.5408
SUMPOI	0.5229	0.7479	0.5866	0.4887	0.4200	0.3712	0.5640	0.5167	0.4880	0.5541	0.5059	0.5087	0.4495	0.5014	0.5477	0.5929
RANK-SUM-IND	0.5334	0.7550	0.5989	0.5010	0.4310	0.3809	0.5767	0.5269	0.4965	0.5673	0.5156	0.5171	0.4596	0.5121	0.5586	0.6032
RANK-SUM	0.5481	0.7590	0.6092	0.5169	0.4512	0.4041	0.5928	0.5414	0.5100	0.5870	0.5296	0.5277	0.4788	0.5279	0.5716	0.6139
RANK-PROD	0.5478	0.7589	0.6090	0.5167	0.4509	0.4038	0.5926	0.5412	0.5098	0.5867	0.5293	0.5275	0.4786	0.5277	0.5714	0.6137
RANK-MM	0.5307	0.7557	0.5993	0.4991	0.4261	0.3736	0.5777	0.5238	0.4907	0.5715	0.5113	0.5093	0.4584	0.5097	0.5554	0.5994
REL-SUM	0.5336	0.7545	0.5980	0.5009	0.4318	0.3829	0.5765	0.5271	0.4973	0.5672	0.5161	0.5175	0.4609	0.5126	0.5583	0.6026
REL-PROD	0.5336	0.7545	0.5980	0.5009	0.4318	0.3829	0.5765	0.5271	0.4973	0.5672	0.5161	0.5175	0.4609	0.5126	0.5583	0.6026
REL-MM	0.5336	0.7545	0.5980	0.5009	0.4318	0.3829	0.5765	0.5271	0.4973	0.5672	0.5161	0.5175	0.4609	0.5126	0.5583	0.6026
REL-SUM-IND	0.5331	0.7540	0.5975	0.5004	0.4313	0.3824	0.5754	0.5268	0.4972	0.5663	0.5157	0.5173	0.4607	0.5122	0.5578	0.6018
REL-PROD-IND	0.5331	0.7540	0.5975	0.5005	0.4313	0.3824	0.5754	0.5268	0.4972	0.5663	0.5157	0.5173	0.4607	0.5122	0.5578	0.6018
REL-MM-IND	0.5330	0.7540	0.5975	0.5003	0.4311	0.3820	0.5752	0.5267	0.4971	0.5662	0.5156	0.5172	0.4605	0.5121	0.5577	0.6017
DOWN-DIST	0.5152	0.7420	0.5793	0.4802	0.4115	0.3628	0.5553	0.5091	0.4811	0.5453	0.4985	0.5017	0.4414	0.4935	0.5402	0.5856
UP-DIST	0.5152	0.7420	0.5794	0.4807	0.4112	0.3627	0.5553	0.5092	0.4811	0.5456	0.4984	0.5016	0.4415	0.4936	0.5400	0.5857

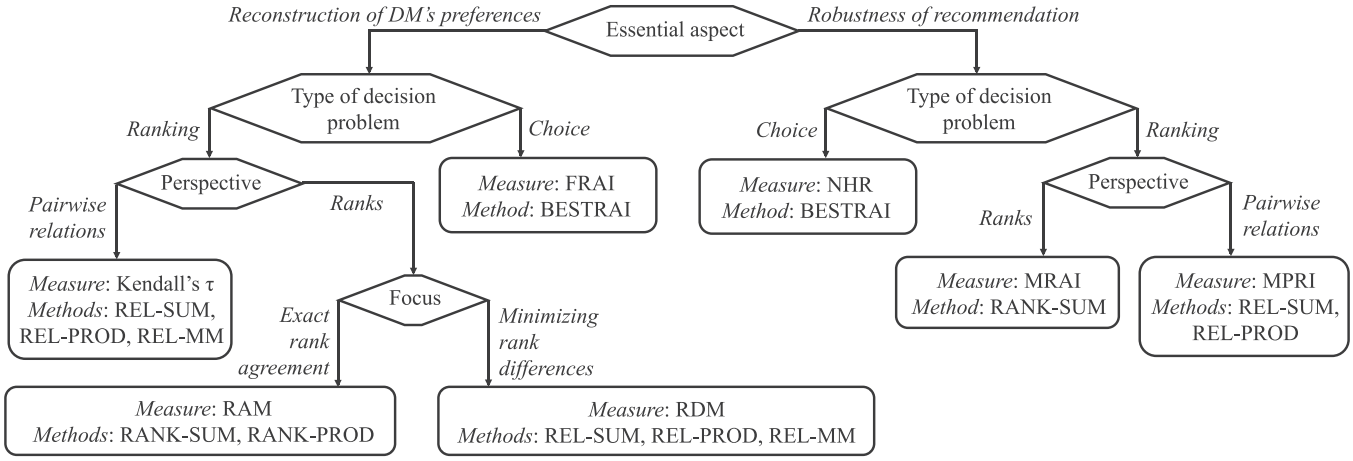


Fig. 5. A decision tree supporting the selection of the measure and the best performing method depending on the characteristics of the problem.

tion. In the methods from the first three groups, the scale leading the recommendation is cardinal, whereas for the last group – it is ordinal. All methods except a few oriented toward selecting a representative model account for the robustness of results obtained in the set of compatible value functions. However, these outcomes are different, referring to the extreme value differences (dominance intensities), stability of pairwise relations, or robustness of ranks attained by alternatives. Moreover, even if the general aim of some methods is the same, they differ in terms of how this objective is implemented. For example, in the case of selecting a representative model, one may opt for the most discriminant, average, central, benevolent, aggressive, parsimonious, or robust value function. In turn, constructing a robust ranking may be conducted using upward or downward distillations, or mathematical programming models maximizing the sum, product, or minimal support quantified with stochastic acceptabilities.

The performance of all methods was compared on problem instances with different complexities. The outcomes of an extensive study were quantified in terms of seven measures. By referring to their specificity, we can design a decision tree facilitating the selection of the best performing procedure based on the experimental outcomes (see Fig. 5). The tree refers to four characteristics important for the practice of decision aiding. The major one points out to maximizing either the similarity between the DM's true preferences and the recommendation suggested by the method or the robustness of the delivered recommendation in terms of the support it is given in the set of compatible value functions. The other essential feature distinguishes between the choice and ranking problems. When dealing with choice, the focus is always on identifying the most preferred alternative. In the case of ranking problems, we can differentiate between the analysis of pairwise relations or ranks assigned to alternatives. The former is focused on one-against-one comparisons, whereas the latter refers to the performance of individual alternatives derived from their comparisons with all remaining options. Finally, given the rank-oriented perspective in the context of reconstruction of the DM's preferences, it may be interesting to consider if alternatives attain precisely the same ranks in the true and predicted rankings or account for the differences between the ranks associated with the alternatives in the two rankings.

There is a consistency in indicating the best performing procedures for the three groups of measures. First, when it comes to identifying the DM's true most preferred alternative (NHR) and the support given to the top-ranked alternative by all feasible models (FRAI), the best results are attained by BESTRAI. It investigates the best ranks attained by all alternatives along with the shares of compatible value functions confirming such a favorable result. Thus, BESTRAI is recommended for use in the context of choice problems. Second, concerning the reproduction of the relations in the DM's true ranking (Kendall's τ and RDM) and the support given to the pairwise relations by all compatible models (MPRI), the best outcome are attained by the REL procedures. They construct a ranking by emphasizing the relations confirmed by the greatest share of value functions consistent with the DM's preferences. Hence, when dealing with ranking problems with the interest of either reconstructing the DM's pairwise preferences or maximizing the robustness of recommendation for all pairs of alternatives, we advise employing REL-SUM, REL-PROD, or REL-MM. Third, as far as reconstructing the positions in the DM's true ranking (RAM) and the support given to the assigned ranks in the feasible polyhedron (MRAI) are concerned, the most advantageous performance is observed for RANK-SUM and RANK-PROD. These methods construct a ranking by placing the alternatives in these positions that they attained most frequently in the set of compatible preference model instances. The above conclusions are constrained by the considered experimental setting (e.g., relatively small MCDA problems and additively rational DMs). Such a limitation concerns all experimental studies whose feasibility requires making arbitrary assumptions and fixing a finite number of parameter values specifying the relevant problem instances. Nevertheless, the likelihood of the conclusions presented in this paper is strengthened by a broad spectrum of considered problem characteristics and analyzing a significant number of 180,000 problem instances.

Even though the best performers differ from one measure to another, the subsets of methods attaining favorable results are consistent given all measures. They include the approaches that construct a robust ranking based on stochastic acceptabilities, methods that exhibit the expected results, and procedures selecting representative value functions that emphasize the robustness preoccupation (REPROC), average the indications of all feasible models (UTAAVE) or can be deemed as central in the feasible polyhedron (ACUTA and UTACHEB). The exception in this regard can be noted for BESTRAI and MINPOI. They deal exceptionally well in the context of choice but fail to provide satisfactory results when considering the entire rankings. The least favorable outcomes were attained by simple decision rules based on the extreme outcomes in the set of feasible models, procedures selecting a representative value function that is benevolent, parsimonious, or the most discriminant, and methods exploiting the dominance intensities. This confirms that an increased computational effort at the stage of constructing a decision recommendation pays off to increase both its

robustness and the chances for reconstructing the DM's true preferences.

The experimental study indicated that the consistency between the reference and resulting models, as well as the robustness of recommendations delivered by all methods, decreased with greater numbers of alternatives and criteria, lesser number of characteristic points, and when moving from linear to piecewise linear marginal value functions. Such trends can be explained given a greater complexity of problems involving more alternatives and criteria, higher flexibility of additive value functions and greater variability of rankings with more criteria and characteristic points, and more constrained space of feasible models when additional pairwise comparisons are available. More significant absolute differences in the performance measures were observed in the lower scale range of different parameters of a decision problem or a ranking model (e.g., when passing from 6 to 8 alternatives, from 3 to 4 criteria, from 2 to 3 characteristic points, of from 4 to 6 pairwise comparisons). In turn, the differences in the upper parts of the parameter scales were lesser (e.g., when passing from 12 to 14 alternatives, from 4 to 5 criteria, from 3 to 4 characteristic points, or from 8 to 10 pairwise comparisons). In general, the relative performances of all procedures were the same irrespective of the problems' and models' parameterization. Some slight differences were related to the operational steps of different methods. For example, the approaches that do not build their recommendations on the robustness of results tend to perform worse for greater problem instances or with more flexible preference models.

The universal setting adopted in the UTA-like methods makes the findings of this paper of interest to researchers in other fields, including choice modeling (CM) [43] and preference learning (PL) [16]. On the one hand, CM, an essential subfield of economics and marketing, employs consumers' revealed or stated preferences in the form of pairwise comparisons (discrete choice). These are used to construct a preference model – often via linear programming – that is typically a utility function [55,57]. The model reveals the importance of various attributes and trade-offs between characteristics, allowing to value products, goods, or services that the consumers have not directly judged [27]. In this perspective, both its form and usefulness are similar to those learned by the UTA-like methods, even though the typical contexts of use (e.g., refining new product development, estimating the willingness to pay, or testing product viability) are different. On the other hand, PL, an important subfield of machine learning and artificial intelligence, also uses holistic observed preferences to infer models predicting the preference for previously unseen items, objects, or instances. The prevailing PL methods search for utility functions by solving regression problems. However, unlike UTA-like methods, they are typically used in the context of large sets of preference statements, e.g., in the search engine or recommender system environments [11].

We envisage the following directions for future research. First, it is possible to develop more procedures for constructing a univocal recommendation. In this regard, an exciting idea consists in incorporating the robust optimization objectives from [58] into procedures selecting a representative value function. Such procedures may be applicable to scenarios where the Decision Maker needs to consider a concrete instance of the preference model along with the recommended decision. Also, the pairwise stochastic acceptabilities can be exploited with other approaches as, e.g., proposed in [36], where the eigenvector method was applied in the context of robust efficiency results. Second, a limitation of our study consisted of assuming that the set of compatible value functions was non-empty. While some approaches (e.g., those exploiting the outcomes of robustness analysis) are applicable only under such a setting, others can also be used when an additive value model cannot perfectly reproduce all DM's pairwise comparisons. Therefore, it would be interesting to adjust some methods to the case of incompatibility and conduct a simulation study focusing specifically on such a context, similarly to what has been in [3]. Third, some methods presented in this paper are universal, being applicable to results obtained with any preference model. However, the selection of procedures for deriving a representative parameter set in the context of outranking methods or representative set of rules applicable with the Rough Set Theory is scarce. Hence novel approaches can be elaborated and subsequently compared in terms of their predictive accuracy or robustness preoccupation. Finally, an appealing direction for future research consists of elaborating preference learning algorithms for constructing a recommendation based on large sets of inconsistent pairwise comparisons. In this paper, we considered problem instances with sizes typical for MCDA. However, adjusting the preference disaggregation algorithms to the era of big data becomes more and more critical given an increasing range of applications where the extensive collections of preferences are already available or observed from the users' behavior rather than directly elicited from the DMs.

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

Miłosz Kadziński: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Visualization, Supervision, Project administration, Funding acquisition. **Michał Wójcik:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Krzysztof Ciomek:** Conceptualization, Methodology, Software, Writing – review & editing, Visualization.

Acknowledgments

Miłosz Kadziński and Michał Wójcik acknowledge financial support from the Polish National Science Center under the SONATA BIS project (grant no. DEC-2019/34/E/HS4/00045). In addition, we recognize the programming contribution of Hanna Kuczka and Paulina Jankowska, who implemented some of the discussed methods in R as part of their Master theses defended in 2016.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.omega.2022.102715

References

[1] Ahn BS, Park KS. Comparing methods for multiattribute decision making with ordinal weights. Computers & Operations Research 2008;35(5):1660–70.
[2] Beuthe M, Eeckhoudt L, Scannella G. A practical multicriteria methodology for assessing risky public investments. Socio-Economic Planning Sciences 2000;34(2):121–39.
[3] Beuthe M, Scannella G. Comparative analysis of UTA multicriteria methods. European Journal of Operational Research 2001;130(2):246–62.
[4] Bous G, Fortemps P, Glineur F, Pirlot M. ACUTA: A novel method for eliciting additive value functions on the basis of holistic preference statements. European Journal of Operational Research 2010;206(2):435–44.
[5] Branke J, Greco S, Słowiński R, Zielniewicz P. Learning value functions in interactive evolutionary multiobjective optimization. IEEE Transactions on Evolutionary Computation 2015;19(1):88–102.
[6] Branke J, Greco S, Słowiński R, Zielniewicz P. Learning value functions in interactive evolutionary multiobjective optimization. IEEE Transactions on Evolutionary Computation 2015;19(1):88–102.
[7] Ciomek K, Kadziński M. Polyrun: A java library for sampling from the bounded convex polytopes. SoftwareX 2021;13:100659.

[8] Ciomek K, Kadziński M, Tervonen T. Heuristics for prioritizing pair-wise elicitation questions with additive multi-attribute value models. Omega 2017;71:27–45.
[9] Ciomek K, Kadziński M, Tervonen T. Heuristics for selecting pair-wise elicitation questions in multiple criteria choice problems. European Journal of Operational Research 2017;262(2):693–707.
[10] Corrente S, Greco S, Kadziński M, Słowiński R. Inducing probability distributions on the set of value functions by Subjective Stochastic Ordinal Regression. Knowledge-Based Systems 2016;112:26–36.
[11] Corrente S, Greco S, Kadziński M, Słowiński R. Robust ordinal regression in preference learning and ranking. Machine Learning 2013;93(2):381–422.
[12] Corrente S, Greco S, Słowiński R. Multiple criteria hierarchy process in robust ordinal regression. Decision Support Systems 2012;53(3):660–74.
[13] Doumpos M, Zopounidis C. Disaggregation Approaches for Multicriteria Classification: An Overview, pages 77–94. Cham: Springer International Publishing; 2018.
[14] Doumpos M, Zopounidis C, Galariotis E. Inferring robust decision models in multicriteria classification problems: An experimental analysis. European Journal of Operational Research 2014;236(2):601–11.
[15] Figueira J, Greco S, Słowiński R. Building a set of additive value functions representing a reference preorder and intensities of preference: GRIP method. European Journal of Operational Research 2009;195(2):460–86.
[16] Fürnkranz J, Hüllermeier E. Preference Learning. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. Berlin: Springer; 2010. p. 789–95.
[17] Ghaderi M, Ruiz F, Agell N. A linear programming approach for learning non–monotonic additive value functions in multiple criteria decision aiding. European Journal of Operational Research 2017;259(3):1073–84.
[18] Greco S, Ehrgott M, Figueira J. Multiple criteria decision analysis–state of the art surveys. International Series in Operations Research & Management Science. New York: Springer; 2016.
[19] Greco S, Kadziński M, Mousseau V, Słowiński R. Robust ordinal regression for multiple criteria group decision problems: UTA^{GMS}-GROUP and UTADIS^{GMS}-GROUP. Decision Support Systems 2012;52(3):549–61.
[20] Greco S, Matarazzo B, Słowiński R. Rough sets theory for multicriteria decision analysis. European Journal of Operational Research 2001;129(1):1–47.
[21] Greco S, Mousseau V, Słowiński R. Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions. European Journal of Operational Research 2008;191(2):416–36.
[22] Greco S, Mousseau V, Słowiński R. Parsimonious preference models for robust ordinal regression. Yverdon, Switzerland: EURO Working Group on MCDA; 2011.
[23] Greco S, Mousseau V, Słowiński R. Robust ordinal regression for value functions handling interacting criteria. European Journal of Operational Research 2014;239(3):711–30.
[24] Hanley N, Mourato S, Wright RE. Choice modelling approaches: a superior alternative for environmental valuation? OECD Economic Surveys 2001;15(3):435–62.
[25] Jacquet-Lagréze E, Siskos Y. Assessing a set of additive utility functions for multicriteria decision making: the UTA method. European Journal of Operational Research 1982;10:151–64.
[26] Jacquet-Lagréze E, Siskos Y. Preference disaggregation: 20 years of MCDA experience. European Journal of Operational Research 2001;130(2):233–45.
[27] Johnson RM. Trade-off analysis of consumer values. Journal of Marketing Research 1974;11(2):121–7.
[28] Kadziński M, Corrente S, Greco S, Słowiński R. Preferential reducts and constructs in robust multiple criteria ranking and sorting. OR Spectrum 2014;36(4):1021–53.
[29] Kadziński M, Ghaderi M, Wasikowski J, Agell N. Expressiveness and robustness measures for the evaluation of an additive value function in multiple criteria preference disaggregation methods: An experimental analysis. Computers & Operations Research 2017;87:146–64.
[30] Kadziński M, Greco S, Słowiński R. Extreme ranking analysis in robust ordinal regression. Omega 2012;40(4):488–501.
[31] Kadziński M, Greco S, Słowiński R. Selection of a representative value function in robust multiple criteria ranking and choice. European Journal of Operational Research 2012;217(3):541–53.
[32] Kadziński M, Greco S, Słowiński R. RUTA: A framework for assessing and selecting additive value functions on the basis of rank related requirements. Omega 2013;41(4):735–51.
[33] Kadziński M, Michalski M. Scoring procedures for multiple criteria decision aiding with robust and stochastic ordinal regression. Computers & Operations Research 2016;71:54–70.
[34] Kadziński M, Tervonen T. Robust multi-criteria ranking with additive value models and holistic pair-wise preference statements. European Journal of Operational Research 2013;228(1):169–80.
[35] Keeney R, Raiffa H. Decisions with Multiple Objectives: Preferences and Value Trade-Offs. Cambridge University Press; 1993.
[36] Labijak-Kowalska A, Kadziński M. Experimental comparison of results provided by ranking methods in data envelopment analysis. Expert Systems with Applications 2021;173:114739.
[37] Lahdelma R, Salminen P. SMAA-2: Stochastic Multicriteria Acceptability Analysis for Group Decision Making. Operations Research 2001;49(3):444–54.
[38] Mastorakis K, Siskos E. Value focused pharmaceutical strategy determination with multicriteria decision analysis techniques. Omega 2016;59:84–96.
[39] Mateos A, Jiménez A, Blanco JF. Ranking methods based on dominance measures accounting for imprecision. In: Rossi F, Tsoukias A, editors. Algorithmic Decision Theory. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. p. 328–39.
[40] Matsatsinis NF, Grigoroudis E, Siskos E. Disaggregation approach to value elicitation. In: Dias L, Morton A, Quigley J, editors. Elicitation. International Series in Operations Research & Management Science, vol. 261. Berlin: Springer; 2018. p. 313–48.
[41] Matsatsinis NF, Samaras AP. MCDA and preference disaggregation in group decision support systems. European Journal of Operational Research 2001;130(2):414–29.
[42] Matsatsinis NF, Siskos Y. MARKEX: An intelligent decision support system for product development decisions. European Journal of Operational Research 1999;113(2):336–54.
[43] McFadden D. Conditional logit analysis of qualitative choice behaviour. In: Zarembkam P, editor. Frontiers in Econometrics. Academic Press New York; 1973. p. 105–42.
[44] Nikas A, Doukas H, Siskos E, Psarras J. International Cooperation for Clean Electricity: A UTASTAR Application in Energy Policy, pages 163–186. Cham: Springer International Publishing; 2018.
[45] Rezaei J. Piecewise linear value functions for multi-criteria decision-making. Expert Systems with Applications 2018;98:43–56.
[46] Roy B. ELECTRE III : Un algorithme de classements fondé sur une représentation floue des préférences en présence de critères multiples. Cahiers du CERO 1978;20(1):3–24.
[47] Roy B, Present M, Silhol D. A programming method for determining which Paris metro stations should be renovated. European Journal of Operational Research 1986;24:318–34.
[48] Salo A, Hämäläinen RP. Preference Programming–Multicriteria Weighting Models under Incomplete Information. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. 167–187
[49] Siskos E, Malafekas M, Askounis D, Psarras J. E-government benchmarking in european union: A multicriteria extreme ranking approach. In: Douligeris C, Polemi N, Karantjias A, Lamersdorf W, editors. Collaborative, Trusted and Privacy-Aware e/m-Services. Berlin, Heidelberg: Springer; 2013. p. 338–48.
[50] Siskos Y, Assimakopoulos N. Multicriteria highway planning: A case study. Mathematical and Computer Modelling 1989;12:1401–10.
[51] 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, Dordrecht, London: Springer Verlag; 2005. p. 297–344.
[52] Siskos Y, Yannacopoulos D. UTASTAR: An ordinal regression method for building additive value functions. Investigacao Operacional 1985;5(1):39–53.
[53] Sobrie O, Gillis N, Mousseau V, Pirlot M. UTA-poly and UTA-splines: Additive value functions with polynomial marginals. European Journal of Operational Research 2018;264(2):405–18.
[54] Spyridakos A, Tsotsolas N, Siskos Y, Yannakopoulos D, Vryzidis I. A visualization approach for robustness analysis in multicriteria disaggregation–aggregation approaches. Operational Research 2020;20(3):1841–61.
[55] Srinivasan V, Shocker A. Linear programming techniques for multidimensional analysis of preferences. Psychometrika 1973;38(3):337–69.
[56] Tervonen T, van Valkenhoef G, Basturk N, Postmus D. Hit-And-Run enables efficient weight generation for simulation-based multiple criteria decision analysis. European Journal of Operational Research 2013;224(3):552–9.
[57] Toubia O, Duncan S, Hauser J, Dahan E. Fast polyhedral adaptive conjoint estimation. Marketing Science 2001;22(3):273–303.
[58] Vetschera R. Deriving rankings from incomplete preference information: A comparison of different approaches. European Journal of Operational Research 2017;258(1):244–53.
[59] Vryzidis I, Spyridakos A, Tsotsolas N. Projects Portfolio Selection Framework Combining MCDA UTASTAR Method with 0–1 Multi-Objective Programming, pages 125–146. Cham: Springer International Publishing; 2018.

Supplementary material [P1]

Review and experimental comparison of ranking and choice procedures for constructing a univocal recommendation in a preference disaggregation setting – eAppendix

Miłosz Kadziński[a], Michał Wójcik[a], Krzysztof Ciomek[a]

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

1. Illustrative study

In this section, we illustrate the use of 35 methods for constructing a univocal recommendation. For this purpose, we consider a problem of ranking six cars that are evaluated in terms of the following five criteria: price (g_1; cost), power (g_2; gain), acceleration (g_3; cost); fuel consumption (g_4, cost), and CO_2 emission (g_5; cost). For the performances, see Table 1. To simulate the DM's policy, we have drawn Marginal Value Functions (MVFs) for the reference model (see Figure 1). The ranks, marginal and comprehensive values attained by all alternatives are provided in Table 1. To perform the analysis, we assume the DM provides the following four randomly selected pairwise comparisons derived from the reference ranking: $a_3 \succ a_5$, $a_6 \succ a_1$, $a_3 \succ a_2$, and $a_4 \succ a_2$.

Table 1: Performances of six cars on five criteria and their marginal and comprehensive values according to the reference model.

	Performances					Reference model						
Alternative	$\boldsymbol{g_1}$	$\boldsymbol{g_2}$	$\boldsymbol{g_3}$	$\boldsymbol{g_4}$	$\boldsymbol{g_5}$	Rank	$\boldsymbol{u_1}$	$\boldsymbol{u_2}$	$\boldsymbol{u_3}$	$\boldsymbol{u_4}$	$\boldsymbol{u_5}$	$\boldsymbol{U}$
a_1 (Audi A3)	22080	105	11.4	5.8	119	4	0.4191	0.0000	0.0000	0.0337	0.0775	0.5304
a_2 (Audi A4)	28100	160	8.6	9.6	164	6	0.1201	0.2095	0.1395	0.0000	0.0000	0.4691
a_3 (BMW 118)	24650	143	9.0	4.5	119	1	0.2834	0.1526	0.1213	0.0555	0.0775	0.6903
a_4 (BMW 320)	32700	177	7.9	6.7	128	5	0.0000	0.2665	0.1713	0.0186	0.0626	0.5190
a_5 (Volvo C30)	22750	136	9.4	7.6	151	2	0.3837	0.1247	0.1032	0.0100	0.0233	0.6450
a_6 (Volvo S40)	27350	180	7.9	8.4	164	3	0.1407	0.2765	0.1713	0.0060	0.0000	0.5946

In what follows, we use MVFs with $\gamma_j = 3$ characteristic points. The set of compatible Additive Value Functions (AVFs) is non-empty. Let us first discuss the intermediate results exploited by some procedures that construct a univocal recommendation. In Table 2, we present the dominance intensities, i.e., minimal value differences for all pairs of alternatives. In case $D(a,b) \geq 0$, a is necessarily weakly preferred to b. Such a robust relation holds for the following pairs: (a_3, a_2), (a_3, a_5), (a_4, a_2), (a_6, a_1), (a_6, a_2), and (a_i, a_i) for $i = 1, \ldots, 6$.

Table 2: Dominance intensities for all pairs of alternatives and scores of alternatives according to AP1, AP2, DME1, and DME2.

Dominance intensities							Scores according to four procedures			
D(a, b)	a_1	a_2	a_3	a_4	a_5	a_6	AP1	AP2	DME1	DME2
a_1	0.0000	-0.9998	-0.9999	-0.9999	-0.9998	-0.9998	-4.9992	-3.5140	-6.73E-05	1.0535
a_2	-0.4922	0.0000	-0.6710	-0.8107	-0.5128	-0.5138	-3.0005	-1.6288	-0.0001	1.0906
a_3	-0.1744	0.0001	0.0000	-0.6238	0.0001	-0.5137	-1.3116	2.3957	0.0002	3.9161
a_4	-0.3188	0.0001	-0.6193	0.0000	-0.3771	-0.3621	-1.6771	2.0774	5.96E-05	3.5251
a_5	-0.5000	-0.3721	-0.7462	-0.8202	0.0000	-0.6976	-3.1360	-0.8788	-4.43E-05	1.9062
a_6	0.0001	5.52E-17	-0.6710	-0.5000	-0.3676	0.0000	-1.5385	1.5485	6.50E-05	3.5085

The set of compatible AVFs can also be exploited with the Monte Carlo simulation to obtain stochastic acceptabilities. In Tables 3 and 4, we present PWIs and RAIs, respectively, estimated based on 10,000 value functions. Clearly, for all pairs $a, b \in A$ such that $a \succsim^N b$, $PWI(a,b) = 1$. However, the share of feasible models confirming the advantage of some alternatives over others is also great for other pairs (see, e.g., (a_3, a_1), (a_4, a_1), and (a_4, a_5)). For another pairs, the preference probabilities are more balanced (see, e.g., (a_1, a_2) and (a_3, a_4)). We distinguish in bold PWIs corresponding

Email addresses: milosz.kadzinski@cs.put.poznan.pl (Miłosz Kadziński), michal.wojcik@cs.put.poznan.pl (Michał Wójcik), k.ciomek@gmail.com (Krzysztof Ciomek)

to the relations observed in the complete reference ranking. Similarly, in Table 4, we emphasize *RAI*s for the positions attained by each alternative in the DM's ranking. Some alternatives (see a_2, a_3, and a_6) have a relatively strong support for their true positions in the set of compatible AVFs. However, the distribution of ranks probabilities is less focused for other alternatives, with a_4 and a_5 attaining five different positions depending on the selected model. Limited support offered to some elements of the true DM's ranking is related to the fact that it is unknown to the method, being supplied with incomplete preference information concerning only four pairs of alternatives.

Table 3: Pairwise winning indices (*PWI*s) for all pairs of alternatives and scores of alternatives according to SUMPOI, MAXPOI, and MINPOI.

Pairwise Winning Indices							Scores according to three procedures		
Alternative	a_1	a_2	a_3	a_4	a_5	a_6	SUMPOI	MAXPOI	MINPOI
a_1	0.000	**0.484**	0.000	**0.034**	0.300	0.000	-3.364	-0.032	-1
a_2	0.516	0.000	0.000	0.000	0.327	0.000	-3,314	0.032	-1
a_3	**1.000**	**1.000**	0.000	**0.587**	**1.000**	**0.769**	3.712	1	0.174
a_4	0.966	**1.000**	0.413	0.000	0.902	0.777	3.116	1	-0.174
a_5	**0.700**	**0.673**	0.000	**0.098**	0.000	**0.169**	-1.72	0.4	-1
a_6	**1.000**	**1.000**	0.231	**0.223**	0.831	0.000	1.57	1	-0.554

Table 4: Rank Acceptability Indices (*RAI*s) for all alternatives and ranks, and scores of alternatives according to EXPRANK, BESTRAI, and WORSTRAI.

Rank Acceptability Indices							Scores according to three procedures		
Rank	1	2	3	4	5	6	EXPRANK	BESTRAI	WORSTRAI
a_1	0.000	0.000	0.016	**0.259**	0.252	0.473	-5.182	-2.984	-6.473
a_2	0.000	0.000	0.000	0.249	0.345	**0.406**	-5.157	-3.751	-6.406
a_3	**0.555**	0.246	0.199	0.000	0.000	0.000	-1.644	-0.445	-3.199
a_4	0.363	0.436	0.119	0.060	**0.022**	0.000	-1.942	-0.637	-5.022
a_5	0.000	**0.051**	0.161	0.286	0.381	0.121	-4.360	-1.949	-6.121
a_6	0.082	0.267	**0.505**	0.146	0.000	0.000	-2.715	-0.918	-4.146

The rankings obtained with all 35 methods are provided in Table 5. UTAMSCVF was the only approach that reproduced the DM's true ranking. The majority of methods (19 out of 35; e.g., ACUTA, EXPRANK, and RANK-SUM) returned the same ranking: $a_3 \succ a_4 \succ a_6 \succ a_5 \succ a_2 \succ a_1$. Only a few methods (e.g., UTAMSVF, DOWN-DIST, and UP-DIST) admitted indifference relation, e.g., by ranking both a_3 and a_4 at the top.

Table 5: Rankings attained by six alternatives in the reference ranking and recommendations provided by the 35 considered methods.

Method	a_1	a_2	a_3	a_4	a_5	a_6	**Method**	a_1	a_2	a_3	a_4	a_5	a_6
REFERENCE	**4**	**6**	**1**	**5**	**2**	**3**	AP1	6	4	1	3	5	2
UTAMP1	6	4	2	1	5	3	AP2	6	5	1	2	4	3
UTAMP2	6	5	1	2	4	3	DME1	5	6	1	3	4	2
UTAMSCVF	4	6	1	5	2	3	DME2	6	5	1	2	4	3
UTAMSVF	6	5	1	1	4	3	MAXPOI	6	5	2	1	4	3
UTAJLS	5	6	3	1	4	2	MINPOI	6	5	1	2	4	3
UTAAVE	6	5	1	2	4	3	SUMPOI	6	5	1	2	4	3
UTACHEB	4	5	3	1	6	2	RANK-SUM-IND	5	5	1	2	4	3
ACUTA	6	5	1	2	4	3	RANK-SUM	6	5	1	2	4	3
UTAROB	6	5	3	1	4	2	RANK-PROD	6	5	1	2	4	3
REPROC	6	5	1	2	4	3	RANK-MM	6	5	1	2	4	3
MAXIMAX	6	5	1	2	4	3	REL-SUM	6	5	1	2	4	3
MAXIMIN	5	4	3	2	6	1	REL-PROD	6	5	1	2	4	3
MM-REGRET	6	4	2	1	5	3	REL-MM	6	5	1	2	4	3
EXPRANK	6	5	1	2	4	3	REL-SUM-IND	6	5	1	2	4	3
BESTRAI	5	6	1	2	4	3	REL-PROD-IND	6	5	1	2	4	3
WORSTRAI	6	5	1	3	4	2	REL-MM-IND	6	5	1	2	4	3
DOWN-DIST	5	5	1	1	4	3	UP-DIST	5	5	1	1	4	3

The MVFs and the respective comprehensive values derived with the ten methods that select a representative value function are shown in Figure 1 and Table 6. The MVFs returned by UTAMP1 and UTAMP2 are very similar, leading to the same rankings. This is understandable given their objective functions that maximize the minimal value difference for pairs of reference alternatives compared by the DM. For UTAMP1, the optimal δ is equal to 0.361, whereas for UTAMP2 – it is only slightly lower (0.3596). This is because the latter approach compromised this objective against a slightly greater minimal slope ρ of MVFs.

The UTAMSCVF method aims at inferring a parsimonious model. This was perfectly attained because MVFs are

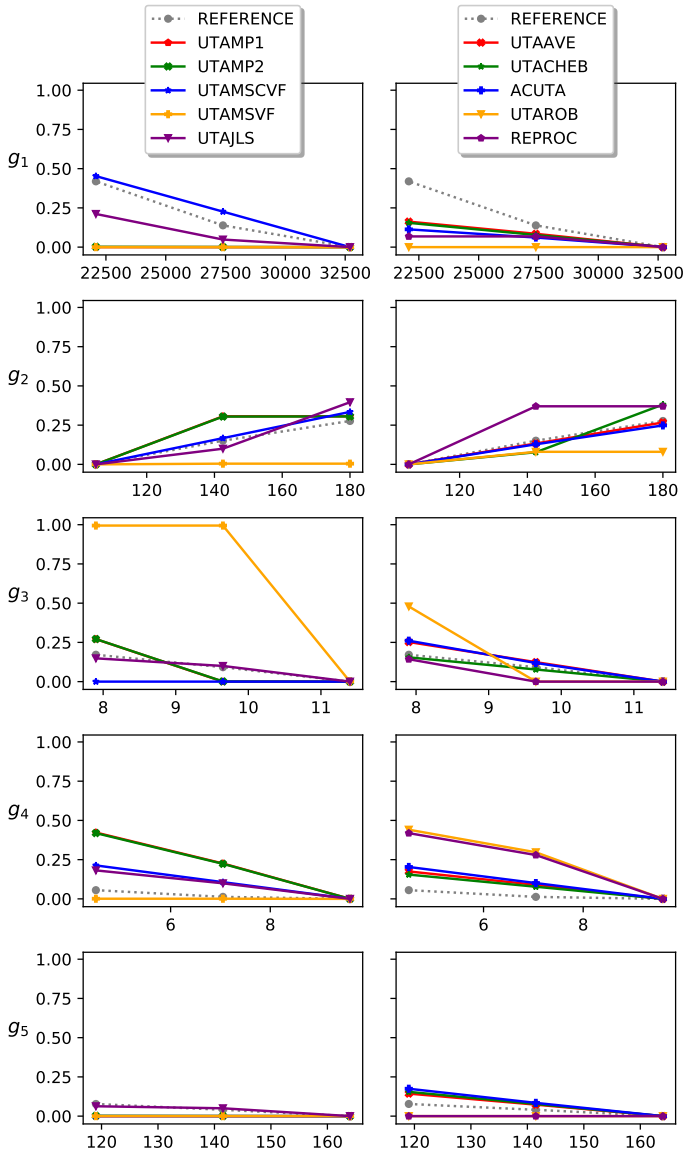


Figure 1: Reference model and representative marginal value functions obtained with ten methods.

linear (see Figure 1), assigning the greatest share in the comprehensive value to g_1, and neglecting the impact of both g_3 and g_5. In turn, UTAMSVF maximized the sum of scores assigned to all reference alternatives. As a result, they are close to one, and the discrimination between alternatives is poor. This was attained by assigning the greatest impact to g_3 and almost nullifying the importance of the remaining four criteria.

Even though UTAJLS and UTAAVE apply a similar idea of selecting an average model, their results differ. Due to considering only the extreme models, UTAJLS assigned slightly greater impact to g_1, g_2, and g_4, whereas UTAAVE – considering a large sample of uniformly distributed models – attributed significant shares in the comprehensive value to g_2 and g_3. Moreover, for UTAAVE, the MVFs are closer to being linear, whereas for UTAJLS – they are either concave or convex on all criteria. For this study, the average model built on Stochastic Ordinal Regression (SOR) is similar to an analytic center of the feasible polyhedron determined by ACUTA. The rankings of these two methods are the same, and the comprehensive scores for six alternatives differ by at most 0.04. The MVFs corresponding to the Chebyshev center are very characteristic in the sense of attributing the same marginal values to the middle breakpoints on all criteria and four out of five best performances. As a result, the impacts of various criteria are more balanced, which is consistent with the objective of finding the center of the hypersphere inscribed in the simplex.

The remaining two procedures for selecting a representative value function emphasize the robustness of results. UTAROB exploits the necessary relation by making the value differences for alternatives compared in the same way by all feasible functions as large as possible. Indeed, such a minimal difference is large, being equal to 0.3314 (see, e.g.,

Table 6: Comprehensive values for all alternatives according to the reference model and the representative value functions selected by ten methods.

Method	a_1	a_2	a_3	a_4	a_5	a_6
REFERENCE	**0.5304**	**0.4591**	**0.6903**	**0.5190**	**0.6450**	**0.5946**
UTAMP1	0.3223	0.4688	0.8298	0.8298	0.4688	0.6833
UTAMP2	0.3234	0.4690	0.8286	0.8286	0.4690	0.6830
UTAMSCVF	0.6114	0.4411	0.7256	0.4421	0.6457	0.6124
UTAMSVF	0.0010	0.9990	1.0000	1.0000	0.9990	0.9995
UTAJLS	0.4137	0.4088	0.5985	0.6888	0.4868	0.6404
UTAAVE	0.4373	0.4701	0.7499	0.7251	0.5177	0.6468
UTACHEB	0.4247	0.4099	0.6142	0.7233	0.4027	0.6502
ACUTA	0.4387	0.4410	0.7668	0.7527	0.4792	0.6179
UTAROB	0.3675	0.3675	0.6989	0.8758	0.3675	0.6989
REPROC	0.4160	0.5149	0.9103	0.8114	0.6137	0.7126

(a_3, a_5), (a_4, a_2), and (a_6, a_2)). The other objective aiming at neglecting the difference for pairs which are not related by $\succsim^N$ led to very similar comprehensive scores for two subsets of alternatives: (a_1, a_2, a_5) and (a_3, a_6). A more detailed information captured by PWIs is handled by REPROC. It maximizes the values difference for pairs $a, b \in A$ such that $PWI(a, b) > PWI(b, a)$. The number of such pairs is greater than those related by the necessary preference. As a result, the optimal value of the objective function is lesser (0.0988) (see, e.g., (a_2, a_1), (a_5, a_2), and (a_3, a_4)). Each alternative that proved to be better than other for a larger share of compatible AVFs is ranked better in the representative ranking determined by REPROC (see, e.g., $PWI(a_5, a_2) = 0.673 > PWI(a_2, a_5) = 0.327$).

The results of three value-based decision rules are presented in Table 7. They provide ambiguous recommendations. In particular, MAXIMAX ranks a_3 at the top because of its excellent comprehensive value equal to one in the most advantageous scenario. Furthermore, according to MINIMAX, a_6 is the most favorable option as it has the greatest minimal comprehensive value (0.3290). Finally, MM-REGRET indicates a_4 as the best because its greatest regret to the most preferred alternative is the least (0.6193).

Table 7: Scores attained by six alternatives according to three value-based decision rules.

Method	a_1	a_2	a_3	a_4	a_5	a_6
MAXIMAX	0.8211	0.9999	1.0000	1.0000	0.9999	0.9999
MINIMAX	7.14E-05	0.1535	0.2727	0.3088	1.35E-16	0.3290
MM-REGRET	-0.9999	-0.8107	-0.6238	-0.6193	-0.8202	-0.6710

The three rank-based decision rules agree concerning the most preferred alternatives. The scores assigned to a_3 by EXPRANK, BESTRAI, and WORSTRAI are -1.644, -0.445, and -3.199, respectively. For the scores of all alternatives, see Table 4. Such favorable evaluations of a_3 derive from attaining:

- the best position (1.644) in an average case (the expected rank for the second-best alternative a_4 is 1.942);
- the first rank in the best case with $RAI(a_3, 1) = 0.555$, being greater than for other potentially optimal alternatives a_4 ($RAI(a_4, 1) = 0.363$) and a_6 ($RAI(a_6, 1) = 0.082$);
- the third rank in the worst case (with $RAI(a_3, 3) = 0.199$), which is better than for the five remaining alternatives attaining positions from fourth (see a_6) to sixth (see a_1, a_2, and a_5) in the least advantageous scenario.

The outcomes for scoring procedures exploiting dominance intensities are provided in Table 2. AP1 assigns the best score to a_3 because its regrets to other alternatives are relatively small, ranging from -0.6238 to 0.0001. On the contrary, a_1 is vastly worse than all other alternatives in the worst-case scenario, with regrets ranging from -0.0998 to -0.9999. AP1 just sums up such regrets over comparisons with all remaining alternatives and favors those for which such a sum of regrets is the least. In turn, AP2 puts together this information with the regrets of all remaining alternatives to a given one. Hence it combines the arguments in favor of each alternative's strength and weakness captured by dominance intensities. Since, in general, other alternatives lose more to a_3 than vice versa, the AP2 score for a_3 is highly positive, putting its ahead of a_4 and a_6 for which the balance is also greater than zero. Note that the sum of AP2 scores for all alternatives is zero because the strength of some alternatives counts as a weakness of some other option.

DME1 considers the same intermediate results as AP2. However, it aggregates them into ratios between positive and negative values in the row and column corresponding to a given alternative in the dominance intensity matrix (see

Table 2). Since there are only a few positive values in this matrix, the ratios and final scores are close to zero. Similar to AP1 and AP2, DME1 ranks a_3 at the top with the score of 0.0002, but the positions of the remaining alternative differ. When it comes to DME2, it considers preference intensities for all pairs $a, b \in A$, deriving them from the comparison of $D(a, b)$ and $D(b, a)$. Such comparisons are again the most advantageous for a_3 (3.9161), which is necessarily preferred to a_2 and a_5 (here, preference intensities are equal to one) while being marginally less favorable in terms of regrets only when compared to a_4. On the other extreme, a_1 attains the lowest DME2 score (1.0535) because it is necessarily outranked by a_6 and loses more in terms of dominance intensities to all remaining alternatives than each of them loses to a_1.

The scores derived with three POI-based procedures are provided in Table 3. They consider the differences between $POI(a, b)$ and $POI(b, a)$ for all pairs $a, b \in A$, while transforming them into scores differently using sum, min, or max operators. According to SUMPOI and MINPOI, a_3 is ranked first, because the comprehensive balance of its POIs is strongly positive (3.712) and its minimal POI advantage over some other alternative is 0.174. For MAXPOI, a_3, a_4, and a_6 attain the same maximal POI difference of one. However, when reducing the POI-based pairwise comparisons only to these three alternatives, a_4 proves to the best, because of its large advantage over a_6 ($POI(a_4, a_6) - POI(a_6, a_4) = 0.777 - 0.223 = 0.554$ is greater than $POI(a_3, a_4) - POI(a_4, a_3) = 0.174$ and $POI(a_3, a_6) - POI(a_6, a_3) = 0.538$).

The remaining methods for constructing a univocal recommendation exploit stochastic acceptabilities. RANK-SUM, RANK-PROD, and RANK-MM derived the same ranking ($a_3 \succ a_4 \succ a_6 \succ a_5 \succ a_2 \succ a_1$) by maximizing the support given to the assignments of alternatives to ranks based on RAIs. For RANK-SUM, the sum of such supports is 2.6 ($RAI(a_3, 1) + RAI(a_4, 2) + RAI(a_6, 3) + RAI(a_5, 4) + RAI(a_2, 5) + RAI(a_1, 6) = 2.6$), for RANK-PROD – their product is 0.0057, and for RANK-MM – the minimal support is 0.286 corresponding to the assignment of a_5 to fourth rank ($RAI(a_5, 4) = 0.286$). The order for RANK-SUM-IND is different in terms of assigning a_1 and a_2 to the same position. This is because such an indifference allowed to maximize the sum of RAI-based supports due to high values of $RAI(a_1, 6) = 0.473$ and $RAI(a_2, 6) = 0.406$.

All mathematical programming models constructing a univocal ranking based on PWIs led to the same solution as RANK-SUM. This means that for this particular problem, it was equivalent to consider the supports provided by stochastic acceptabilities to the assignments of alternatives to ranks and pairwise relations. None PWI-based procedure opted for an indifference relation for any pair of alternatives as such shared ranks were not observed for any compatible value function in the sample of $10,000$ compatible AVFs. Let us just note that the sum of supports to the ranking derived with REL-SUM was 12.721 ($PWI(a_3, a_4) + PWI(a_3, a_6) + \ldots + PWI(a_2, a_1) = 12.721$), whereas the minimal support optimized by REL-MM was 0.516 corresponding to $PWI(a_2, a_1)$. Consequently, similar to REPROC, each alternative preferred to others for a greater share of compatible AVFs is ranked better in the constructed ranking.

The upward and downward distillations applied to the POI matrix constructed the same ranking, admitting indifference relations between a_3 and a_4 at the very top and a_1 and a_2 at the very bottom. When DOWN-DIST considers all alternatives jointly in the first iteration, a_3 and a_4 have the same greatest qualities equal to three. This is because their POIs over a_1, a_2, and a_5 are very high and significantly greater than the inverse POIs. However, in the internal distillation, it is not possible to distinguish among a_3 and a_4, because $POI(a_3, a_4)$ is not significantly greater than $POI(a_4, a_3)$. In the same spirit, when UP-DIST considers all alternatives, a_1 and a_2 have the same least quality equal to -3. This is because the POIs of a_3, a_4, and a_6 over these these two alternatives are very high. Again, in the internal distillation, it is impossible to differentiate a_1 and a_2, because $POI(a_1, a_2)$ and $POI(a_2, a_1)$ are alike.

2. Boxplots

In this section, we present boxplots for the performance measures. For NHR and MRAI, they are provided and discussed in the main paper. Figures 2–6 represent such boxplots for the remaining five measures (Kendall's τ, RDM, RAM, MPRI, and FRAI). Their discussion would be similar to this provided in the main paper for other measures.

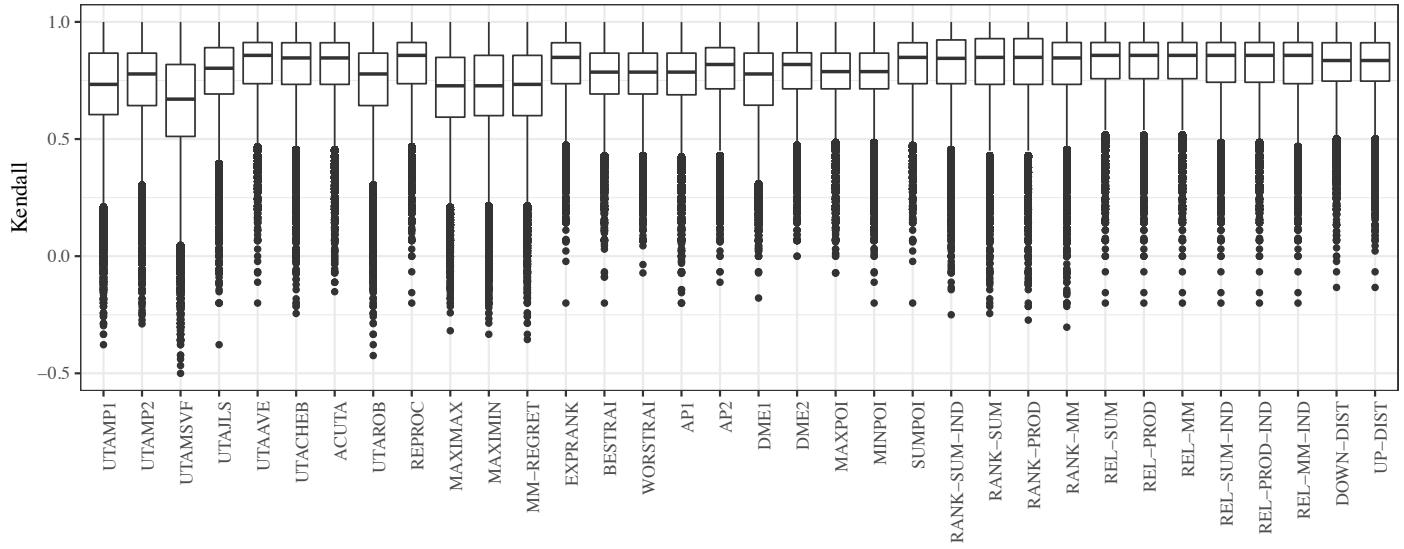


Figure 2: Boxplot for Kendall's τ.

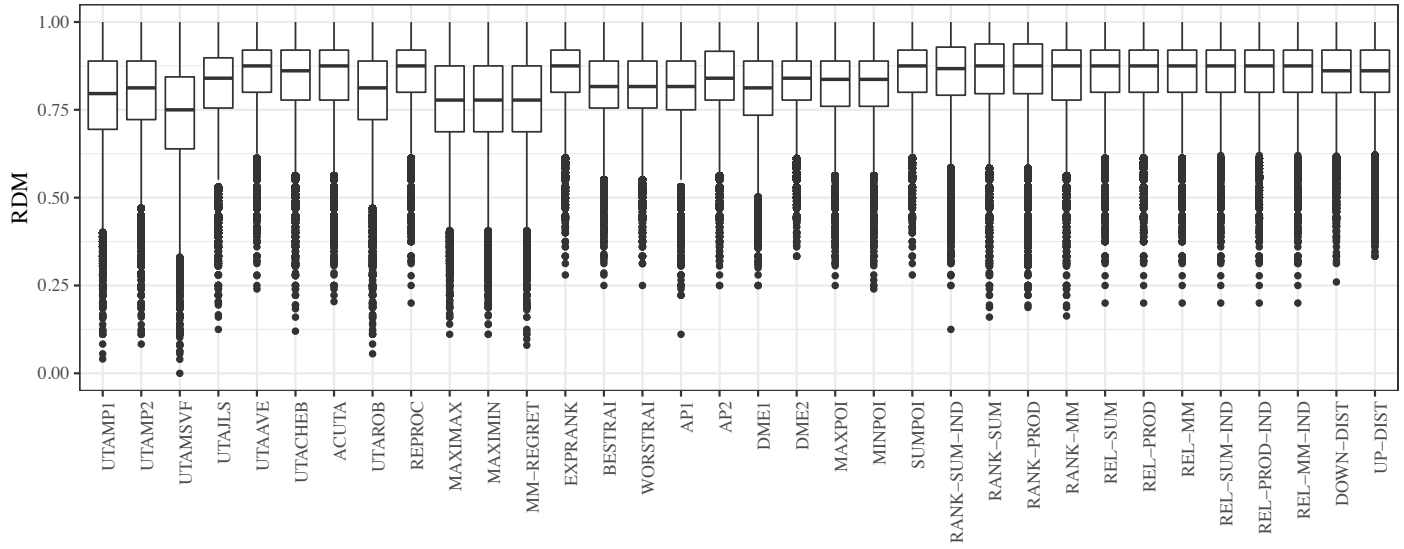


Figure 3: Boxplot for Rank Difference Measure.

3. The Hasse diagrams representing statistically significant differences

In this section, we present the Hasse diagrams representing statistically significant differences between the considered methods. For NHR and MRAI, they are provided and discussed in the main paper. Figures 7–11 represent such diagrams for the remaining five measures (Kendall's τ, RDM, RAM, MPRI, and FRAI).

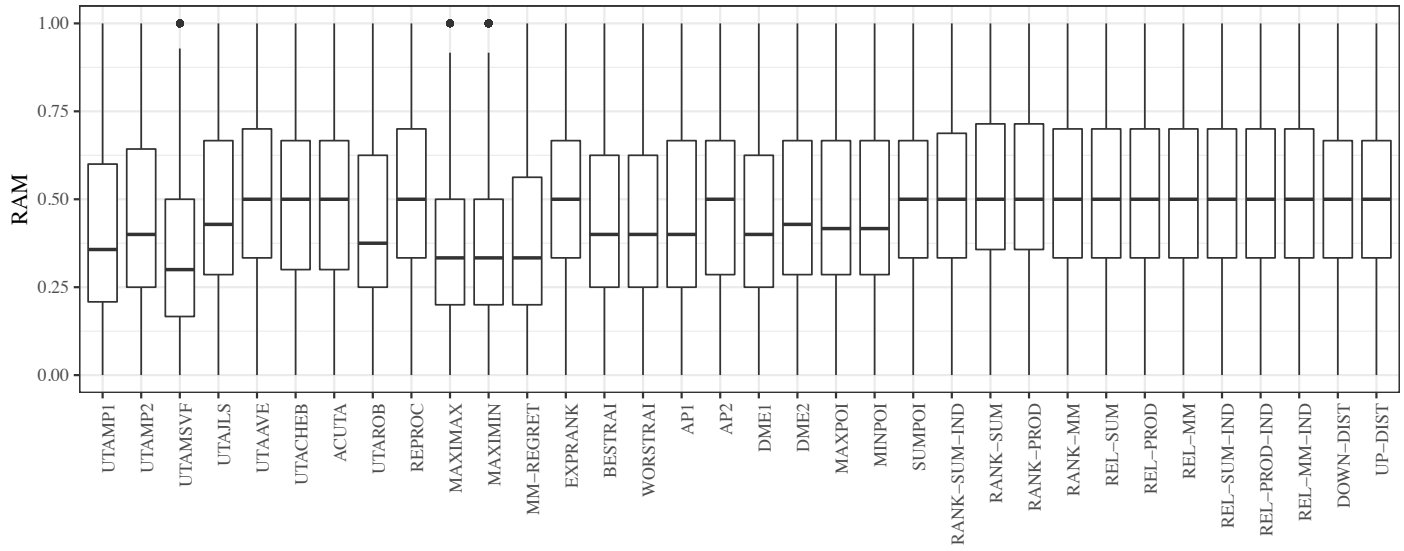


Figure 4: Boxplot for Rank Agreement Measure.

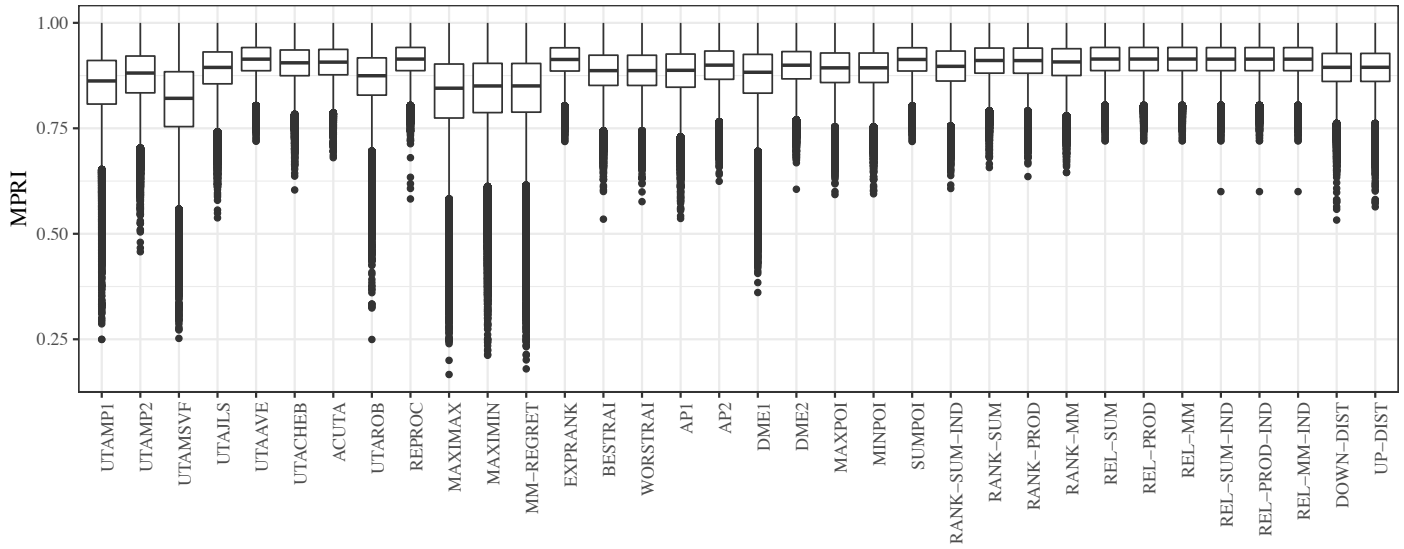


Figure 5: Boxplot for Mean Pairwise Relation Acceptability Index.

4. Experimental results – the extreme performance of the procedures exploiting incomplete preference information

In the main paper, we discussed the experimental results while focussing on the average performance across all considered problem instances. In this section, we refer to the best and the worst performances given the seven measures as well as their relation with the mean attainments. The extreme performances can be observed in the respective boxplots presented in the main paper for NHR and MRAI and in the e-Appendix for Kendall's τ, RDM, RAM, MPRI, and FRAI.

4.1. Similarity between the DM's simulated model and the derived recommendation

In this section, we discuss the similarities in recommendations provided by the reference model and the procedures exploiting incomplete preference information. When it comes to NHR, all methods attained the worst (i.e., 0) and the best (i.e., 1) possible values for some problem instances. In fact, twelve out of 35 methods (e.g., UTAAVE, RANK-SUM, or REL-PROD) obtained only the extreme values for all generated instances. This can be explained as follows. According to the DM's models, there was a unique true most preferred alternative for all considered scenarios. The high discrimination capacity of some methods implied that they generated rankings without shared positions. Then, when the top-ranked alternative was the true most preferred option, NHR was equal to one, whereas if the two alternatives did not align, NHR was zero. However, some other methods tend to rank at least two alternatives at the very top. If one of them was the true most preferred option, then NHR took values between zero and one. For example, for some problems

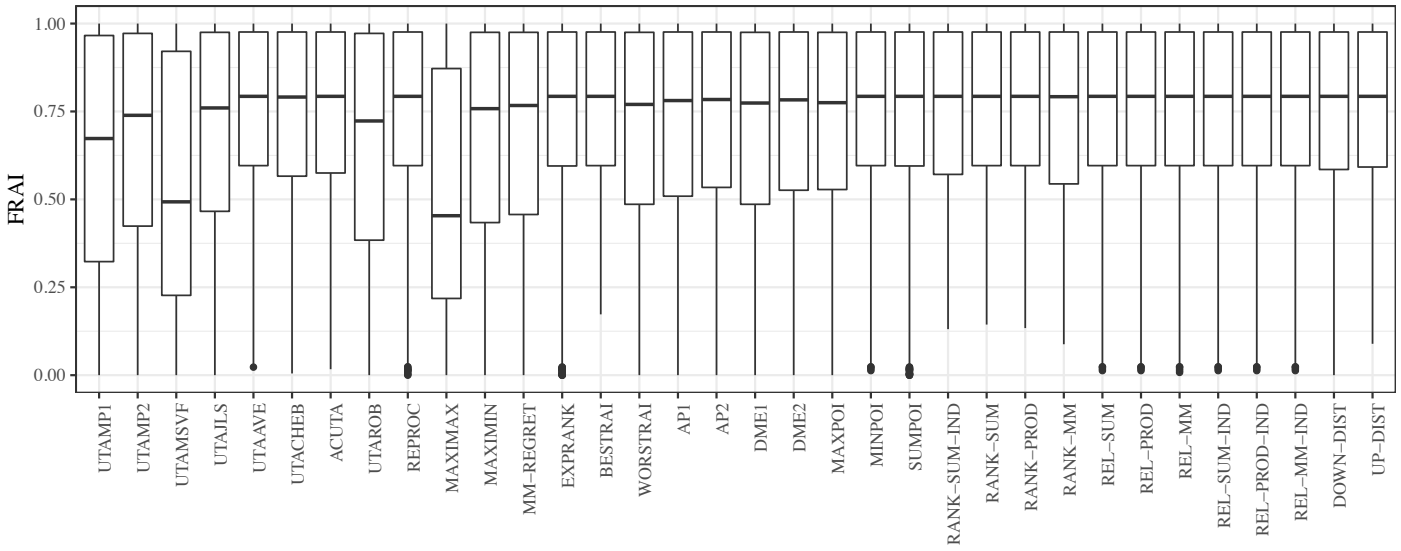


Figure 6: Boxplot for First Rank Acceptability Index.

instances, MAXIMAX and UTAMSVF – characterized by low discrimination power – ranked twelve alternatives at the top. Moreover, the outlying observations presented on the NHR boxplot (see main paper) for REPROC confirm that for some problems, the method ranked the true most preferred alternative first as one out of two, three, four, five, or six alternatives (leading to NHR ranging from 1/2 to 1/6).

The method for which $NHR = 0$ was observed least frequently (i.e., for 16.24% of the considered instances) is UP-DIST. In this regard, it was followed by RANK-SUM (18%), DOWN-DIST (20.99%), and BESTRAI (23.27%). However, the latter approach accomplished the best average NHR value. This is mainly due to attaining NHR$\in (0, 1)$ only for 0.05% considered instances. For other well-performing procedures, such ambiguous hit was observed for a more significant share of problems (for UP-DIST – around 15%, for RANK-SUM – 13%, and for DOWN-DIST – 7%). This, in turn, lowered their average performance in terms of NHR. The procedures that most often failed to identify the true most preferred alternatives were UTAMSCVF (almost 40%), UTAMP1 (over 35%), and UTMSVF (over 34%).

The analysis of results for 35 procedures confirms that the shares of problem instances with NHR equal to zero or one strongly correlate with the average NHR values (see Figure 12). The Pearson's correlation coefficients are -0.748 for the shares of problems leading to $NHR = 0$ and 0.963 for problems with $NHR = 1$. This is an expected result given a low share of problems (2.84%) for which NHR values between zero and one were observed. Hence, in general, the greater the ratio of instances for which the method attained $NHR = 1$ and the lesser the share of problems with $NHR = 0$, the better the average performance of the method in terms of this choice-oriented measure.

The extreme results for ranking-oriented measures quantifying the similarity between the reference and resulting models will be discussed jointly. Each method reproduced the entire true DM's ranking for some considered problem instances. Then, they attained the maximal values for Kendall's τ, RDM, and RAM. Such a scenario was observed most often (from 13.87% to 14.03% of the considered instances) for the RANK methods that emphasize the highest RAI values. On the contrary, the reproduction of the complete ranking was the least frequent for UTAMSVF (6.20%) and the decision rules such as MAXIMAX (6.33%), MAXIMIN (6.62%), and MM-REGRET (6.84%). The correlation between the shares of instances for which the DM's ranking is fully reproduced and the average values of the ranking-oriented performance measures is high (see Figure 13). The precise values of the Pearson's correlation coefficients are 0.856 for Kendall's τ, 0.854 for RDM, and 0.907 for RAM with p-values lesser than 10^{-9} for each of the three measures.

The relation between the minimal and average values for Kendall's τ, RDM, and RAM is presented in Figure 14. For Kendall's τ, the Pearson's correlation coefficient between these values is 0.667 with p-value of $1.22 \cdot 10^{-5}$. The respective coefficient for RDM is 0.688 (p-value $= 4.88 \cdot 10^{-6}$). The worst minimal values of Kendall's τ and RDM were attained by UTAMSVF (-0.5 and 0, respectively) and UTAMSCVF (-0.467 and 0, respectively). The highest most pessimistic values of Kendall's τ were observed for DME2 (0.0), MAXPOI (-0.071), WORSTRAI (-0.071), and AP2 (-0.111), whereas the greatest values of RDM in the worst-case scenario were noted for DME2 and UP-DIST (0.333) followed by SUMPOI (0.28), EXPRANK (0.28), and DOWN-DIST (0.26). The REL methods – which proved to be the best in the

average case given the two measures – attained -0.2 for Kendall's τ and 0.2 for RDM in the most pessimistic scenario.

In the case of RAM, all methods obtained the lowest value of zero for some considered problem instances. The respective shares of scenarios for which no single alternative was assigned to its position in the DM's true ranking are presented in Figure 15. These shares are marginal for RANK-SUM (0.21% of problems), UP-DIST, DOWN-DIST (0.26%), RANK-SUM-IND (0.69%), RANK-PROD (0.72%), and the REL methods (0.81%). Interestingly, even if MAXIMAX attained the least average RAM values, the measure was equal to zero for this method only for 1.42% of instances. On the contrary, UTAJLS – a clearly better average performer in terms of RAM – obtained $RAM = 0$ for nearly 2% of problem instances. The lowest possible RAM values were again noted most often for UTAMSCVF (5.07%) and UTAMSVF (4.16%). Also, there is a strong negative correlation between the shares of instances with the lowest possible RAM values and the mean values of this measure (the Pearson's correlation coefficient is -0.8333 with p-value of $5.41 \cdot 10^{-10}$).

The performances of UP-DIST and DOWN-DIST deserve special attention. When it comes to the average case, they were always placed in the upper half of the ranking for the three measures. However, they reconstructed the entire ranking relatively rarely, only for 8.2% of considered problem instances. At the same time, they belong to the best procedures when it comes to the worst attained values of Kendall's τ and RDM. In general, their results are the most stable, which is confirmed by the lowest standard deviations among all methods given the three ranking-oriented measures.

4.2. Robustness of provided recommendations

In this section, we discuss the robustness of recommendations provided by the considered procedures understood in terms of the support all compatible value functions give them. Attaining the highest possible value for MRAI and MPRI is possible only if all compatible value functions confirm the same ranks (in the case of MRAI) and pairwise relations (in the case of MPRI). Hence, the possibility of attaining MRAI or MPRI equal to one depends on the problem characteristics. Due to the incompleteness of DM's preference information, the methods are rarely given a chance to do so. The highest share of problem instances with the maximal MRAI or MPRI values is equal to 1.08%, observed for 22 out of 35 procedures (including, e.g., UTAAVE, UTACHEB, ACUTA, and the REL and RANK procedures; see Figure 16). The sole outlier for which this share is significantly lower (0.77%) is UTAMSCVF.

The correlation between the lowest and mean values of MRAI (0.951) and MPRI (0.922) is high. It is confirmed by Figure 17, which exhibits the worst value attained by 35 procedures given the two measures. UTAMSVF is the only method that attained the lowest possible MRAI of zero. Conversely, RANK-SUM-IND and RANK-PROD achieved the highest minimal MRAI scores (0.17), followed by the REL procedures (0.14). The REL methods that excluded indifference in the delivered ranking achieved the highest minimum MPRI values (0.720), followed by UTA-AVE (0.719). Noteworthy, REPROC – placed just before the REL methods in the average case – attained a relatively poor lowest value of MPRI (0.583), placing it in the lower half of the ranking in this regard.

The results related to the extreme performances given FRAI are presented in Figure 18. It confirms that all methods achieved the highest possible value of FRAI for some problem instances. For most procedures, this happened for 17.06% of considered instances (i.e., all problems for which all sampled value functions ranked the same alternative at the top). For some other methods, mainly including procedures leading to the selection of some extreme models such as UTAMSCVF (13.38%), MAXIMAX (14.99%), and UTAMSVF (15.66%), these shares are slightly lower. The Pearson's correlation coefficient between the maximal and average values of FRAI is 0.699. The worst possible value of FRAI was attained for at least one problem instance by 19 out of 35 procedures, including, e.g., UTAMP, MAXIMIN, AP, and DME methods. Then, no sampled compatible value function confirmed the alternative selected by these approaches as the most preferred one, hence leading to $FRAI = 0$. This occurred most commonly for MAXIMAX (2.03% of the considered problem instances), UTAMSVF (1.46%), and MM-REGRET (1.14%). The remaining methods always recommended the alternative that appeared at the top for at least one sampled compatible value function. The best results in the most pessimistic scenario given FRAI were attained by BESTRAI (0.173) and the RANK methods (from 0.131 to 0.144).

5. Performance trends

In this section, we present detailed results for performance trends. Such trends for NHR and MRAI were provided and discussed in the main paper. In Tables 8–12, we report them for the remaining five measures (Kendall's τ, RDM, RAM, MPRI, and FRAI). However, their discussion would be similar to this provided in the main paper.

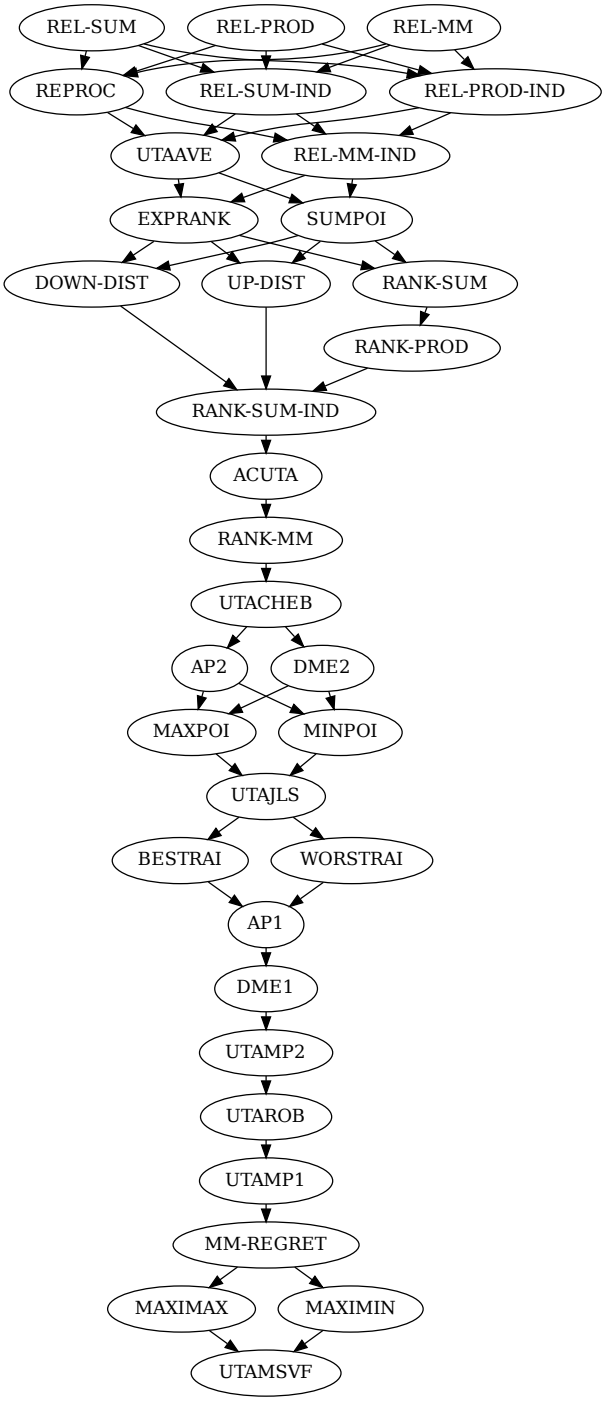


Figure 7: The Hasse diagram indicating the statistically significant differences in terms of Kendall's τ based on the Wilcoxon test with p-value equal to 0.05.

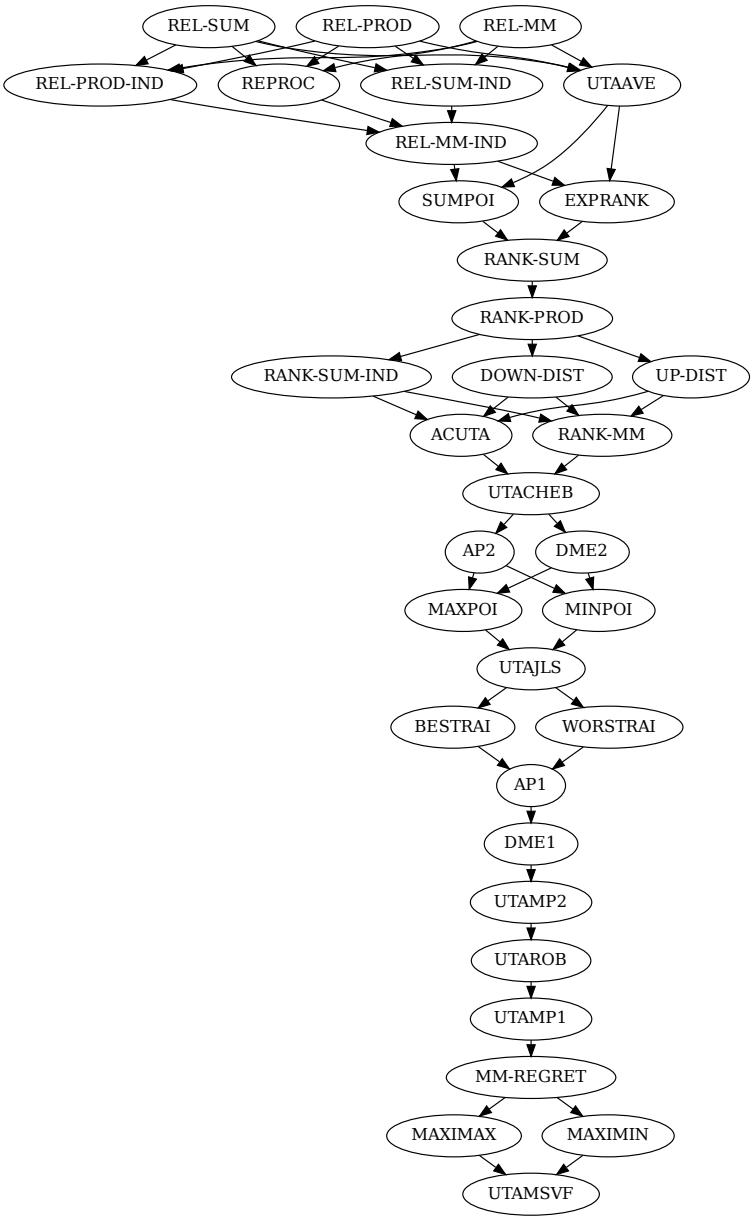


Figure 8: The Hasse diagram indicating the statistically significant differences in terms of RDM based on the Wilcoxon test with p-value equal to 0.05.

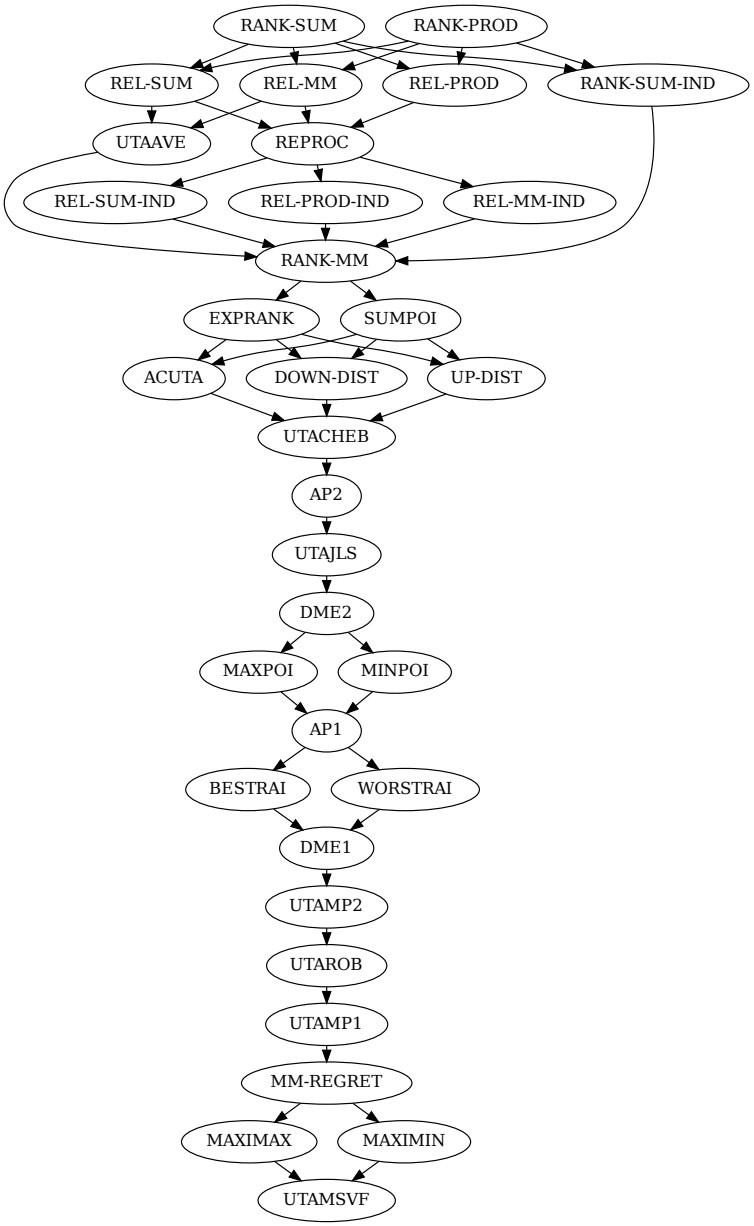


Figure 9: The Hasse diagram indicating the statistically significant differences in terms of RAM based on the Wilcoxon test with p-value equal to 0.05.

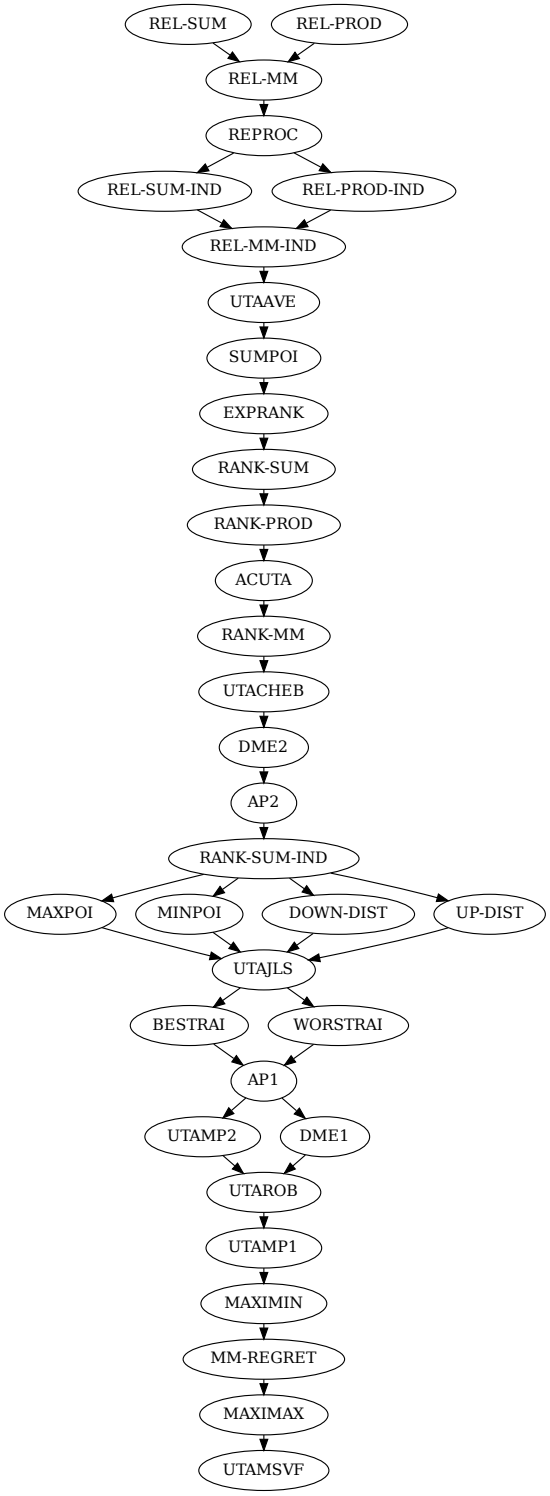


Figure 10: The Hasse diagram indicating the statistically significant differences in terms of $MPRI$ based on the Wilcoxon test with p-value equal to 0.05.

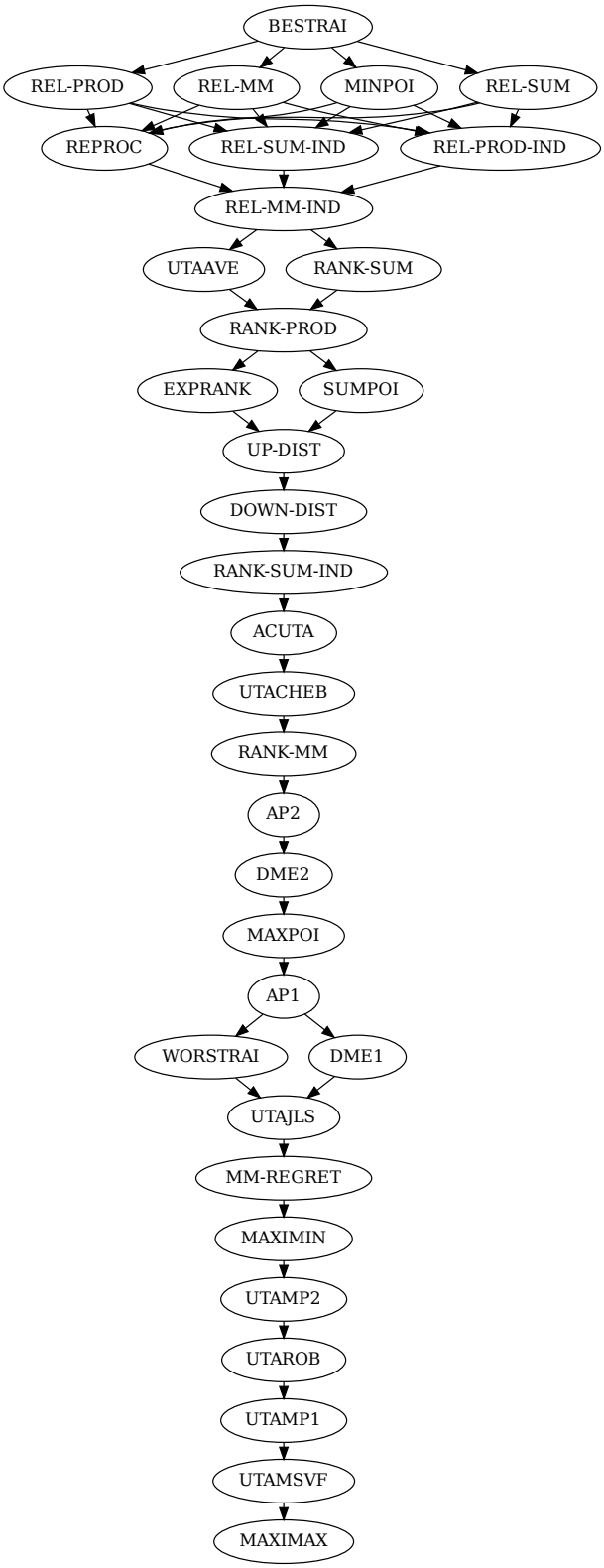


Figure 11: The Hasse diagram indicating the statistically significant differences in terms of $FRAI$ based on the Wilcoxon test with p-value equal to 0.05.

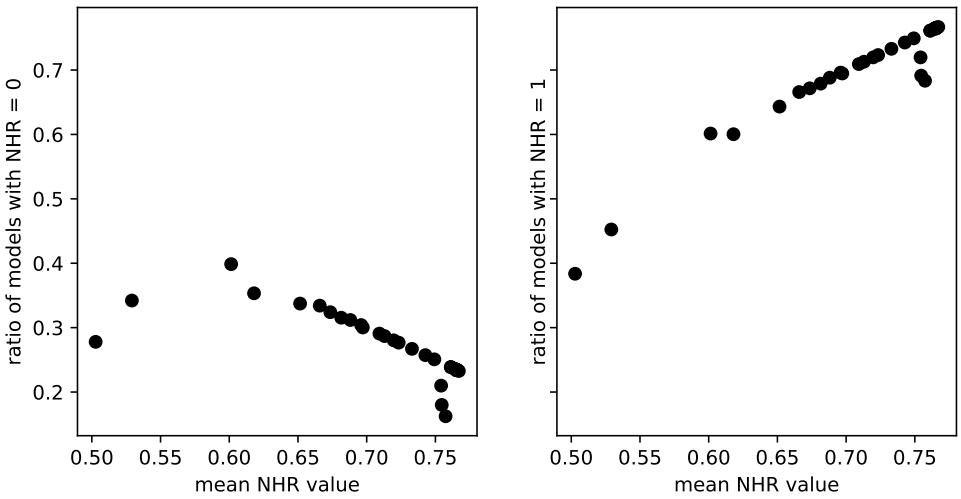


Figure 12: The relation between the average NHR values and the share of models leading to NHR equal to either 0 or 1 based on the performance of 35 procedures.

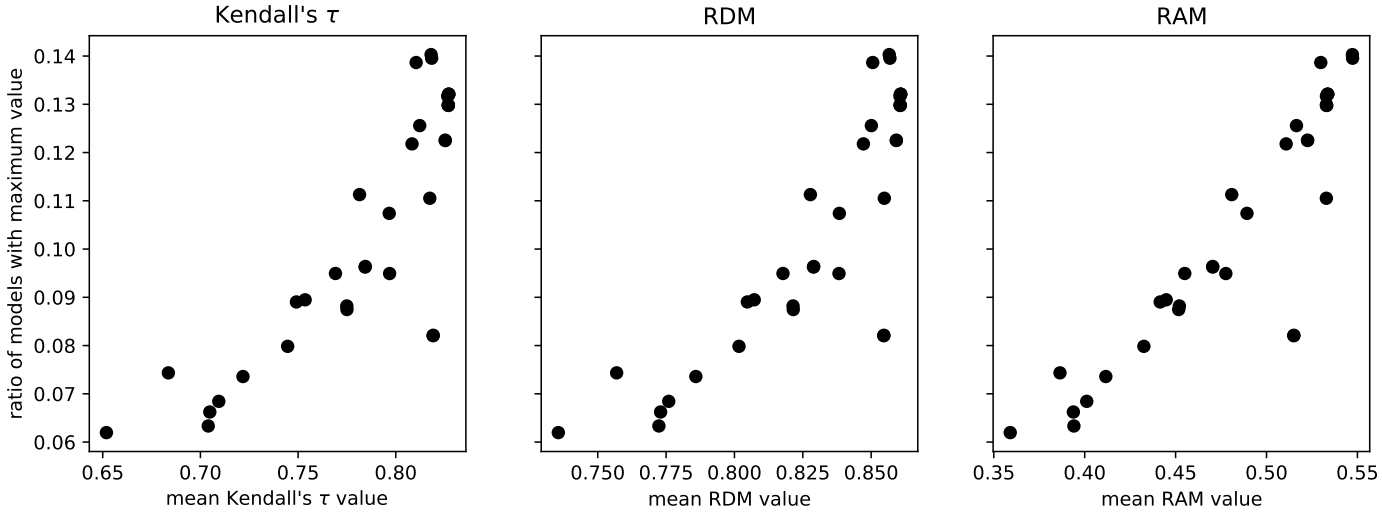


Figure 13: The relation between the average Kendall's τ, RDM, and RAM values and the share of models leading to the maximal values of these performance measures based on the results attained by 35 procedures.

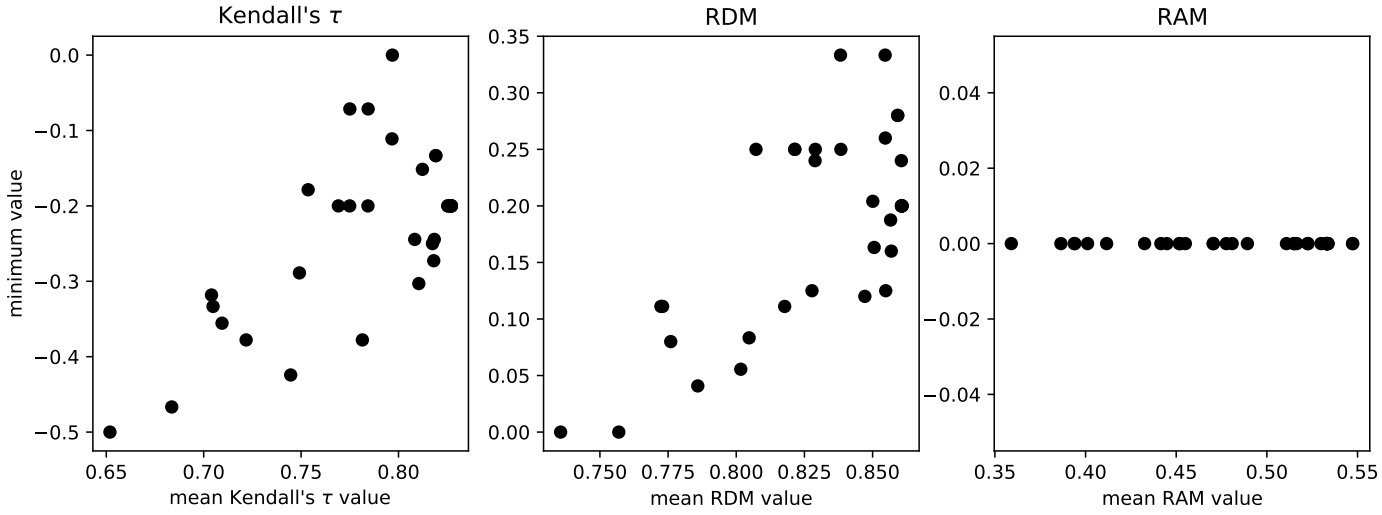


Figure 14: The relation between the average Kendall's τ, RDM, and RAM values and the minimum values of these performance measures based on the results attained by 35 procedures.

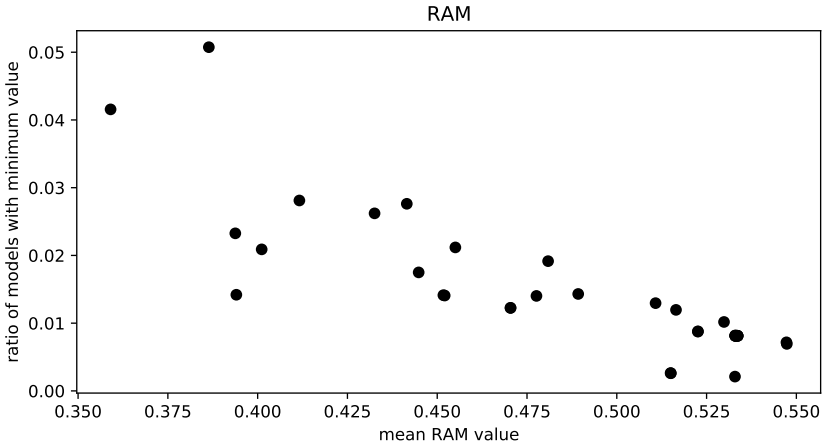


Figure 15: The relation between the average RAM values and the share of models leading to the least possible value of this measure based on the results attained by 35 procedures.

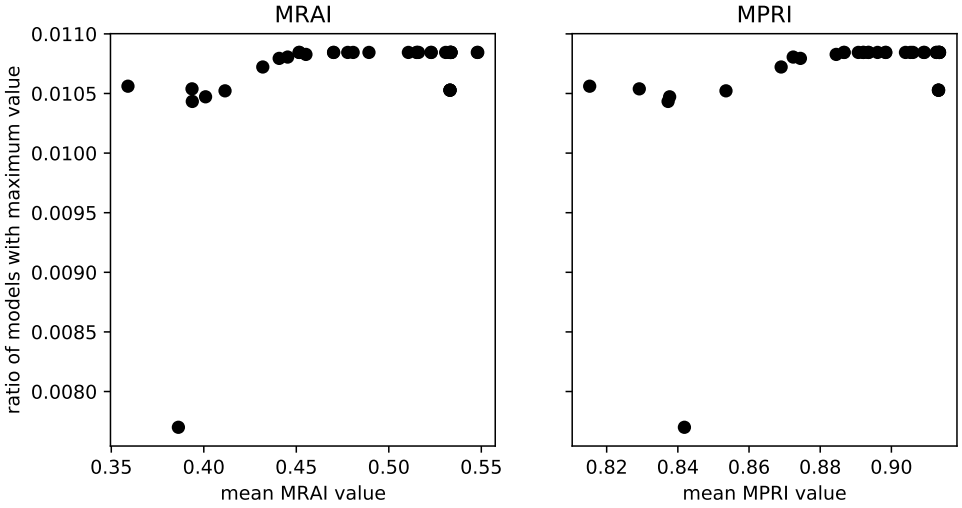


Figure 16: The relation between the average MRAI and MPRI values and the share of models leading to the maximal values of these performance measures based on the results attained by 35 procedures.

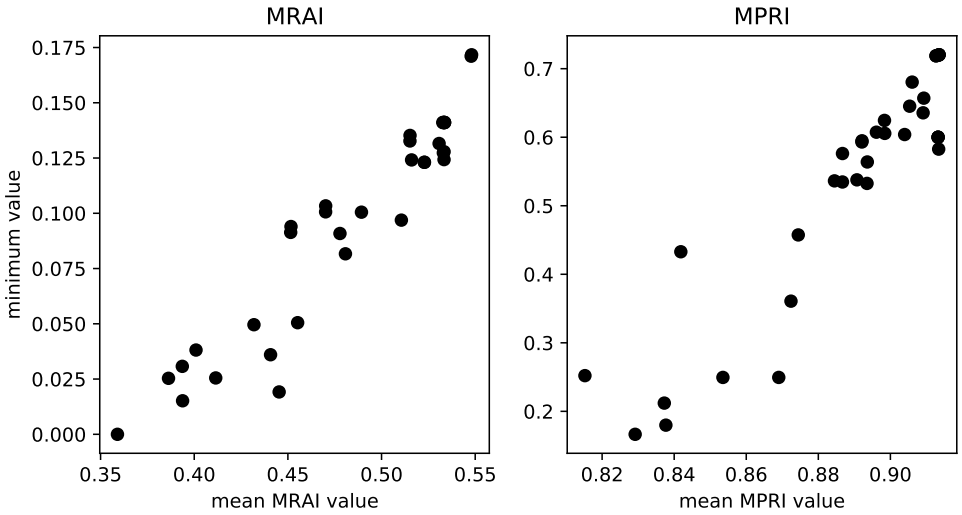


Figure 17: The relation between the average MRAI and MPRI values and the minimum values of these performance measures based on the results attained by 35 procedures.

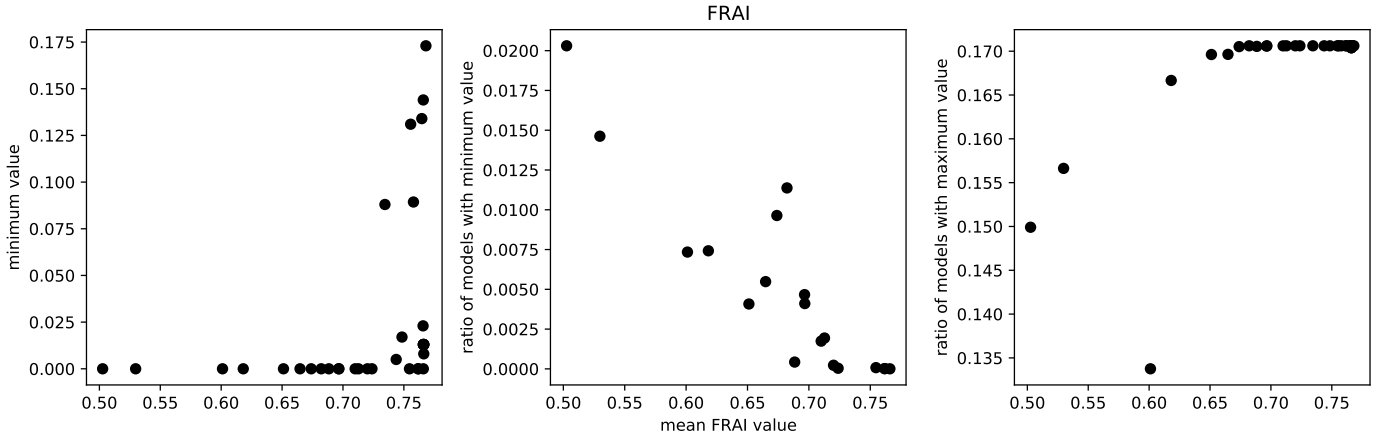


Figure 18: The minimal FRAI values and the shares of models leading to the minimal or maximal FRAIs based on the results attained by 35 procedures.

Table 8: Average Kendall's τ for different numbers of alternatives, criteria, characteristic points, and pairwise comparisons.

		Alternatives					Criteria			Characteristic points			Pairwise comparisons			
METHOD	AVG	6	8	10	12	14	3	4	5	2	3	4	4	6	8	10
UTAMP1	0.7218	0.8082	0.7375	0.7047	0.6851	0.6734	0.7612	0.7172	0.6870	0.7752	0.7067	0.6835	0.6195	0.7030	0.7607	0.8040
UTAMP2	0.7492	0.8294	0.7636	0.7337	0.7140	0.7050	0.7930	0.7444	0.7100	0.7961	0.7385	0.7129	0.6638	0.7321	0.7811	0.8196
UTAMSCVF	0.6836	0.7964	0.7063	0.6591	0.6345	0.6217	0.7430	0.6767	0.6311		0.6802	0.6870	0.5815	0.6619	0.7220	0.7689
UTAMSVF	0.6519	0.7779	0.6765	0.6278	0.5974	0.5798	0.6996	0.6469	0.6091	0.6867	0.6400	0.6290	0.5468	0.6284	0.6910	0.7413
UTAJLS	0.7815	0.8539	0.7953	0.7658	0.7515	0.7408	0.8181	0.7794	0.7470	0.8199	0.7736	0.7509	0.7125	0.7665	0.8067	0.8401
UTAAVE	0.8266	0.8840	0.8368	0.8154	0.8021	0.7947	0.8504	0.8242	0.8052	0.8401	0.8182	0.8214	0.7768	0.8144	0.8450	0.8701
UTACHEB	0.8083	0.8722	0.8191	0.7956	0.7813	0.7735	0.8364	0.8060	0.7827	0.8238	0.7979	0.8033	0.7566	0.7954	0.8270	0.8544
ACUTA	0.8123	0.8751	0.8239	0.7999	0.7850	0.7774	0.8362	0.8096	0.7909	0.8260	0.8031	0.8076	0.7622	0.7993	0.8306	0.8569
UTAROB	0.7447	0.8182	0.7548	0.7282	0.7145	0.7076	0.7881	0.7410	0.7049	0.7870	0.7309	0.7160	0.6638	0.7269	0.7748	0.8132
REPROC	0.8269	0.8844	0.8373	0.8157	0.8024	0.7949	0.8508	0.8246	0.8055	0.8406	0.8186	0.8216	0.7772	0.8146	0.8452	0.8707
MAXIMAX	0.7040	0.8064	0.7245	0.6837	0.6604	0.6451	0.7490	0.7003	0.6627	0.7649	0.6930	0.6541	0.6020	0.6841	0.7426	0.7872
MAXIMIN	0.7048	0.8066	0.7251	0.6860	0.6599	0.6465	0.7487	0.7007	0.6650	0.7658	0.6952	0.6535	0.6032	0.6852	0.7434	0.7875
MM-REGRET	0.7094	0.8099	0.7299	0.6900	0.6654	0.6519	0.7501	0.7045	0.6736	0.7592	0.6936	0.6754	0.6124	0.6911	0.7457	0.7885
EXPRANK	0.8253	0.8827	0.8354	0.8140	0.8008	0.7936	0.8490	0.8229	0.8041	0.8388	0.8170	0.8202	0.7753	0.8134	0.8436	0.8690
BESTRAI	0.7749	0.8533	0.7861	0.7583	0.7429	0.7341	0.8062	0.7716	0.7471	0.7957	0.7639	0.7652	0.7154	0.7595	0.7966	0.8283
WORSTRAI	0.7750	0.8531	0.7867	0.7594	0.7421	0.7339	0.8060	0.7716	0.7475	0.7958	0.7639	0.7654	0.7151	0.7599	0.7968	0.8282
AP1	0.7691	0.8458	0.7834	0.7537	0.7364	0.7264	0.8068	0.7660	0.7347	0.8080	0.7567	0.7428	0.7010	0.7529	0.7945	0.8281
AP2	0.7967	0.8588	0.8074	0.7838	0.7703	0.7630	0.8303	0.7937	0.7659	0.8317	0.7849	0.7734	0.7397	0.7823	0.8176	0.8470
DME1	0.7536	0.8354	0.7664	0.7371	0.7186	0.7106	0.8062	0.7517	0.7029	0.8077	0.7429	0.7102	0.6694	0.7364	0.7851	0.8235
DME2	0.7969	0.8565	0.8050	0.7842	0.7724	0.7662	0.8314	0.7944	0.7648	0.8280	0.7874	0.7752	0.7390	0.7829	0.8183	0.8472
MAXPOI	0.7844	0.8597	0.7959	0.7690	0.7528	0.7447	0.8161	0.7817	0.7555	0.8085	0.7725	0.7722	0.7261	0.7696	0.8055	0.8364
MINPOI	0.7843	0.8595	0.7958	0.7683	0.7529	0.7451	0.8164	0.7814	0.7551	0.8082	0.7725	0.7723	0.7267	0.7695	0.8050	0.8360
SUMPOI	0.8253	0.8828	0.8354	0.8140	0.8008	0.7936	0.8490	0.8229	0.8041	0.8388	0.8170	0.8202	0.7753	0.8134	0.8437	0.8690
RANK-SUM-IND	0.8174	0.8805	0.8302	0.8052	0.7902	0.7811	0.8419	0.8148	0.7956	0.8304	0.8087	0.8132	0.7643	0.8044	0.8369	0.8642
RANK-SUM	0.8184	0.8817	0.8313	0.8062	0.7906	0.7822	0.8427	0.8154	0.7970	0.8302	0.8102	0.8147	0.7651	0.8057	0.8377	0.8650
RANK-PROD	0.8181	0.8815	0.8310	0.8058	0.7904	0.7818	0.8424	0.8153	0.7966	0.8298	0.8100	0.8144	0.7648	0.8052	0.8375	0.8649
RANK-MM	0.8104	0.8775	0.8252	0.7983	0.7807	0.7703	0.8359	0.8076	0.7878	0.8230	0.8018	0.8066	0.7547	0.7971	0.8308	0.8591
REL-SUM	0.8271	0.8844	0.8374	0.8159	0.8026	0.7953	0.8509	0.8248	0.8057	0.8408	0.8188	0.8218	0.7776	0.8149	0.8453	0.8707
REL-PROD	0.8271	0.8844	0.8374	0.8159	0.8026	0.7953	0.8509	0.8248	0.8057	0.8408	0.8188	0.8218	0.7775	0.8149	0.8453	0.8707
REL-MM	0.8271	0.8844	0.8374	0.8159	0.8026	0.7953	0.8509	0.8248	0.8057	0.8408	0.8188	0.8218	0.7776	0.8149	0.8453	0.8707
REL-SUM-IND	0.8269	0.8842	0.8372	0.8157	0.8024	0.7951	0.8505	0.8247	0.8056	0.8405	0.8186	0.8217	0.7775	0.8147	0.8451	0.8703
REL-PROD-IND	0.8269	0.8842	0.8372	0.8157	0.8024	0.7952	0.8506	0.8247	0.8056	0.8405	0.8186	0.8217	0.7775	0.8147	0.8451	0.8703
REL-MM-IND	0.8268	0.8843	0.8372	0.8157	0.8022	0.7949	0.8504	0.8246	0.8055	0.8404	0.8185	0.8216	0.7775	0.8146	0.8451	0.8702
DOWN-DIST	0.8192	0.8784	0.8299	0.8078	0.7937	0.7862	0.8431	0.8166	0.7979	0.8321	0.8107	0.8147	0.7672	0.8069	0.8382	0.8645
UP-DIST	0.8191	0.8785	0.8294	0.8075	0.7937	0.7864	0.8430	0.8168	0.7975	0.8325	0.8105	0.8143	0.7675	0.8068	0.8377	0.8644

Table 9: Average Rank Difference Measures (RDMs) for different numbers of alternatives, criteria, characteristic points, and pairwise comparisons.

		Alternatives					Criteria			Characteristic points			Pairwise comparisons			
METHOD	AVG	6	8	10	12	14	3	4	5	2	3	4	4	6	8	10
UTAMP1	0.7859	0.8491	0.7951	0.7722	0.7600	0.7529	0.8136	0.7825	0.7616	0.8245	0.7748	0.7584	0.7155	0.7718	0.8123	0.8439
UTAMP2	0.8047	0.8646	0.8134	0.7921	0.7794	0.7741	0.8358	0.8012	0.7773	0.8389	0.7971	0.7782	0.7448	0.7919	0.8270	0.8551
UTAMSCVF	0.7569	0.8388	0.7705	0.7376	0.7224	0.7153	0.7987	0.7518	0.7203		0.7547	0.7592	0.6869	0.7410	0.7829	0.8169
UTAMSVF	0.7356	0.8254	0.7495	0.7170	0.6982	0.6878	0.7686	0.7318	0.7064	0.7607	0.7265	0.7196	0.6648	0.7185	0.7615	0.7976
UTAJLS	0.8278	0.8832	0.8364	0.8150	0.8054	0.7989	0.8543	0.8261	0.8030	0.8567	0.8214	0.8052	0.7786	0.8164	0.8456	0.8705
UTAAVE	0.8605	0.9064	0.8674	0.8508	0.8414	0.8367	0.8783	0.8586	0.8447	0.8714	0.8540	0.8563	0.8239	0.8512	0.8740	0.8931
UTACHEB	0.8472	0.8971	0.8542	0.8364	0.8264	0.8217	0.8679	0.8453	0.8283	0.8596	0.8391	0.8428	0.8097	0.8373	0.8606	0.8811
ACUTA	0.8501	0.8996	0.8577	0.8396	0.8291	0.8245	0.8679	0.8480	0.8343	0.8612	0.8430	0.8460	0.8137	0.8402	0.8633	0.8831
UTAROB	0.8017	0.8564	0.8074	0.7883	0.7800	0.7762	0.8325	0.7988	0.7736	0.8329	0.7914	0.7807	0.7454	0.7883	0.8224	0.8506
REPROC	0.8607	0.9066	0.8677	0.8508	0.8414	0.8367	0.8785	0.8587	0.8447	0.8715	0.8541	0.8564	0.8239	0.8513	0.8740	0.8935
MAXIMAX	0.7724	0.8471	0.7848	0.7564	0.7414	0.7321	0.8039	0.7696	0.7436	0.8144	0.7645	0.7382	0.7026	0.7576	0.7985	0.8308
MAXIMIN	0.7730	0.8473	0.7850	0.7581	0.7414	0.7332	0.8037	0.7701	0.7451	0.8151	0.7661	0.7378	0.7033	0.7584	0.7991	0.8311
MM-REGRET	0.7760	0.8497	0.7884	0.7604	0.7448	0.7366	0.8045	0.7725	0.7509	0.8107	0.7648	0.7524	0.7091	0.7622	0.8007	0.8319
EXPRANK	0.8592	0.9053	0.8659	0.8493	0.8399	0.8356	0.8768	0.8573	0.8435	0.8698	0.8527	0.8551	0.8222	0.8501	0.8725	0.8920
BESTRAI	0.8214	0.8821	0.8282	0.8077	0.7973	0.7919	0.8440	0.8190	0.8013	0.8365	0.8133	0.8144	0.7784	0.8098	0.8368	0.8606
WORSTRAI	0.8215	0.8820	0.8287	0.8082	0.7968	0.7919	0.8439	0.8190	0.8017	0.8365	0.8135	0.8146	0.7784	0.8100	0.8371	0.8606
AP1	0.8178	0.8767	0.8268	0.8047	0.7932	0.7872	0.8448	0.8154	0.7931	0.8463	0.8086	0.7984	0.7689	0.8055	0.8357	0.8609
AP2	0.8384	0.8868	0.8449	0.8275	0.8184	0.8143	0.8630	0.8361	0.8160	0.8648	0.8294	0.8209	0.7970	0.8274	0.8535	0.8756
DME1	0.8072	0.8694	0.8151	0.7939	0.7814	0.7764	0.8447	0.8057	0.7714	0.8458	0.7992	0.7767	0.7479	0.7942	0.8293	0.8576
DME2	0.8382	0.8850	0.8430	0.8275	0.8195	0.8161	0.8636	0.8363	0.8148	0.8616	0.8309	0.8222	0.7962	0.8278	0.8536	0.8753
MAXPOI	0.8290	0.8874	0.8361	0.8159	0.8051	0.8004	0.8522	0.8270	0.8078	0.8471	0.8202	0.8197	0.7871	0.8178	0.8440	0.8670
MINPOI	0.8289	0.8871	0.8359	0.8156	0.8053	0.8007	0.8524	0.8267	0.8076	0.8469	0.8200	0.8198	0.7873	0.8178	0.8437	0.8669
SUMPOI	0.8592	0.9053	0.8659	0.8493	0.8399	0.8356	0.8768	0.8573	0.8436	0.8698	0.8527	0.8551	0.8222	0.8501	0.8725	0.8920
RANK-SUM-IND	0.8548	0.9040	0.8633	0.8443	0.8341	0.8284	0.8730	0.8528	0.8387	0.8654	0.8481	0.8510	0.8163	0.8450	0.8687	0.8892
RANK-SUM	0.8569	0.9053	0.8653	0.8466	0.8361	0.8311	0.8751	0.8546	0.8409	0.8675	0.8503	0.8528	0.8189	0.8473	0.8705	0.8908
RANK-PROD	0.8566	0.9051	0.8649	0.8463	0.8359	0.8308	0.8748	0.8544	0.8406	0.8672	0.8501	0.8526	0.8186	0.8469	0.8703	0.8906
RANK-MM	0.8506	0.9020	0.8603	0.8403	0.8283	0.8220	0.8697	0.8483	0.8337	0.8618	0.8436	0.8463	0.8110	0.8405	0.8648	0.8859
REL-SUM	0.8608	0.9067	0.8678	0.8510	0.8416	0.8370	0.8786	0.8589	0.8449	0.8716	0.8543	0.8565	0.8241	0.8515	0.8741	0.8935
REL-PROD	0.8608	0.9066	0.8678	0.8510	0.8415	0.8370	0.8786	0.8589	0.8449	0.8716	0.8542	0.8565	0.8241	0.8515	0.8741	0.8935
REL-MM	0.8608	0.9066	0.8678	0.8510	0.8416	0.8370	0.8786	0.8589	0.8449	0.8716	0.8543	0.8565	0.8241	0.8515	0.8741	0.8935
REL-SUM-IND	0.8606	0.9065	0.8676	0.8508	0.8414	0.8369	0.8783	0.8588	0.8448	0.8714	0.8541	0.8564	0.8241	0.8513	0.8739	0.8932
REL-PROD-IND	0.8606	0.9065	0.8676	0.8508	0.8414	0.8369	0.8783	0.8588	0.8448	0.8714	0.8541	0.8564	0.8241	0.8513	0.8739	0.8932
REL-MM-IND	0.8606	0.9065	0.8676	0.8508	0.8412	0.8367	0.8781	0.8587	0.8448	0.8713	0.8541	0.8563	0.8241	0.8512	0.8739	0.8931
DOWN-DIST	0.8547	0.9019	0.8619	0.8446	0.8348	0.8302	0.8724	0.8527	0.8389	0.8648	0.8481	0.8511	0.8164	0.8452	0.8685	0.8885
UP-DIST	0.8546	0.9019	0.8614	0.8445	0.8347	0.8304	0.8723	0.8528	0.8387	0.8651	0.8479	0.8508	0.8166	0.8452	0.8681	0.8884

Table 10: Average Rank Agreement Measures (RAMs) for different numbers of alternatives, criteria, characteristic points, and pairwise comparisons.

		Alternatives					Criteria			Characteristic points			Pairwise comparisons			
METHOD	AVG	6	8	10	12	14	3	4	5	2	3	4	4	6	8	10
UTAMP1	0.4116	0.6365	0.4631	0.3710	0.3144	0.2731	0.4585	0.4038	0.3725	0.4858	0.3889	0.3602	0.3188	0.3834	0.4439	0.5003
UTAMP2	0.4415	0.6671	0.4952	0.4038	0.3414	0.3001	0.4970	0.4326	0.3950	0.5065	0.4263	0.3918	0.3564	0.4155	0.4706	0.5237
UTAMSCVF	0.3864	0.6330	0.4415	0.3382	0.2796	0.2397	0.4421	0.3762	0.3410		0.3850	0.3878	0.3010	0.3597	0.4158	0.4691
UTAMSVF	0.3590	0.6066	0.4075	0.3123	0.2540	0.2149	0.4007	0.3517	0.3248	0.4036	0.3447	0.3289	0.2777	0.3323	0.3854	0.4408
UTAJLS	0.4809	0.7089	0.5403	0.4421	0.3795	0.3336	0.5318	0.4750	0.4358	0.5438	0.4653	0.4335	0.4003	0.4569	0.5085	0.5578
UTAAVE	0.5330	0.7547	0.5967	0.5003	0.4316	0.3815	0.5757	0.5269	0.4963	0.5661	0.5157	0.5171	0.4598	0.5106	0.5585	0.6030
UTACHEB	0.5108	0.7352	0.5717	0.4756	0.4090	0.3625	0.5568	0.5050	0.4706	0.5489	0.4908	0.4926	0.4425	0.4889	0.5343	0.5775
ACUTA	0.5165	0.7399	0.5783	0.4826	0.4136	0.3680	0.5591	0.5103	0.4800	0.5517	0.4984	0.4994	0.4479	0.4945	0.5401	0.5834
UTAROB	0.4326	0.6497	0.4824	0.3930	0.3384	0.2993	0.4884	0.4241	0.3852	0.4999	0.4102	0.3876	0.3523	0.4055	0.4599	0.5125
REPROC	0.5334	0.7553	0.5977	0.5006	0.4317	0.3817	0.5760	0.5277	0.4964	0.5664	0.5163	0.5174	0.4597	0.5114	0.5585	0.6040
MAXIMAX	0.3941	0.6378	0.4506	0.3498	0.2865	0.2455	0.4406	0.3876	0.3540	0.4560	0.3802	0.3460	0.3058	0.3666	0.4244	0.4795
MAXIMIN	0.3938	0.6384	0.4491	0.3500	0.2866	0.2448	0.4397	0.3870	0.3546	0.4571	0.3807	0.3435	0.3055	0.3661	0.4243	0.4793
MM-REGRET	0.4011	0.6447	0.4577	0.3570	0.2946	0.2516	0.4452	0.3937	0.3644	0.4584	0.3836	0.3613	0.3132	0.3748	0.4309	0.4856
EXPRANK	0.5226	0.7484	0.5858	0.4887	0.4190	0.3710	0.5632	0.5171	0.4875	0.5534	0.5056	0.5087	0.4490	0.5011	0.5471	0.5931
BESTRAI	0.4521	0.6982	0.5080	0.4095	0.3453	0.2995	0.4939	0.4462	0.4162	0.4836	0.4354	0.4373	0.3803	0.4283	0.4763	0.5235
WORSTRAI	0.4517	0.6973	0.5084	0.4088	0.3442	0.2998	0.4927	0.4457	0.4168	0.4827	0.4359	0.4366	0.3804	0.4278	0.4758	0.5228
AP1	0.4551	0.6923	0.5158	0.4141	0.3493	0.3038	0.5060	0.4483	0.4108	0.5135	0.4356	0.4160	0.3764	0.4298	0.4816	0.5324
AP2	0.4893	0.7085	0.5469	0.4531	0.3908	0.3470	0.5427	0.4826	0.4425	0.5521	0.4684	0.4473	0.4139	0.4651	0.5149	0.5631
DME1	0.4448	0.6763	0.4979	0.4051	0.3438	0.3011	0.5079	0.4374	0.3891	0.5127	0.4286	0.3932	0.3579	0.4179	0.4751	0.5284
DME2	0.4776	0.6982	0.5324	0.4412	0.3797	0.3366	0.5311	0.4716	0.4302	0.5319	0.4593	0.4417	0.4029	0.4548	0.5029	0.5500
MAXPOI	0.4704	0.7107	0.5278	0.4300	0.3641	0.3195	0.5157	0.4651	0.4303	0.5098	0.4524	0.4490	0.3987	0.4471	0.4946	0.5412
MINPOI	0.4704	0.7104	0.5282	0.4288	0.3647	0.3199	0.5158	0.4645	0.4309	0.5103	0.4513	0.4497	0.3983	0.4474	0.4942	0.5419
SUMPOI	0.5226	0.7485	0.5858	0.4887	0.4190	0.3710	0.5632	0.5171	0.4875	0.5535	0.5056	0.5087	0.4491	0.5011	0.5471	0.5932
RANK-SUM-IND	0.5329	0.7552	0.5993	0.4998	0.4306	0.3796	0.5757	0.5272	0.4958	0.5662	0.5159	0.5166	0.4586	0.5111	0.5582	0.6037
RANK-SUM	0.5473	0.7592	0.6095	0.5155	0.4499	0.4026	0.5922	0.5408	0.5090	0.5862	0.5293	0.5265	0.4774	0.5261	0.5711	0.6148
RANK-PROD	0.5472	0.7595	0.6094	0.5153	0.4499	0.4021	0.5921	0.5405	0.5091	0.5860	0.5292	0.5266	0.4773	0.5256	0.5714	0.6147
RANK-MM	0.5299	0.7546	0.5992	0.4978	0.4249	0.3729	0.5769	0.5230	0.4896	0.5701	0.5111	0.5084	0.4573	0.5079	0.5546	0.5996
REL-SUM	0.5337	0.7554	0.5978	0.5009	0.4318	0.3824	0.5762	0.5281	0.4968	0.5667	0.5166	0.5176	0.4601	0.5117	0.5587	0.6042
REL-PROD	0.5336	0.7554	0.5978	0.5009	0.4318	0.3824	0.5761	0.5280	0.4968	0.5667	0.5166	0.5176	0.4600	0.5117	0.5587	0.6042
REL-MM	0.5336	0.7554	0.5978	0.5009	0.4318	0.3823	0.5762	0.5281	0.4967	0.5667	0.5166	0.5176	0.4600	0.5117	0.5587	0.6042
REL-SUM-IND	0.5330	0.7550	0.5971	0.5002	0.4310	0.3819	0.5750	0.5276	0.4966	0.5657	0.5161	0.5173	0.4600	0.5112	0.5580	0.6030
REL-PROD-IND	0.5331	0.7550	0.5971	0.5002	0.4310	0.3819	0.5750	0.5276	0.4966	0.5657	0.5162	0.5173	0.4600	0.5112	0.5580	0.6030
REL-MM-IND	0.5329	0.7551	0.5972	0.5001	0.4307	0.3815	0.5748	0.5275	0.4965	0.5655	0.5160	0.5173	0.4601	0.5111	0.5580	0.6026
DOWN-DIST	0.5151	0.7423	0.5793	0.4806	0.4108	0.3623	0.5552	0.5098	0.4801	0.5443	0.4986	0.5023	0.4407	0.4929	0.5406	0.5860
UP-DIST	0.5150	0.7421	0.5790	0.4805	0.4105	0.3628	0.5552	0.5096	0.4801	0.5452	0.4982	0.5015	0.4411	0.4926	0.5397	0.5864

Table 11: Average Mean Pairwise Relation Acceptability Indices (MPRIs) for different numbers of alternatives, criteria, characteristic points, and pairwise comparisons.

		Alternatives					Criteria			Characteristic points			Pairwise comparisons			
METHOD	AVG	6	8	10	12	14	3	4	5	2	3	4	4	6	8	10
UTAMP1	0.8535	0.8954	0.8612	0.8452	0.8353	0.8304	0.8727	0.8514	0.8365	0.8876	0.8471	0.8259	0.7934	0.8447	0.8764	0.8997
UTAMP2	0.8744	0.9144	0.8820	0.8666	0.8568	0.8524	0.8966	0.8721	0.8546	0.8977	0.8690	0.8566	0.8315	0.8663	0.8904	0.9095
UTAMSCVF	0.8418	0.8983	0.8526	0.8305	0.8171	0.8107	0.8713	0.8388	0.8154		0.8398	0.8438	0.7909	0.8312	0.8608	0.8845
UTAMSVF	0.8152	0.8801	0.8280	0.8026	0.7871	0.7783	0.8392	0.8123	0.7942	0.8433	0.8086	0.7938	0.7536	0.8030	0.8383	0.8661
UTAJLS	0.8907	0.9267	0.8976	0.8836	0.8751	0.8705	0.9090	0.8895	0.8736	0.9100	0.8869	0.8753	0.8562	0.8834	0.9035	0.9198
UTAAVE	0.9132	0.9418	0.9185	0.9076	0.9008	0.8974	0.9251	0.9119	0.9028	0.9202	0.9090	0.9106	0.8882	0.9076	0.9224	0.9348
UTACHEB	0.9040	0.9356	0.9098	0.8977	0.8901	0.8865	0.9181	0.9025	0.8913	0.9118	0.8986	0.9014	0.8776	0.8978	0.9135	0.9269
ACUTA	0.9060	0.9375	0.9119	0.8999	0.8923	0.8886	0.9184	0.9045	0.8952	0.9130	0.9015	0.9037	0.8808	0.9001	0.9151	0.9281
UTAROB	0.8690	0.9026	0.8735	0.8614	0.8549	0.8527	0.8919	0.8674	0.8478	0.8933	0.8629	0.8509	0.8255	0.8604	0.8853	0.9050
REPROC	0.9134	0.9420	0.9187	0.9078	0.9009	0.8975	0.9253	0.9120	0.9029	0.9204	0.9091	0.9107	0.8883	0.9078	0.9225	0.9349
MAXIMAX	0.8292	0.8852	0.8404	0.8185	0.8048	0.7969	0.8564	0.8270	0.8042	0.8814	0.8215	0.7846	0.7581	0.8184	0.8564	0.8838
MAXIMIN	0.8373	0.8892	0.8472	0.8274	0.8146	0.8079	0.8607	0.8348	0.8163	0.8824	0.8329	0.7965	0.7683	0.8283	0.8637	0.8887
MM-REGRET	0.8377	0.8893	0.8477	0.8279	0.8151	0.8084	0.8599	0.8351	0.8181	0.8785	0.8297	0.8049	0.7696	0.8290	0.8637	0.8884
EXPRANK	0.9126	0.9414	0.9178	0.9070	0.9002	0.8969	0.9245	0.9113	0.9022	0.9195	0.9084	0.9101	0.8874	0.9070	0.9218	0.9343
BESTRAI	0.8867	0.9262	0.8930	0.8786	0.8699	0.8658	0.9026	0.8849	0.8726	0.8974	0.8811	0.8815	0.8563	0.8792	0.8976	0.9137
WORSTRAI	0.8867	0.9262	0.8929	0.8786	0.8699	0.8659	0.9025	0.8847	0.8729	0.8976	0.8809	0.8816	0.8564	0.8791	0.8978	0.9135
AP1	0.8845	0.9229	0.8914	0.8766	0.8681	0.8635	0.9035	0.8829	0.8671	0.9039	0.8785	0.8711	0.8503	0.8768	0.8970	0.9139
AP2	0.8984	0.9297	0.9036	0.8920	0.8849	0.8816	0.9153	0.8969	0.8829	0.9160	0.8927	0.8864	0.8696	0.8917	0.9089	0.9233
DME1	0.8724	0.9174	0.8812	0.8636	0.8522	0.8477	0.9025	0.8719	0.8429	0.9030	0.8668	0.8475	0.8222	0.8647	0.8916	0.9111
DME2	0.8984	0.9284	0.9027	0.8920	0.8859	0.8831	0.9158	0.8971	0.8824	0.9140	0.8938	0.8875	0.8692	0.8918	0.9091	0.9236
MAXPOI	0.8921	0.9295	0.8979	0.8842	0.8762	0.8727	0.9080	0.8905	0.8778	0.9043	0.8860	0.8860	0.8629	0.8848	0.9028	0.9179
MINPOI	0.8921	0.9295	0.8981	0.8843	0.8761	0.8726	0.9082	0.8906	0.8777	0.9040	0.8862	0.8862	0.8631	0.8850	0.9026	0.9178
SUMPOI	0.9127	0.9414	0.9179	0.9070	0.9002	0.8969	0.9245	0.9113	0.9022	0.9195	0.9084	0.9101	0.8875	0.9070	0.9218	0.9343
RANK-SUM-IND	0.8961	0.9315	0.9021	0.8889	0.8808	0.8772	0.9102	0.8945	0.8837	0.9037	0.8909	0.8937	0.8656	0.8894	0.9072	0.9223
RANK-SUM	0.9093	0.9407	0.9157	0.9034	0.8953	0.8912	0.9212	0.9078	0.8987	0.9152	0.9051	0.9076	0.8824	0.9033	0.9191	0.9322
RANK-PROD	0.9091	0.9406	0.9154	0.9032	0.8952	0.8910	0.9211	0.9076	0.8985	0.9150	0.9049	0.9074	0.8822	0.9031	0.9189	0.9321
RANK-MM	0.9053	0.9391	0.9128	0.8993	0.8902	0.8852	0.9180	0.9039	0.8942	0.9117	0.9009	0.9035	0.8772	0.8992	0.9156	0.9294
REL-SUM	0.9135	0.9420	0.9188	0.9079	0.9011	0.8977	0.9254	0.9121	0.9030	0.9205	0.9093	0.9108	0.8885	0.9079	0.9226	0.9350
REL-PROD	0.9135	0.9420	0.9188	0.9079	0.9011	0.8977	0.9254	0.9121	0.9030	0.9205	0.9093	0.9108	0.8885	0.9079	0.9226	0.9350
REL-MM	0.9135	0.9420	0.9188	0.9079	0.9011	0.8977	0.9254	0.9121	0.9030	0.9205	0.9093	0.9108	0.8885	0.9079	0.9226	0.9350
REL-SUM-IND	0.9132	0.9418	0.9186	0.9076	0.9008	0.8975	0.9248	0.9120	0.9030	0.9200	0.9091	0.9107	0.8884	0.9077	0.9223	0.9346
REL-PROD-IND	0.9132	0.9418	0.9186	0.9076	0.9008	0.8975	0.9248	0.9120	0.9030	0.9200	0.9091	0.9107	0.8884	0.9077	0.9223	0.9346
REL-MM-IND	0.9132	0.9418	0.9185	0.9076	0.9008	0.8974	0.9248	0.9119	0.9029	0.9200	0.9090	0.9106	0.8883	0.9076	0.9223	0.9346
DOWN-DIST	0.8935	0.9233	0.8974	0.8869	0.8811	0.8789	0.9075	0.8919	0.8811	0.9018	0.8885	0.8903	0.8639	0.8868	0.9042	0.9191
UP-DIST	0.8936	0.9234	0.8974	0.8870	0.8812	0.8789	0.9076	0.8920	0.8812	0.9019	0.8885	0.8904	0.8641	0.8869	0.9042	0.9191

Table 12: Average First Rank Acceptability Indices ($FRAI$s) for different numbers of alternatives, criteria, characteristic points, and pairwise comparisons.

		Alternatives					Criteria			Characteristic points			Pairwise comparisons			
METHOD	AVG	6	8	10	12	14	3	4	5	2	3	4	4	6	8	10
UTAMP1	0.6181	0.7687	0.6574	0.5926	0.5489	0.5230	0.6629	0.6126	0.5789	0.6957	0.5951	0.5636	0.5078	0.5956	0.6584	0.7106
UTAMP2	0.6647	0.7963	0.6976	0.6421	0.6057	0.5819	0.7203	0.6583	0.6156	0.7308	0.6463	0.6172	0.5760	0.6450	0.6960	0.7419
UTAMSCVF	0.6011	0.7644	0.6472	0.5687	0.5257	0.4994	0.6666	0.5930	0.5436		0.5959	0.6063	0.5038	0.5774	0.6362	0.6870
UTAMSVF	0.5298	0.7214	0.5763	0.4971	0.4454	0.4086	0.5780	0.5222	0.4891	0.5817	0.5162	0.4914	0.4196	0.5023	0.5681	0.6290
UTAJLS	0.6884	0.8130	0.7213	0.6663	0.6329	0.6087	0.7390	0.6843	0.6420	0.7627	0.6782	0.6244	0.6109	0.6702	0.7162	0.7564
UTAAVE	0.7658	0.8631	0.7929	0.7497	0.7222	0.7009	0.7956	0.7613	0.7404	0.7863	0.7531	0.7579	0.7136	0.7525	0.7840	0.8129
UTACHEB	0.7437	0.8487	0.7725	0.7262	0.6967	0.6745	0.7791	0.7393	0.7127	0.7672	0.7277	0.7362	0.6911	0.7298	0.7619	0.7919
ACUTA	0.7485	0.8537	0.7764	0.7306	0.7015	0.6801	0.7796	0.7436	0.7221	0.7683	0.7348	0.7423	0.6979	0.7349	0.7659	0.7951
UTAROB	0.6512	0.7805	0.6799	0.6248	0.5932	0.5774	0.7068	0.6444	0.6023	0.7140	0.6295	0.6099	0.5678	0.6305	0.6818	0.7246
REPROC	0.7660	0.8635	0.7933	0.7500	0.7224	0.7005	0.7958	0.7615	0.7406	0.7868	0.7532	0.7579	0.7138	0.7527	0.7842	0.8132
MAXIMAX	0.5027	0.7156	0.5554	0.4652	0.4074	0.3696	0.5526	0.4981	0.4573	0.6317	0.4667	0.4096	0.3833	0.4714	0.5447	0.6111
MAXIMIN	0.6739	0.8099	0.7125	0.6505	0.6133	0.5833	0.7232	0.6677	0.6308	0.7115	0.6722	0.6381	0.6017	0.6568	0.6989	0.7382
MM-REGRET	0.6822	0.8185	0.7227	0.6605	0.6220	0.5872	0.7338	0.6750	0.6377	0.7396	0.6679	0.6390	0.6073	0.6640	0.7081	0.7492
EXPRANK	0.7617	0.8619	0.7898	0.7453	0.7175	0.6942	0.7919	0.7572	0.7361	0.7817	0.7491	0.7544	0.7084	0.7482	0.7805	0.8099
BESTRAI	0.7681	0.8638	0.7946	0.7521	0.7253	0.7046	0.7981	0.7636	0.7425	0.7893	0.7555	0.7594	0.7170	0.7549	0.7860	0.8145
WORSTRAI	0.6967	0.8198	0.7272	0.6764	0.6423	0.6176	0.7378	0.6896	0.6627	0.7209	0.6816	0.6874	0.6334	0.6791	0.7189	0.7553
AP1	0.7100	0.8263	0.7408	0.6892	0.6580	0.6357	0.7563	0.7047	0.6689	0.7467	0.6986	0.6847	0.6483	0.6931	0.7316	0.7671
AP2	0.7238	0.8316	0.7515	0.7039	0.6765	0.6554	0.7672	0.7198	0.6844	0.7751	0.7074	0.6889	0.6676	0.7086	0.7434	0.7757
DME1	0.6965	0.8157	0.7275	0.6750	0.6442	0.6200	0.7519	0.6917	0.6458	0.7469	0.6853	0.6572	0.6222	0.6777	0.7229	0.7630
DME2	0.7201	0.8281	0.7469	0.6997	0.6729	0.6528	0.7666	0.7155	0.6782	0.7624	0.7070	0.6908	0.6611	0.7039	0.7411	0.7743
MAXPOI	0.7127	0.8321	0.7422	0.6918	0.6594	0.6380	0.7569	0.7071	0.6741	0.7451	0.6962	0.6969	0.6524	0.6953	0.7340	0.7691
MINPOI	0.7664	0.8635	0.7935	0.7504	0.7229	0.7014	0.7962	0.7619	0.7410	0.7870	0.7538	0.7583	0.7143	0.7532	0.7845	0.8135
SUMPOI	0.7618	0.8619	0.7898	0.7453	0.7175	0.6942	0.7919	0.7572	0.7361	0.7817	0.7491	0.7544	0.7084	0.7482	0.7805	0.8100
RANK-SUM-IND	0.7556	0.8603	0.7863	0.7393	0.7085	0.6834	0.7874	0.7509	0.7284	0.7771	0.7419	0.7477	0.6992	0.7414	0.7753	0.8064
RANK-SUM	0.7659	0.8627	0.7926	0.7497	0.7228	0.7019	0.7960	0.7615	0.7403	0.7870	0.7533	0.7576	0.7143	0.7526	0.7841	0.8128
RANK-PROD	0.7647	0.8621	0.7916	0.7485	0.7212	0.7001	0.7949	0.7602	0.7390	0.7857	0.7520	0.7565	0.7127	0.7512	0.7830	0.8120
RANK-MM	0.7345	0.8572	0.7772	0.7194	0.6757	0.6427	0.7713	0.7294	0.7027	0.7606	0.7196	0.7232	0.6731	0.7187	0.7561	0.7900
REL-SUM	0.7664	0.8635	0.7935	0.7504	0.7229	0.7014	0.7962	0.7619	0.7410	0.7870	0.7538	0.7583	0.7143	0.7532	0.7845	0.8135
REL-PROD	0.7664	0.8635	0.7935	0.7504	0.7229	0.7014	0.7961	0.7619	0.7410	0.7870	0.7538	0.7583	0.7143	0.7532	0.7845	0.8135
REL-MM	0.7663	0.8635	0.7935	0.7504	0.7229	0.7014	0.7961	0.7619	0.7410	0.7870	0.7538	0.7583	0.7143	0.7531	0.7845	0.8135
REL-SUM-IND	0.7660	0.8633	0.7933	0.7501	0.7226	0.7009	0.7955	0.7617	0.7409	0.7864	0.7535	0.7582	0.7142	0.7529	0.7842	0.8129
REL-PROD-IND	0.7660	0.8633	0.7933	0.7501	0.7226	0.7009	0.7955	0.7617	0.7409	0.7864	0.7535	0.7582	0.7142	0.7529	0.7842	0.8129
REL-MM-IND	0.7660	0.8633	0.7933	0.7500	0.7224	0.7008	0.7954	0.7616	0.7409	0.7863	0.7534	0.7581	0.7140	0.7528	0.7842	0.8128
DOWN-DIST	0.7546	0.8569	0.7837	0.7375	0.7092	0.6859	0.7847	0.7498	0.7293	0.7739	0.7421	0.7479	0.6999	0.7405	0.7739	0.8042
UP-DIST	0.7579	0.8582	0.7859	0.7413	0.7132	0.6907	0.7887	0.7532	0.7317	0.7791	0.7448	0.7497	0.7042	0.7441	0.7767	0.8065

Publication [P2]

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*, vol. 278, p. 110871, 2023

DOI: 10.1016/j.knosys.2023.110871.

Number of citations[2]:

- according to Web of Science: 2
- according to Google Scholar: 5

[2]as of September 23, 2024

Knowledge-Based Systems 278 (2023) 110871

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Selection of a representative sorting model in a preference disaggregation setting: A review of existing procedures, new proposals, and experimental comparison

Michał Wójcik, Miłosz Kadziński *, Krzysztof Ciomek

Poznan University of Technology, Faculty of Computing and Telecommunications, Piotrowo 2, 60-965 Poznań, Poland

ARTICLE INFO

Article history:
Received 4 February 2023
Received in revised form 27 July 2023
Accepted 28 July 2023
Available online 1 August 2023

Keywords:
Multiple criteria decision aiding
Preference disaggregation
Sorting
Representative model
Robustness analysis

ABSTRACT

We consider preference disaggregation in the context of multiple criteria sorting. The value function parameters and thresholds separating the classes are inferred from the Decision Maker's (DM's) assignment examples. Given the multiplicity of sorting models compatible with indirect preferences, selecting a single, representative one can be conducted differently. We review several procedures for this purpose, aiming to identify the most discriminant, average, central, parsimonious, or robust models. Also, we present three novel procedures that implement the robust assignment rule in practice. They exploit stochastic acceptabilities and maximize the support given to the resulting assignments by all feasible sorting models. The performance of fourteen procedures is verified on problem instances with different complexities. The results of an experimental study indicate the most efficient procedures in terms of classification accuracy, reproducing the DM's model, and delivering the most robust assignments. These include approaches identifying differently interpreted centers of the feasible polyhedron and robust methods introduced in this paper. Moreover, we discuss how the performance of all procedures is affected by different numbers of classes, criteria, characteristic points, and reference assignments.

1. Introduction

In multiple criteria sorting problems, alternatives need to be assigned to preference-ordered classes [1]. Each of them is predefined and associated with a precise semantic, implying the same subsequent treatment of alternatives placed in a given category. The presence of multiple, potentially conflicting criteria makes such ordinal classification problems challenging. For this reason, the field of Multiple Criteria Decision Aiding (MCDA) offers a variety of methods that support the Decision Makers (DMs) in carrying forward the solution process (see, e.g., [2]). They are helpful in problem structuring, preference elicitation, construction and exploitation of the preference model, and explaining the recommended assignments [3]. In recent years, the approaches adopting a preference disaggregation perspective have been prevailing [4]. They construct a sorting model using a regression-like scheme based on the DM's decision examples. Such approaches facilitate the solution process by lowering the cognitive effort on the part of DMs and not requiring specialized knowledge required when directly specifying values of decision model parameters.

The most popular preference disaggregation sorting method is UTADIS [5]. It accepts indirect preference information in the form of assignment examples, specifying the desired classification for a subset of reference alternatives [6]. Such holistic statements are translated into compatible parameters of an additive value function and thresholds separating the classes on a scale of a comprehensive value [7]. UTADIS has been appreciated in the MCDA community for using an intuitive sorting procedure with highly interpretable alternatives' scores and class thresholds, while at the same time being free of statistic hypotheses and restrictions [8,9]. Also, it handles both qualitative and quantitative criteria, differentiates between inter- and intra-criteria attractiveness, and provides means for interaction with the DMs who might review the model by changing or enriching their preferences [10]. Such appealing features have motivated the practical use of UTADIS for solving real-world decision problems concerning, e.g., credit risk assessment [11], supplier classification [12], sorting activities in civil construction [13], and adoption of green chemistry principles in nanotechnology [10].

The basic variant of UTADIS has been extended in numerous ways. In particular, it was generalized to an example-based procedure where the classes are delimited implicitly by decision

* Correspondence to: Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland.
E-mail addresses: michal.wojcik@cs.put.poznan.pl (M. Wójcik), milosz.kadzinski@cs.put.poznan.pl (M. Kadziński), k.ciomek@gmail.com (K. Ciomek).

https://doi.org/10.1016/j.knosys.2023.110871

examples rather than class thresholds [14]. Furthermore, a sequential classification technique, called M.H.DIS, was introduced in [15] to consider the assignment of not yet classified alternatives to the most preferred class in a stepwise fashion. Moreover, UTADIS was advanced to a robustness analysis framework, where a multiplicity of compatible sorting models are exploited to verify the stability of classification. In Robust Ordinal Regression (ROR), all such models are translated into the necessary and possible results using mathematical programming [14,16]. In turn, in Stochastic Ordinal Regression (SOR) – the Monte Carlo simulations derive a large, representative set of such models whose results are summarized in the form of stochastic acceptabilities [17]. Then, in Bayesian Ordinal Regression (BOR), a posterior distribution over a set of all potential sorting models is derived to emphasize the differences in the models' abilities to reconstruct the DM's assignment examples [18]. Also, many works proposed dedicated techniques for dealing with the inconsistency of assignment examples. They aim at restoring the consistency [19], minimize a misclassification error [7], use preference models compatible with different preferential reducts [10], or incorporate contingent, inter-related models that altogether reconstruct the holistic preference information [20,21]. In the same spirit, some optimization techniques were devised for handling large sets of assignment examples [22–24].

Further methodological advancements have been devoted to supporting preference elicitation, tolerating uncertain performances, enriching incorporated models, and addressing various structures and types of handled decision problems. In [25], one proposed active learning strategies that minimize the number of assignment examples needed for arriving at a sufficiently robust recommendation. Also, [26] tolerated hesitancy regarding assignment examples and the performance of alternatives using probability linguistic term sets. Moreover, [27] introduced a unified framework handling preference information in the form of assignment-based pairwise comparisons and constraints on the category sizes [28,29] along with assignment examples, whereas [22] accounted for valued desired classifications and [18] handled potentially uncertain assignments. An additive value function used in UTADIS was extended to admit interactions between criteria [30], non-monotonicity [23,31–33] or polynomial character [34] of marginal value functions. Furthermore, [35] adapted the method to a hierarchical structure of criteria, whereas [33,36] considered a multi-decision classification problem with many inter-related decision attributes. Finally, dedicated group decision methods were devised for handling preferences of multiple DMs and either arriving at a collective recommendation [37,38] or investigating the spaces of consensus and disagreement observed in the group [39].

Various real-world applications and methodological developments confirm the status of UTADIS as one of the essential methods in MCDA. This paper deals with procedures for selecting a single instance of the threshold-based value-driven sorting model. Since the polyhedron of all functions and thresholds compatible with the stated indirect preference information can be quite large [14,16], such a selection can be performed differently. Whichever the choice or construction procedure, exhibiting a single representative sorting model allows the DM to analyze the shapes of marginal functions, the trade-offs between criteria, the dispersion of class thresholds, the comprehensive values of individual alternatives, and margins of safety in the recommended univocal assignments [40]. A single model is, therefore, a synthetic and intuitive solution to the sorting problem, supporting the validity of the derived recommendation or motivating reactions from DMs.

We contribute to the literature in a three-fold way. First, we review different concepts underlying the selection of a representative sorting model in the context of UTADIS. The primary idea consists of choosing the most discriminant model in terms of the differences between comprehensive values of reference alternatives assigned to different classes and/or marginal values associated with consecutive characteristic points of per-criterion functions [41]. Furthermore, we discuss the concept of controlling the slope of marginal functions [42]. Another methodological stream is oriented toward identifying a central model with the proviso that the concept of centrality is interpreted in various ways [43,44]. Moreover, we refer to the mean models obtained by averaging either the extreme models compatible with the DM's preferences [45] or a large sample of uniformly distributed ones [10]. The last postulate builds on the outcomes of robustness analysis by making use of necessary, possible [40] or stochastic results [17] to define the targets that should be emphasized in the representative case.

Our second contribution consists of proposing novel procedures for selecting a single sorting model representative in the sense of robustness preoccupation. Specifically, we refer to the outcomes of Stochastic Ordinal Regression in the form of Class Acceptability Indices (CAIs) and Assignment-based Pairwise Outranking Indices (APOIs) [17]. They quantify the shares of compatible sorting models, confirming a given alternative's assignment to a particular class or supporting one alternative being assigned to a class that is at least as good as another. The representative model emphasizes the most frequent classifications of all alternatives, the most common assignment-based preference relations for all pairs of alternatives, or both of these objectives at once. Similar to [40], we refer to the "one for all, all for one" motto by representing all compatible sorting models, which contribute to the definition of a representative one. However, we build on more informative and detailed outcomes in the form of stochastic acceptabilities [46] rather than the possible and necessary assignments that need to be confirmed by at least one or all compatible models, respectively. We illustrate all procedures, including the existing and newly introduced ones, on a single decision problem to clarify their operational steps.

The third contribution consists of a thorough experimental evaluation of the fourteen discussed procedures. The problem of choosing the "best" sorting model in the preference disaggregation methods is ill-defined. However, one can consider some objective criteria for the meaningful comparison of various procedures. In particular, we account for five measures that make sense in the context of both using incomplete preference information concerning a subset of reference alternatives and the multiplicity of sorting models compatible with the DM's assignments examples. They concern (i) the ability to reconstruct reference classification for all alternatives, (ii) the robustness of derived assignments in terms of the support they are given by all compatible models, and (iii) the capability of restoring the preference model in terms of per-criterion marginal value functions, alternatives' comprehensive values, and class thresholds. The experiment involves problems with different numbers of classes, criteria, characteristic points of marginal functions, and reference alternatives assigned by the DM to each class. We discuss the average results attained over all considered settings and the trends observed with the increasing model's complexity and availability of preference information.

Our study can be seen as a significant extension of the experiments discussed in [44], where only four procedures have been collated in a similar context. Compared to [44], we include additional methods that account for the shape of Marginal Value Functions (MVFs), determine an average model, or emphasize the robustness of recommendations obtained with a set of all feasible models in six different ways. Also, we consider a more comprehensive set of parameter values characterizing decision problems. In this way, we provide richer insights into the performance

trends. Finally, we refer to a more extensive set of measures capturing the capability of restoring the preference model.

The paper's remainder is organized in the following. Section 2 reminds UTADIS and its robust counterparts. In Section 3, we discuss various procedures for selecting a representative sorting model. The eAppendix (supplementary material available online) illustrates their use in a didactic example. In Section 4, we present the results of an extensive experimental comparison of different approaches. The last section concludes the paper.

2. Reminder on UTADIS and robustness analysis

The following notation is used in the paper:

- $A = \{a_1, a_2, \ldots, a_i, \ldots, a_n\}$ – a finite set of n alternatives; each of them is evaluated in terms of m criteria;
- $A^R = \{a_1^*, a_2^*, \ldots, a_r^*\}$ – a finite set of r reference alternatives; $A^R \subseteq A$;
- $G = \{g_1, g_2, \ldots, g_j, \ldots, g_m\}$ – a finite set of m evaluation criteria, $g_j : A \to \mathbb{R}$ for all $j \in J = \{1, \ldots, m\}$; without loss of generality, we assume that all of the criteria in G are of gain type;
- $X_j = \{g_j(a_i), a_i \in A\}$ – a finite set of performances of all alternatives in A on criterion g_j;
- $x_j^1, x_j^2, \ldots, x_j^{n_j(A)}$ – the ordered values of X_j, $x_j^{k-1} < x_j^k$, $k = 2, \ldots, n_j(A)$, where $n_j(A) = |X_j|$ and $n_j(A) \leq n$; thus, $X = \prod_{j=1}^{m} X_j$ is the performance space; note that X_j can also be enriched with the extreme values of the performance scale that are not attained by any alternative;
- $C_1, C_2, \ldots, C_p$ – p pre-defined and preference-ordered classes so that C_l is preferred to C_{l-1} for $l = 2, \ldots, p$.

To compute the desirability of each alternative $a \in A$, UTADIS [5] considers an Additive Value Function (AVF) [47]:

$$U(a) = \sum_{j=1}^{m} u_j(g_j(a)), \forall a \in A, \tag{1}$$

where u_j, $j = 1, \ldots, m$, are MVFs being piece-wise linear monotonic and defined by a pre-defined number γ_j of equally distributed characteristic points $\beta_j^1, \beta_j^2, \ldots, \beta_j^{\gamma_j}$, with the extreme points set to $\beta_j^1 = x_j^1$ and $\beta_j^{\gamma_j} = x_j^{n_j(A)}$, and:

$$\beta_j^s = x_j^1 + (x_j^{n_j(A)} - x_j^1)\frac{s-1}{\gamma_j - 1}, \; j = 1, \ldots, m, s = 1, \ldots, \gamma_j. \tag{2}$$

A comprehensive value is normalized in the $[0, 1]$ range by assuming that $u_j(\beta_j^1) = 0$, for $j = 1, \ldots, m$, and $\sum_{j=1}^{m} u_j(\beta_j^{\gamma_j}) = 1$. To enable control over the difference between marginal values assigned to the subsequent characteristic points, we consider variable $\rho \geq 0$ introduced as follows:

$$u_j(\beta_j^s) - u_j(\beta_j^{s-1}) \geq \rho, \; j = 1, \ldots, m, s = 2, \ldots, \gamma_j. \tag{3}$$

In the basic setting, ρ is set to 0, which implies fulfillment of the weak monotonicity constraints. The marginal value for performance $x_j^k \in [\beta_j^s, \beta_j^{s+1}]$ can be computed using a linear interpolation:

$$u_j(x_j^k) = u_j(\beta_j^s) + (u_j(\beta_j^{s+1}) - u_j(\beta_j^s))\frac{x_j^k - \beta_j^s}{\beta_j^{s+1} - \beta_j^s}, \quad j = 1, \ldots, m, \; k = 1, \ldots, n_j(A). \tag{4}$$

UTADIS incorporates a threshold-based sorting procedure, where each class C_l is delimited by the lower t_{l-1} and upper t_l thresholds defined on a scale of a comprehensive value U. For simplicity, we do not consider the lower threshold of the least preferred class C_1 and the upper threshold of the most preferred class C_p, which could be arbitrarily fixed to $t_0 = 0$ and $t_p > 1$. Hence, to derive the assignment for alternative $a \in A$, $U(a)$ is compared with a vector of $p-1$ thresholds $t = [t_1, \ldots, t_{l-1}, t_l, \ldots, t_{p-1}]$ such that $t_1 \geq \varepsilon$, $t_l - t_{l-1} \geq \varepsilon$, for $l = 2, \ldots, p-1$, and $t_{p-1} + \varepsilon \leq 1$, where ε is an arbitrarily small positive value. Due to the threshold-based division of the scale of possible comprehensive values U into disjoint class intervals, each value and, thus, each alternative is assigned to exactly one class in a given model.

In UTADIS, the parameters of an assumed sorting model are inferred from the DM's indirect preference information. It consists of the desired class assignments for reference alternatives in A^R. Assigning an alternative a to class C_l can be written as $a \to^{DM} C_l$, where $l \in \{1, \ldots, p\}$. It can also be defined with function I, which indicates to which of the p classes a given alternative is assigned:

$$\forall a^* \in A^R, \; a^* \to^{DM} C_l : \; I_{DM}(a^*) = l. \tag{5}$$

In this paper, we consider only precise assignment examples. They are reproduced via preference disaggregation that ensures a comprehensive value of each reference alternative $a^* \in A^R$ is within the range $[t_{l-1}, t_l)$ delimited by the lower and upper thresholds corresponding to its desired class C_l, i.e.:

$$\forall a^* \in A^R : \; I_{DM}(a^*) = l \in \{1, \ldots, p-1\} \implies t_l - U(a^*) \geq \delta + \varepsilon, \tag{6}$$

$$\forall a^* \in A^R : \; I_{DM}(a^*) = l \in \{2, \ldots, p\} \implies U(a^*) - t_{l-1} \geq \delta, \tag{7}$$

where $\delta \geq 0$. Overall, a set $\mathcal{U}^R$ of compatible AVFs and class thresholds is defined by the following set E^{A^R} of linear constraints:

$$\left.\begin{array}{l} \left.\begin{array}{l} u_j(\beta_j^1) = 0, j = 1, \ldots, m, \\ \sum_{j=1}^{m} u_j(\beta_j^{\gamma_j}) = 1, \end{array}\right\} (E^N) \\ u_j(\beta_j^s) - u_j(\beta_j^{s-1}) \geq \rho, \; j = 1, \ldots, m, s = 2, \ldots, \gamma_j, \\ \left.\begin{array}{l} t_1 \geq \varepsilon, \\ t_l - t_{l-1} \geq \varepsilon, \; l = 2, \ldots, p-1, \\ t_{p-1} + \varepsilon \leq 1, \end{array}\right\} (E^T) \\ \forall a^* \in A^R : \; I_{DM}(a^*) = l \in \{1, \ldots, p-1\} \implies t_l - U(a^*) \geq \delta + \varepsilon, \\ \forall a^* \in A^R : \; I_{DM}(a^*) = l \in \{2, \ldots, p\} \implies U(a^*) - t_{l-1} \geq \delta, \end{array}\right\} (E^{A^R}) \tag{8}$$

where $\rho \geq 0$, $\delta \geq 0$, and ε is a small positive constant that transforms non-strict inequalities into strict ones (in our implementation, we set its value to 10^{-6}). Please note that constraints ensuring the monotonicity of the thresholds ($t_l - t_{l-1} \geq \varepsilon$) are redundant when the DM assigns at least one alternative to each class. However, we keep them for the clarity of presentation and comprehensiveness of the model under all scenarios.

Since the DM's preference information is incomplete, when E^{A^R} is feasible, $\mathcal{U}^R$ typically consists of infinitely many sorting models. To choose one of them, one needs to optimize an objective function. In Section 3, we discuss fourteen procedures that differ mainly with respect to considering various objectives and/or incorporating additional variables and constraints. Some methods optimize ρ and/or δ, which are then treated as variables. In general, ρ and δ allow for controlling the difference between marginal values assigned to the successive characteristic points and the distances of alternatives' comprehensive values from the class limits, respectively. If ρ or δ are not optimized, they are

treated as constants and set to zero. In this way, we ensure non-strict monotonicity of marginal value functions or reproduction of the assignment examples.

In what follows, we discuss the approaches for robustness analysis, whose results will be exploited by some procedures. ROR verifies the possibility or necessity of certain relationships based on a set of all compatible sorting models. This requires checking the consistency of the basic constraint set E^{A^R} with additional constraints representing a verified hypothesis. For this purpose, the relations between the pairs of alternatives in individual models were specified. One of them is the weak assignment-based preference $\succsim^{\rightarrow}$, defined as follows:

$$\forall a, b \in A \text{ and } U \in \mathcal{U}^R \ : \ a \succsim_U^{\rightarrow} b \iff I_U(a) \geq I_U(b), \tag{9}$$

where $I_U(a)$ is the index of class to which a is assigned by function U. When $I_U(a) = l$, then $a \rightarrow^U C_l$. Hence $a \succsim_U^{\rightarrow} b$ means that, according to U, the class of a is at least as good as the class of b. If $\succsim_U^{\rightarrow}$ is confirmed for all compatible sorting models, then this relation is *necessary* according to ROR, i.e.:

$$\forall a, b \in A : \ a \succsim^{\rightarrow,N} b \iff \forall U \in \mathcal{U}^R : \ a \succsim_U^{\rightarrow} b. \tag{10}$$

Note that if $\neg(a \succsim^{\rightarrow,N} b)$, there exists at least one compatible sorting model in $\mathcal{U}^R$ that assigns b to a more preferred class than a. Relation $\sim_U^{\rightarrow}$ for a single model U along with its robust counterpart $\sim^{\rightarrow,N}$ based on a set of models $\mathcal{U}^R$ can be defined analogously by checking whether one alternative is assigned to the same class as another ($I_U(a) = I_U(b)$). In the same spirit, relations $\succ_U^{\rightarrow}$ and $\succ^{\rightarrow,N}$ reflect the assignment to a strictly more preferred class ($I_U(a) > I_U(b)$). The truth of these relations is verified using linear programming [17,40].

In SOR, $\mathcal{U}^R$ is exploited with the Monte Carlo simulations to derive a set $S \subseteq \mathcal{U}^R$ of uniformly distributed compatible sorting models. Specifically, we sample from a convex polyhedron defined by constraint set E^{A^R} with $\delta = \rho = 0$ and ε set to a small positive value (in our case, 10^{-6}). In practice, $S \subset \mathcal{U}^R$ and $|S| \ll |\mathcal{U}^R|$. The results obtained for these models are summarized in the form of four stochastic acceptabilities: Class Acceptability Indices (*CAI*s) and Assignment-based Pair-wise Winning (*APWI*), Outranking (*APOI*) and Equality (*APEI*) Indices. $CAI \in [0, 1]$ quantifies the share of compatible sorting models assigning $a \in A$ to class C_l. Its approximation CAI' is defined as follows, i.e.:

$$\forall a \in A \ \forall l \in \{1, \ldots, p\} : \ CAI'(a, C_l) = \frac{|\{U \in S : I_U(a) = l\}|}{|S|}. \tag{11}$$

Furthermore, $APWI : A \times A \rightarrow [0, 1]$ is defined as the share of all models in $\mathcal{U}^R$, which classify one alternative into a more preferred class than another alternative. Its approximation $APWI'$ is computed in the following way:

$$\forall a, b \in A : \ APWI'(a, b) = \frac{|\{U \in S : I_U(a) > I_U(b)\}|}{|S|}. \tag{12}$$

The remaining pairwise indices, i.e., *APOI*s and *APEI*s, are defined analogously by referring to the shares of models confirming that one alternative is assigned to a class, respectively, at least as good ($I_U(a) \geq I_U(b)$) or the same ($I_U(a) = I_U(b)$) as another. In this paper, we sample from set $\mathcal{U}^R$ using the Hit-And-Run (HAR) algorithm [48] implemented in [49]. Note that even if the stochastic acceptability indices are defined in the range between 0 and 1, for clarity of presentation, they can also be expressed in percentages between 0% and 100%.

3. Procedures for selecting a representative sorting model

In this section, we review different concepts underlying the selection of a representative sorting model in the context of UTADIS. Their most distinctive feature is optimizing a unique objective function subject to the constraint set E^{A^R} that defines a set of all compatible value functions and class thresholds. Some procedures focus only on selecting a value function. In this case, the thresholds are set in equal distances between extreme comprehensive values of reference alternatives assigned to each class. This is consistent with selecting the most discriminating function, emphasizing the derived assignments as sharply as possible given a value function model.

The three methods introduced in this paper are distinguished as CAI, APOI, and COMB. They exploit stochastic acceptability indices of class assignments and assignment-based pairwise relations. Moreover, UTADISMP3 and MSCVF adjust to the sorting context the procedures that have been so far used for ranking problems [50]. The remaining nine methods have been proposed before in the preference disaggregation literature, and a suitable reference is provided for each of them.

3.1. The most discriminant models

Let us start with the max–min formulations that seek the most discriminant model parameters. In the context of multiple criteria ranking, this idea was first implemented in UTAMP1 [41]. When it comes to sorting, **UTADISMP1** [40,44] postulates maximizing the minimal difference between comprehensive values of reference alternatives and their respective class thresholds, i.e.:

$$\text{Maximize } \delta, \ \text{s.t. } E^{A^R}. \tag{13}$$

In this way, the gap between all consecutive classes is maximized, yielding a model that is away from the boundaries of the polyhedron of all compatible sorting models [43]. As a result, the DM's assignment examples are reproduced in a bold and robust way [44].

Another procedure is motivated by the ranking method, called UTAMP2 [41]. Apart from optimizing δ, i.e., the distances between the comprehensive values and class thresholds, it maximizes the difference between marginal values assigned to all pairs of consecutive characteristic points. The problem solved by **UTADISMP2** is the following:

$$\text{Maximize } \delta + \rho, \ \text{s.t. } E^{A^R}, \tag{14}$$

where $\rho \geq 0$. The method has similar features to UTADISMP1, while favoring steeper linear components of marginal value functions. This prevents weakly monotonic functions with level parts or even neglecting some criteria whose marginal functions take zero values for all performances. The ρ component is considered alone in **UTADISMP3**, highlighting the differences in values of marginal functions even more:

$$\text{Maximize } \rho, \ \text{s.t. } E^{A^R}. \tag{15}$$

3.2. Parsimonious decision model

UTADISMP3 impacts the shape of marginal value functions by desiring the most discriminant ones. In turn, [42] postulated selecting as linear MVFs as possible, i.e., functions minimally deviating from the linearity. The model corresponding to this idea is called a Minimal Slope Change Value Function, in short, **MSCVF**. It can be obtained by solving the following Linear Programming

(LP) model:

$$\text{Minimize } \phi,$$

$$s.t. \quad \left.\begin{array}{l} E^{A^R} \\ \frac{u_j(\beta_j^k)-u_j(\beta_j^{k-1})}{\beta_j^k-\beta_j^{k-1}} \\ -\frac{u_j(\beta_j^{k-1})-u_j(\beta_j^{k-2})}{\beta_j^{k-1}-\beta_j^{k-2}} \le \phi \\ \frac{u_j(\beta_j^{k-1})-u_j(\beta_j^{k-2})}{\beta_j^{k-1}-\beta_j^{k-2}} \\ -\frac{u_j(\beta_j^k)-u_j(\beta_j^{k-1})}{\beta_j^k-\beta_j^{k-1}} \le \phi \end{array}\right\} \begin{array}{l} \text{for } j=1,\ldots,m, \\ k=3,\ldots,\gamma^j \end{array} \Bigg\} (E^{A^R}_{MSCVF}) \tag{16}$$

Note that using MSCVF makes sense when at least three characteristic points are considered on a given criterion. The above idea can be interpreted as favoring a parsimonious decision model consistent with the Occam razor principle. It says that "entities must not be multiplied beyond necessity", which can be intuitively interpreted as: "the simplest explanation is most likely the correct one" [51].

Both UTADISMP3 and MSCVF aim for specific shapes of marginal value functions. The former tries to maximize the difference between consecutive points, while functions are as linear as possible in the latter. However, it is not always possible to achieve this goal due to other constraints. In that case, the procedure looks for the closest acceptable solution. Therefore, these methods can be considered flexible and adaptable to the existing conditions.

3.3. Average models

Another appealing idea consists of conducting a post-optimality analysis, deriving a set of representative sorting models, and averaging them to form an approximation of the polyhedron's centroid model [44]. It has been implemented in two different ways.

UTADIS-JLS was motivated by the system of $2m$ extreme solutions originally considered in the context of ranking problems [45]. Each of them is obtained by minimizing or maximizing the greatest value attained by MVF for each criterion, i.e.:

$$\text{Maximize / Minimize } u_j(\beta_j^{\gamma_j}), \text{ s.t. } E^{A^R}. \tag{17}$$

Note that $u_j(\beta_j^{\gamma_j})$ can be interpreted as a weight or a trade-off constant of criterion g_j.

A disadvantage of UTADIS-JLS consists of accounting only for the extreme models. In [10], the concept of finding an "average" model was generalized by considering a large sample $S = \{U^1, U^2, \ldots, U^{|S|}\}$ of models considered in SOR. The **CENTROID** procedure is not based on optimization. It derives an average of all samples that can be considered a stochastic approximation of the central solution. This applies to both characteristic points of MVFs and class threshold values:

$$\begin{aligned} t_l &= \frac{1}{|S|}\sum_{i=1}^{|S|} t_l^i, \ l=1,\ldots,p-1, \\ u_j(\beta_j^s) &= \frac{1}{|S|}\sum_{i=1}^{|S|} u_j^i(\beta_j^s), \ j=1,\ldots,m, s=1,\ldots,\gamma^j. \end{aligned} \tag{18}$$

It is worth noting that since the space of possible solutions ($\mathcal{U}^R$) is convex, the average model also satisfies all constraints [14]. In [52], it is called a barycenter solution. Such average models are claimed to be more robust and less vulnerable to changes in the DM's assignment examples [44]. Moreover, procedures based on an analogous idea exhibit favorable performance in the context of multiple criteria ranking [50,52].

3.4. Central models

Opting for an average model can be seen as a particular implementation of selecting a central model. However, the concept of centrality can be interpreted in different ways, two of which – denoted **CHEBYSHEV** and **ACUTADIS** – are discussed in this subsection. The Chebyshev center of a polyhedron is a midpoint of the largest Euclidean ball that fits in a polyhedron. A model corresponding to such a center was proposed in [44]. To determine it, one needs to maximize variable r that is inscribed in each monotonicity and assignment-based constraint:

$$\text{Maximize } r,$$

$$s.t. \quad \left.\begin{array}{l} E^N,\ E^T, \\ u_j(\beta_j^s) - u_j(\beta_j^{s-1}) - \sqrt{2}r \ge 0, \\ \quad j=1,\ldots,m, s=2,\ldots,\gamma_j, \\ \forall a_i^* \in A^R:\ I_{DM}(a_i^*) = l \in \{1,\ldots,p-1\} \\ \quad \Longrightarrow\ t_l - U(a_i^*) - b_i r \ge \varepsilon, \\ \forall a_i^* \in A^R:\ I_{DM}(a_i^*) = l \in \{2,\ldots,p\} \\ \quad \Longrightarrow\ U(a_i^*) - t_{l-1} - c_i r \ge 0, \end{array}\right\} (E^{A^R}_{CC}) \tag{19}$$

where $\sqrt{2}$, b_i, and c_i are the Euclidean norms of the decision variables' coefficients (except r) in constraint in which they occur [44]. For example, consider the inequality that ensures the monotonicity of the MVFs at the consecutive characteristic points. In each of these inequalities, only $u_j(\beta_j^s)$ and $u_j(\beta_j^{s-1})$ have non-zero coefficients of 1 and -1, respectively. Thus, $\|(0,\ldots,0,1,-1,0,\ldots,0)\|_2 = \sqrt{1^2+(-1)^2} = \sqrt{2}$. The values of b_i and c_i in the remaining inequalities can be determined analogously. Let us emphasize that since these norms are constants, the constraints remain linear. Such a solution can be considered central because it is equally distant from all essential inequality constraints.

ACUTADIS postulates selecting an analytic center rather than the Chebyshev one. It was initially proposed for ranking problems and adjusted to the scope of sorting in [40]. It corresponds to the model maximizing the logarithmic barrier function of the slacks $(d_{i^-}, d_{i^+}, d_{js})$ involved in the essential constraints of E^{A^R} [44]:

$$\text{Maximize } \sum_{a_i^* \in A^R} (\log d_{i^-} + \log d_{i^+}) + \sum_{j=1}^{m}\sum_{s=2}^{\gamma_j} \log d_{js},$$

$$s.t. \quad \left.\begin{array}{l} E^N,\ E^T, \\ u_j(\beta_j^s) - u_j(\beta_j^{s-1}) = d_{js},\ j=1,\ldots,m, s=2,\ldots,\gamma_j, \\ \forall a_i^* \in A^R: I_{DM}(a_i^*) = l \in \{1,\ldots,p-1\} \\ \quad \Longrightarrow\ t_l - U(a^*) - d_{i^-} = \varepsilon, \\ \forall a_i^* \in A^R: I_{DM}(a_i^*) = l \in \{2,\ldots,p\} \\ \quad \Longrightarrow\ U(a_i^*) - t_{l-1} - d_{i^+} = 0. \end{array}\right\} (E^{A^R}_{AC}) \tag{20}$$

The above non-linear problem can be solved using Newton's method [43], always leading to a unique solution, if any exists.

3.5. Robust models based on exact outcomes

The methods for robustness analysis were developed to exploit a set of all compatible models [14,16]. The derived outcomes reflect the stability of the sorting recommendation. However, their use for real-world decision aiding indicated that it is not easy for some users to comprehend such robust results and an abstract concept of infinitely many compatible models. This motivated the development of procedures for selecting a representative

Maximize ι,

$$s.t. \quad \left.\begin{array}{l} E^{A^R}, \\ U(a) - U(b) - \iota \geq U(c) - U(d), \\ U(a) - U(b) - \iota \geq U(d) - U(c), \end{array}\right\} \forall a, b, c, d \in A: \ (a \succsim^{\rightarrow,N} b) \wedge \neg(b \succsim^{\rightarrow,N} a) \wedge (c \sim^{\rightarrow,N} d). \left.\right\} (E^{A^R}_{R-compromise}) \quad (23)$$

Box I.

sorting model that can be exhibited to the DMs. The primary idea consisted of representing all compatible sorting models that contribute to the definition of a representative one. In this way, one does not lose the advantage of knowing all compatible ones while gaining a model instance that can be used to analyze the impact of different criteria, separation of decision classes, and robustness in the sense of distances of alternatives' values from class thresholds.

In [40], two objectives were defined to emphasize the robustness concerns. They are based on exact robust outcomes computed with mathematical programming. On the one hand, for all pairs of alternatives such that one of them is assigned to a class at least as good as another for all feasible models and for at least one of them it is assigned to a class strictly better, the difference between their comprehensive values should be maximized:

Maximize ω,

$$s.t. \quad \left.\begin{array}{l} E^{A^R}, \\ U(a) - U(b) \geq \omega \\ \forall a, b \in A: \ (a \succsim^{\rightarrow,N} b) \wedge \neg(b \succsim^{\rightarrow,N} a). \end{array}\right\} (E^{A^R}_{R-iterative_I}) \quad (21)$$

On the other hand, the value difference should be minimized for all pairs of alternatives necessarily assigned to the same class. This can be conducted while respecting the optimization of the previous target (i.e., setting $\omega = \omega^*$):

Minimize λ,

$$s.t. \quad \left.\begin{array}{ll} E^{A^R}_{R-iterative_I}, & \\ \omega = \omega^*, & \\ U(c) - U(d) \leq \lambda & \forall c, d \in A: \ (c \sim^{\rightarrow,N} d), \\ U(d) - U(c) \leq \lambda & \forall c, d \in A: \ (c \sim^{\rightarrow,N} d). \end{array}\right\} (E^{A^R}_{R-iterative_{II}}) \quad (22)$$

In the above *iterative* procedure, called **ROBUST-ITER**, the main objective is to maximize the value differences for those pairs of alternatives where there is a one-sided weak preference necessary relation. Once this is achieved, the secondary goal is to minimize the differences among those pairs where alternatives are assigned to the same class by all $U \in \mathcal{U}^R$.

An alternative approach is **ROBUST-COMP**, where a *compromise* solution is selected to attain both objectives simultaneously by maximizing the difference between the first and the second goal is given in Box I. The results of **ROBUST-ITER** and **ROBUST-COMP** will typically be different because they attain the two targets in various ways.

3.6. Robust models based on stochastic outcomes

A sorting model that is representative in terms of the robustness preoccupation can also be selected based on the stochastic outcomes computed by SOR. The idea implemented in **REPDIS** consists of emphasizing the advantage of these alternatives, which are assigned to a better class than others for a greater share of compatible sorting models, i.e., $APWI'(a, b) > APWI'(b, a)$. This can be attained by maximizing the minimal value difference for pairs of alternatives satisfying the above condition:

Maximize ω,

$$s.t. \quad \left.\begin{array}{l} E^{A^R}, \\ U(a) - U(b) \geq \omega(a, b), \\ \forall a, b \in A: \ APWI'(a, b) > APWI'(b, a), \\ \omega(a, b) \geq \omega, \\ \forall a, b \in A: \ APWI'(a, b) > APWI'(b, a). \end{array}\right\} (E^{A^R}_{APWI'_I}) \quad (24)$$

In the second stage, one can optimize the sum of elementary value differences $\omega(a, b)$, while respecting the results of the first stage by setting $\omega = \omega^*$, i.e.: Maximize

$$\text{Maximize} \sum_{\forall a,b \in A:\ APWI'(a,b) > APWI'(b,a)} \omega(a, b) \text{ s.t. } E^{A^R}_{APWI'_I} \cup \{\omega = \omega^*\}.$$

In what follows, we discuss three novel approaches that exploit the stochastic acceptability indices for selecting a single, robust sorting model. These models are inspired by the procedures for deriving the robust rankings proposed in [53]. Apart from handling ordinal classification problems and suitably exploiting sorting-specific results, the notable differences include (a) inferring a representative, feasible sorting model rather than constructing only the most robust recommendation and (b) ensuring the DM's preference information is reproduced.

The first method, called **CAI**, aims at maximizing the $CAI'(a_i, C_l)$ corresponding to the class assignment C_l suggested for each alternative $a \in A$ by a given sorting model $U \in \mathcal{U}^R$, denoted by $a \rightarrow^U C_l$. Due to the intrinsic nature of *CAI*s, maximization involves the product of values for individual alternatives instead of a sum. The main reason is that the relationships between *CAI*s should be compared in terms of a ratio rather than a difference. For example, $CAI'(a, C_1) = 0.25$ and $CAI'(a, C_2) = 0.75$ indicate that $a \rightarrow C_2$ occurred three times more often than $a \rightarrow C_1$ in the space of compatible sorting models. The objective function can be formulated as follows:

$$U^* = \arg\max_{U \in \mathcal{U}^R} \prod_{\forall a_i \in A:\ a_i \rightarrow^U C_l} CAI'(a_i, C_l).$$

We will replace the above non-linear form with its linear counterpart. Specifically, we replace the product of numbers with the sum of their logarithms (note that CAI' values are computed beforehand). The objective function needs to build on *CAI*s that correspond to the class assignments of alternatives suggested by the selected model. This is ensured by introducing binary variables x_{il} that should be equal to one when $a_i \rightarrow^U C_l$ is satisfied. After these transformations, the following problem is

obtained:

$$\left.\begin{array}{l}\text{Maximize } \kappa_{cai} = \sum\limits_{a_i \in A}\sum\limits_{l=1}^{p} x_{il} \log CAI'(a_i, C_l), \\ \\ s.t. \begin{array}{l} E^{A^R}, \\ \forall a_i \in A\ \forall l \in \{2, \ldots, p\}:\ U(a_i) - t_{l-1} - \delta - M x_{il} \geq -M, \\ \forall a_i \in A\ \forall l \in \{1, \ldots, p-1\}:\ t_l - U(a_i) - \delta - M x_{il} \geq \varepsilon - M, \\ \forall a_i \in A:\ \sum\limits_{l=1}^{p} x_{il} = 1, \\ \forall a_i \in A\ \forall l \in \{1, \ldots, p\}:\ x_{il} \in \{0, 1\}, \\ \forall a_i \in A\ \forall l \in \{1, \ldots, p\}:\ CAI'(a_i, C_l) = 0 \implies x_{il} = 0, \\ \forall a_i \in A\ \forall l \in \{1, \ldots, p\}:\ a_i \rightarrow^{DM} C_l \implies x_{il} = 1, \end{array} \end{array}\right\} (E^{A^R}_{CAI_I}) \tag{25}$$

where $M \gg 1$ is a large constant. Note that for any inequality in the form: $X - Mb \geq -M$, where X is an expression whose value can be determined and b is a binary variable, b may be equal to one only if $X \geq 0$. Hence, variable x_{il} will be equal to one when the conditions justifying $I_U(a_i) = l$, i.e., $U(a_i) \geq t_{l-1}$ and $t_l > U(a_i)$, are met. Then, other variables x_{ih}, $h \neq l$, will be set to zero, hence satisfying the following constraint $\sum\limits_{l=1}^{p} x_{il} = 1$. To avoid situations where $CAI'(a_i, C_l) = 0$ is included in the objective function, and thus the logarithm value is undefined, we forbid the corresponding class assignments by setting $x_{il} = 0$. Finally, we ensure that all DM's classification examples are reproduced.

Solving the above Mixed-Integer Linear Programming (MILP) problem allows identifying a sorting model that best represents the entire space in terms of assignments of alternatives to classes, measured with *CAI*s. As a secondary objective, we will regularize the model to balance the maximal shares of all criteria in the comprehensive value, hence advocating for a more central function. Specifically, we will minimize the deviations between the greatest marginal values for all pairs of criteria:

$$\left.\begin{array}{l}\text{Minimize } \xi, \\ \\ s.t. \begin{array}{ll} E^{A^R}_{CAI_I}, & \\ \kappa_{cai} = \kappa^*_{cai}, & \\ \forall i, j \in \{1, \ldots, m\} \wedge i \neq j: & u_i(\beta_i^{\gamma_i}) - u_j(\beta_j^{\gamma_j}) \leq \xi, \\ \forall i, j \in \{1, \ldots, m\} \wedge i \neq j: & u_j(\beta_j^{\gamma_j}) - u_i(\beta_i^{\gamma_i}) \leq \xi. \end{array} \end{array}\right\} (E^{A^R}_{CAI_{II}}) \tag{26}$$

This secondary target will also be considered in the context of the following two procedures. Since the model used for this purpose will be the same, we will not repeat it to save space.

An analogous approach, called **APOI**, can be formulated based on the analysis of the stability of assignment-based relations for all pairs of alternatives rather than class assignments of individual alternatives. In particular, we will consider the following stochastic acceptabilities for all $(a_i, a_j) \in A \times A$:

- $APWI'(a_i, a_j)$ indicating the share of models for which a_i is assigned to a more preferred class than a_j, i.e., $I_U(a_i) > I_U(a_j)$;
- $APEI'(a_i, a_j)$ indicating the share of models for which a_i is assigned to the same class as a_j, i.e., $I_U(a_i) = I_U(a_j)$;
- $APWI'(a_j, a_i)$ indicating the share of models for which a_i is assigned to a less preferred class than a_j, i.e., $I_U(a_i) < I_U(a_j)$.

Overall, we aim at identifying the model emphasizing the assignment-based pairwise relations captured with *APWI*′s and *APEI*′s in the best way, i.e.:

$$U^* = \arg\max_{U \in \mathcal{U}^R} \prod_{(a_i, a_j) \in A \times A: i \neq j} \begin{cases} APWI'(a_i, a_j) & \text{if } I_U(a_i) > I_U(a_j), \\ APEI'(a_i, a_j) & \text{if } I_U(a_i) = I_U(a_j), \\ APWI'(a_j, a_i) & \text{if } I_U(a_i) < I_U(a_j). \end{cases}$$

Similar to the CAI procedure, we introduce the binary variables corresponding to the three possible relations for each pair of alternatives $(a_i, a_j) \in A \times A$, $i \neq j$: v_{ij} corresponding to a scenario with a_i being assigned to a more preferred class than a_j (for the inverse situation, we consider v_{ji}) and e_{ij} standing for a_i and a_j being assigned to the same class. After transforming the product of elementary objectives into the sum of respective logarithms, the following model can be formulated:

$$\begin{aligned} \text{Maximize } \kappa_{apoi} = & \sum_{(a_i, a_j) \in A \times A: i \neq j} v_{ij} \log APWI'(a_i, a_j) \\ & + \sum_{(a_i, a_j) \in A \times A: i \neq j} e_{ij} \log APEI'(a_i, a_j) \\ & + \sum_{(a_i, a_j) \in A \times A: i \neq j} v_{ji} \log APWI'(a_j, a_i), \end{aligned}$$

$$\left.\begin{array}{l} E^{A^R}_{CAI_I} \\ \forall (a_i, a_j) \in A \times A \wedge i \neq j:\ \sum\limits_{l=1}^{p} l x_{il} - \sum\limits_{l=1}^{p} l x_{jl} - M v_{ij} \geq 0.5 - M, \\ s.t. \quad \forall (a_i, a_j) \in A \times A \wedge i \neq j:\ \sum\limits_{l=1}^{p} l x_{il} - \sum\limits_{l=1}^{p} l x_{jl} - M v_{ij} \leq 0.5, \\ \forall (a_i, a_j) \in A \times A \wedge i \neq j:\ v_{ij} + e_{ij} + v_{ji} = 1, \\ \forall (a_i, a_j) \in A \times A \wedge i \neq j:\ v_{ij}, e_{ij}, v_{ji} \in \{0, 1\}, \\ \forall (a_i, a_j) \in A \times A \wedge i \neq j:\ APWI'(a_i, a_j) = 0 \implies v_{ij} = 0, \\ \forall (a_i, a_j) \in A \times A \wedge i \neq j:\ APEI'(a_i, a_j) = 0 \implies e_{ij} = 0, \\ \forall a_i \in A\ \forall l \in \{1, \ldots, p\}:\ a_i \rightarrow^{DM} C_l \implies x_{il} = 1. \end{array}\right\} (E^{A^R}_{APOI_I}) \tag{27}$$

The role of M is the same as in the CAI procedure. The first three constraints mentioned above enforce $v_{ij} = 0$ when a_i is not assigned to a class better than a_j. However, if a_i is assigned to a more preferred class than a_j, then the second constraint enforces $v_{ij} = 1$. In case both $v_{ij} = 0$ and $v_{ji} = 0$, the third constraint implies $e_{ij} = 1$. The three variables are used to select the factor in the maximization function for each pair of alternatives. In this way, the optimization focuses on assigning alternatives to classes to reflect as closely as possible the relationships between pairs of alternatives in the entire set of sorting models compatible with DM's preferences. Again, we incorporate constraints that prohibit relations confirmed by none compatible model in the stochastic analysis.

The joint focus on reproducing the most frequent assignments of individual alternatives and the most supported assignment-based preference relations is reflected in the **COMB** procedure. It combines the objective functions considered in **CAI** and **APOI** under a unified framework, hence reconciling the two perspectives:

$$\begin{aligned} \text{Maximize } \kappa_{comb} = & \sum_{a_i \in A}\sum_{l=1}^{p} x_{il} \log CAI'(a_i, C_l) \\ & + \sum_{(a_i, a_j) \in A: i \neq j} v_{ij} \log APWI'(a_i, a_j) \\ & + \sum_{(a_i, a_j) \in A: i \neq j} e_{ij} \log APEI'(a_i, a_j) \\ & + \sum_{(a_i, a_j) \in A: i \neq j} v_{ji} \log APWI'(a_j, a_i), \ \text{s.t. } E^{A^R}_{APOI_I}. \end{aligned} \tag{28}$$

Still, the idea of reflecting the outcomes of SOR in a single model that can be exhibited to the DM is maintained.

To illustrate how the procedures for selecting a representative sorting model work, we consider an example problem concerning the evaluation of 30 major European cities in implementing green policy [54]. Such an illustration aims to remind the operational steps and emphasize the peculiar features of the typical solutions returned by all procedures in a specific study. Also, we demonstrate that they may lead to various models and recommendations even if operating on the same data and incorporating the same preference model. Moreover, such an example, exhibiting a deeper analysis of a single problem, makes the results of the subsequent experimental comparison (in particular, measure values) more understandable. In turn, the experimental section focuses on the average case performance across all considered problem instances. The results of the illustrative study are presented in eAppendix 1.

4. Computational experiments

This section is devoted to the computational experiments performed to examine the quality and characteristics of procedures for selecting a representative sorting model. We define the measures used to compare the 14 approaches and the features of problem instances considered during the tests. The results obtained for each measure are discussed in detail, given the average outcomes across all considered settings and performance trends observed when changing some parameter values. Moreover, we assess the statistical significance of the differences observed between the results attained by various pairs of methods using appropriate tests. Finally, we use the linear regression model to identify the average impact of changes in individual problem parameters on the measured values.

4.1. Comparative measures

The performance of the procedures for selecting a single sorting model will be quantified in terms of five measures. The results will be used to compare individual methods in three main aspects. First, they show how the assignments used to simulate the decision-making policy are reflected in the results based on incomplete preference information. In this way, we quantify the predictive accuracy. Second, we consider how representative the recommended assignments are for the compatible sorting models. Thus, we refer to the robustness of the recommendation delivered by each approach given the multiplicity of outcomes obtained with a set of models consistent with the DM's preferences. The third goal is to compare the structure of models, i.e., marginal value functions, comprehensive values of alternatives, and thresholds. This provides conclusions about the similarity between the DM's decision policy and the preference model that attempts to capture it.

Let us denote a set of all non-reference alternatives that the DM has not classified by $A^T = A \setminus A^R$. The reference model composed of marginal value functions, comprehensive values, and class thresholds is denoted by U^{REF}, and the analogous model returned by procedure P is U^P.

Classification accuracy. To determine the quality of the sorting model, we can verify how far the solution proposed by the procedure is from the comprehensive DM's preferences in terms of recommended assignments. We focus only on the non-reference alternatives because all procedures reproduce the assignments of reference solutions. Therefore, the classification accuracy captures the proportion of alternatives in set A^T for which the recommended and reference assignments agree [44], i.e.:

$$accuracy(U^P) = \frac{|a : a \in A^T \wedge I_{U^P}(a) = I_{U^{REF}}(a)|}{|A^T|}. \tag{29}$$

Assignment acceptability. Another measure compares the assignments recommended by different procedures with the classification obtained in the entire set of sorting models. The assignment acceptability reflects average support given to the assignments recommended by a particular procedure for all non-reference alternatives in terms of class acceptability indices CAI's derived from the analysis of all feasible solutions, i.e.:

$$MCAI(U^P) = \frac{1}{|A^T|} \sum_{a \in A^T : a \to^{U^P} C_{l_a}} CAI'(a, C_{l_a}) \in [0, 1]. \tag{30}$$

The maximal $MCAI$ value can be obtained when each non-reference alternative is assigned to the class with the highest CAI value. As noted in [44], this approach to classification is based on the *robust assignment rule*. The value defined in this way is marked as $MCAI_{max}$:

$$MCAI_{max} = \frac{1}{|A^T|} \sum_{a \in A^T} \max_{l \in \{1,\ldots,p\}} CAI'(a, C_l) \in [0, 1]. \tag{31}$$

In what follows, we consider an *Absolute MCAI* ($MCAI_{abs} = MCAI(U^P)$), and the *Relative MCAI*, which makes the measure values more interpretable by referring them to the best possible solution that could be obtained for a given problem:

$$MCAI_{rel}(U^P) = \frac{MCAI(U^P) - MCAI_{max}}{MCAI_{max}} = \frac{MCAI(U^P)}{MCAI_{max}} - 1. \tag{32}$$

The advantage of accounting for the average value is its high interpretability. For example, $MCAI(U^P) = 0.7$ indicates that the assignment obtained with U^P for each alternative is, on average, supported by 70% of feasible models. Recommending precise assignments confirmed by a large share of all feasible models is a desirable property of sorting methods.

Clearly, other approaches to aggregating individual CAIs are also possible. The basic ones include the product of all CAIs or a minimum of CAIs confirming the assignment to a given class. The obtained value may be challenging to interpret in the first case, especially when the model assigns several alternatives to classes with low CAIs. The product of several low values could obscure the high certainty for the remaining assignments. In turn, the measure based on the minimum does not reflect the distribution of all CAIs. Then, models with significantly different acceptability indices on most alternatives may be characterized with the same measure value, failing to capture various robustness levels of the delivered recommendations.

The following three measures focus on the similarity between models rather than assignments. Hence they focus on the proximity of models and their various components. Such a perspective is complementary to the predictive performance and robustness of the delivered recommendation.

Differences between marginal values. To capture the agreement between the shapes of MVFs, we compare the marginal values assigned to all characteristic points except the least preferred. The latter ones are, by definition, always assigned values equal to zero. Such a measure – summarizing absolute value differences – can be considered as the comprehensive distance between the reference model U^{REF} and U^P obtained with procedure P:

$$\Delta^{REF}_{TO}(U^P) = \frac{1}{\sum_{j=1}^{m}(n_j^A - 1)} \sum_{j=1}^{m} \sum_{k=2}^{n_j^A} |u_j^{U^P}(\beta_j^k) - u_j^{U^{REF}}(\beta_j^k)|. \tag{33}$$

Another perspective concerns the distance of MVFs from a sorting model that represents well the feasible space of all models. For this purpose, we adopt the outcomes of the CENTROID procedure (denoted with an upper script *CENT*, e.g., U^{CENT}), which is an average of a large sample of uniformly distributed value functions and class thresholds. It can be defined in the following way:

$$\Delta_{TO}^{CENT}(U^P) = \frac{1}{\sum_{j=1}^{m}(n_j^A - 1)} \sum_{j=1}^{m} \sum_{k=2}^{n_j^A} |u_j^{U^P}(\beta_j^k) - u_j^{U^{CENT}}(\beta_j^k)|. \quad (34)$$

Such a measure indicates to what extent the solutions generated by a given procedure deviate from the average solution. In this case, the results need to be interpreted as a specific characteristic of the models returned by various methods rather than a performance measure indicating some good or bad approaches.

Differences between comprehensive values. Another measure refers to the aggregated results at the level that considers all criteria jointly. Instead of comparing the MVFs, it summarizes the differences between comprehensive values attained by all non-reference alternatives for the reference and resulting models:

$$\Delta_{CV}^{REF}(U^P) = \frac{1}{|A^T|} \sum_{a \in A^T} |U^{U^P}(a) - U^{U^{REF}}(a)|. \quad (35)$$

Differences between threshold values. The last measure concerns the similarity between separating class thresholds in the reference and resulting models:

$$\Delta_{TH}^{REF}(U^P) = \frac{1}{p-1} \sum_{l=1}^{p-1} |t_l^{U^P} - t_l^{U^{REF}}|. \quad (36)$$

It captures if the method can reproduce the range width of comprehensive values that justify an assignment to a given class and their positions on the scale of AVF. The values of the above measures for selected procedures based on the results of an illustrative case study are available in eAppendix 2.

In our view, the predictive accuracy and the recommendation robustness are more important than the similarity with the reference model. Accuracy – reflecting the fraction of predictions a given model got right – is a fundamental metric for evaluating classifiers in the Machine Learning (ML) context. In a preference disaggregation setting, it captures how well sparse and incomplete preference information on a subset of reference alternatives is used to reconstruct the DM's comprehensive decision policy on the entire set of alternatives. In turn, the robustness of outcomes delivered by a given method indicates how well these results represent the entire space of models compatible with incomplete assignment examples. This is more important in the MCDA context when addressing uncertainty related to the existence of multiple consistent sorting models, out of which one is used to derive final recommendations. However, from the experimental perspective, the robustness measures can also be interpreted in relation to classification accuracy. Indeed, they build on the set of all models compatible with the DM's preference information. The fact that only assignments of reference alternatives are available to the method implies that each compatible model could have served as the true one, and the same input preferences on the reference set would be derived from it. In this perspective, *MCAI* can be seen as an average classification accuracy while assuming that each compatible sorting model was used as the true one.

The model similarity measures represent another perspective, referring to the distances in the space of parameter values. In a way, they capture if incomplete preferences are sufficient for reconstructing the complete form of a model from which they were derived. However, this aspect is less relevant for the practice of decision-aiding. In the end, it is the recommendation that matters most to the DM. In this regard, even if the differences between parameter values are minor, the differences in provided classifications can be substantial. Conversely, large differences in parameter values may not imply inconsistencies in the suggested assignments. Furthermore, preferences of real DMs are not derived from a pre-defined additive value model combined with precise class thresholds. Hence, the true reference model typically does not exist in practical applications, even if the reference decisions serving as the benchmark are often known. However, in the experimental setting, the model similarity measures can still support understanding the characteristics of various procedures in view of measures focusing on other perspectives. For example, small differences in model parameter values can allow explaining a favorable predictive accuracy or robustness. In turn, some other methods may attain decent results despite significant differences in derived value functions and thresholds.

4.2. Experimental setting

When generating instances of test problems, we considered various settings for the dimensionality of data:

- the number of classes – $p \in \{2, 3, 4, 5\}$;
- the number of criteria – $m \in \{3, 5, 7, 9\}$;
- the number of equally distributed characteristic points for each criterion g_j – $\gamma_j \in \{2, 4, 6\}$;
- the number of reference alternatives assigned by the DM to each of p classes – $R \in \{3, 5, 7, 10\}$.

In this way, we covered relatively simple problems with three linear criteria and six reference alternatives in two classes, and complex problems with 9 criteria associated with marginal functions with $6 - 1 = 5$ linear pieces and up to 50 reference alternatives in five decision classes. The number of non-reference alternatives from set A^T is ten for each class. In this way, we represent the realistic scenarios in which the set of reference alternatives is at least as large as the test (non-reference) set. Consequently, the greatest problem instances involved up to 100 alternatives. It is a high value when considering the typical MCDA setting, which nevertheless still makes feasible the execution of robustness analysis methods incorporated by some of the considered procedures. For each combination of parameter values, we averaged the results over 100 problem instances. Hence we considered $4 \times 4 \times 3 \times 4 \times 100 = 19{,}200$ instances in total.

For each instance, we followed the procedure described in [44]. Hence, two pools of 1,000 alternatives were drawn, each with m criteria values. The alternatives' performances on each criterion were drawn using the uniform distribution. This does not exclude dominated alternatives. However, in the case of a sorting problem, even if the DM assigns the dominating alternative to some class, it is usually impossible to determine to which class the dominated alternative will be assigned precisely. It only allows us to delimit the range of possible assignments.

The reference alternatives were randomly selected from the first pool, and the test (non-reference) alternatives were chosen from the other pool. The alternatives in these two pools were evaluated with a randomly generated AVF serving as the DM's reference model. For simplicity, we assumed that the number of equidistant characteristic points for the respective MVFs was equal to γ_j in the considered problem setting. Then, the separating class thresholds $t = [t_1, \ldots, t_{p-1}]$ were set to respect the following proportions of alternatives from the first (reference) pool being assigned to particular classes: for $p = 2$ – 50-50, for $p = 3$ – 30-40-30, for $p = 4$ – 20-30-30-20, and for $p = 5$ – 15-20-30-20-15. Such divisions correspond to realistic

scenarios in which extreme classes are less common than intermediate ones. The threshold values were determined using the interpolated values of the respective percentiles in the set of all scores obtained in the reference set so as to divide it into the given proportions. Specifically, they were set randomly in the value range, guaranteeing pre-defined proportions. That is, for 2-class problems, t_1 was determined as the 50-th percentile of all scores; for 3-class problems, t_1 is the 30-th percentile, and t_2 is the 70-th percentile; for 4-class problems, thresholds take the values of the 20-th, 50-th, and 80-th percentiles, and for 5-class problems, the thresholds are the 15-th, 35-th, 65-th, and 85-th percentiles, respectively. Such thresholds were used to derive class assignments for alternatives contained in both pools. Finally, depending on the considered setting, a pre-defined number of alternatives were randomly selected for each class to construct sets of reference and test alternatives. When put together, these two sets (A^R and A^T) formed a set of alternatives A that would be normally considered by the DM facing a particular decision problem.

The 14 methods were run for all problem instances except for MSCVF for problems with $\gamma_j = 2$ characteristic points. In this case, the marginal value functions for all methods are linear. For each problem instance, the values of stochastic acceptability indices were estimated based on 10,000 sorting models generated with HAR.

4.3. Results

In this section, we discuss the results of an experimental comparison of the 14 procedures for selecting a single, representative sorting model. For each measure, we consider the outcomes averaged over all problem instances and the mean values of the performance measures obtained for different values of each problem dimension (p, m, γ_j, and R).

4.3.1. Classification accuracy

The accuracy of the classification is important and, in many cases, the main evaluation criterion in the context of choosing the best method. Average classification accuracies over all problem instances are provided in Table 1. To check the significance of the relationships between the results of each method, the Wilcoxon signed-rank test [55] for paired samples with a p-value of 0.05 was performed. The test results are reflected in the Hasse diagram in Fig. 1, which shows if there is a statistically significant difference between the results of different approaches.

The difference between the best and worst performers is substantial (over 12%). The best accuracy was obtained by ACUTADIS (0.8313), which identifies an analytic center of the polyhedron using non-linear optimization. In general, seeking the central solution is an excellent strategy to increase classification accuracy. This is confirmed by the results attained by other approaches implementing this concept, i.e., CENTROID (0.8134) and CHEBYSHEV (0.8099). Highly favorable results (between 0.8113 and 0.8119) are obtained by the approaches exploiting the stochastic acceptability indices: CAI, APOI, and COMB. The advantageous performance of these methods, along with the high position of CENTROID, confirms the usefulness of conducting robustness analysis with the Monte Carlo simulations. Slightly lesser classification accuracies were attained with the traditional procedures, which are most often used in the context of UTADIS due to their simplicity, i.e., UTADISMP1, UTADISMP2, and UTADIS-JLS. They choose either the most discriminant model or an average model, though, based on the analysis of extreme ones only.

One of the worst average accuracies was achieved by procedures focusing on the shape of the MVFs, i.e., MSCVF and UTADISMP3. Note that the comparison of the mean value for the MSCVF is not appropriate because it has not been assessed for problems with linear MVFs, so the mean value may be underestimated. In the context of this method, more significant observations can be made by comparing the values for the problems with $\gamma_j \in \{4, 6\}$. The unfavorable performance of UTADISMP3 in terms of predictive accuracy and other measures is partially due to penalizing non-linear MVFs, while the models of simulated DMs are drawn randomly, typically involving non-linear marginal functions. Nonetheless, it is interesting to note that UTADISMP2 – additionally involving the discriminating component – performs slightly better.

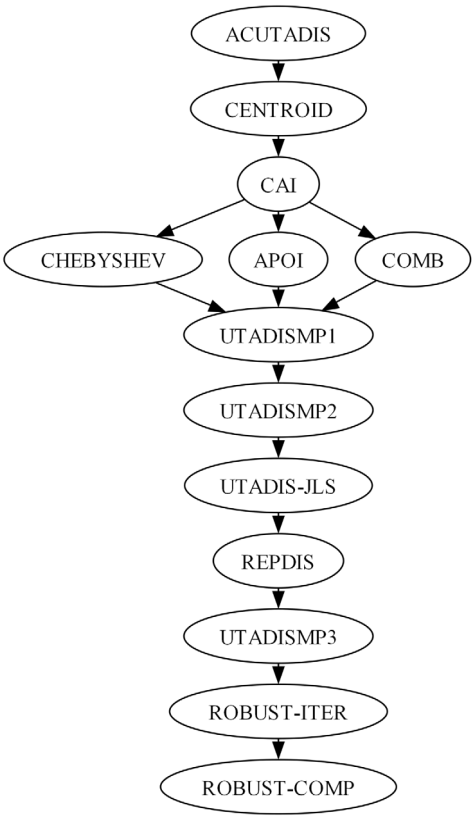


Fig. 1. The Hasse diagram indicating the statistically significant differences in terms of the classification accuracy based on the Wilcoxon test with p-value equal to 0.05.

Both the mean values and the Hasse diagram shown in Fig. 1 confirm that procedures exploiting the exact outcomes of robustness analysis achieved significantly worse results than UTADIS MP3. ROBUST-ITER and ROBUST-COMP allowed for reproducing the correct assignment for over 10% less non-reference alternatives than ACUTADIS. The objectives considered by these approaches differ vastly from the best-performing methods. A general conclusion from the experiment is that when one aims to maximize classification accuracy, a sorting model should be selected by exploiting the feasible polyhedron or considering the robustness of shapes or recommendations delivered with a large subset of all compatible models.

The number of classes significantly impacts the classification accuracy attained by different approaches. Table 1 confirms that the accuracies decrease for a greater number of classes. For example, for UTADISMP2 – the accuracy ranges between 0.8108 for $p = 2$ and 0.7553 for $p = 5$. It is intuitive because, with more classes, the sorting problem becomes more challenging, the sub-spaces of feasible models corresponding to different classes are more constrained, and the class thresholds become closer to each other. As

Table 1
Mean values and standard deviations of classification accuracy for all considered problem settings, different numbers of classes and criteria.

Procedure	All settings		Number of classes				Number of criteria			
	mean	std	2	3	4	5	3	5	7	9
UTADISMP1	0.7897	0.1252	0.8094	0.7852	0.7814	0.7829	0.8503	0.8023	0.7686	0.7376
UTADISMP2	0.7756	0.1210	0.8108	0.7760	0.7604	0.7553	0.8294	0.7872	0.7567	0.7292
UTADISMP3	0.7483	0.1231	0.8101	0.7494	0.7207	0.7132	0.7989	0.7569	0.7287	0.7088
UTADIS-JLS	0.7703	0.1345	0.7917	0.7621	0.7606	0.7669	0.8237	0.7859	0.7506	0.7211
CHEBYSHEV	0.8099	0.1124	0.8423	0.8074	0.7956	0.7942	0.8623	0.8204	0.7918	0.7650
MSCVF	0.7100	0.1304	0.7521	0.7097	0.6956	0.6828	0.7860	0.7234	0.6811	0.6496
ACUTADIS	0.8313	0.1040	0.8548	0.8288	0.8214	0.8203	0.8794	0.8424	0.8136	0.7899
CENTROID	0.8134	0.1140	0.8434	0.8129	0.7993	0.7979	0.8714	0.8255	0.7915	0.7650
REPDIS	0.7571	0.1207	0.7841	0.7545	0.7444	0.7456	0.8138	0.7684	0.7349	0.7114
CAI	0.8119	0.1145	0.8426	0.8118	0.7975	0.7958	0.8705	0.8242	0.7897	0.7632
APOI	0.8113	0.1151	0.8412	0.8114	0.7971	0.7952	0.8695	0.8234	0.7890	0.7631
COMB	0.8113	0.1150	0.8414	0.8115	0.7971	0.7953	0.8696	0.8236	0.7891	0.7631
ROBUST-ITER	0.7294	0.1330	0.7554	0.7207	0.7190	0.7226	0.8005	0.7445	0.7026	0.6703
ROBUST-COMP	0.7238	0.1348	0.7518	0.7179	0.7126	0.7130	0.7967	0.7381	0.6966	0.6639

a result, the comprehensive values of non-reference alternatives have lower chances of fitting in the value range corresponding to their expected class. The greatest decrease in performance is observed between problems with 2 and 3 classes (from 2.42% for UTADISMP1 to 6.07% for UTADISMP3). However, with the increasing number of classes, these differences become lesser, and when comparing the results for 4 and 5 for some procedures – they are negligible, and for five procedures (e.g., UTADISMP1, REPDIS, and ROBUST-ITER) – there is even a marginal increase in performance. This suggests that, at this point, an increased problem complexity implied by a higher number of classes is well balanced by an added information value offered by more assignment examples.

Compared to other methods, a marginal decrease in accuracy with an increasing p is an additional advantage of ACUTADIS. This procedure proves to be more robust to modifying p, increasing its relative advantage over the remaining methods when more classes are considered. In the same spirit, the underperformance of MSCVF is more evident in instances involving more classes.

The number of criteria and characteristic points affect the accuracies similarly to the number of classes. With the increase in m and γ_j, the performance of all procedures deteriorates (see Tables 1 and 2). For example, for CHEBYSHEV, an average accuracy ranges between 0.8623 and 0.7650 for 3 and 9 criteria, respectively, and between 0.8526 and 0.7790 for 2 and 6 characteristic points. Again, this is intuitive because, with more criteria and characteristic points, the space of feasible models becomes more significant, and MVFs become more flexible. The only exception is observed in the improved performance of MSCVF when passing from $\gamma_j = 4$ to 6. However, this can be attributed to an extremely poor classification accuracy attained by this procedure already for less flexible MVFs.

The average differences between accuracies for problems with three and nine criteria range from 8.95 to 13.64% (see Table 1). Hence, they are more substantial than between the extreme numbers of classes (e.g., for UTADISMP1 and UTADIS-JLS – even four times greater). The sole exception in this regard is UTADISMP3.

The decrease in accuracy is visible between all subsequent numbers of criteria. For all fourteen procedures, it is on average 4.68% between 3 and 5 criteria, 3.44% between 5 and 7 criteria, and 2.74% between 7 and 9 criteria. As for the number of characteristic points (see Table 2), there is a clear difference in the accuracy of methods between linear and piecewise-linear MVFs. The scores attained for MVFs with 2 and 6 characteristic points differ from 4.53% for UTADISMP2 up to 14.73% for UTADIS-JLS.

Noteworthy, UTADIS-JLS achieved relatively high results (85.02% compared to 86.96% accuracy achieved by the best method – ACUTADIS) when using linear value functions. However, when employing six characteristic points, the difference between these two methods increased to over 10%. Such a difference is associated with optimizing values assigned to the last characteristic points for each MVF. For the linear functions, this contributes to controlling their entire shapes and selecting more central value functions. In turn, with greater γ_j, the marginal values of intermediate characteristic points are not directly affected by the optimized model. For MSCVF, the differences in accuracies attained for MVFs with 4 and 6 points are negligible. This is due to the characteristic of the method, which – regardless of the number of points – tries to linearize the marginal functions as much as possible.

The increase in the number of reference alternatives per class positively affects the classification accuracy (see Table 2). For example, for UTADISMP1, the accuracy ranges between 0.7069 and 0.8547 for, respectively, $R = 3$ and 10. A greater number of assignment examples makes the knowledge available to the methods more complete, offering additional arguments on the DM's sorting policy. From a mathematical viewpoint, additional indirect statements constrain the space of feasible models, leaving lesser freedom to the procedures for selecting a representative model.

With limited preference information (see $R = 3$), ACUTADIS has a clear advantage over the remaining methods (more than 2.5% over CENTROID). Generally, the margin between the stochastic- (CENTROID, CAI, APOI, COMB) or centralization-based (ACUTADIS, CENTROID, CHEBYSHEV) and the remaining approaches is greater with more sparse DM's preferences. For example, for $R = 3$ – the difference in accuracies of APOI and UTADISMP1 is 3.52%, whereas, for $R = 10$, it drops to 1.18%. This emphasizes the usefulness of the best-performing approaches when only a few assignment examples are available.

To investigate the impact of individual parameters on the results more holistically and to compare the strength of the influence of particular problem features on the measured values, we conducted a linear regression analysis. In the definition of the regression problem, the explanatory variables were the values of four parameters determining the problem, and the expected value was the average accuracy value for each method. The slope coefficients of individual parameters and the intercept value shown in Table 3 were determined separately for each method based on the results obtained during the experiment.

Taking into account the methods' specificities and the previously observed features, one group of approaches for which the slope coefficients are very similar can be distinguished. The regression models in Table 3 confirm the high similarity of the results obtained by the methods exploiting stochastic analysis outcomes. REPDIS, CAI, APOI, and COMB use acceptability indices

Table 2
Average classification accuracy for different numbers of characteristic points and reference alternatives per class.

Procedure	Number of ch. points			Number of reference assignments			
	2	4	6	3	5	7	10
UTADISMP1	0.8446	0.7739	0.7507	0.7069	0.7772	0.8200	0.8547
UTADISMP2	0.7993	0.7735	0.7540	0.6948	0.7631	0.8042	0.8405
UTADISMP3	0.7869	0.7363	0.7217	0.6695	0.7340	0.7753	0.8145
UTADIS-JLS	0.8502	0.7579	0.7029	0.6695	0.7559	0.8061	0.8498
CHEBYSHEV	0.8526	0.7980	0.7790	0.7423	0.7987	0.8341	0.8644
MSCVF		0.7096	0.7105	0.6113	0.6986	0.7442	0.7861
ACUTADIS	0.8696	0.8184	0.8059	0.7713	0.8232	0.8520	0.8788
CENTROID	0.8673	0.7979	0.7748	0.7459	0.8021	0.8380	0.8674
REPDIS	0.8126	0.7376	0.7212	0.6763	0.7437	0.7848	0.8237
CAI	0.8662	0.7962	0.7733	0.7439	0.8007	0.8365	0.8665
APOI	0.8649	0.7958	0.7731	0.7421	0.8003	0.8362	0.8665
COMB	0.8650	0.7960	0.7731	0.7422	0.8004	0.8363	0.8665
ROBUST-ITER	0.7927	0.7098	0.6858	0.6397	0.7107	0.7588	0.8085
ROBUST-COMP	0.7905	0.7036	0.6774	0.6352	0.7057	0.7543	0.8001

Table 3
Coefficients of solutions obtained for the linear regression problem for average accuracy depending on the defined dimensions for individual procedures.

Procedure	No. of classes	No. of criteria	No. of ch. points	No. of ref. alt.	Intercept
UTADISMP1	−0.008322	−0.018596	−0.023477	0.020602	0.895558
UTADISMP2	−0.018225	−0.016563	−0.011330	0.020305	0.857199
UTADISMP3	−0.031941	−0.014931	−0.016284	0.020284	0.888062
UTADIS-JLS	−0.007605	−0.017163	−0.036815	0.025072	0.890492
CHEBYSHEV	−0.015584	−0.016025	−0.018392	0.017054	0.927554
MSCVF	−0.022204	−0.022567	0.000432	0.024155	0.770027
ACUTADIS	−0.011083	−0.014865	−0.015927	0.014911	0.929816
CENTROID	−0.015004	−0.017669	−0.023124	0.016995	0.958154
REPDIS	−0.012565	−0.017034	−0.022829	0.020548	0.866217
CAI	−0.015500	−0.017815	−0.023215	0.017133	0.958832
APOI	−0.015236	−0.017681	−0.022958	0.017357	0.954017
COMB	−0.015265	−0.017689	−0.022991	0.017333	0.954538
ROBUST-ITER	−0.010021	−0.021621	−0.026704	0.023687	0.853010
ROBUST-COMP	−0.012159	−0.021994	−0.028258	0.023165	0.866598

for class assignments or pairwise comparisons. Apart from them, the CENTROID approach also uses raw sampling results.

For the five methods mentioned above, adding more dimensions to the problem decreases accuracy. An additional attribute reduces it by an average of 1.7–1.8%, and another characteristic point in MVFs by 2.3%. The REPDIS method is better than the other approaches when changing the other two parameters. This method is more robust in the case of an increased number of classes (−1.2% vs. −1.5% for other methods), and its accuracy increases more evidently when the number of available assignments increases (accuracy increases by over 2% vs. 1.7% for other methods). However, this may be because, on average, this method performs much worse than the others, so it is easier to make progress in correctly assigning alternatives when the problem becomes more straightforward.

In the context of obtaining additional preference information, the above regularity is confirmed for all other methods. Indication of additional assignments to each class has the greatest positive impact on the average accuracy of the weakest approaches: UTADIS-JLS (increase by 2.5%), MSCVF (2.4%), and ROBUST methods (2.3–2.4%). On the contrary, increased availability of preferences has the most negligible impact on the method with the best average value – ACUTADIS (1.5%).

The stability of the results of MSCVF is noticeable when the number of characteristic points changes. The pursuit of this method to obtain the functions that are *as linear as possible* makes its results practically insensitive to changing this parameter. It is entirely different from UTADIS-JLS, which does not consider the values for the internal MVFs at all, so in this case, the impact is most significant (a 3.6% decrease in accuracy for each next point).

4.3.2. Assignment acceptability

Average assignment acceptabilities over all problem instances are provided in Table 4. The ranking of procedures resulting from the Wilcoxon paired test with p-value equal to 0.05 for both absolute and relative MCAI are presented as the Hasse diagram in Fig. 2. The difference between the best and worst-performing procedures is enormous (almost 0.25). The procedures exploiting stochastic acceptabilities attained the highest absolute MCAIs. In particular, the CAI procedure emphasizes the most frequent assignments when selecting a representative model. It successfully attains this target with an absolute MCAI equal to 0.8979 and its relative counterpart being close to zero. This means that the CAI method identifies a model that classifies all alternatives according to the *robust assignment rule* [44], i.e., it assigns each alternative to a class associated with the highest CAI. Since, for some problem instances, no model optimized such an objective in a perfect way (in the experiment – this happened for 4 out of 19,200 instances), the average relative MCAI for this method is slightly above zero. The APOI and COMB methods are only marginally worse in this regard (absolute MCAI equal to 0.8975). This means that considering stochastic acceptabilities for the assignment-based pairwise preference relations led to different assignments for very few problem instances. This confirms that the two perspectives are highly consistent in guiding the methods to the most robust assignments.

Another group of methods that perform well in terms of assigning alternatives to their most frequent classes in the set of all feasible models is composed of CENTROID (0.8968), CHEBYSHEV (0.8620), and ACUTADIS (0.8449). Note that CAI, APOI, and COMB have a competitive advantage over these methods in considering

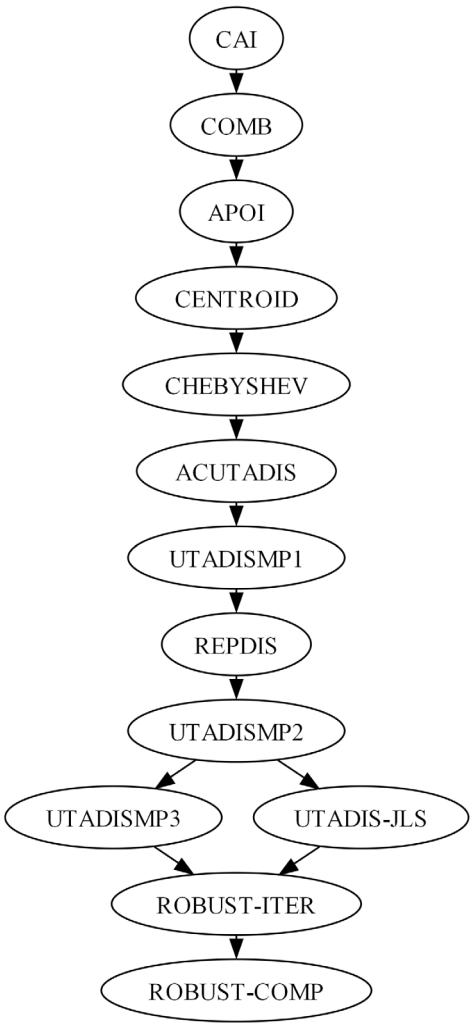


Fig. 2. The Hasse diagram indicating the statistically significant differences in terms of the absolute and relative MCAI based on the Wilcoxon test with p-value equal to 0.05.

the assignments of all alternatives, including non-reference ones, already at the stage of identifying a representative model. When such an approach is too costly in terms of required computational effort, one can opt for methods selecting a central model that exploit only the information provided by the DM. Interestingly, unlike for the classification accuracy, ACUTADIS performs slightly worse than procedures selecting an average model or the Chebyshev center.

The worst performers in terms of assignment acceptability are the same as those for classification accuracy. The least robustness of recommended assignments is observed for MSCVF (0.7161). It is understandable, given that the objective function optimized by this approach has nothing in common with alternatives' assignments or the robustness of results. Moreover, this procedure omitted the problems with linear MVFs. Surprisingly low MCAIs are attained by ROBUST-ITER (0.7400) and ROBUST-COMP (0.7272). These procedures build on the outcomes of robustness analysis. However, they focus on the necessary and possible relations derived from mathematical programming. This proves that such extreme, robust outcomes are often too scarce to provide valuable insights and guide the procedures to select a model that would be representative in terms of robustness preoccupation. Since the space between the necessary and the possible may be quite large, using stochastic acceptabilities computed with the Monte Carlo simulation and filling this gap is more beneficial for most problem instances.

Table 4
Mean values and standard deviations of assignment acceptabilities for all considered problem settings.

Procedure	Absolute		Relative	
	mean	std	mean	std
UTADISMP1	0.8034	0.1058	−0.1060	0.1010
UTADISMP2	0.7902	0.1046	−0.1208	0.0998
UTADISMP3	0.7814	0.1005	−0.1312	0.0887
UTADIS-JLS	0.7766	0.1191	−0.1357	0.1193
CHEBYSHEV	0.8620	0.0669	−0.0407	0.0396
MSCVF	0.7161	0.1209	−0.2095	0.1202
ACUTADIS	0.8449	0.0760	−0.0594	0.0622
CENTROID	0.8968	0.0491	−0.0012	0.0025
REPDIS	0.7980	0.0902	−0.1123	0.0782
CAI	0.8979	0.0485	−1.4E−07	1.4E−05
APOI	0.8975	0.0491	−0.0005	0.0020
COMB	0.8975	0.0490	−0.0005	0.0018
ROBUST-ITER	0.7400	0.1251	−0.1773	0.1228
ROBUST-COMP	0.7272	0.1261	−0.1913	0.1256

In Tables 5–8, we provide the average assignment acceptabilities for different values of particular dimensions. In general, the robustness of recommended assignments increases with fewer classes, criteria, and characteristic points and a greater number of reference alternatives per class. Hence, these trends are the same as for the classification accuracy. They can be attributed to the same reasons. Less flexible models and greater information load lead to more constrained space of feasible models and more robust sorting results. For example, for UTADISMP2 – the difference between extreme values for each dimension are as follows: for $p \in \{2, 5\}$ – 0.0359, for $m \in \{3, 9\}$ – 0.0908, for $\gamma_j \in \{2, 6\}$ – 0.0221, and for $R \in \{3, 10\}$ – 0.1539. This indicates that the number of reference alternatives per class has the greatest impact on the robustness of recommended assignments. In contrast, the influence of the number of characteristic points is the least.

When it comes to absolute MCAIs attained for different numbers of classes (see Table 5), the greatest differences are observed for problems with 2 and 3 classes. The deviation from the general trend is noted for some procedures when comparing the results for problems with 4 and 5 classes. As far as various procedures are concerned, the performance of CAI, APOI, COMB, and CENTROID is the most stable (e.g., for the last approach, absolute MCAI is 0.9099 for $p = 2$ and 0.8931 for $p = 5$). All these methods share the component of performing the stochastic acceptability analysis. The highest decrease in the quality of the generated solutions can be seen for UTADISMP3. In the case of absolute values, it falls from 0.8421 for 2 classes to 0.7446 for 5 classes. This decrease is also the highest in the case of relative values. For the problems with 2 classes, the deviation from the optimal MCAI value for all alternatives was 0.0757, and with 5 classes, it was over 2 times higher and equal to 0.1695. For the remaining approaches, these relative values are more stable, and for some of them (see, e.g., UTADISMP1 and REPDIS), they tend to perform even slightly better when moving from three to five classes.

When the number of criteria increases, the trends for absolute and relative MCAIs are more consistent (see Table 6). For all procedures, the robustness of recommended assignments decreases in terms of absolute values and their distances from the best possible solution. Interestingly, a slight decrease in absolute values is also noticeable for the CAI method, which, apart from a few exceptions, always obtains the solution with the highest absolute MCAI value possible. This may lead to the conclusion that a greater number of attributes results in greater model flexibility. Herefore, the recommendations resulting from the stochastic

Table 5
Average assignment acceptability for different numbers of classes.

Procedure	Absolute				Relative			
	2	3	4	5	2	3	4	5
UTADISMP1	0.8155	0.7949	0.7982	0.8048	−0.1048	−0.1120	−0.1060	−0.1013
UTADISMP2	0.8158	0.7857	0.7793	0.7799	−0.1039	−0.1223	−0.1274	−0.1295
UTADISMP3	0.8421	0.7840	0.7549	0.7446	−0.0757	−0.1247	−0.1551	−0.1695
UTADIS-JLS	0.8050	0.7675	0.7642	0.7698	−0.1154	−0.1424	−0.1442	−0.1410
CHEBYSHEV	0.8830	0.8592	0.8525	0.8535	−0.0310	−0.0403	−0.0449	−0.0464
MSCVF	0.7645	0.7175	0.6993	0.6831	−0.1728	−0.2058	−0.2210	−0.2385
ACUTADIS	0.8744	0.8410	0.8319	0.8321	−0.0399	−0.0602	−0.0676	−0.0702
CENTROID	0.9099	0.8936	0.8906	0.8931	−0.0009	−0.0013	−0.0014	−0.0013
REPDIS	0.8132	0.7910	0.7918	0.7959	−0.1079	−0.1168	−0.1131	−0.1112
CAI	0.9107	0.8947	0.8918	0.8943	0	−1.9E−08	−4.1E−07	−1.5E−07
APOI	0.9104	0.8942	0.8913	0.8939	−0.0003	−0.0006	−0.0006	−0.0005
COMB	0.9105	0.8943	0.8914	0.8939	−0.0003	−0.0006	−0.0005	−0.0004
ROBUST-ITER	0.7597	0.7298	0.7313	0.7390	−0.1661	−0.1854	−0.1818	−0.1757
ROBUST-COMP	0.7563	0.7205	0.7167	0.7153	−0.1698	−0.1954	−0.1978	−0.2021

Table 6
Average assignment acceptability for different numbers of criteria.

Procedure	Absolute				Relative			
	3	5	7	9	3	5	7	9
UTADISMP1	0.8593	0.8163	0.7830	0.7549	−0.0540	−0.0909	−0.1248	−0.1545
UTADISMP2	0.8379	0.8030	0.7727	0.7471	−0.0774	−0.1058	−0.1365	−0.1634
UTADISMP3	0.8163	0.7867	0.7686	0.7541	−0.1018	−0.1243	−0.1419	−0.1570
UTADIS-JLS	0.8231	0.7886	0.7607	0.7340	−0.0943	−0.1215	−0.1494	−0.1776
CHEBYSHEV	0.8821	0.8635	0.8549	0.8476	−0.0283	−0.0380	−0.0446	−0.0517
MSCVF	0.7896	0.7280	0.6876	0.6592	−0.1272	−0.1932	−0.2417	−0.2761
ACUTADIS	0.8794	0.8509	0.8323	0.8169	−0.0313	−0.0518	−0.0694	−0.0853
CENTROID	0.9065	0.8960	0.8930	0.8918	−0.0009	−0.0012	−0.0013	−0.0015
REPDIS	0.8301	0.8019	0.7853	0.7745	−0.0863	−0.1069	−0.1225	−0.1334
CAI	0.9073	0.8970	0.8941	0.8931	0	−1.7E−07	0	−4.1E−07
APOI	0.9069	0.8966	0.8937	0.8927	−0.0005	−0.0005	−0.0005	−0.0006
COMB	0.9069	0.8967	0.8937	0.8927	−0.0004	−0.0004	−0.0005	−0.0005
ROBUST-ITER	0.8081	0.7561	0.7134	0.6823	−0.1114	−0.1586	−0.2031	−0.2359
ROBUST-COMP	0.7957	0.7405	0.7009	0.6717	−0.1250	−0.1758	−0.2167	−0.2477

analysis give less robust conclusions. In the case of CAI, however, these changes are much smaller than in the case of, e.g., ROBUST-COMP or MSCVF. For these two approaches and problems with 9 criteria, the mean relative loss to the optimum is nearly twice as large as for the problems with 3 criteria.

The general trend of decreasing absolute MCAIs with a greater number of characteristic points is visible in Table 7. However, it is not valid for all procedures. For the best-performing methods, including CAI, APOI, COMB, CHEBYSHEV, and CENTROID, it is inverse. For example, the absolute MCAIs for CAI are 0.8849, 0.8942, and 0.9145 for $\gamma_j = 2, 4, 6$. For the approaches exploiting the stochastic acceptabilities, one may interpret that more flexible MVFs offer greater chances for better fitting the models to reflect the CAIs and APOIs. Even though for the procedures identifying the Chebyshev and analytic centers, the absolute MCAIs increased when moving from linear to piecewise linear MVFs, their relative counterparts marginally deteriorated.

With a more significant number of reference alternatives per class, the trends of increasing absolute MCAI and decreasing loss to the most robust assignment are unanimously confirmed for all procedures (see Table 8). For example, for CHEBYSHEV, its absolute assignment acceptability increases from 0.8105 to 0.9044 when moving from $R = 3$ to 10, and its relative loss decreases from 0.0572 to 0.0277. With additional preference information, the entropy of class acceptability indices gets lower, and hence the feasible models become more similar in terms of the suggested sorting recommendations [25]. Consequently, irrespective of the applied procedure, the chances of selecting a model whose assignments are highly robust get higher.

The results of the linear regression analysis for the relative and absolute assignment acceptabilities are available in eAppendix 3.

4.3.3. Differences between marginal and comprehensive values and class thresholds

In this section, we discuss the results for the remaining three measures jointly because the underlying rankings are similar to a large extent. This is understandable because all measures concern the similarity between the models derived with different approaches and the reference model, even if they refer to its various components. We present the average differences between marginal and comprehensive values and class thresholds in Table 9. In addition, for the marginal values, we report the difference to an average solution obtained with CENTROID, in the space of all models consistent with DM's preferences.

For all these measures, we performed Wilcoxon signed-rank tests to investigate the statistical significance of the observed differences with a p-value of 0.05. Moreover, the coefficients of the influence of individual problem features on the results were determined by solving the linear regression problem. The tables with regression coefficients and the Hasse diagrams reflecting the relationships resulting from the statistical test outcomes are available in eAppendix 4.

Given all three measures, the most significant similarity to the reference model is observed for the outcomes of ACUTADIS, CENTROID, and CHEBYSHEV. For example, for ACUTADIS, the distance from the reference model in terms of marginal values is 0.0417; for comprehensive values – it is 0.0502, and for class thresholds – 0.0409. For the procedures identifying an average solution and the Chebyshev center, the measure values are only slightly higher. The distances of the function returned by ACUTADIS and CHEBYSHEV from the centroid solution are very low, suggesting that the three procedures return similar models. For UTADIS-JLS, which implements an analogous selection rule to

Table 7
Average assignment acceptability for different numbers of characteristic points.

Procedure	Absolute			Relative		
	2	4	6	2	4	6
UTADISMP1	0.8484	0.7966	0.7650	−0.0428	−0.1108	−0.1645
UTADISMP2	0.7949	0.8029	0.7728	−0.1032	−0.1034	−0.1557
UTADISMP3	0.7828	0.7711	0.7904	−0.1175	−0.1393	−0.1369
UTADIS-JLS	0.8460	0.7667	0.7171	−0.0457	−0.1444	−0.2171
CHEBYSHEV	0.8568	0.8540	0.8753	−0.0328	−0.0457	−0.0434
MSCVF		0.7160	0.7162		−0.2013	−0.2178
ACUTADIS	0.8646	0.8326	0.8374	−0.0237	−0.0698	−0.0848
CENTROID	0.8842	0.8931	0.9132	−0.0009	−0.0014	−0.0014
REPDIS	0.8199	0.7847	0.7893	−0.0753	−0.1235	−0.1379
CAI	0.8849	0.8942	0.9145	−3.2E−07	−1.1E−07	0
APOI	0.8844	0.8938	0.9142	−0.0007	−0.0005	−0.0003
COMB	0.8844	0.8939	0.9142	−0.0006	−0.0005	−0.0003
ROBUST-ITER	0.7972	0.7262	0.6965	−0.1020	−0.1902	−0.2396
ROBUST-COMP	0.7883	0.7105	0.6828	−0.1119	−0.2075	−0.2545

Table 8
Average assignment acceptability for different numbers of reference alternatives per class.

Procedure	Absolute				Relative			
	3	5	7	10	3	5	7	10
UTADISMP1	0.7170	0.7914	0.8344	0.8707	−0.1641	−0.1119	−0.0843	−0.0639
UTADISMP2	0.7040	0.7776	0.8211	0.8579	−0.1794	−0.1274	−0.0987	−0.0776
UTADISMP3	0.7110	0.7695	0.8052	0.8399	−0.1735	−0.1375	−0.1168	−0.0971
UTADIS-JLS	0.6785	0.7624	0.8102	0.8553	−0.2085	−0.1439	−0.1104	−0.0802
CHEBYSHEV	0.8105	0.8535	0.8797	0.9044	−0.0572	−0.0430	−0.0348	−0.0277
MSCVF	0.6232	0.7041	0.7479	0.7893	−0.2865	−0.2164	−0.1828	−0.1524
ACUTADIS	0.7939	0.8367	0.8618	0.8870	−0.0756	−0.0614	−0.0543	−0.0464
CENTROID	0.8575	0.8903	0.9102	0.9293	−0.0019	−0.0013	−0.0010	−0.0007
REPDIS	0.7288	0.7875	0.8213	0.8542	−0.1517	−0.1167	−0.0989	−0.0818
CAI	0.8591	0.8914	0.9111	0.9299	−5.3E−07	−5.3E−08	0	0
APOI	0.8580	0.8911	0.9110	0.9299	−0.0013	−0.0004	−0.0002	−0.0001
COMB	0.8581	0.8911	0.9110	0.9299	−0.0012	−0.0004	−0.0002	−0.0001
ROBUST-ITER	0.6480	0.7213	0.7704	0.8202	−0.2452	−0.1908	−0.1547	−0.1183
ROBUST-COMP	0.6378	0.7087	0.7581	0.8042	−0.2570	−0.2048	−0.1680	−0.1354

CENTROID, such a distance is higher. This is also reflected in more substantial differences from the DM's reference model. Therefore, it is apparent that averaging only extreme models does not lead to obtaining an *average* solution.

Favorable results in terms of differences between marginal and comprehensive values are attained with REPDIS. However, when considering the class thresholds, these differences are higher. It is intuitive because REPDIS does not optimize the threshold values, focusing only on selecting a representative value function. Still, REPDIS proves better in terms of the three measures than the remaining methods exploiting the stochastic acceptabilities. For example, for CAI, the distance from the reference model in terms of marginal values is 0.0642; for comprehensive values – it is 0.0916, and for class thresholds – 0.0818, being 1.5–2 times higher than for the best-rated ACUTADIS. This confirms that aiming to reproduce the most common results attained in the set of all compatible sorting models does not ideally allow replicating a single reference model.

The group of UTADISMP methods and ROBUST-ITER achieve intermediate results. In the case of UTADISMP1 and ROBUST-ITER, both procedures attain the same values in terms of distances built on marginal and comprehensive values. This is because they aim to identify the most discriminant models. While UTADISMP1 exploits only the DM's preference information, ROBUST-ITER refers to the necessary assignment-based preference relation in the set of all alternatives. However, this relation is heavily influenced by the DM's assignment examples because all reference alternatives from the more preferred classes are necessarily preferred to the alternatives assigned to the least preferred classes. Still, the necessary relation can be richer, involving pairs that are compared in the same way by all feasible models, even if this is not a direct consequence of the DM's statements. Also, the differences between these methods can be typically observed for the measure values related to class thresholds. This is because UTADISMP1 directly optimizes their values, while ROBUST-ITER is focused only on the parameters of the AVF. Furthermore, UTADISMP2 constructed models that are, on average, slightly more similar to the reference ones than UTADISMP1, whereas the similarity results for ROBUST-COMP are marginally worse than for its iterative counterpart.

UTADIS-JLS achieves a relatively good approximation of the marginal values compared to the reference and the centroid models. Therefore, averaging the extreme models can be beneficial if the primary aim is to understand how the DM evaluates particular criteria. However, this approach achieves the worst results for the similarity measures based on comprehensive values and thresholds.

Finally, the worst-performing methods include CAI, APOI, COMB, MSCVF, and ROBUST-COMP. For the three methods exploiting the acceptability indices, it can be concluded that although they usually obtain models that differ significantly from the reference one, they create consistent and representative solutions in the context of preference information provided by DM. In turn, MSCVF and ROBUST-COMP fail to offer satisfactory results given all analyzed aspects.

The differences between the reference and resulting models obtained for different values of each problem dimension (p, m, γ_j, and R) together with linear regression models and Hasse diagrams resulting from the Wilcoxon signed-rank tests are discussed in the eAppendix.

Table 9
Average value and standard deviation of differences between marginal values, comprehensive values, and class thresholds from the reference model (in case of marginal values, also the differences from the centroid model).

Procedure	Marginal values				Comprehensive values		Class thresholds	
	Reference		Centroid					
	mean	std	mean	std	mean	std	mean	std
UTADISMP1	0.0582	0.0439	0.0493	0.0412	0.0811	0.0709	0.0594	0.0602
UTADISMP2	0.0606	0.0412	0.0516	0.0396	0.0803	0.0679	0.0588	0.0572
UTADISMP3	0.0607	0.0362	0.0424	0.0269	0.0655	0.0367	0.0526	0.0440
UTADIS-JLS	0.0558	0.0403	0.0426	0.0330	0.0997	0.0887	0.1004	0.1024
CHEBYSHEV	0.0459	0.0313	0.0207	0.0174	0.0553	0.0375	0.0461	0.0436
MSCVF	0.0628	0.0384	0.0547	0.0312	0.0870	0.0421	0.0758	0.0551
ACUTADIS	0.0417	0.0286	0.0239	0.0158	0.0502	0.0337	0.0409	0.0395
CENTROID	0.0445	0.0291	0	0	0.0545	0.0380	0.0461	0.0449
REPDIS	0.0492	0.0357	0.0256	0.0262	0.0622	0.0519	0.0684	0.0680
CAI	0.0642	0.0375	0.0463	0.0256	0.0916	0.0801	0.0818	0.0877
APOI	0.0644	0.0377	0.0465	0.0256	0.0908	0.0790	0.0812	0.0868
COMB	0.0644	0.0377	0.0465	0.0256	0.0908	0.0790	0.0812	0.0868
ROBUST-ITER	0.0582	0.0439	0.0493	0.0412	0.0811	0.0709	0.0958	0.1023
ROBUST-COMP	0.0614	0.0454	0.0541	0.0425	0.0853	0.0741	0.0978	0.1035

5. Summary and future research

We considered preference disaggregation in the context of multiple criteria sorting. We assumed the classification is driven by an additive value function and thresholds separating the categories. The parameters of such a model are inferred from the Decision Maker's assignment examples. Using such indirect and incomplete preference information leads to infinitely many compatible sorting models, potentially implying different assignments for the non-reference alternatives. Given the multiplicity of feasible models, selecting a single, representative one can be conducted in different ways.

We reviewed several procedures for such a selection. They aim to identify the most discriminant, average, central, parsimonious, or robust model. These ideas differ regarding the exploited information and aspects to be emphasized that translate into the relevant constraints and an objective function. Our core contribution is proposing three novel procedures that assign the alternatives according to the robust classification rule. For this purpose, they exploit class acceptability indices and/or assignment-based pairwise acceptabilities and maximize the support given to the resulting assignments by all feasible sorting models. The use of all approaches, including the existing and novel ones, was illustrated in a study concerning the green performance assessment of European cities.

In the extensive experimental study, we compared the performance of all procedures on problem instances with different complexities. The results were quantified in terms of five measures. When it comes to reproducing the assignments generated by a simulated Decision Maker's model and the parameters of this model, involving marginal and comprehensive values as well as class thresholds, the best performers are the same. They include the procedures that determine a central sorting model with the proviso that it can be an analytic center, the Chebyshev center, or an average determined based on a large sample of compatible models. When it comes to approximating the unknown model parameters, favorable results were also attained by the most discriminant procedures.

We performed a related experiment whose initial results – derived from the analysis of a smaller set of problems – were not included in the paper due to their extremely high correlation in terms of the ranking of methods imposed by the classification accuracy. Namely, for each simulated DM's value function, we drew different sets of reference alternatives and verified how well each procedure performs, predicting the classification for the non-reference alternatives across different reference sets. Such a measure reflects how strongly the arbitrary choice of reference alternatives influences each method's performance. An observed high similarity of outcomes was expected. The results reported in the main paper for each problem setting were averaged across 100 instances with various performances and simulated DMs. With so many repetitions, whether we derive the mean predictive performance from analyzing various instances or the same instances with different reference sets does not influence the methods' average performance.

The novel approaches exploiting stochastic acceptabilities proved to be the best in emphasizing the robustness of results in a univocal recommendation. This is, however, at the increased computational cost related to conducting robustness analysis for all alternatives and solving a more challenging optimization problem. Overall, the results returned by the three procedures were highly similar, with CAI attaining only slightly better results than APOI and COMB. Even if the idea underlying these methods is very alike, they exploit various results and optimize different objectives. Hence such a high similarity or even the same outcomes delivered for most problems could not be predicted before conducting the experiments. Based on the obtained experimental results, we recommend using CAI when the DM focuses on robustness.

The favorable performance of CAI, APOI, and COMB in ensuring high robustness could have been anticipated. Their aim consists in providing the best possible representation of the recommendations feasible in the entire space of compatible sorting models. Given that each of them is consistent with the DM's incomplete preference information and could have served as the reference model, high robustness is also related to average high classification accuracy. Indeed, the experiments confirmed the novel procedures attained favorable results in terms of classification accuracy.

The center-oriented procedures also achieved high robustness of results. Moreover, CENTROID and ACUTADIS attained higher predictive accuracy than CAI. Among them, CENTROID ensures lower computational costs, requiring no optimization. Finally, focusing only on the shape of marginal value functions, as in the UTADISMP methods, or exploiting the exact, necessary results, as in the ROBUST approaches, did not lead to favorable outcomes given any considered measure.

The favorable performance of center-oriented procedures regarding predictive accuracy and robustness can be explained by referring to the model similarity measures. These methods returned models very close to the reference ones in terms of comprehensive and marginal values and class thresholds. In this perspective, they can be considered the best candidates for default choices in the applications of UTADIS-like methods when the DM cares about all measures. On the contrary, the models constructed by CAI, APOI, and COMB were among the least similar to the reference ones. Specifically, they were in the bottom four for value-based measures and the bottom half of the ranking for threshold-oriented metrics. This aspect is less relevant for the practice of decision-aiding. However, it confirms that even if the parameter values of various models differ significantly, they may nevertheless provide similar recommendations based on the incomplete DM's preferences. This emphasizes the importance of informativeness and trustworthiness of the information supplied at the method's input. Also, when solving real-world problems, the form of the abstract DM's model is unknown, making the comparison with its values infeasible. The above speaks in favor of using the novel procedures proposed in this paper as they ensure high predictive capabilities and results' robustness despite not reproducing the reference parameter values closely.

When limiting the analysis of results only to the four approaches considered in [44], our findings are largely consistent. The statistically significant rankings reported in [44] given both the classification accuracy and the assignments' acceptability indicate the following order: ACUTADIS (called analytic center in [44]), CHEBYSHEV (Chebyshev center), UTADISMP1 (max–min), and UTADIS-JLS (post-optimality). We confirmed the same ranking given the prediction performance and observed inverse positions, with marginal differences, only for the center-based approaches when it comes to the recommendation robustness.

The experimental study indicated that the classification accuracy of procedures and assignment acceptability of their recommendation decreased with more classes, criteria, and characteristic points and fewer reference alternatives per class. These outcomes can be justified given a more significant challenge posed by the classification problems with more classes, higher flexibility of a preference model with more criteria and breakpoints, and greater information gain offered by additional assignment examples. This is consistent with the findings reported in [44,56] given both performance trends and the positive association between predictive performance and recommendation robustness.

The average differences between the reference and delivered models given values of parameters such as marginal values assigned to particular characteristic points, alternatives' comprehensive values, or class thresholds exhibit slightly different trends. They become lower with more classes (also implying more assignment examples) and reference assignments per class and higher with more characteristic points. Regarding the impact of the number of criteria, the observed regularities were unclear and differed from one approach to another.

Due to a broader range of parameter values considered in this paper compared to [44], we gained more insights into how they affect the reported measures. In most cases, the trend of change in values of all measures was non-linear with respect to considered values of different dimensions. Specifically, greater modifications were observed in the lower scale range of different parameters of a decision problem or a sorting model (e.g., when passing from 2 to 3 classes, from 3 to 5 criteria, from 2 to 4 characteristic points, of from 3 to 5 reference assignments per class). In turn, the differences in the upper parts of the parameter scales were lesser (e.g., when passing from 4 to 5 classes, from 7 to 9 criteria, from 4 to 6 characteristic points, of from 7 to 10 reference assignments per class).

We envisage the following directions for future research. Firstly, in this paper, we focused only on analyzing procedures for selecting a representative sorting model in case of compatibility with the DM's preference information. However, it would be useful to extend the study in terms of both simulating artificial DMs' policies with the models that do not ensure such a compatibility (e.g., by allowing some errors in making the classifications suggested by a simulated model) as well as considering procedures that are specifically oriented toward selecting a representative model in case of inconsistency [7,41]. This would reflect how much each method is influenced by the errors and inconsistencies and even quantify if it can correct them. Nonetheless, this requires studying a completely different set of methods handling inconsistency between the provided preferences and an assumed model. In the same spirit, it would be interesting to verify the conclusions for other preference models (e.g., the Choquet integral [57,58]) or uncertain preference information [18,59].

Second, when generating the assignment examples in the experimental study, we simulated realistic scenarios in which extreme classes were less common than intermediate ones. We could also consider other distributions, e.g., assuming that all classes are represented by the same number of reference alternatives or just simulating a certain number of assignments without influencing their distribution and hence tolerating very unbalanced ones. However, each setting requires a separate report spanning tens of pages. Also, our initial experiments on a limited set of instances confirmed that even if the absolute values of classification accuracy or robustness-oriented measure differ, the relative rankings of methods are not influenced.

Third, it would be interesting to design the procedures compromising between deriving central and robust models. This would allow them to score well in classification accuracy, reproducing the unknown DM's model, and in the support given to their recommendation in the set of all compatible models. Also, we may construct the most robust recommendation without exhibiting any feasible model and including constraints related to using a threshold-based value-driven sorting procedure as in the approaches proposed in this paper. In this way, we can investigate how restrictive the underlying model assumptions are in constructing a robust recommendation.

Finally, a similar review and results of an experimental study in the context of multiple criteria ranking can be found in [50]. The variety of procedures applicable in this context is wider, including procedures to construct a representative value function, decision rules, scoring methods, and approaches for constructing the most robust recommendation without exhibiting the model.

CRediT authorship contribution statement

Michał Wójcik: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Miłosz Kadziński:** Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Krzysztof Ciomek:** Writing – review & editing, Methodology, Conceptualization.

Data availability

Data is available in the paper.

Acknowledgments

Miłosz Kadziński and Michał Wójcik acknowledge financial support from the Polish National Science Center under the SONATA BIS project (grant no. DEC-2019/34/E/HS4/00045).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.knosys.2023.110871.

References

[1] C. Zopounidis, M. Doumpos, Multicriteria classification and sorting methods: a literature review, European J. Oper. Res. 138 (2002) 229–246.
[2] S. Greco, M. Ehrgott, J. Figueira, Multiple Criteria Decision Analysis: State of the Art Surveys, Springer New York, NY, 2016.
[3] P.A. Alvarez, A. Ishizaka, L. Martinez, Multiple-criteria decision-making sorting methods: A survey, Expert Syst. Appl. 183 (2021) 115368.
[4] M. Doumpos, C. Zopounidis, Disaggregation approaches for multicriteria classification: An overview, in: N. Matsatsinis, E. Grigoroudis (Eds.), Preference Disaggregation in Multiple Criteria Decision Analysis: Essays in Honor of Yannis Siskos, Springer, Cham, 2018, pp. 77–94.
[5] J. Devaud, G. Groussaud, E. Jacquet-Lagrèze, UTADIS: Une méthode de construction de fonctions d'utilité additives rendant compte de jugements globaux, in: EURO Working Group on MCDA, Bochum, Germany, 1980.
[6] E. Jacquet-Lagrèze, Y. Siskos, Preference disaggregation: 20 years of MCDA experience, European J. Oper. Res. 130 (2) (2001) 233–245.
[7] C. Zopounidis, M. Doumpos, PREFDIS: a multicriteria decision support system for sorting decision problems, Comput. Oper. Res. 27 (7–8) (2000) 779–797.
[8] A. Dimitras, Evaluation of greek construction companies' securities using UTADIS method, Eur. Res. Stud. J. 5 (1–2) (2002) 1–95.
[9] Y. Siskos, E. Grigoroudis, N. Matsatsinis, UTA methods, in: J. Figueira, S. Greco, M. Ehrgott (Eds.), Multiple Criteria Decision Analysis: State of the Art Surveys, Springer Verlag, Boston, Dordrecht, London, 2005, pp. 297–344.
[10] M. Kadziński, M. Cinelli, K. Ciomek, S. Coles, M. Nadagouda, R. Varma, K. Kirwan, Co-constructive development of a green chemistry-based model for the assessment of nanoparticles synthesis, European J. Oper. Res. 264 (2) (2018) 472–490.
[11] C. Zopounidis, M. Doumpos, A preference disaggregation decision support system for financial classification problems, European J. Oper. Res. 130 (2) (2001) 402–413.
[12] E.D. Manshadi, M.R. Mehregan, H. Safari, Supplier classification using UTADIS method based on performance criteria, Int. J. Acad. Res. Bus. Soc. Sci. 5 (2) (2015) 31–45.
[13] R.P. Palha, A.T. de Almeida, L.H. Alencar, A model for sorting activities to be outsourced in civil construction based on ROR-UTADIS, Math. Probl. Eng. 2016 (2016) 9236414.
[14] S. Greco, V. Mousseau, R. Słowiński, Multiple criteria sorting with a set of additive value functions, European J. Oper. Res. 207 (4) (2010) 1455–1470.
[15] M. Doumpos, K. Kosmidou, G. Baourakis, C. Zopounidis, Credit risk assessment using a multicriteria hierarchical discrimination approach: A comparative analysis, European J. Oper. Res. 138 (2) (2002) 392–412.
[16] M. Köksalan, S. Bilgin Özpeynirci, An interactive sorting method for additive utility functions, Comput. Oper. Res. 36 (9) (2009) 2565–2572.
[17] M. Kadziński, T. Tervonen, Stochastic ordinal regression for multiple criteria sorting problems, Decis. Support Syst. 55 (1) (2013) 55–66.
[18] Z. Ru, J. Liu, M. Kadziński, X. Liao, Probabilistic ordinal regression methods for multiple criteria sorting admitting certain and uncertain preferences, European J. Oper. Res. 311 (2) (2023) 596–616.
[19] V. Mousseau, L.C. Dias, J. Figueira, Dealing with inconsistent judgments in multiple criteria sorting models, 4OR 4 (2) (2006) 145–158.
[20] M. Kadziński, M. Ghaderi, M. Dabrowski, Contingent preference disaggregation model for multiple criteria sorting problem, European J. Oper. Res. 281 (2) (2020) 369–387.
[21] J. Liu, M. Kadziński, X. Liao, Modeling contingent decision behavior: A Bayesian nonparametric preference-learning approach, INFORMS J. Comput. 34 (4) (2023) 764–785.
[22] J. Liu, M. Kadziński, X. Liao, X. Mao, Y. Wang, A preference learning framework for multiple criteria sorting with diverse additive value models and valued assignment examples, European J. Oper. Res. 286 (3) (2020) 963–985.
[23] J. Liu, X. Liao, M. Kadziński, R. Słowiński, Preference disaggregation within the regularization framework for sorting problems with multiple potentially non-monotonic criteria, European J. Oper. Res. 276 (3) (2019) 1071–1089.
[24] K. Martyn, M. Kadziński, Deep preference learning for multiple criteria decision analysis, European J. Oper. Res. 305 (2) (2023) 781–805.
[25] M. Kadziński, K. Ciomek, Active learning strategies for interactive elicitation of assignment examples for threshold-based multiple criteria sorting, European J. Oper. Res. 293 (2) (2021) 658–680.
[26] S. Sun, H. Liao, Value-driven multiple criteria sorting with probabilistic linguistic information considering uncertain assignment examples, Int. J. Inf. Technol. Decis. Mak. 21 (01) (2022) 83–107.
[27] M. Kadziński, K. Ciomek, R. Słowiński, Modeling assignment-based pairwise comparisons within integrated framework for value-driven multiple criteria sorting, European J. Oper. Res. 241 (3) (2015) 830–841.
[28] V. Mousseau, L. Dias, J. Figueira, On the notion of category size in multiple criteria sorting models, in: Cahier Du LAMSADE 205, Université Paris-Dauphine, Paris, France, 2003.
[29] S. Özpeynirci, Ö. Özpeynirci, V. Mousseau, An interactive algorithm for multiple criteria constrained sorting problem, Ann. Oper. Res. 267 (1) (2018) 447–466.
[30] J. Liu, M. Kadziński, X. Liao, X. Mao, Data-driven preference learning methods for value-driven multiple criteria sorting with interacting criteria, INFORMS J. Comput. 33 (2) (2021) 586–606.
[31] M. Guo, X. Liao, J. Liu, A progressive sorting approach for multiple criteria decision aiding in the presence of non-monotonic preferences, Expert Syst. Appl. 123 (2019) 1–17.
[32] M. Kadziński, K. Martyn, M. Cinelli, R. Słowiński, S. Corrente, S. Greco, Preference disaggregation for multiple criteria sorting with partial monotonicity constraints: Application to exposure management of nanomaterials, Internat. J. Approx. Reason. 117 (2020) 60–80.
[33] M. Kadziński, K. Martyn, M. Cinelli, R. Słowiński, S. Corrente, S. Greco, Preference disaggregation method for value-based multi-decision sorting problems with a real-world application in nanotechnology, Knowl.-Based Syst. 218 (2021) 106879.
[34] M. Esmaelian, H. Shahmoradi, F. Nemati, P-UTADIS: A multi criteria classification method, Curr. Future Dev. Artif. Intell. 1 (1) (2017) 214–267.
[35] S. Corrente, M. Doumpos, S. Greco, R. Słowiński, C. Zopounidis, Multiple criteria hierarchy process for sorting problems based on ordinal regression with additive value functions, Ann. Oper. Res. 251 (1) (2017) 117–139.
[36] A. Ulucan, K. Atici, A multiple criteria sorting methodology with multiple classification criteria and an application to country risk evaluation, Technol. Econ. Dev. Econ. 19 (1) (2013) 93–124.
[37] F.-L. Cai, X. Liao, K.-L. Wang, An interactive sorting approach based on the assignment examples of multiple decision makers with different priorities, Ann. Oper. Res. 197 (1) (2012) 87–108.
[38] J. Liu, X. Liao, J.-B. Yang, A group decision-making approach based on evidential reasoning for multiple criteria sorting problem with uncertainty, European J. Oper. Res. 246 (3) (2015) 858–873.
[39] S. Greco, M. Kadziński, V. Mousseau, R. Słowiński, Robust ordinal regression for multiple criteria group decision problems: UTA^{GMS}-GROUP and UTADIS^{GMS}-GROUP, Decis. Support Syst. 52 (3) (2012) 549–561.
[40] S. Greco, M. Kadziński, R. Słowiński, Selection of a representative value function in robust multiple criteria sorting, Comput. Oper. Res. 38 (11) (2011) 1620–1637.
[41] M. Beuthe, G. Scannella, Comparative analysis of UTA multicriteria methods, European J. Oper. Res. 130 (2) (2001) 246–262.
[42] J. Branke, S. Greco, R. Słowiński, P. Zielniewicz, Learning value functions in interactive evolutionary multiobjective optimization, IEEE Trans. Evol. Comput. 19 (1) (2015) 88–102.
[43] G. Bous, P. Fortemps, F. Glineur, M. Pirlot, ACUTA: A novel method for eliciting additive value functions on the basis of holistic preference statements, European J. Oper. Res. 206 (2) (2010) 435–444.
[44] M. Doumpos, C. Zopounidis, E. Galariotis, Inferring robust decision models in multicriteria classification problems: An experimental analysis, European J. Oper. Res. 236 (2) (2014) 601–611.
[45] E. Jacquet-Lagrèze, Y. Siskos, Assessing a set of additive utility functions for multicriteria decision making: the UTA method, European J. Oper. Res. 10 (1982) 151–164.
[46] T. Tervonen, J. Figueira, R. Lahdelma, J. Almeida-Dias, P. Salminen, A stochastic method for robustness analysis in sorting problems, European J. Oper. Res. 192 (1) (2009) 236–242.
[47] R.L. Keeney, H. Raiffa, Decisions with Multiple Objectives: Preferences and Value Trade-Offs, Cambridge University Press, 1993.
[48] R.L. Smith, Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions, Oper. Res. 32 (6) (1984) 1296–1308.
[49] K. Ciomek, M. Kadziński, Polyrun: A Java library for sampling from the bounded convex polytopes, SoftwareX 13 (2021) 100659.
[50] M. Kadziński, M. Wójcik, K. Ciomek, Review and experimental comparison of ranking and choice procedures for constructing a univocal recommendation in a preference disaggregation setting, Omega 113 (2022) 102715.
[51] S. Greco, V. Mousseau, R. Słowiński, Parsimonious preference models for robust ordinal regression, in: EURO Working Group on MCDA, Yverdon, Switzerland, 2011.
[52] S.G. Arcidiacono, S. Corrente, S. Greco, Scoring from pairwise winning indices, Comput. Oper. Res. 157 (2023) 106268.
[53] R. Vetschera, Deriving rankings from incomplete preference information: A comparison of different approaches, European J. Oper. Res. 258 (1) (2017) 244–253.
[54] EIU, Democracy index 2019, A Year of Democratic Setbacks and Popular Protest, Economist Intelligence Unit, London, 2019.

[55] M. Hollander, D.A. Wolfe, E. Chicken, Nonparametric Statistical Methods, Volume 751, John Wiley & Sons, 2013.
[56] R. Vetschera, Y. Chen, K.W. Hipel, D. Marc Kilgour, Robustness and information levels in case-based multiple criteria sorting, European J. Oper. Res. 202 (3) (2010) 841–852.
[57] S.G. Arcidiacono, S. Corrente, S. Greco, Robust stochastic sorting with interacting criteria hierarchically structured, European J. Oper. Res. 292 (2) (2021) 735–754.
[58] H. Liao, Q. Yang, X. Wu, Customer preference analysis from online reviews by a 2-additive choquet integral-based preference disaggregation model, Technol. Econ. Dev. Econ. 29 (2023) 411–437.
[59] X. Wu, H. Liao, Value-driven preference disaggregation analysis for uncertain preference information, Omega 115 (2023) 102793.

Supplementary material [P2]

Selection of a representative sorting model in a preference disaggregation setting: a review of existing procedures, new proposals, and experimental comparison eAppendix

Michał Wójcik[a], Miłosz Kadziński[a], Krzysztof Ciomek[a]

[a]*Poznan University of Technology, Faculty of Computing and Telecommunications, Piotrowo 2, 60-965 Poznań, Poland*

Abstract

The eAppendix contains additional material not included in the main paper. First, we report the results of the illustrative study demonstrating the use of fourteen procedures on the same problem. Then, we discuss the values of the performance measures for selected procedures based on the results of an illustrative study. Furthermore, we present the results of the linear regression analysis for the relative and absolute assignment acceptabilities. Finally, we elaborate on the results concerning all measures quantifying the similarity between the models derived with different approaches and the reference model.

Keywords: Multiple criteria decision aiding, Preference disaggregation, Sorting, Representative model, Robustness analysis, Computational study

1. Illustrative study

To illustrate how the procedures for selecting a representative sorting model work, we consider an example problem concerning the evaluation of 30 major European cities in implementing green policy. In the considered study, each city is rated in terms of the following four criteria: CO_2 emissions (g_1), energy consumption (g_2), water management (g_3), and waste and land use (g_4). The performances on the scale between 0 and 10 were determined by considering various indicators. They are given in Table 1. We will employ UTADIS to assign the cities to three classes: C_1, C_2, and C_3, where C_3 is the most preferred category. We assume that a marginal function for each criterion has three characteristic points ($\gamma_j = 3$ for $j = 1, \ldots, 4$). Moreover, they are defined over the $[0, 10]$ range, and thus $\beta_j^1 = 0$, $\beta_j^2 = 5$, and $\beta_j^3 = 10$.

Then, we drew three reference alternatives for each class to form the DM's indirect preference supplied as the input for UTADIS: $a_{15}, a_{20}, a_{27} \rightarrow C_1$, $a_7, a_{18}, a_{19} \rightarrow C_2$, and $a_1, a_8, a_{10} \rightarrow C_3$. To simulate the DM's policy, we randomly selected an additive value function with marginal functions depicted in Figure 1. All alternatives were assessed given this function (see Table 1 for marginal and comprehensive values). Subsequently, a pair of thresholds ($t_1 = 0.3977$ and $t_2 = 0.6543$) was selected to delimit the three preference-ordered classes, and derive the assignments with the DM's reference model (see Table 1). The threshold values were chosen randomly in such value ranges, which guaranteed that each class received ten alternatives. The reference alternatives are marked in red, and their labels are provided under the axis in Figure 2. Some procedures discussed in the previous section make use of robust results. In particular, we employed Hit-And-Run (HAR) for deriving CAI's (see Table 1) and $APWI'$s (see Table 2). They were computed based on $10,000$ compatible sorting models.

In what follows, we discuss the results obtained with 14 procedures for selecting a representative sorting model. The respective MVFs are illustrated in Figure 3. For precise marginal values assigned to the characteristic points and class thresholds, see Table 3. Tables 4 and 5 show the comprehensive values and class assignments determined with all approaches. To save space, we provide detailed results only for nine non-reference cities, for which at least one method recommended a class that differed from the one assigned by the reference model. For the remaining twelve alternatives, all 14 procedures recommended an assignment compatible with the indication of the reference model.

Email addresses: michal.wojcik@cs.put.poznan.pl (Michał Wójcik), milosz.kadzinski@cs.put.poznan.pl (Miłosz Kadziński), k.ciomek@gmail.com (Krzysztof Ciomek)

Table 1: Evaluation of decision alternatives (cities) on four criteria, their marginal and comprehensive values according to a reference model, and Class Acceptability Indices $CAI'(a_i, C_l)$ for all alternatives and classes.

	Performances				Reference values					Class acceptabilities		
Alternative	g_1	g_2	g_3	g_4	u_1	u_2	u_3	u_4	$U(a)$	C_1	C_2	C_3
a_1 (Oslo)	9.58	8.71	6.85	8.23	0.0691	0.2384	0.1896	0.1992	0.6963	0.000	0.000	1.000
a_2 (Stockholm)	8.99	7.61	7.14	7.99	0.0616	0.2240	0.2074	0.1872	0.6802	0.000	0.029	0.971
a_3 (Zurich)	8.48	6.92	8.88	8.82	0.0552	0.2149	0.3142	0.2286	0.8129	0.000	0.000	1.000
a_4 (Copenhagen)	8.35	8.69	8.88	8.05	0.0535	0.2381	0.3142	0.1902	0.7960	0.000	0.000	1.000
a_5 (Brussels)	8.32	6.19	9.05	7.26	0.0532	0.2054	0.3246	0.1508	0.7339	0.000	0.000	1.000
a_6 (Paris)	7.81	4.66	8.55	6.72	0.0467	0.1769	0.2939	0.1239	0.6414	0.000	0.056	0.944
a_7 (Rome)	7.57	6.40	6.88	5.96	0.0437	0.2081	0.1915	0.0859	0.5292	0.000	1.000	0.000
a_8 (Vienna)	7.53	7.76	9.13	8.60	0.0432	0.2259	0.3295	0.2177	0.8163	0.000	0.000	1.000
a_9 (Madrid)	7.51	5.52	8.59	5.85	0.0429	0.1966	0.2964	0.0804	0.6163	0.000	0.124	0.876
a_{10} (London)	7.34	5.64	8.58	7.16	0.0408	0.1982	0.2958	0.1458	0.6805	0.000	0.000	1.000
a_{11} (Helsinki)	7.30	4.49	7.92	8.69	0.0403	0.1704	0.2553	0.2222	0.6881	0.000	0.200	0.800
a_{12} (Amsterdam)	7.10	7.08	9.21	8.98	0.0378	0.2170	0.3344	0.2366	0.8258	0.000	0.000	1.000
a_{13} (Berlin)	6.75	5.48	9.12	8.63	0.0334	0.1961	0.3289	0.2192	0.7775	0.000	0.000	1.000
a_{14} (Ljubljana)	6.67	2.23	4.19	5.95	0.0323	0.0846	0.0638	0.0854	0.2662	1.000	0.000	0.000
a_{15} (Riga)	5.55	3.53	6.43	5.72	0.0182	0.1340	0.1639	0.0739	0.3900	1.000	0.000	0.000
a_{16} (Istanbul)	4.86	5.55	5.59	4.86	0.0110	0.1970	0.1124	0.0370	0.3573	1.000	0.000	0.000
a_{17} (Athens)	4.85	4.94	7.26	5.33	0.0109	0.1875	0.2148	0.0545	0.4677	0.088	0.912	0.000
a_{18} (Budapest)	4.85	2.43	6.97	6.27	0.0109	0.0922	0.1970	0.1014	0.4016	0.000	1.000	0.000
a_{19} (Dublin)	4.77	4.55	7.14	6.38	0.0108	0.1727	0.2074	0.1069	0.4978	0.000	1.000	0.000
a_{20} (Warsaw)	4.65	5.29	4.90	5.17	0.0105	0.1936	0.0747	0.0465	0.3252	1.000	0.000	0.000
a_{21} (Bratislava)	4.54	4.19	7.65	5.60	0.0102	0.1590	0.2387	0.0680	0.4759	0.001	0.999	0.000
a_{22} (Lisbon)	4.05	5.77	5.42	5.34	0.0091	0.1999	0.1019	0.0550	0.3659	1.000	0.000	0.000
a_{23} (Vilnius)	3.91	2.39	7.71	7.31	0.0088	0.0907	0.2424	0.1533	0.4952	0.000	0.935	0.065
a_{24} (Bucharest)	3.65	3.42	4.07	3.62	0.0082	0.1298	0.0620	0.0275	0.2276	1.000	0.000	0.000
a_{25} (Prague)	3.44	3.26	8.39	6.30	0.0078	0.1237	0.2841	0.1029	0.5185	0.000	0.849	0.151
a_{26} (Tallinn)	3.40	1.70	7.90	6.15	0.0077	0.0645	0.2541	0.0954	0.4216	0.008	0.991	0.001
a_{27} (Zagreb)	3.20	4.34	4.43	4.04	0.0072	0.1647	0.0675	0.0307	0.2701	1.000	0.000	0.000
a_{28} (Belgrade)	3.15	4.65	3.90	4.30	0.0071	0.1765	0.0594	0.0327	0.2757	1.000	0.000	0.000
a_{29} (Sofia)	2.95	2.16	1.83	3.32	0.0067	0.0820	0.0279	0.0252	0.1418	1.000	0.000	0.000
a_{30} (Kiev)	2.49	1.50	5.96	1.43	0.0056	0.0569	0.1351	0.0109	0.2085	1.000	0.000	0.000

Table 2: Part of the matrix with the $APWI'$ values.

a_i \ a_j	...	a_9	a_{10}	a_{11}	...	a_{24}	a_{25}	a_{26}	...
...	...	...	...	...	...	...	...	...	...
a_9	...	0.000	0.000	0.171	...	1.000	0.725	0.880	...
a_{10}	...	0.124	0.000	0.200	...	1.000	0.849	0.999	...
a_{11}	...	0.095	0.000	0.000	...	1.000	0.649	0.800	...
...	...	...	...	...	...	...	...	...	...
a_{24}	...	0.000	0.000	0.000	...	0.000	0.000	0.000	...
a_{25}	...	0.000	0.000	0.000	...	1.000	0.000	0.158	...
a_{26}	...	0.000	0.000	0.000	...	0.992	0.000	0.000	...
...	...	...	...	...	...	...	...	...	...

The UTADISMP1 method aims at reproducing the DM's preferences by maximizing the difference between comprehensive values of reference alternatives and the thresholds of their desired classes. This objective implies that due to the existence of reference alternatives with comprehensive values close to the thresholds (e.g., a_{15} and a_{18}), the resulting marginal functions and class thresholds differ from the reference ones. Indeed, maximizing the difference between these values and thresholds often requires assigning positive marginal values only in the last segments (see Figure 3). Moreover, the operational procedure underlying UTADISMP1 implies that non-reference alternatives with very similar performance profiles to reference alternatives are assigned to the same class. This can be observed for, e.g., a_{19} and a_{17} or a_6 and a_{10}.

The models obtained with UTADISMP2 and UTADISMP3 are the same. This is understandable since both procedures account for maximizing the minimal slope of MVFs, while UTADISMP2 additionally considers the same objective as UTADISMP1. The evidence of maximizing the differences between marginal values assigned to successive characteristic points is visible in Table 3. For all criteria, $u_j(5)$ has the same value (0.10763), and in three cases, $u_j(10)$ is exactly twice as large (0.21526), hence satisfying the monotonicity constraints with a large margin (ρ). In this case, the slacks for other constraints were rather marginal. For example, comprehensive values of two reference alternatives $U(a_{15}) = 0.4770076$ and $U(a_{18}) = 0.4770096$ are very close to threshold $t_1 = 0.4770086$, though being assigned to different classes: C_1 and C_2, respectively. A characteristic consequence of maximizing ρ is that for many problems, the solutions obtained by these two methods have a relatively even distribution of the maximal values of MVFs and their curvatures are close to being

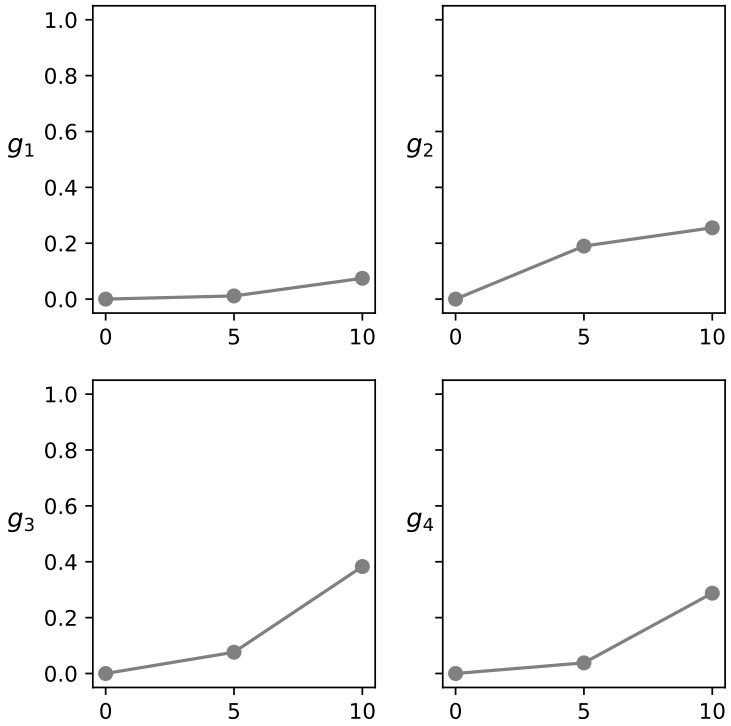


Figure 1: Marginal value functions in the reference model.

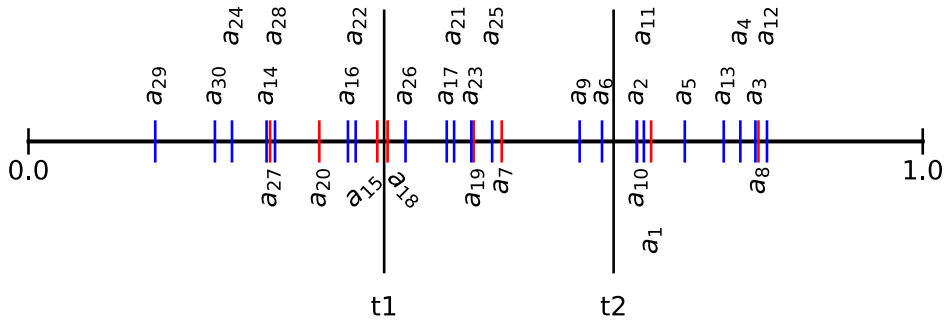


Figure 2: Comprehensive values for all alternatives in relation to class thresholds in the reference model.

linear.

An explicit mechanism for deriving the marginal functions which minimally deviate from linearity is implemented in MSCVF. For the considered problem, it obtained an ideal model, satisfying the following condition: $\forall j \in \{1,2,3,4\}$: $\frac{u_j(10)-u_j(5)}{10-5} = \frac{u_j(5)-u_j(0)}{5-0}$ for all criteria, which translated to the lowest possible objective function's value ($\phi = 0$). The linear MVFs are visible in Figure 3. Obviously, attaining such parsimony is not possible for all problems as it depends on the alternatives' performances and reference assignments.

UTADIS-JLS is a heuristic approach that constructs a representative model by averaging the extreme compatible ones that maximize and minimize the greatest value of the individual MVFs. For the considered problem, this led to non-negligible maximal shares of all MVFs with the predominant role of g_3 ($u_3(10) = 0.3571$) and g_4 ($u_4(10) = 0.3779$) and well-distributed class thresholds ($t_1 = 0.3999624$ and $t_2 = 0.603032$). Interestingly, in the final model, $u_4(5) = 0.0$, which means that this marginal value was equal to zero in the eight intermediate models. Analyzing the results obtained with other methods, many solutions repeat this pattern. This suggests that low scores (below 5.0) w.r.t. waste and land use (g_4) may have no or negligible impact on the recommended class assignments. As a result, Kiev ($g_4(a_{30}) = 1.43$) and Istanbul ($g_4(a_{16}) = 4.86$) are often scored equally on u_4, despite a noticeable difference in their performances.

The CENTROID method is similar to UTADIS-JLS in terms of deriving an average model. However, when doing so, it considers a large sample of uniformly distributed models. The marginal value functions obtained with CENTROID confirm that the extreme models considered by UTADIS-JLS are not representative of the entire feasible polyhedron. In

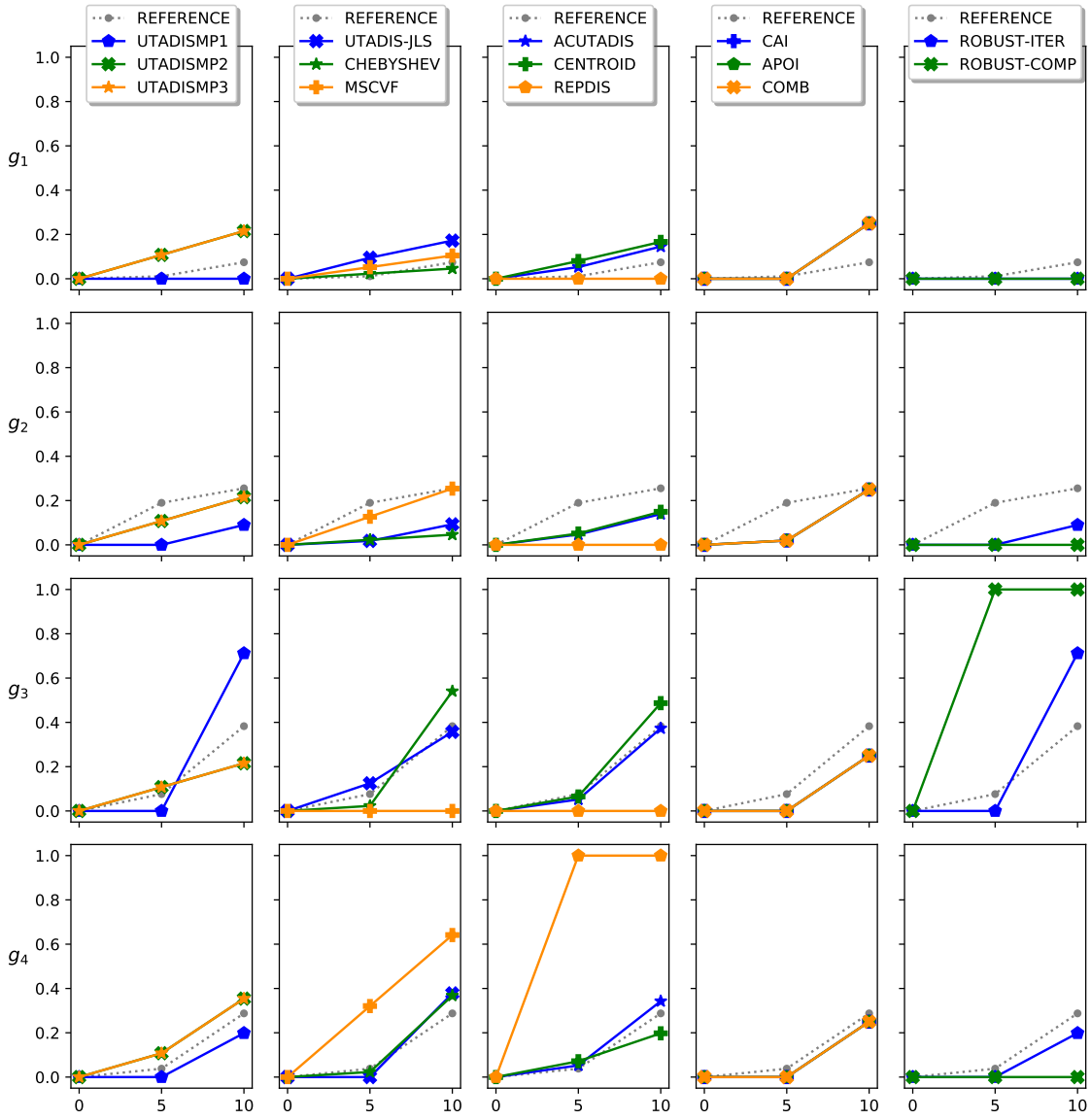


Figure 3: Marginal value function obtained with 14 procedures for selecting a representative sorting model.

particular, the maximal shares of u_2 and u_3 are greater, whereas the impact of u_4 is reduced (even though $u_4(5)$ is not zeroed in this case). A detailed analysis of the derived model confirms that incomplete indirect preference information (in this case, concerning 9 out of 30 alternatives) does not allow for reproducing the reference model accurately, even if the assignment examples are perfectly reproduced. When comparing the two models in Figure 3, one can observe the overestimation of the maximum value for g_1 ($u_{1_{REFERENCE}}(10) = 0.0744$ and $u_{1_{CENTROID}}(10) = 0.1663$) and g_3 ($u_{3_{REFERENCE}}(10) = 0.3829$ and $u_{3_{CENTROID}}(10) = 0.4868$) and the underestimation for g_2 ($u_{2_{REFERENCE}}(10) = 0.2553$ and $u_{2_{CENTROID}}(10) = 0.1494$) and g_4 ($u_{4_{REFERENCE}}(10) = 0.2875$ and $u_{4_{CENTROID}}(10) = 0.1975$).

In the CHEBYSEV method, the "central" model is determined in a more formalized way as the center of the hypersphere inscribed in the polyhedron defining the set of all compatible sorting models. For this purpose, the constraints incorporate variable r representing the value of the hypersphere radius. The obtained MVFs are similar to those obtained with UTADISMP2 in the sense of assigning the same marginal values to mid-points on all criteria ($u_j(5)$). Also, the values assigned to the end points $u_1(10)$ and $u_2(10)$ are exactly twice as large. This is due to optimizing variable r, which is responsible for maximizing the minimal differences between marginal values assigned to successive characteristic points. In addition, this variable is also used in constraints reproducing the class assignments, as the hypersphere radius depends on these constraints too. As a result, the comprehensive values of reference alternatives also highly diverge from the thresholds, which is mainly attained thanks to high maximal shares on u_3 and u_4.

In the same spirit, ACUTADIS derives a central model corresponding to an analytic center of the polyhedron. The

Table 3: Marginal values assigned to characteristic points and class thresholds obtained by 14 procedures for selecting a representative sorting model.

Method	$u_1(5)$	$u_1(10)$	$u_2(5)$	$u_2(10)$	$u_3(5)$	$u_3(10)$	$u_4(5)$	$u_4(10)$	t_1	t_2
REFERENCE	**0.0113**	**0.0744**	**0.1898**	**0.2553**	**0.0762**	**0.3829**	**0.0380**	**0.2875**	**0.397722**	**0.654317**
UTADISMP1	0.0000	0.0000	0.0000	0.0897	0.0000	0.7118	0.0000	0.1985	0.281516	0.408793
UTADISMP2	0.1076	0.2153	0.1076	0.2153	0.1076	0.2153	0.1076	0.3542	0.477009	0.678257
UTADISMP3	0.1076	0.2153	0.1076	0.2153	0.1076	0.2153	0.1076	0.3542	0.477009	0.678257
UTADIS-JLS	0.0948	0.1722	0.0177	0.0927	0.1250	0.3571	0.0000	0.3779	0.399624	0.603032
CHEBYSHEV	0.0229	0.0459	0.0229	0.0459	0.0229	0.5405	0.0229	0.3678	0.328974	0.500873
MSCVF	0.0524	0.1047	0.1270	0.2540	0.0000	0.0000	0.3206	0.6413	0.514602	0.624040
ACUTADIS	0.0527	0.1453	0.0466	0.1390	0.0519	0.3731	0.0527	0.3426	0.352980	0.539950
CENTROID	0.0801	0.1663	0.0511	0.1494	0.0654	0.4868	0.0715	0.1975	0.419456	0.591524
REPDIS	1.2e−5	1.2e−5	0.0000	6.8e−6	0.0000	1.6e−5	0.999959	0.999965	0.999978	0.999985
CAI	0.0000	0.2500	0.0195	0.2500	0.0000	0.2500	0.0000	0.2500	0.148776	0.354549
APOI	0.0000	0.2500	0.0195	0.2500	0.0000	0.2500	0.0000	0.2500	0.148776	0.354549
COMB	0.0000	0.2500	0.0195	0.2500	0.0000	0.2500	0.0000	0.2500	0.148776	0.354549
ROBUST-ITER	0.0000	0.0000	0.0000	0.0897	0.0000	0.7118	0.0000	0.1985	0.232163	0.359440
ROBUST-COMP	0.0000	1.6e−6	0.0000	0.0000	0.999978	0.999987	0.0000	1.1e−5	0.999984	0.999986

Table 4: Comprehensive values for a subset of non-reference alternatives assigned by 14 procedures for selecting a representative sorting model.

Method	a_6	a_9	a_{11}	a_{17}	a_{21}	a_{22}	a_{23}	a_{25}	a_{26}
REFERENCE	**0.6414**	**0.6163**	**0.6881**	**0.4677**	**0.4759**	**0.3659**	**0.4952**	**0.5185**	**0.4216**
UTADISMP1	0.5737	0.5542	0.5622	0.3348	0.4011	0.0871	0.4775	0.5342	0.4585
UTADISMP2	0.6449	0.6149	0.7139	0.4909	0.4898	0.4525	0.5231	0.4966	0.4442
UTADISMP3	0.6449	0.6149	0.7139	0.4909	0.4898	0.4525	0.5231	0.4966	0.4442
UTADIS-JLS	0.5746	0.5151	0.6858	0.3643	0.3943	0.2763	0.5080	0.4574	0.4170
CHEBYSHEV	0.5891	0.5358	0.6567	0.3475	0.4016	0.1578	0.5146	0.5172	0.4488
MSCVF	0.6311	0.5940	0.7478	0.5181	0.5131	0.5314	0.5704	0.5228	0.4732
ACUTADIS	0.5805	0.5399	0.6433	0.3661	0.3965	0.2548	0.4761	0.4644	0.4092
CENTROID	0.6556	0.6456	0.6417	0.4639	0.4909	0.3120	0.5106	0.5438	0.4821
REPDIS	0.999985	0.999984	0.999985	0.999979	0.999979	0.999972	0.999980	0.999980	0.999978
CAI	0.4222	0.3910	0.4630	0.1488	0.1788	0.0930	0.2603	0.2472	0.2091
APOI	0.4222	0.3910	0.4630	0.1488	0.1788	0.0930	0.2603	0.2472	0.2091
COMB	0.4222	0.3910	0.4630	0.1488	0.1788	0.0930	0.2603	0.2472	0.2091
ROBUST-ITER	0.5737	0.5542	0.5622	0.3348	0.4011	0.0871	0.4775	0.5342	0.4585
ROBUST-COMP	0.999989	0.999987	0.999992	0.999983	0.999984	0.999980	0.999988	0.999987	0.999986

underlying optimization model is non-linear, considering the sum of logarithms of the slack variables involved in each inequality. The obtained MVFs are strictly increasing, the class thresholds are well-separated, and u_3 and u_4 have about 2.5 times greater impact on the comprehensive values than u_1 and u_2. ACUTADIS is also one out of only four methods that made only a single mistake in classifying the non-reference alternatives. The incorrectly rated city is Paris (a_6), which is relatively similar to London (a_{10}) assigned by the DM to C_3. The latter alternative is distant from the lower threshold of its desired class ($U(a_{10}) = 0.6143$ and $t_2 = 0.5399$). This implies that the comprehensive value of Paris also fits in the range associated with the most preferred class.

The REPDIS procedure returned a model which builds the comprehensive scores based on just a single criterion (in this case – g_4). Hence, the maximal share of u_4 is equal to one, whereas the marginal value assigned to $u_4(5)$ is very close to one (0.999974). As a result, the differences between comprehensive values of a large set of alternatives, as well as between class thresholds, are extremely small. This is an undesired effect from the viewpoint of the results' interpretability. It suggests that for this particular problem, the objectives built on the analysis of $APWI'$s proved too challenging to let the method emphasize the value differences for all pairs of alternatives simultaneously. To maximize its objective function while considering the conflicting sub-objectives, REPDIS opted to balance the alternatives' comprehensive assessments. The same problem can be observed for ROBUST-COMP with the proviso that in this case, criterion g_3 was used as the sole one from which alternatives derived positive values. A side effect of such minor differences is that when comparing the classification suggested by such models for non-reference alternatives and the ones derived with the DM's simulated model, there is no match for many pairs. In the case of ROBUST-COMP, such mistakes are observed for 5 out of 21 cities.

ROBUST-ITER and ROBUST-COMP take into account the necessary assignment-based preference relations. While ROBUST-COMP attempts to consider the two objectives relevant to this approach at once, ROBUST-ITER optimizes them one after another. For the illustrative study, such an approach led to a more intuitive and interpretable model. In fact, the model obtained after considering the first objective was not modified when subsequently optimizing the other

Table 5: Class assignments for a subset of non-references alternatives determined with 14 procedures for selecting a representative sorting model.

Method	a_6	a_9	a_{11}	a_{17}	a_{21}	a_{22}	a_{23}	a_{25}	a_{26}
REFERENCE	**2**	**2**	**3**	**2**	**2**	**1**	**2**	**2**	**2**
UTADISMP1	3	3	3	2	2	1	3	3	3
UTADISMP2	2	2	3	2	2	1	2	2	1
UTADISMP3	2	2	3	2	2	1	2	2	1
UTADIS-JLS	2	2	3	1	1	1	2	2	2
CHEBYSHEV	3	3	3	2	2	1	3	3	2
MSCVF	3	2	3	2	1	2	2	2	1
ACUTADIS	3	2	3	2	2	1	2	2	2
CENTROID	3	3	3	2	2	1	2	2	2
REPDIS	2	2	2	2	2	1	2	2	2
CAI	3	3	3	2	2	1	2	2	2
APOI	3	3	3	2	2	1	2	2	2
COMB	3	3	3	2	2	1	2	2	2
ROBUST-ITER	3	3	3	2	3	1	3	3	3
ROBUST-COMP	3	3	3	1	2	1	3	3	2

objective. Hence, the resulting model was determined solely by maximizing the value differences for pairs of alternatives related by the necessary assignment-based preference relation (e.g., (a_7, a_{15}) among reference alternatives and (a_{25}, a_{24}) among non-reference alternatives). The value differences for pairs always assigned to the same class were just a side effect of the primary optimization. Clearly, this observation does not hold for all decision problems because the secondary objective can often break ties when selecting among models that optimize the primary objective equally well. When it comes to the assignments of non-reference alternatives, ROBUST-ITER misclassified 6 out of 21 cities compared to the assignments provided by the DM's reference model.

The three novel approaches proposed in this paper (CAI, APOI, and COMB) selected the same model for the considered problem. Putting the objective functions of CAI and APOI together, COMB often returns a result that matches the solution of either model. However, such a perfect agreement between CAI and APOI is less common. Nevertheless, it can be justified because they build their outcomes on the stochastic acceptability indices, even if CAI focuses on the class assignments and APOI considers assignment-based pairwise relations. Still, such a high similarity between the models returned by these methods – confirmed also in the experimental section – could not be predicted beforehand, without verifying how these approaches work in practice.

When it comes to the considered study, the model discovered by these approaches is characterized by equal maximal shares of all criteria (0.25) and a positive marginal value assigned to the mid-point only for u_2. Such a balanced distribution implied relatively low comprehensive values of all alternatives (see Table 4) and low thresholds separating the classes ($t_1 = 0.148776$ and $t_2 = 0.354549$). To explain the operational procedure of CAI and APOI, let us focus on Riga and Athens. According to Stochastic Ordinal Regression (SOR), Riga is assigned to C_1 by all models ($CAI'(a_{15}, C_1) = 1.0$). For Athens, there is an ambiguity in the assignments ($CAI'(a_{17}, C_1) = 0.088$ and $CAI'(a_{17}, C_2) = 0.912$). As a result, they are assigned to a class better than Riga for the vast majority (91.2%) of models ($APOI'(a_{15}, a_{17}) = 0.088$ and $APOI'(a_{17}, a_{15}) = 1$). Hence to optimize the objective functions' values and emphasize the most frequent results in the representative models, the novel procedures opt for assigning Riga to C_1 and Athens to C_2, even if it was challenging to separate these two alternatives ($U(a_{15}) = 0.1487745$, $U(a_{17}) = 0.1487765$, and $t_1 = 0.1487755$).

2. Measure values for selected procedures based on the results of the illustrative study

In this section, we report selected measure values for a few procedures based on the results of the illustrative study.

Classification accuracy. When considering the results reported in Table 5 for 9 non-reference alternatives and remembering that the classification of the remaining 12 test options agreed with the references one, the classification accuracy obtained by UTADISMP2 was $accuracy(U^{UTADISMP2}) = \frac{20}{21} = 0.9524$ and for CHEBYSHEV – it was $accuracy(U^{CHEBYSHEV}) = \frac{17}{21} = 0.8095$. The former procedure misclassified only a_{26}, whereas for the latter – four non-reference alternatives (a_6, a_9, a_{23}, and a_{25}) were classified incorrectly w.r.t. the reference assignment.

Assignment acceptability. When considering CAI's reported in Table 1, for 21 non-reference alternatives, $MCAI_{max}$ is equal to 0.9656. In fact, the maximal CAI' was lesser than one only for 9 alternatives. Four approaches (CENTROID, CAI, APOI, and COMB) identified a solution with $MCAI_{abs} = MCAI_{max}$. Consequently, for these methods, $MCAI_{rel}(U^P) =$

0. Hence these procedures perfectly reflect the most robust assignments. Note that this value is lower for the reference model, which assigns alternatives a_6 and a_9 to class C_2. However, their class acceptabilities for class C_3 are higher than for C_2 (e.g., for a_6 – $CAI'(a_6, C_2) = 0.056$ and $CAI'(a_6, C_3) = 0.944$). As a result, for the reference model, $MCAI_{abs}(U^{REF}) = 0.8875$ and $MCAI_{rel}(U^{REF}) = \frac{0.8875}{0.9656} - 1 = -0.0809$. This example emphasizes that $MCAI$ captures whether a given procedure reconstructs the most common results observed for all compatible sorting models rather than the reconstruction of the reference assignments.

Differences between marginal values. When considering the results reported in Table 3, the model which is the closest to the reference one in terms of Δ_{TO}^{REF} was obtained with ACUTADIS ($\Delta_{TO}^{REF}(U^{ACUTADIS}) = 0.0595$). On the other extreme, REPDIS identified the furthest solution from the reference model (0.3331). As far as the comparison with an average model is concerned, the outcome of the CENTROID procedure is, by definition, the same (i.e., $\Delta_{TO}^{CENT}(U^{CENTROID}) = 0.0$). However, other methods which also aimed to identify a central model attained quite favorable scores too: for ACUTADIS – $\Delta_{TO}^{CENT}(U^{ACUTADIS}) = 0.0443$, for UTADIS-JLS – 0.0690, and for CHEBYSHEV – 0.0781. Again, for REPDIS, the distance was vast (0.3413).

Differences between comprehensive values. Part of the results needed to compute the differences between comprehensive values for the illustrative study is available in Table 4. Taking into account the comprehensive values of 21 non-reference alternatives, the closest model to the reference one was obtained with CENTROID ($\Delta_{CV}^{REF}(U^{CENTROID}) = 0.0265$). In turn, the furthest distance can be attributed to ROBUST-COMP ($\Delta_{CV}^{REF}(U^{ROBUST-COMP}) = 0.4287$).

Differences between threshold values. For the illustrative example, the difference between threshold values can be determined based on Table 3. For UTADIS-JLS, the threshold values are the closest to the reference model ($\Delta_{TH}^{REF}(U^{UTADIS-JLS}) = 0.0266$). On the other extreme, they are the furthest for REPDIS and ROBUST-COMP (for both of them, $\Delta_{TH}^{REF}(U^{P}) = 0.4740$). Indeed, the separation between classes was very poor for these methods, and all thresholds were close to one.

3. Linear regression solutions obtained for the assignment acceptabilities

Linear regression allows for further analysis of trends in the context of individual procedures. First, we focus on the relative assignment acceptabilities. The coefficients of solutions obtained for the relative MCAI are presented in Table 6.

Table 6: Coefficients of solutions obtained for the linear regression problem for average relative MCAI depending on the defined dimensions for individual procedures.

Procedure	No. of classes	No. of criteria	No. of ch. points	No. of ref. alt.	Intercept
UTADISMP1	0.001629	-0.016765	-0.030410	0.013842	0.023974
UTADISMP2	-0.008175	-0.014437	-0.013143	0.014108	-0.041136
UTADISMP3	-0.031191	-0.009152	-0.004839	0.010618	-0.014165
UTADIS-JLS	-0.007858	-0.013888	-0.042860	0.017717	0.035798
CHEBYSHEV	-0.005071	-0.003836	-0.002640	0.004089	-0.014888
MSCVF	-0.021215	-0.024763	-0.008259	0.018436	-0.060628
ACUTADIS	-0.009831	-0.008980	-0.015267	0.004029	0.064738
CENTROID	-0.000146	-0.000110	-0.000115	0.000163	-0.000610
REPDIS	-0.000623	-0.007850	-0.015653	0.009646	-0.060658
CAI	-0.000000	-0.000000	0.000000	0.000000	-0.000000
APOI	-0.000039	-0.000020	0.000096	0.000171	-0.001702
COMB	-0.000041	-0.000021	0.000082	0.000152	-0.001464
ROBUST-ITER	-0.002523	-0.020901	-0.034387	0.017789	-0.016656
ROBUST-COMP	-0.009931	-0.020450	-0.035637	0.017098	0.001854

There is a high similarity in the results obtained for the methods based on stochastic analysis. Due to the objective function converging with the measure definition, CAI reaches its maximum value regardless of the problem size. Hence the coefficients are equal to 0. This regularity is also visible for APOI, COMB, and CENTROID, for which the slope coefficients are very close to 0. The exception is REPDIS, whose average values are sensitive, especially to the change in the number of characteristic points. The decrease in relative MCAI by almost 1.6% may be because REPDIS focuses only on emphasizing the relationship between pairs of alternatives and not on whether they will be assigned to the most common class in sampling models.

Unique among all the methods is the positive effect of increasing the number of classes for UTADISMP1. Although the increase in the regression's slope coefficient is not substantial (0.16%), other methods working similarly, such as UTADISMP2 and UTADISMP3, show significant decreases. For UTADISMP3, we observe the largest decrease among all

methods. It is equal to 3.1% for each newly introduced class. In both cases, these effects are likely caused by more classes introducing more assignments. The crucial aspect for UTADISMP1 is that there is a quadratic growth in the number of constraints highlighting the discriminatory nature of this approach, depending on the number of assignments provided by the DM. The decrease for UTADISMP3 is likely due to focusing solely on maximizing the minimum slope between two consecutive points. The other slopes are not optimized in any way, and the more assignments and classes, the more challenging it becomes to recreate the most robust classification properly.

The ROBUST procedures respond the worst to the increase in the number of characteristic points and almost the worst to the increase in the number of criteria. Both features introduce more flexibility to the models, reducing the number of necessary relations in the set of alternatives. Fewer such relations imply fewer constraints for constructing the objective function and a wider choice of the resulting model. Consequently, the derived model may not reflect the most popular dependencies occurring in the entire space of consistent solutions.

The MCAI-based results confirm the relationship between the strength of the impact of the number of assignments provided and the average results obtained by the methods. The worse the method's average results, the more significant the positive impact of the increase in the number of reference alternatives (see Table 4 and Figure 2). Adding one alternative per class, on average, increases the relative MCAI by around 1.7–1.8% for the weakest ROBUST methods. In contrast, for CHEBYSHEV, which is one of the best procedures, it only increases by 0.4%.

Table 7 presents the slope coefficients obtained from solving the linear regression problem for the absolute MCAI measure. Since the CAI method obtained the maximum MCAI relative value equal to 0 for almost all cases, the absolute values of this method are also the highest achievable ones. For this reason, the slope coefficients for CAI can be viewed as the maximum achievable Absolute coefficients, and they can serve as a benchmark when evaluating other procedures. Thus, it is evident that the increase in the complexity of MVFs and reference assignments positively affects the robustness of the recommendation of the *robust assignment rule*. In turn, an increase in the number of classes and criteria decreases the best possible MCAI values, but not as much as for the other factors. Again, other approaches based on acceptability indices and the CENTROID method perform similarly to CAI.

Regarding the increase in the number of classes, only UTADISMP1 is more resistant to it than the procedures mentioned above. For this approach, the number of classes matters little as the procedure emphasizes the correct separation of classes irrespective of how many categories are considered. UTADISMP1 is more sensitive to changing the number of criteria. Adding one criterion, in this case, results in an average 1.7% drop in Absolute MCAI, which is large compared to the CAI with only a 0.2% drop. Only MSCVF (2.2%) and the group of ROBUST methods (2.1%) recorded more significant declines.

Table 7: Coefficients of solutions obstained for the linear regression problem for average absolute MCAI depending on the defined dimensions for individual procedures.

Procedure	**No. of classes**	**No. of criteria**	**No. of ch. points**	**No. of ref. alt.**	**Intercept**
UTADISMP1	-0.002868	-0.017331	-0.020834	0.021358	0.867229
UTADISMP2	-0.011423	-0.015128	-0.005519	0.021423	0.809100
UTADISMP3	-0.032175	-0.010236	0.001899	0.017977	0.835487
UTADIS-JLS	-0.010897	-0.014758	-0.032217	0.024559	0.878676
CHEBYSHEV	-0.009519	-0.005601	0.004618	0.013091	0.828673
MSCVF	-0.026236	-0.021589	0.000087	0.023000	0.793282
ACUTADIS	-0.013617	-0.010300	-0.006802	0.012953	0.900571
CENTROID	-0.005335	-0.002359	0.007266	0.010007	0.838038
REPDIS	-0.005126	-0.009175	-0.007657	0.017424	0.792688
CAI	-0.005212	-0.002266	0.007381	0.009879	0.838452
APOI	-0.005249	-0.002279	0.007453	0.010017	0.837100
COMB	-0.005250	-0.002281	0.007443	0.010002	0.837287
ROBUST-ITER	-0.006047	-0.021004	-0.025165	0.024151	0.836867
ROBUST-COMP	-0.012680	-0.020583	-0.026393	0.023389	0.854476

Despite the decrease in relative MCAI value with the increase in the number of MVF segments for almost all methods, many show slight increases in the context of absolute MCAI. This is because the increase in the highest achievable values was significant (0.7%), and the relative decreases for, e.g., CHEBYSHEV and UTADISMP3 were not that high. For this reason, these two methods obtained higher absolute results when more characteristic points were considered.

The relationships between the obtained scores and the number of reference alternatives that were observed for other

measures are also present in this case. Again, the best-performing methods obtained the lowest average profit from considering additional information from the DM – the value for CAI increased by only less than 1%. In turn, the UTADISMP methods achieved improvement between 1.8 and 2.1% and the ROBUST methods between 2.3 and 2.4%. These outcomes again underline the superiority of methods incorporating stochastic analysis to obtain more robust results, especially when the availability of DM's preferences is limited.

4. Experimental results concerning differences between marginal and comprehensive values and class thresholds

In this section, we consider the outcomes concerning the differences between the reference and resulting models obtained for different values of each problem dimension (p, m, γ_j, and R).

4.1. Differences between marginal values

An increase in the number of classes has a positive effect on reproducing the DM's preference model in terms of marginal values (see Table 8). This is confirmed for all procedures. The monotonicity of the decrease in differences is the greatest between 2- and 3-class problems. Furthermore, UTADISMP1 and ROBUST-ITER (both attain 0.1009 for 2-class and 0.0313 for 5-class instances) improve most significantly with the increase of p. For 2-class problems, these two perform the worst, whereas, for 5-class problems, their results are only slightly inferior to REPDIS (0.291) and center-based approaches (0.0253 for ACUTADIS, 0.280 for CENTROID, and 0.0287 for CHEBYSHEV).

Table 8: Average differences between marginal values for various numbers of classes.

	Reference				**Centroid**			
Procedure	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
UTADISMP1	0.1009	0.0591	0.0413	0.0313	0.0933	0.0495	0.0318	0.0227
UTADISMP2	0.0971	0.0616	0.0466	0.0370	0.0867	0.0522	0.0381	0.0293
UTADISMP3	0.0819	0.0648	0.0531	0.0430	0.0495	0.0463	0.0401	0.0337
UTADIS-JLS	0.0841	0.0601	0.0446	0.0345	0.0611	0.0457	0.0356	0.0279
CHEBYSHEV	0.0697	0.0484	0.0368	0.0287	0.0320	0.0222	0.0163	0.0123
MSCVF	0.0953	0.0645	0.0496	0.0417	0.0832	0.0550	0.0432	0.0374
ACUTADIS	0.0650	0.0440	0.0326	0.0253	0.0336	0.0255	0.0202	0.0162
CENTROID	0.0672	0.0469	0.0358	0.0280	0	0	0	0
REPDIS	0.0777	0.0519	0.0381	0.0291	0.0471	0.0268	0.0170	0.0117
CAI	0.0883	0.0697	0.0553	0.0434	0.0633	0.0510	0.0399	0.0309
APOI	0.0886	0.0698	0.0556	0.0436	0.0634	0.0511	0.0401	0.0312
COMB	0.0886	0.0698	0.0556	0.0436	0.0634	0.0511	0.0401	0.0312
ROBUST-ITER	0.1009	0.0591	0.0413	0.0313	0.0933	0.0495	0.0318	0.0227
ROBUST-COMP	0.0981	0.0628	0.0470	0.0374	0.0884	0.0546	0.0408	0.0326

A higher number of criteria also reduces a gap between marginal values (see Table 9). This is understandable because, with more criteria, their shares in the comprehensive values decrease, leading to lesser differences between the compared models. The most substantial differences for the extreme numbers of criteria can be observed for UTADISMP3 and UTADIS-JLS. For both approaches, the mean differences between marginal values decreased more than two times when comparing problems with three and nine criteria (for UTADISMP3 - from 0.0873 to 0.0421, and for UTADIS-JLS – from 0.0822 to 0.0390). With a greater number of criteria, their solutions become more similar to the central models, which approximate the DM's preferences up to a satisfactory level.

The increasing number of characteristic points affects the performance of procedures differently (see Table 10). For most methods (including ROBUST-ITER, UTADISMP1, ROBUST-COMP, UTADIS-JLS, REPDIS, CAI, APOI, and COMB), an increase of γ_j leads to a more significant difference between marginal values – and thus a deterioration in the quality of reconstructing the reference AVF (e.g., for UTADIS-JLS – the distance increases from 0.0420 to 0.0686). The center-oriented approaches (CENTROID, ACUTADIS, and CHEBYSHEV) maintain the stability of distances of their models from the reference one for different numbers of characteristic points.

The decrease in the average differences between marginal values can be observed for UTADISMP2 (0.0625 to 0.0600) and UTADISMP3 (0.0645 to 0.0569). Both approaches maximize the differences between marginal values assigned to the consecutive points. The greater the share of such values among variables optimized by the methods, the better the results achieved by these procedures. A similar trend is observed for MSCVF, which is also focused on optimizing the

Table 9: Average differences between marginal values for various numbers of criteria.

	Reference				Centroid			
Procedure	**3**	**5**	**7**	**9**	**3**	**5**	**7**	**9**
UTADISMP1	0.0748	0.0596	0.0516	0.0467	0.0594	0.0509	0.0454	0.0416
UTADISMP2	0.0825	0.0619	0.0519	0.0461	0.0681	0.0524	0.0452	0.0406
UTADISMP3	0.0873	0.0632	0.0502	0.0421	0.0636	0.0438	0.0342	0.0279
UTADIS-JLS	0.0822	0.0566	0.0455	0.0390	0.0644	0.0423	0.0340	0.0295
CHEBYSHEV	0.0621	0.0474	0.0395	0.0346	0.0296	0.0215	0.0172	0.0146
MSCVF	0.0837	0.0645	0.0548	0.0481	0.0670	0.0567	0.0503	0.0448
ACUTADIS	0.0560	0.0430	0.0362	0.0316	0.0328	0.0247	0.0204	0.0176
CENTROID	0.0591	0.0459	0.0388	0.0342	0	0	0	0
REPDIS	0.0681	0.0497	0.0420	0.0371	0.0381	0.0246	0.0208	0.0191
CAI	0.0897	0.0669	0.0542	0.0459	0.0663	0.0478	0.0385	0.0325
APOI	0.0901	0.0670	0.0544	0.0460	0.0668	0.0479	0.0387	0.0325
COMB	0.0901	0.0670	0.0544	0.0460	0.0668	0.0479	0.0387	0.0325
ROBUST-ITER	0.0748	0.0596	0.0516	0.0467	0.0593	0.0509	0.0454	0.0416
ROBUST-COMP	0.0824	0.0627	0.0532	0.0471	0.0711	0.0555	0.0471	0.0427

shape of MVFs. However, since the latter approach is applicable for settings with more than two characteristic points, in this case, the observation is confirmed only for the results obtained for instances with four and six breakpoints.

Table 10: Average differences between marginal values for various numbers of characteristic points.

	Reference			Centroid		
Procedure	**2**	**4**	**6**	**2**	**4**	**6**
UTADISMP1	0.0526	0.0606	0.0613	0.0391	0.0514	0.0576
UTADISMP2	0.0625	0.0593	0.0600	0.0537	0.0466	0.0544
UTADISMP3	0.0645	0.0607	0.0569	0.0571	0.0398	0.0303
UTADIS-JLS	0.0420	0.0570	0.0686	0.0249	0.0453	0.0576
CHEBYSHEV	0.0446	0.0469	0.0462	0.0272	0.0206	0.0144
MSCVF		0.0664	0.0591		0.0574	0.0521
ACUTADIS	0.0416	0.0424	0.0410	0.0225	0.0262	0.0229
CENTROID	0.0393	0.0474	0.0467	0	0	0
REPDIS	0.0426	0.0522	0.0528	0.0214	0.0262	0.0293
CAI	0.0508	0.0690	0.0727	0.0362	0.0494	0.0532
APOI	0.0510	0.0693	0.0729	0.0367	0.0495	0.0532
COMB	0.0510	0.0693	0.0729	0.0367	0.0495	0.0532
ROBUST-ITER	0.0526	0.0606	0.0613	0.0391	0.0514	0.0575
ROBUST-COMP	0.0526	0.0638	0.0677	0.0414	0.0566	0.0643

The impact of different numbers of reference alternatives per class is reported in Table 11. With richer preference information, the differences between marginal values become lesser for all procedures. Again, the change in the number of reference alternatives has the greatest impact on the performance of UTADISMP1 and ROBUST-ITER (compare 8.26% for $R = 3$ and 4.01% for $R = 10$). Interestingly, for CAI, APOI, and COMB, the relative distances from the centroid solution are stable for different values of R. However, their distances from the reference model decrease when additional assignment examples become available.

Table 11: Average differences between marginal values for various numbers of reference alternatives per class.

	Reference				Centroid			
Procedure	**3**	**5**	**7**	**10**	**3**	**5**	**7**	**10**
UTADISMP1	0.0826	0.0609	0.0491	0.0401	0.0749	0.0516	0.0399	0.0309
UTADISMP2	0.0824	0.0629	0.0527	0.0443	0.0738	0.0536	0.0434	0.0355
UTADISMP3	0.0749	0.0641	0.0558	0.0480	0.0503	0.0446	0.0398	0.0349
UTADIS-JLS	0.0745	0.0590	0.0493	0.0405	0.0573	0.0450	0.0373	0.0307
CHEBYSHEV	0.0594	0.0484	0.0410	0.0349	0.0272	0.0217	0.0184	0.0156
MSCVF	0.0843	0.0646	0.0549	0.0473	0.0746	0.0559	0.0473	0.0410
ACUTADIS	0.0535	0.0440	0.0373	0.0320	0.0268	0.0246	0.0228	0.0213
CENTROID	0.0577	0.0467	0.0397	0.0339	0	0	0	0
REPDIS	0.0621	0.0513	0.0449	0.0385	0.0286	0.0265	0.0249	0.0225
CAI	0.0742	0.0668	0.0604	0.0553	0.0492	0.0481	0.0452	0.0426
APOI	0.0747	0.0670	0.0606	0.0553	0.0497	0.0482	0.0454	0.0426
COMB	0.0746	0.0670	0.0606	0.0553	0.0497	0.0481	0.0454	0.0426
ROBUST-ITER	0.0826	0.0609	0.0491	0.0401	0.0749	0.0516	0.0399	0.0309
ROBUST-COMP	0.0821	0.0644	0.0535	0.0454	0.0757	0.0565	0.0459	0.0384

Tables 12 and 13 show the results of solving the linear regression problems in terms of differences with the reference and CENTROID model. When it comes to the former, the most significant differences are visible in the case of an increase in

the number of classes. Across all methods, they are 2 to 5 times more important than adding another criterion, and 3 to 6 times more important than adding another reference alternative to each class. This is likely because the new assignments are provided from different ranges, significantly reducing the solution space. However, the situation is different when the number of reference alternatives per class is increased. Then, additional assignments are similar to the already known ones, which does not reduce the space of feasible models considerably.

Table 12: Coefficients of solutions obtained for the linear regression problems for the average difference from the marginal values of the reference models depending on the defined dimensions for individual procedures.

Procedure	**No. of classes**	**No. of criteria**	**No. of ch. points**	**No. of ref. alt.**	**Intercept**
UTADISMP1	-0.022675	-0.004616	0.002178	-0.005895	0.193353
UTADISMP2	-0.019523	-0.005965	-0.000635	-0.005255	0.200092
UTADISMP3	-0.012856	-0.007416	-0.001889	-0.003792	0.181453
UTADIS-JLS	-0.016422	-0.007032	0.006650	-0.004739	0.158525
CHEBYSHEV	-0.013490	-0.004521	0.000384	-0.003430	0.140136
MSCVF	-0.017546	-0.005818	-0.003662	-0.005094	0.209241
ACUTADIS	-0.013055	-0.003989	-0.000155	-0.003024	0.130845
CENTROID	-0.012886	-0.004084	0.001849	-0.003324	0.127481
REPDIS	-0.015961	-0.005042	0.002545	-0.003288	0.145710
CAI	-0.014912	-0.007195	0.005490	-0.002695	0.154426
APOI	-0.014893	-0.007242	0.005461	-0.002748	0.155309
COMB	-0.014890	-0.007241	0.005464	-0.002745	0.155254
ROBUST-ITER	-0.022672	-0.004616	0.002174	-0.005894	0.193345
ROBUST-COMP	-0.019806	-0.005773	0.003783	-0.005118	0.182168

Table 13: Coefficients of solutions obtained for the linear regression problems for the average difference from the marginal values of the CENTROID models depending on the defined dimensions for individual procedures.

Procedure	**No. of classes**	**No. of criteria**	**No. of ch. points**	**No. of ref. alt.**	**Intercept**
UTADISMP1	-0.022940	-0.002933	0.004617	-0.006064	0.166660
UTADISMP2	-0.018645	-0.004489	0.000177	-0.005278	0.176035
UTADISMP3	-0.005375	-0.005834	-0.006701	-0.002189	0.136704
UTADIS-JLS	-0.010969	-0.005645	0.008170	-0.003709	0.105339
CHEBYSHEV	-0.006511	-0.002461	-0.003195	-0.001621	0.081169
MSCVF	-0.014934	-0.003655	-0.002642	-0.004609	0.170932
ACUTADIS	-0.005781	-0.002485	0.000107	-0.000782	0.063473
CENTROID	0.000000	0.000000	0.000000	0.000000	0.000000
REPDIS	-0.011595	-0.003041	0.001975	-0.000869	0.081989
CAI	-0.010826	-0.005538	0.004251	-0.000981	0.106513
APOI	-0.010776	-0.005600	0.004140	-0.001034	0.107687
COMB	-0.010774	-0.005595	0.004141	-0.001035	0.107645
ROBUST-ITER	-0.022934	-0.002932	0.004615	-0.006065	0.166642
ROBUST-COMP	-0.018109	-0.004693	0.005713	-0.005162	0.155052

The Hasse diagrams presented in Figures 4 and 5 show the results of the Wilcoxon signed-rank tests for paired samples with a p-value equals 0.05. The performances of various methods in terms of these measures were discussed in the main paper. Therefore, we note only the most peculiar features. ACUTADIS returns solutions that are more similar to the reference ones than those obtained with CHEBYSHEV. However, when comparing the similarity with respect to the centroid solution, the order between these procedures is inverse. Then, ROBUST-COMP is the most distant from the average model. Nevertheless, in the context of the reference model, it is better than stochastic methods: CAI, APOI, and COMB. Surprisingly, these approaches perform poorly compared to CENTROID, given that their models are based on the same sampling data. This shows that the average solution can differ significantly from the solution recreating the most popular acceptability indices in the whole space of consistent models.

4.2. Differences between comprehensive values

A greater number of classes has a positive effect on reproducing the original comprehensive values assigned by the DM to alternatives (see Table 14). The ACUTADIS method turns out to be the best irrespective of p (0.0727 for 2-class and 0.0337 for 5-class problem instances). On the contrary, the most considerable relative differences can be observed for UTADISMP1 and ROBUST-ITER (the difference between comprehensive decreases from 0.1434 to 0.0434 when moving from 2 to 5 classes). These two methods have one of the worst results for 2-class problems, but with 5 classes, only ACUTADIS and central-based methods (CENTROID, CHEBYSHEV) achieved better average values. Exactly the

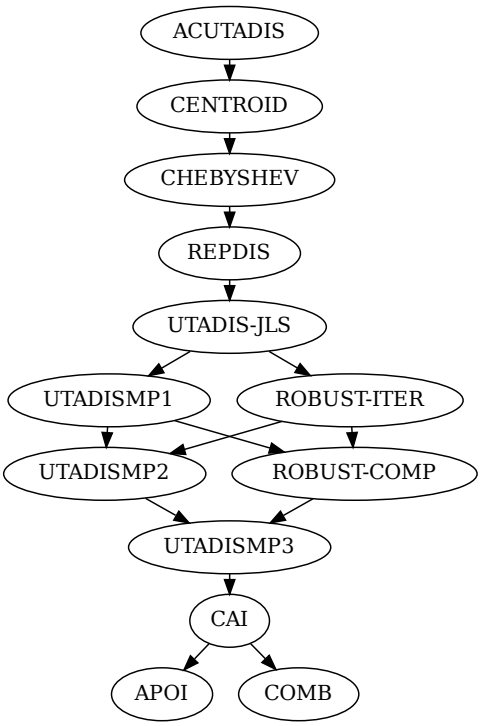


Figure 4: The Hasse diagram indicating the statistically significant differences in terms of the differences from the marginal values of the reference models based on the Wilcoxon test with p-value equal to 0.05.

opposite situation can be observed for UTADISMP3, which gives way only to the three mentioned methods for 2-class dilemmas, and with 5 classes, it is worse than more than half of the approaches.

Table 14: Average differences between comprehensive values for various numbers of classes.

Procedure	**2**	**3**	**4**	**5**
UTADISMP1	0.1434	0.0807	0.0568	0.0434
UTADISMP2	0.1380	0.0791	0.0579	0.0460
UTADISMP3	0.0835	0.0695	0.0587	0.0501
UTADIS-JLS	0.1531	0.1054	0.0790	0.0611
CHEBYSHEV	0.0773	0.0589	0.0470	0.0381
MSCVF	0.1178	0.0887	0.0748	0.0665
ACUTADIS	0.0727	0.0532	0.0412	0.0337
CENTROID	0.0760	0.0580	0.0463	0.0378
REPDIS	0.0935	0.0661	0.0497	0.0394
CAI	0.1553	0.0938	0.0664	0.0511
APOI	0.1541	0.0933	0.0649	0.0508
COMB	0.1541	0.0933	0.0650	0.0508
ROBUST-ITER	0.1434	0.0806	0.0568	0.0434
ROBUST-COMP	0.1383	0.0853	0.0646	0.0529

An increase in the number of criteria differently influences the results of particular methods (see Table 15). A greater number of performance dimensions positively affects the performance of UTADISMP3, MSCVF, CHEBYSHEV, ACUTADIS, CENTROID, and REPDIS. The greatest relative differences between 3- and 9-attribute problems are observed for UTADISMP3 (from 7.22% to 5.98%) and REPDIS (from 7.05% to 5.85%). In general, these methods optimize the shape of MCVFs or exploit the geometry of the polyhedron of all feasible models. On the contrary, with a more significant number of criteria, the average difference from the reference model in terms of comprehensive values increases for UTADISMP1, UTADISMP2, CAI, APOI, COMB, and ROBUST-ITER. These approaches focus on optimizing the comprehensive values of alternatives, usually making them as discriminatory as possible, though based on differently formulated objectives.

Table 16 shows that for the vast majority of procedures, adding characteristic points leads to an increased difference between comprehensive values. The most significant increase is observed between instances with two and four characteristic points (e.g., for UTADIS-JLS – the respective values are 0.0280 and 0.1109). The central-based approaches also record relatively large increases for the above instances, but this increase is already minimal when moving from four to

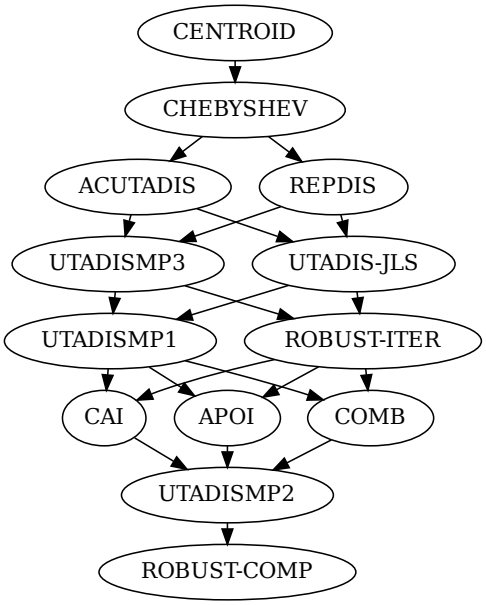


Figure 5: The Hasse diagram indicating the statistically significant differences in terms of the differences from the marginal values of the CENTROID models based on the Wilcoxon test with p-value equal to 0.05.

Table 15: Average differences between comprehensive values for various numbers of criteria.

Procedure	**3**	**5**	**7**	**9**
UTADISMP1	0.0790	0.0801	0.0813	0.0838
UTADISMP2	0.0790	0.0793	0.0802	0.0825
UTADISMP3	0.0722	0.0672	0.0626	0.0598
UTADIS-JLS	0.1017	0.0941	0.0985	0.1043
CHEBYSHEV	0.0588	0.0561	0.0534	0.0530
MSCVF	0.0923	0.0876	0.0850	0.0829
ACUTADIS	0.0541	0.0507	0.0486	0.0475
CENTROID	0.0569	0.0553	0.0532	0.0528
REPDIS	0.0705	0.0611	0.0586	0.0585
CAI	0.0892	0.0907	0.0914	0.0952
APOI	0.0890	0.0899	0.0905	0.0937
COMB	0.0889	0.0899	0.0905	0.0937
ROBUST-ITER	0.0790	0.0801	0.0812	0.0839
ROBUST-COMP	0.0896	0.0851	0.0826	0.0839

six breakpoints. Though slightly greater in terms of absolute values, the same effect can be observed for the methods that select the most discriminant model. In general, these results confirm that the move from linear MVFs to functions with three linear pieces increases the flexibility of the models more substantially than the change from three to five linear pieces.

A greater number of reference alternatives per class lets all procedures construct the models that are more similar to the reference one in terms of comprehensive values (see Table 17). The largest relative decreases in differences – from 0.1165 to 0.0556 – are achieved by UTADISMP1 and ROBUST-ITER. On the contrary, a minor reduction in terms of similarities between comprehensive values is observed for CAI, APOI, and COMB (from 0.1115 to 0.0750 for CAI and from 0.1102 to 0.0745 for APOI and COMB). It is apparent that the narrowing of the space of coherent models with additional preferential information makes solutions of discriminative approaches more and more similar to the reference model. On the other hand, although the methods based on the stochastic analysis of the solution space reproduce preferential information well, the resulting model differs from the one used by DM.

Table 18 shows the slope coefficients resulting from linear regression, indicating the impact of individual parameters on the difference between the comprehensive values. Again, the number of classes is the essential factor. Including one additional class in the problem yields some methods to reduce the distance by as much as 3.3%. Only in the case of UTADIS-JLS, the influence of the number of characteristic points is slightly more significant. In this case, adding another characteristic point and making MVFs more flexible significantly increases the distance from reference comprehensive values. Again, the weakness of this method can be attributed to focusing only on the extreme characteristic points.

UTADIS-JLS is also the worst procedure in this context, whereas ACUTADIS, CENTROID, and CHEBYSHEV

Table 16: Average differences between comprehensive values for various numbers of characteristic points.

Procedure	**2**	**4**	**6**
UTADISMP1	0.0359	0.0960	0.1113
UTADISMP2	0.0404	0.0917	0.1087
UTADISMP3	0.0413	0.0781	0.0770
UTADIS-JLS	0.0280	0.1109	0.1601
CHEBYSHEV	0.0300	0.0676	0.0684
MSCVF		0.0900	0.0840
ACUTADIS	0.0279	0.0611	0.0617
CENTROID	0.0262	0.0678	0.0696
REPDIS	0.0280	0.0761	0.0824
CAI	0.0329	0.1003	0.1417
APOI	0.0331	0.1003	0.1389
COMB	0.0331	0.1003	0.1389
ROBUST-ITER	0.0359	0.0960	0.1113
ROBUST-COMP	0.0355	0.1009	0.1195

Table 17: Average differences between comprehensive values for various numbers of reference alternatives per class.

Procedure	**3**	**5**	**7**	**10**
UTADISMP1	0.1165	0.0839	0.0682	0.0556
UTADISMP2	0.1130	0.0826	0.0689	0.0566
UTADISMP3	0.0788	0.0680	0.0615	0.0536
UTADIS-JLS	0.1365	0.1056	0.0862	0.0704
CHEBYSHEV	0.0689	0.0573	0.0508	0.0443
MSCVF	0.1091	0.0888	0.0794	0.0706
ACUTADIS	0.0615	0.0521	0.0464	0.0409
CENTROID	0.0680	0.0566	0.0500	0.0436
REPDIS	0.0754	0.0642	0.0584	0.0507
CAI	0.1115	0.0962	0.0839	0.0750
APOI	0.1102	0.0950	0.0833	0.0745
COMB	0.1102	0.0950	0.0833	0.0745
ROBUST-ITER	0.1165	0.0839	0.0682	0.0556
ROBUST-COMP	0.1156	0.0883	0.0739	0.0632

perform most favorably. This relation is visible in the Hasse diagram shown in Figure 6. It is worth noting the high position of UTADISMP3 in the ranking, as it performs worse only than the central models and REPDIS. Apparently, maximizing slopes gives a relatively good approximation of comprehensive values, despite the different shapes of the MVFs and poor achievements in terms of classification accuracy. However, this aspect is not crucial in most problems. In turn, it is more important to reconstruct the preferences and DM's classification policy than to correctly reproduce comprehensive values.

4.3. Differences between class thresholds

For all considered procedures, the average difference between class thresholds decreases as the number of classes increases (see Table 19). However, the level of this reduction ranges between methods. The greatest discrepancies from around 0.18 to around 0.05 are observed for methods based on robustness analysis (ROBUST-ITER and ROBUST-COMP). This is related to the fact that we consider more assignment examples with a greater number of classes. These, in turn, imply additional constraints, leading to enriched necessary inference that leaves lesser flexibility to the class thresholds when optimized by the methods. Though slightly less substantial in absolute terms, a similar trend can be observed for the CAI-based approaches. The least improvement with the increase of p can be observed for UTADISMP3 and MSCVF. These approaches do not optimize the threshold values, adhering instead to a default procedure that sets the thresholds at equal distances from the extremely evaluated reference alternatives for each class.

Analogously, the average difference between class thresholds for most procedures decreases as more criteria are considered (see Table 20). However, some methods, such as UTADIS-JLS, CAI, APOI, COMB, and ROBUST-ITER, do not follow this trend for all analyzed values of m. In general, a greater number of attributes makes the trade-offs between the criteria smaller. This, in turn, implies that selecting class thresholds and MVFs that ensure the reconstruction of DM's preferences is more challenging due to lesser flexibility.

The results reported in Table 21 reveal a significant impact of the number of characteristic points on deviation in threshold values of the reference model. Again, the most significant differences can be observed between instances with two and four breakpoints. However, when collating the outcomes for four and six characteristic points, for some approaches

Table 18: Coefficients of solutions obtained for the linear regression problems for the average difference from the comprehensive values of the reference models depending on the defined dimensions for individual procedures.

Procedure	**No. of classes**	**No. of criteria**	**No. of ch. points**	**No. of ref. alt.**	**Intercept**
UTADISMP1	-0.032393	0.000785	0.018870	-0.008374	0.166579
UTADISMP2	-0.029732	0.000569	0.017078	-0.007731	0.160913
UTADISMP3	-0.011100	-0.002104	0.008918	-0.003521	0.103271
UTADIS-JLS	-0.030218	0.000618	0.033025	-0.009231	0.127304
CHEBYSHEV	-0.012960	-0.001004	0.009598	-0.003416	0.089665
MSCVF	-0.016752	-0.001532	-0.002988	-0.005272	0.202681
ACUTADIS	-0.012876	-0.001095	0.008455	-0.002864	0.085938
CENTROID	-0.012622	-0.000723	0.010858	-0.003396	0.080843
REPDIS	-0.017890	-0.001929	0.013600	-0.003420	0.103332
CAI	-0.033990	0.000945	0.027208	-0.005178	0.128469
APOI	-0.033838	0.000744	0.026463	-0.005049	0.130438
COMB	-0.033827	0.000756	0.026460	-0.005049	0.130338
ROBUST-ITER	-0.032386	0.000783	0.018865	-0.008372	0.166564
ROBUST-COMP	-0.027688	-0.000983	0.020991	-0.007240	0.149364

Table 19: Average differences between class thresholds for various numbers of classes.

Procedure	**2**	**3**	**4**	**5**
UTADISMP1	0.0884	0.0634	0.0475	0.0385
UTADISMP2	0.0836	0.0622	0.0486	0.0409
UTADISMP3	0.0595	0.0569	0.0497	0.0443
UTADIS-JLS	0.1459	0.1079	0.0827	0.0650
CHEBYSHEV	0.0583	0.0497	0.0411	0.0352
MSCVF	0.0928	0.0783	0.0693	0.0628
ACUTADIS	0.0547	0.0440	0.0346	0.0304
CENTROID	0.0592	0.0496	0.0407	0.0349
REPDIS	0.1066	0.0707	0.0533	0.0431
CAI	0.1390	0.0837	0.0586	0.0458
APOI	0.1389	0.0835	0.0570	0.0454
COMB	0.1388	0.0835	0.0571	0.0455
ROBUST-ITER	0.1828	0.0904	0.0624	0.0477
ROBUST-COMP	0.1720	0.0944	0.0685	0.0563

such as CHEBYSHEV, ACUTADIS, CENTROID, UTADISMP3, and MSCVF, the differences are negligible, or even the trend becomes inverse. For the first three methods mentioned above, this observation confirms that for various levels of flexibility of MVFs, the center-oriented methods are stable and reproduce the DM's preference model well. Furthermore, the stability of threshold-based similarity values for MSCVF and UTADISMP3 is likely because these methods emphasize the shape of the MVFs. MSCVF aims for functions that are as linear as possible, and UTADISMP3 opts for the highest possible and, therefore, equal distances between consecutive points.

The impact of the number of reference alternatives per class on the difference between class thresholds in the reference and resulting models is reported in Table 22. Clearly, these differences decrease for all methods with additional assignment examples. On the one hand, the most considerable reduction between instances with three and ten reference alternatives per class is observed for the ROBUST methods. This corresponds with the trend already explained for different numbers of classes. On the other hand, the least reductions between the extreme R values are noted for the stochastic and central-based procedures. The former approaches perform rather poorly when few reference alternatives are available, and the space of compatible sorting models is large. In turn, the latter ones achieve stable, good performance regardless of the number of reference alternatives, and the enriched preference information helps them reproduce the reference model even more faithfully.

Table 23 indicates that the performance of procedures concerning changes in individual parameters is similar. The distance from the reference thresholds decreases most when more classes and fewer characteristic points are considered. Modifying these dimensions implies additional restrictions resulting from more assignments or reducing the model's flexibility. The most sensitive to the changes mentioned above are the ROBUST methods, the stochastic approaches, and the UTADIS-JLS. In turn, the slope coefficients are low for the procedures that aim at central or the most discriminating models. Hence they are more resilient to changes, at least in the context of restoring the threshold values. The stability of results attained by these methods allowed them to stay ahead of the other approaches regardless of the problem setting. This is confirmed by Figure 7, exhibiting the statistically significant differences derived from the Wilcoxon signed-rank test.

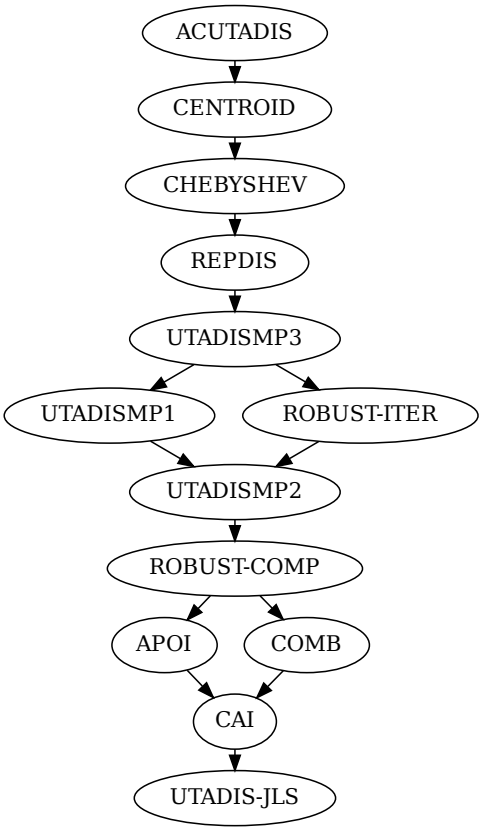


Figure 6: The Hasse diagram indicating the statistically significant differences in terms of the differences from the comprehensive values of the reference models based on the Wilcoxon test with p-value equal to 0.05.

Table 20: Average differences between class thresholds for various numbers of criteria.

Procedure	**3**	**5**	**7**	**9**
UTADISMP1	0.0681	0.0597	0.0553	0.0547
UTADISMP2	0.0671	0.0592	0.0548	0.0542
UTADISMP3	0.0661	0.0546	0.0467	0.0431
UTADIS-JLS	0.1143	0.0938	0.0944	0.0990
CHEBYSHEV	0.0581	0.0477	0.0402	0.0383
MSCVF	0.0914	0.0780	0.0698	0.0640
ACUTADIS	0.0513	0.0422	0.0363	0.0340
CENTROID	0.0573	0.0478	0.0406	0.0388
REPDIS	0.0850	0.0674	0.0617	0.0596
CAI	0.0831	0.0814	0.0792	0.0835
APOI	0.0833	0.0808	0.0787	0.0822
COMB	0.0833	0.0808	0.0786	0.0822
ROBUST-ITER	0.0990	0.0947	0.0942	0.0954
ROBUST-COMP	0.1113	0.0971	0.0927	0.0901

Table 21: Average differences between class thresholds for various numbers of characteristic points.

Procedure	**2**	**4**	**6**
UTADISMP1	0.0238	0.0761	0.0785
UTADISMP2	0.0264	0.0733	0.0767
UTADISMP3	0.0282	0.0676	0.0620
UTADIS-JLS	0.0234	0.1157	0.1620
CHEBYSHEV	0.0214	0.0602	0.0566
MSCVF		0.0803	0.0712
ACUTADIS	0.0180	0.0540	0.0508
CENTROID	0.0181	0.0615	0.0587
REPDIS	0.0386	0.0807	0.0860
CAI	0.0221	0.0914	0.1319
APOI	0.0227	0.0917	0.1292
COMB	0.0227	0.0918	0.1292
ROBUST-ITER	0.0507	0.1092	0.1275
ROBUST-COMP	0.0462	0.1120	0.1352

Table 22: Average differences between class thresholds for various numbers of reference alternatives per class.

Procedure	**3**	**5**	**7**	**10**
UTADISMP1	0.0771	0.0623	0.0536	0.0447
UTADISMP2	0.0755	0.0614	0.0535	0.0449
UTADISMP3	0.0657	0.0537	0.0485	0.0425
UTADIS-JLS	0.1380	0.1063	0.0868	0.0704
CHEBYSHEV	0.0565	0.0472	0.0430	0.0376
MSCVF	0.0934	0.0764	0.0702	0.0632
ACUTADIS	0.0498	0.0421	0.0380	0.0337
CENTROID	0.0568	0.0475	0.0424	0.0377
REPDIS	0.0865	0.0714	0.0623	0.0536
CAI	0.1016	0.0862	0.0736	0.0658
APOI	0.1009	0.0853	0.0732	0.0655
COMB	0.1009	0.0853	0.0732	0.0655
ROBUST-ITER	0.1396	0.1009	0.0791	0.0636
ROBUST-COMP	0.1400	0.1013	0.0825	0.0674

Table 23: Coefficients of solutions obtained for the linear regression problems for the average difference from the thresholds of the reference models depending on the defined dimensions for individual procedures.

Procedure	**No. of classes**	**No. of criteria**	**No. of ch. points**	**No. of ref. alt.**	**Intercept**
UTADISMP1	-0.016560	-0.002237	0.013671	-0.004507	0.104303
UTADISMP2	-0.014172	-0.002145	0.012584	-0.004244	0.097485
UTADISMP3	-0.005290	-0.003843	0.008430	-0.003177	0.080309
UTADIS-JLS	-0.026804	-0.002264	0.034660	-0.009425	0.128048
CHEBYSHEV	-0.007782	-0.003346	0.008778	-0.002597	0.074495
MSCVF	-0.009914	-0.004530	-0.004550	-0.004089	0.185982
ACUTADIS	-0.008207	-0.002878	0.008198	-0.002219	0.067990
CENTROID	-0.008175	-0.003141	0.010150	-0.002636	0.069441
REPDIS	-0.020774	-0.004104	0.011866	-0.004586	0.146961
CAI	-0.030461	-0.000043	0.027457	-0.005076	0.110568
APOI	-0.030681	-0.000271	0.026619	-0.005001	0.115014
COMB	-0.030655	-0.000265	0.026625	-0.004996	0.114836
ROBUST-ITER	-0.043315	-0.000578	0.019208	-0.010544	0.239967
ROBUST-COMP	-0.037290	-0.003400	0.022253	-0.009983	0.222106

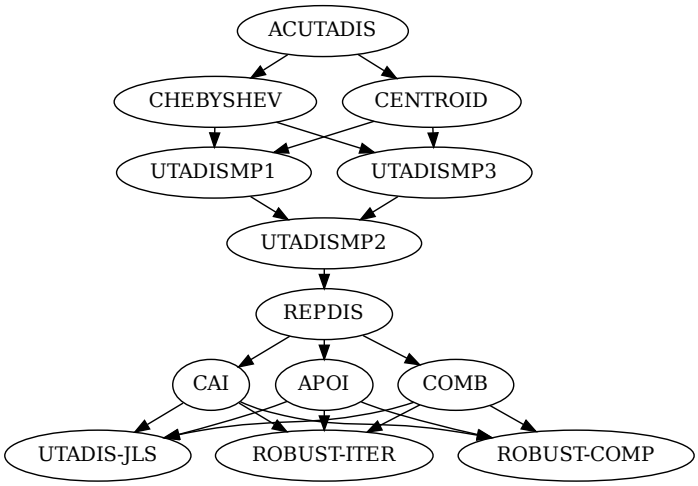


Figure 7: The Hasse diagram indicating the statistically significant differences in terms of the differences from the thresholds of the reference models based on the Wilcoxon test with p-value equal to 0.05.

Publication [P3]

M. Wójcik and M. Kadziński, "Nature-inspired Preference Learning Algorithms Using the Choquet Integral", in *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '24, (New York, NY, USA), p. 440–448, Association for Computing Machinery, 2024

DOI: 10.1145/3638529.3654054.

Nature-inspired Preference Learning Algorithms Using the Choquet Integral

Michał Wójcik*
michal.wojcik@cs.put.poznan.pl
Faculty of Computing and Telecommunication,
Poznań University of Technology
Poznań, Poland

Miłosz Kadziński
milosz.kadzinski@cs.put.poznan.pl
Faculty of Computing and Telecommunication,
Poznań University of Technology
Poznań, Poland

ABSTRACT

We introduce various algorithms for learning the parameters of a threshold-based sorting procedure powered by the Choquet integral. This model accounts for interactions between monotonic criteria and facilitates categorizing decision alternatives into predefined, preferentially ordered classes. We focus on developing heuristic preference learning methods capable of efficiently processing large datasets of classification examples. Specifically, we utilize Local Search, Simulated Annealing, and nature-inspired approaches such as Genetic Algorithm, Fish School Search, and Particle Swarm Optimization. We demonstrate the effectiveness of the proposed model through a case study. Additionally, we present an experimental comparison of the recommendation accuracy achieved by these algorithms on a suite of benchmark sorting problems.

CCS CONCEPTS

• **Applied computing** → **Multi-criterion optimization and decision-making**; • **Theory of computation** → **Evolutionary algorithms**.

KEYWORDS

Preference learning, Choquet integral, Evolutionary algorithm, Particle swarm optimization, Fish school search

ACM Reference Format:
Michał Wójcik and Miłosz Kadziński. 2024. Nature-inspired Preference Learning Algorithms Using the Choquet Integral. In *Genetic and Evolutionary Computation Conference (GECCO '24), July 14–18, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3638529.3654054

1 INTRODUCTION

Preference learning is a subfield of machine learning focused on predicting or inferring preferences [10]. It primarily addresses instance ranking or sorting challenges [1]. They involve allocating alternatives to preference-ordered categories based on multiple criteria. The significant challenges in preference learning include dealing with incomplete, inconsistent preference information in the form of desired assignments for a subset of alternatives, scalability to large preference sets, and ensuring high interpretability of the applied models [16].

This study explores the parameterization of a sorting procedure using the Choquet integral [7]. This model facilitates the aggregation of individual criteria while accounting for complex interactions (e.g., complementarity or redundancy) via non-additive measures. For this purpose, it incorporates meaningful preference parameters, called capacities, assigning weights to individual criteria and their combinations [11]. Its flexibility and general applicability make it a powerful tool in various application fields across economics, operations research, and decision theory. For example studies, see [3, 5, 25, 31].

However, applying the Choquet integral is challenging due to the need to determine capacities for all criteria subsets [30]. To mitigate the cognitive load of specifying numerous parameters, various preference learning techniques have been proposed [15]. They include logistic regression adaptations [29], convex quadratic programming [22], and neural network approaches [23]. Yet, these methods primarily incorporate a two-additive Choquet integral, capturing only pairwise criteria interactions. Other recent works focus on more advanced representations of interactions [9, 14, 20], focusing on sparse, compact, or contextual model variants.

This paper presents a comprehensive suite of algorithms for learning the Choquet integral model parameters within a threshold-based sorting framework. We focus on learning from large sets of assignment examples and deriving compatible capacities for all subsets of criteria. The proposed methods are based on linear programming and adapted metaheuristics, including local search variants, simulated annealing, genetic algorithms, fish school search, and particle swarm optimization. The model's interpretability is showcased through parameter illustration for a real-world problem. Then, predictive performance is evaluated using a set of monotonic learning benchmarks, with classification accuracy and Area Under the Curve as metrics. The influence of learning and testing set sizes on outcomes is also examined.

2 PROBLEM DEFINITION

This section delineates the Choquet integral, the scope of sorting, and the form of Decision Maker's (DM's) preferences.

Let us consider a multiple criteria sorting problem involving assigning n alternatives in set $A = \{a_1, \ldots, a_n\}$, evaluated on a family of m criteria $G = \{g_1, \ldots, g_m\}$, to one of predefined, preferentially ordered classes from set $C = \{C_1, \ldots, C_p\}$, where C_{l+1} is preferred

*Corresponding author.

GECCO '24, July 14–18, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0494-9/24/07…$15.00
https://doi.org/10.1145/3638529.3654054

to C_l for each $l \in \{1, \ldots, p-1\}$. The performance of an alternative a_i on criteria $g_j : A \to \mathbb{R}$ is denoted by $g_j(a_i)$. Without loss of generality, we assume that $g_j(a_i) \in [0, 1]$ and all criteria are of gain type, i.e., the greater the performance $g_j(a_i)$, the more preferred it is for the DM.

The Choquet integral. The Choquet integral is based on the concept of fuzzy or non-additive measure (also called *capacity*) [7]. Given a set of criteria G, a fuzzy measure incorporated into preference model M is defined for each subset of G, as a set function $\mu_M : 2^G \to [0, 1]$ with the following normalization and monotonicity assumptions:

$$\mu_M(\emptyset) = 0, \qquad \mu_M(G) = 1, \tag{1}$$

$$\mu_M(G_1) \leq \mu_M(G_2), \text{ for each } G_1 \subseteq G_2 \subseteq G. \tag{2}$$

Then, the Choquet integral is defined as a function $Ch_M : A \to \mathbb{R}$:

$$Ch_M(a) = \sum_{j=1}^{m} [g_{(j)}(a) - g_{(j-1)}(a)] \cdot \mu_M(G_{(j)}), \tag{3}$$

where $(\cdot)$ is a permutation of $\{1, \ldots, m\}$, such that: $g_{(0)}(a) = 0 \leq g_{(1)}(a) \leq \ldots \leq g_{(m)}(a)$, and $G_{(j)} = \{g_{(1)}(a), \ldots, g_{(j)}(a)\}$. When $g_j(a) \in [0, 1]$ for each $j \in \{1, \ldots, m\}$, then $Ch_M(a) \in [0, 1]$. To support the understanding of the above notation, the supplementary material illustrates calculating the Choquet integral value for an example alternative.

Threshold-based sorting procedure. To perform the assignment, we use a score-driven threshold-based sorting procedure [12]. In this study, the score is expressed as the Choquet integral $Ch_M(a) : A \to \mathbb{R} \in [0, 1]$. Moreover, $p + 1$ separating class thresholds $(t_0, t_1, \ldots, t_p)$ complete model M, such that:

$$t_0 = 0, \qquad t_l - t_{l-1} \geq \varepsilon, \; l \in 1, \ldots, p, \qquad t_{p-1} \leq 1 - \varepsilon, \qquad t_p > 1, \tag{4}$$

where ε value is an arbitrarily small positive value. The assignment of alternative a_i to class C_l is implied by the following conditions:

$$I_M(a_i) = l \iff t_{l-1} \leq Ch_M(a_i) < t_l. \tag{5}$$

Hence $I_M : A \to \mathbb{N} \in \{1, \ldots, p\}$ is the index of a class to which a_i is assigned using model M.

Preference information. We assume the DM provides desired assignments for a subset of reference alternatives $A^R \subseteq A$. This can be modelled using function $I_{DM} : A^R \to \mathbb{N} \in \{1, \ldots, p\}$.

Model evaluation. When learning the parameters of model M, some approaches need to evaluate M using a loss function. For this purpose, we use the *regret* function $r_M : A^R \to \mathbb{R} \in [0, 1]$ [23]:

$$r_M(a_i^*) = max\{t_{I_{DM}(a_i^*)-1} - Ch_M(a_i^*), \; 0, \; Ch_M(a_i^*) - t_{I_{DM}(a_i^*)}\}. \tag{6}$$

It is equal to zero when the score of a_i^* falls into the range associated with the class specified by the DM. Otherwise, it captures the distance from the nearest threshold of the desired class. Then, the loss function $L(M) : M \to \mathbb{R} \in [0, 1]$ for model M aggregates the regrets for all reference alternatives:

$$L(M) = \frac{1}{|A^R|} \sum_{a_i^* \in A^R} r_M(a_i^*). \tag{7}$$

The lesser $L(M)$, the more preferred model M as it better fits the DM's indirect preferences, i.e.:

$$M_1 \succ M_2 \iff L(M_1) < L(M_2). \tag{8}$$

3 PREFERENCE LEARNING APPROACHES

This section describes novel preference learning approaches that aim to find sorting model M that best reflects the DM's preferences, thus minimizing $L(M)$. They represent various streams of algorithms. In particular, we consider a) local and global search methods, b) single solution and population-based techniques as well as ensemble approaches, or c) linear programming and nature-inspired techniques. The interest in mathematical programming derives from their prevailing role in decision analysis [2]. In turn, the remaining methods proved their ability to efficiently search a continuous, constrained space of solutions. They also make use of convexity, ensuring that a linear combination of two feasible solutions leads to a valid solution (e.g., crossover in genetic algorithm or motion direction combination in particle swarm optimization).

3.1 Mathematical Programming

The first group of methods is based on Linear Programming (LP). They differ in terms of the optimized objective function.

*3.1.1 Minimize maximum regret [**MMR**].* The first approach minimizes the highest value of regret $r_M(a_i^*)$ across all reference alternatives, marked as e, i.e.:

$$\begin{aligned} & \text{Minimize } e, \\ s.t. \quad & \left.\begin{array}{ll} \textit{eqs. (1)-(6)}, & \\ t_{I_{DM}(a_i^*)-1} - Ch_M(a_i^*) \leq e, & \forall a_i^* \in A^R, \\ Ch_M(a_i^*) - t_{I_{DM}(a_i^*)} + \varepsilon \leq e, & \forall a_i^* \in A^R, \\ e \geq 0. & \end{array}\right\} (E^{A^R}_{MMR}) \end{aligned} \tag{9}$$

*3.1.2 Minimize the number of misclassified alternatives [**MNR**].* The other approach minimizes the number of alternatives for which the regret is positive, i.e., $r_M(a_i^*) > 0$. For this, we introduce binary variables b_{e_i} indicating the misclassification of $a_i^* \in A^R$:

$$\begin{aligned} & \text{Minimize } \sum_{a_i^* \in A^R} b_{e_i}, \\ s.t. \quad & \left.\begin{array}{ll} \textit{eqs. (1)-(6)}, & \\ t_{I_{DM}(a_i^*)-1} - Ch_M(a_i^*) \leq b_{e_i}, & \forall a_i^* \in A^R, \\ Ch_M(a_i^*) - t_{I_{DM}(a_i^*)} + \varepsilon \leq b_{e_i}, & \forall a_i^* \in A^R, \\ b_{e_i} \in \{0, 1\}, & \forall a_i^* \in A^R. \end{array}\right\} (E^{A^R}_{MNR}) \end{aligned} \tag{10}$$

3.1.3 General approach. The solution to each of the above mathematical programming problems requires determining 2^m capacities μ_M, $p + 1$ threshold values, and 1 real or $|A^R|$ binary values capturing variously defined misclassification errors. Their values must satisfy the desired properties of capacities μ_M for all subsets of G and constraints implied by the assignments of reference alternatives. When the number of variables and constraints is high, finding an optimal solution within a limited time may be impossible.

To address this problem, we apply a bagging-inspired approach [6]. The procedure incorporating [**MMR**] is described as Algorithm 1. It requires two parameters – $S \in \mathbb{N}_+$ indicating the number of models to be aggregated and $p \in (0, 1]$ indicating the fraction of reference alternatives that should be used for training a

Algorithm 1 MMR-based search [**MMR**]

Input: S – no. of models, p – proportion of ref. alts. to select
m^* – best model found
$M[0 \dots S-1]$ – empty array of S models
$t \leftarrow |A^R| \cdot p$
$i \leftarrow 0$
while not stopping condition **do**
  $T^R \leftarrow \lfloor t \rfloor$ randomly selected alternatives from A^R
  $m \leftarrow MMR(T^R)$
  if $i < S$ **then**
    $M[i] \leftarrow m$
  else if $m \succ M[S-1]$ **then**
    $M[S-1] \leftarrow m$
  end if
  $M \leftarrow sort(M)$
  if $M[0] \succ m^*$ **then**
    $m^* \leftarrow M[0]$
  end if
  if $getWeightedAverageModel(M) \succ m^*$ **then**
    $m^* \leftarrow getWeightedAverageModel(M)$
  end if
  $i \leftarrow i + 1$
end while
return m^*

single model. The algorithm iteratively creates subsequent models and then checks whether the found models are better than the existing ones based on the loss function value. The following symbols appear in the algorithm: T^R is a subset of randomly selected alternatives from A^R without replacement, $MMR(T^R)$ returns the model obtained by solving the LP problem defined in Section 3.1.1, $sort(M)$ sorts the array of models M according to values $L(M[i])$ for each $i \in \{0, \dots, S-1\}$ in the ascending order, and $getWeightedAverageModel(M)$ determines a model that is the weighted average of all models in M, where the weight of model $M[i]$ is $w_i = (L(M[i]) + \varepsilon)^{-1}$:

$$\mu_{M_{avg}}(G') = \frac{\sum_{i=0}^{S-1} w_i \cdot \mu_{M[i]}(G')}{\sum_{i=0}^{S-1} w_i}, \text{ for each } G' \subseteq G,$$
$$t_{l_{M_{avg}}} = \frac{\sum_{i=0}^{S-1} w_i \cdot t_{l_{M[i]}}}{\sum_{i=0}^{S-1} w_i}, \text{ for each } l \in 0, \dots, p. \tag{11}$$

This way, higher-quality models have a more significant impact on the average solution. At the same time, the algorithm can also choose one of the solutions directly produced by $MMR(T^R)$. A similar algorithm can be defined for [**MNR**] when using $MNR(T^R)$ rather than $MMR(T^R)$.

3.2 Local Search

Another suite of approaches performs an optimization through effective exploration and exploitation of the solution space consistent with Eqs. (1)-(6). To sample models, we use the Hit-And-Run (HAR) algorithm [26] implemented in [8]. It generates a sequence of feasible models that asymptotically approach a uniform distribution. At this stage, we do not impose any constraints implied by the DM's assignment example. To define the following algorithms, we assume that function $getRandomModel()$ uses HAR to generate a random model respecting Eqs. (1)-(6).

Let us start with defining the neighborhood relation between models M_1 and M_2 representing the underlying parameter values:

$$v_1 = [\mu_{M_1}(G_0), \dots, \mu_{M_1}(G_{2^{|G|}-1}), t_{0_{M_1}}, \dots, t_{p_{M_1}}],$$
$$v_2 = [\mu_{M_2}(G_0), \dots, \mu_{M_2}(G_{2^{|G|}-1}), t_{0_{M_2}}, \dots, t_{p_{M_2}}]. \tag{12}$$

They are deemed *neighbors* ($N_r(M_1, M_2, r) = 1$) when the Euclidean distance between v_1 and v_2 is not greater than r (radius), being the algorithm's parameter, i.e.:

$$N_r(M_1, M_2, r) = \begin{cases} 1, & \text{if } ||v_1, v_2|| \leq r, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Function $generateNeighbor(m, r)$ creates neighbor m' of model m with the maximum distance of r. First, it generates a random vector (point) from the n-dimensional unit sphere and scales its length by a random factor $k \in (0, r]$. Then, it is added to the initial solution vector v_m to obtain the neighboring solution $v_{m'}$. Such a solution is not guaranteed to be feasible. In this case, we use the *crop* strategy to reduce the vector's length so that the resulting solution is located on or close to the feasible solution space boundary.

The subsequently presented variants of local search attempt to find the optimal solution by iteratively searching neighboring solutions of the currently selected one.

*3.2.1 Greedy Local Search [**GLS**].* In the *greedy* version of local search [28], we accept neighbor m_N of the current model m whenever $L(m_N) < L(m)$. To prevent getting stuck in the local optimum in the case of no improvement for a long time during the search, we use the parameter S_N specifying the maximum number of neighbors to be verified. If the method fails to find a better model in S_N iterations, the best solution found m^* is updated (if necessary), and the algorithm continues by starting with a new, randomly selected model. The pseudocode of [**GLS**] is described as Algorithm 2.

*3.2.2 Steep Local Search [**SLS**].* The *steep* variant of local search [28] generates S_N neighbors of the current solution and selects the one providing the most significant improvement in model quality. Similar to **GLS**, if no newly generated model provides a lower loss function value, the best-found solution m^* is updated, and the search is repeated from the randomly generated solution. This strategy is expected to converge to the local optima faster than **GLS**. However, it needs to examine S_N neighbors in each iteration. This number is usually smaller in **GLS**, which can facilitate the search process.

*3.2.3 Simulated Annealing [**SAN**].* Simulated Annealing [18] applies a different acceptance criterion. Specifically, neighbor m_N is accepted as a new solution when its quality is better than the initial solution m. However, m_N can also be accepted with a certain probability even when being worse than m. Then, the probability depends on the quality difference between m and m_N as well as parameter t, called *temperature*. Overall, the acceptance probability can be defined as follows:

$$P(m \leftarrow m_N) = \begin{cases} 1, & \text{if } L(m_N) < L(m), \\ e^{-\frac{L(m_N)-L(m)}{t}}, & \text{otherwise.} \end{cases} \tag{14}$$

Algorithm 2 Greedy Local Search [**GLS**]

Input: S_N – number of neighbors, r – neighborhood radius

$m \leftarrow getRandomModel()$
$m^* \leftarrow m$
while not stopping condition **do**
  $p \leftarrow 0$
  while $p < S_N$ **do**
    $m_N \leftarrow generateNeigbor(m, r)$
    if $m_N \succ m$ **then**
      $m \leftarrow m_N$
      break
    end if
    $p \leftarrow p + 1$
  end while
  if $m \succ m^*$ **then**
    $m^* \leftarrow m$
  end if
  if $p == S_N$ **then**
    $m \leftarrow getRandomModel()$
  end if
end while
return m^*

The value of t is initially set to *initial temperature* t_s, and then successively decreased. This reduces the probability of accepting a worse solution and increases the pressure to obtain better ones. After each iteration, the value of t is multiplied by t_{dr}, indicating the *temperature decrease ratio*. When t reaches a value lesser than *minimum temperature* t_{min}, the search re-starts with a random model and $t = t_s$. Note that for this algorithm, unlike the previous two approaches, we consider only one neighbor model in each iteration. Therefore, the S_N parameter is not used here. The algorithm returns the best solution found during the search.

3.3 Genetic Algorithm

The genetic algorithm [**GEN**], a pioneering nature-inspired metaheuristic [19], is employed for preference learning, as described in Algorithm 3. It comprises a series of steps executed iteratively, simulating successive generations of individuals and applying evolutionary processes. In what follows, we detail functions and parameters integrated into [**GEN**]:

- $generatePopulation(S)$ returns S random models using the $getRandomModel()$ function;
- $selection(M_p, s_S, t_S)$ chooses 2 out of S individuals from array M_p using one of the following procedures:
  - **RWS** *Roulette wheel selection* [21] with the probability of selecting individual m_i being proportional to $\frac{1}{L(m_i)+\varepsilon}$;
  - **TS** *Tournament selection* [24] with t_S individuals drawn and the one with the lowest $L(m)$ among them being selected.
- $crossover(m_1, m_2, s_C, p_C)$ returns an individual resulting from $getWeightedAverageModel([m_1, m_2])$ with probability p_C and weights determined using one of the two schema:
  - **AC** *Average Crossover* with both weights equal to 0.5;
  - **RC** *Random Crossover* with weights equal to α and $1 - \alpha$, where $\alpha \in [0, 1]$ is a random value;

  otherwise, the function returns an individual identical to m_1.
- $mutation(m_o, s_M, p_M, r_M)$ makes small changes to an individual to increase diversity in the solution pool. With probability p_M, it applies one of two strategies to individual m_o:
  - **NM** *Neighbor Mutation* generating a neighbor model, using $generateNeighbor(m_o, r_M)$;
  - **SWM** *Single Weight Mutation* changing only one parameter $\mu_M(G')$ for a randomly selected $G' \subset G \neq \emptyset$, by checking lower ($\mu_M^{LB}(G')$) and upper ($\mu_M^{UB}(G')$) bounds that meet the constraints, and selecting a random value in $[\mu_M^{LB}(G'), \mu_M^{UB}(G')]$;

  otherwise, the function returns m_o without modifying it.
- $elitismSelection(M_p, M_o)$ returns S models with the lowest $L(m)$ values among $2 \cdot S$ individuals from M_p and M_o;
- $bestIndividual(M_p)$ returns the model with the lowest $L(m)$ value among S individuals from M_p;
- st indicates how many generations without improvement are acceptable; if this happens, the algorithm restarts from a random population of individuals (models).

Algorithm 3 Genetic Algorithm [**GEN**]

Input: S – population size, st – max. stagnation, s_S – selection strategy, t_S – tournament size, p_C – crossover probability, s_C – crossover strategy, p_M – mutation probability, s_M – mutation strategy, r_M – mutation range

m^* – best model found
while not stopping condition **do**
  $M_p[0, \ldots, S-1] \leftarrow generatePopulation(S)$
  $st_{ctr} \leftarrow 0$
  while $st_{ctr} < st$ **do**
    $i \leftarrow 0$
    $M_o[0 \ldots S-1]$ – empty array of S models
    while $i < S$ **do**
      $m_1, m_2 \leftarrow selection(M_p, s_S, t_S)$
      $m_o \leftarrow crossover(m_1, m_2, s_C, p_C)$
      $M_o[i] \leftarrow mutation(m_o, s_M, p_M, r_M)$
      $i \leftarrow i + 1$
    end while
    $M_p \leftarrow elitismSelection(M_p, M_o)$
    $m \leftarrow bestIndividual(M_p)$
    if $m \succ m^*$ **then**
      $m^* \leftarrow m$; $st_{ctr} \leftarrow 0$
    else
      $st_{ctr} \leftarrow st_{ctr} + 1$
    end if
  end while
end while
return m^*

3.4 Fish School Search

Fish School Search draws inspiration from the social behavior observed in certain fish species, with solutions evolving through a simulation of school dynamics. The algorithm searches for the best model by repeating a sequence of fish movements, which can be divided into:

- *individual* – here, each fish independently explores its immediate surroundings in search of "food", metaphorically representing

an opportunity to enhance the model's quality. Success in this endeavor leads to an increase in the fish's weight;
- *collective-instinctive* – this phase mimics the successful *individual* movements of other fish, specifically those that resulted in fitness improvements.
- *collective-volitive* – the school's radius is adjusted based on the collective performance, expanding or contracting in response to the overall increase or decrease in the school's mass, respectively. This adjustment reflects the school's aggregate success in finding better solutions, aiming to boost the group's exploratory capabilities.

Our implementation of [**FSS**] is largely based on [4], and the detailed procedure is outlined in Algorithm 4. In what follows, we delve into the algorithm's key components and discuss the adaptations made to enhance its performance:

- The $generateFishPopulation(S)$ function creates models similarly to $generatePopulation(S)$, but with the addition of generating S fishes. Each fish is assigned a weight W_i, initialized following the approach outlined in [4].
- The three *Movement* operators are implemented in a manner consistent with those described in [4]. After each movement, m^* is updated if the new solution obtained is better. Additionally, for all three operators, if a model movement leads to a constraint violation, a *crop* strategy is employed to rectify this.
- For $individualMovement$, the fitness function is $f(m) = -L(m)$, implying that a reduction in $L(m)$ results in increased fitness. To compute the *movement vector*, a new model m' is generated via $m' = getNeighbor(m, s_{ind})$, and the displacement vector is calculated as $v_d = v_{m'} - v_m$.
- The $feeding$ operator has been slightly adapted with respect to [4]. While fish weight increases following model improvement after $individualMovement$, it decreases otherwise, being multiplied by w_{dr}, where $w_{dr} \in (0, 1)$.
- Given the absence of a fixed number of iterations, adjustments were made to how the values of input parameters s_{ind} and s_{vol} diminish over time. The decay rates $s_{ind_{dr}}$ and $s_{vol_{dr}}$ are chosen from within the interval $(0, 1)$ to ensure a gradual reduction in both parameters' values.

Algorithm 4 Fish School Search [**FSS**]

Input: S – population size, $s_{ind_{start}}$ – initial individual movement step, $s_{ind_{dr}}$ – individual movement decrease ratio, $s_{vol_{start}}$ – initial volitive movement step, $s_{vol_{dr}}$ – volitive movement decrease ratio, w_{scale} – weights scale, w_{dr} – weight decrease ratio

m^* – best model found
$F[0, \ldots, S-1] \leftarrow generateFishPopulation(S)$
$s_{ind} \leftarrow s_{ind_{start}}, \quad s_{vol} \leftarrow s_{vol_{start}}$
while not stopping condition **do**
  $individualMovement(F, s_{ind})$
  $feeding(F, w_{dr})$
  $collectiveInstinctiveMovement(F)$
  $colectiveVolitiveMovement(F, s_{vol})$
  $s_{ind} \leftarrow s_{ind} \cdot s_{ind_{dr}}; \quad s_{vol} \leftarrow s_{vol} \cdot s_{vol_{dr}}$
end while
return m^*

3.5 Particle Swarm Optimization

Particle Swarm Optimization, initially introduced in [17], draws inspiration from the dynamics observed within large flocks of birds. It posits that each particle in the swarm adjusts its trajectory based on a velocity vector, which is influenced by cognitive and social components. The procedural framework of [**PSO**] is detailed in Algorithm 5 and encompasses the following steps:

- $generateParticlePopulation(S, v_r)$ returns S particles that contain the current model m, the best found model m_p, the best globally found model m_g, and a velocity vector v whose length does not exceed v_r;
- $updateBestParticleModel(P)$ updates the best model found so far by a given particle – if $P[i]_m \succ P[i]_{m_p}$, then $P[i]_{m_p} \leftarrow P[i]_m$; this model is also assigned to m^*, if $P[i]_m \succ m^*$;
- $updateBestGlobalModel(P)$ updates the best model found so far by all of the particles – if $m_p^* \succ P[i]_{m_g}$, then $P[i]_{m_g} \leftarrow m_p^*$, where m_p^* is the best model among $P[i]_{m_p}$;
- $moveParticles(P, w, c_1, c_2)$ updates the velocity vector and then the model parameters. We assume that $P[i]_m$, $P[i]_{m_p}$ $P[i]_{m_g}$ contain a vector representation of the individual models. For random values $r_1, r_2 \in [0, 1]$, the update formulas are as follows:
  – $P[i]_v \leftarrow w \cdot P[i]_v + c_1 \cdot r_1 \cdot (P[i]_{m_p} - P[i]_m) + c_2 \cdot r_2 \cdot (P[i]_{m_g} - P[i]_m)$;
  – $P[i]_m \leftarrow P[i]_m + P[i]_v$;
  As in [**FSS**], when the model fails to meet the constraints after movement, we use the *crop* strategy to shorten the model's shift-vector.

Algorithm 5 Particle Swarm Optimization [**PSO**]

Input: S – population size, v_r – initial velocity radius, w, c_1, c_2 – velocity vector modification coefficients

m^* – best model found
$P[0, \ldots, S-1] \leftarrow generateParticlePopulation(S, v_r)$
while not stopping condition **do**
  $updateBestParticleModel(P)$
  $updateBestGlobalModel(P)$
  $moveParticles(P, w, c_1, c_2)$
end while
return m^*

4 POST-OPTIMIZATION TECHNIQUES

In this section, we present two post-optimization strategies designed to: (a) enhance the model's accuracy by identifying optimal threshold values, and (b) minimize $L(m)$ using the backpropagation algorithm. These techniques are applied across all discussed algorithms – the former is utilized during the optimization phase for each model obtained, and the latter is employed for the final, optimal model m^* produced by a particular method.

4.1 Heuristic to optimize threshold values

The values of separating class thresholds t_l, $l = 1, \ldots, p-1$, can be optimized by preference learning algorithms similarly to capcities $\mu_(G')$. Then, they are modified as an inherent part of the optimized solutions in line with the operations of the specific algorithm (e.g., mutation and crossover for [**GEN**]). Alternatively, the

methods can learn only the Choquet integral parameters, while the following approach can determine the thresholds:

- Given capacities‘μ_M, for each combination of $\{t'_1, \ldots, t'_{p-1}\}$, calculate the classification accuracy over $a_i^* \in A^R$;
- Select the combination $\{t_1'^*, \ldots, t_{p-1}'^*\}$ which provides the highest accuracy and assign it to model M.

Even though each threshold t'_l can take any value in $[0, 1]$, we can limit the search space to values $Ch(a_i^*) \pm \epsilon$ for each $a_i^* \in A^R$. Searching such a space for small p and reasonable sizes of A^R is fast and guarantees the best possible outcome regarding the model accuracy. We applied this technique to each algorithm during experiments whenever a new solution appeared.

4.2 Backpropagation

For model M and alternative a_i^*, we can determine the prediction loss function $l_M(a_i^*) = \frac{r_M^S(a_i^*)^2}{2}$, where:

$$r_M^S(a_i^*) = \begin{cases} t_{I_{DM}(a_i^*)-1} - Ch_M(a_i^*), & \text{if } Ch_M(a_i^*) < t_{I_{DM}(a_i^*)-1}, \\ Ch_M(a_i^*) - t_{I_{DM}(a_i^*)}, & \text{if } Ch_M(a_i^*) > t_{I_{DM}(a_i^*)}, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

This approach facilitates the calculation of the gradient for the model parameters $\mu_M(G')$, enabling the application of the backpropagation algorithm commonly utilized in optimizing neural networks [13]. Consistent with its principles, the model parameters are updated in the following manner:

$$\mu_M(G')' = \mu_M(G') - \sum_{a_i^* \in A^R} \eta \frac{\delta l_M(a_i^*)}{\delta \mu_M(G')}, \quad \forall G' \subset G \neq \emptyset. \tag{16}$$

When updating individual parameters, there is a risk of constraint violations. To address this, we adopt a strategy where parameter η is progressively decreased to ensure model validity. This adjustment process continues until either a predefined processing time limit is reached or η diminishes to a minimal threshold of 10^{-12}, suggesting that further model enhancements are infeasible without breaching constraints. Notably, this methodology is equally applicable to threshold adjustments, although it is employed post-optimization of $\mu_M(G')$.

Our preliminary investigations revealed that allocating 5% of each algorithm's optimization duration to implementing the backpropagation scheme yields superior solutions. The supplementary material includes details.

5 EXPERIMENTAL ANALYSIS

This section presents the experimental analysis framework and the results that highlight the performance of the evaluated preference learning methods.

5.1 Benchmark problems

We evaluated the proposed algorithms using five benchmark datasets: two (Breast Cancer [BCC] ($|A| = 286$, $m = 7$) and Computer Processing Units [CPU] ($|A| = 209$, $m = 6$)) sourced from the UCI repository, and three (Employee Selection [ESL] ($|A| = 488$, $m = 4$), Employee Rejection/Acceptance [ERA] ($|A| = 1000$, $m = 4$), and Lecturers Evaluation [LEV] ($|A| = 1000$, $m = 4$)) obtained from the WEKA machine learning framework.

Table 1: Analyzed hyperparameter values.

Algorithms	Hyperparameters values
MMR, MNR	$S \in \{1, 5, 10\}, \quad p \in \{0.05, 0.1, 0.25, 0.5, 1\}$
GLS, SLS	$S_N \in \{10, 25, 100, 250\}$ $r \in \{0.01, 0.025, 0.1, 0.25, 1.0\}$
SAN	$t_{min} = 10^{-9}, \quad r \in \{0.01, 0.025, 0.1, 0.25, 1.0\}$ $t_s \in \{10, 100, 1000\}, \quad t_{dr} \in \{0.95, 0.99, 0.995\}$
GEN	$S \in \{25, 100, 250\}, \quad st = 10,$ $(s_S, t_S) \in \{(\textbf{RWS}, 0), (\textbf{TS}, 2), (\textbf{TS}, 5)\},$ $p_C \in \{0.8, 1.0\}, \quad s_C \in \{\textbf{AC}, \textbf{RC}\},$ $p_M \in \{0.2, 0.5, 0.8\}$ $(s_M, r_M) \in \{(\textbf{NM}, 0.01), (\textbf{SWM}, 0.0)\}$
FSS	$S \in \{25, 100, 250\},$ $s_{ind_{start}} \in \{0.01, 0.05\}, \quad s_{vol_{start}} = 2 \cdot s_{ind_{start}}$ $s_{ind_{dr}} = s_{vol_{dr}} \in \{0.95, 0.99\}$ $w_{scale} \in \{10, 50\}, \quad w_{dr} \in \{0.01, 0.02, 0.05\}$
PSO	$S \in \{25, 100, 250\}, v_r \in \{0.01, 0.1\}, (w, c_1, c_2) \in$ $\{(0.8, 0.1, 0.1), (0.6, 0.2, 0.2), (0.4, 0.3, 0.3),$ $(0.2, 0.4, 0.4), (0, 0.25, 0.75), (0, 0.75, 0.25)\}$

Consistent with previous studies in [23, 27, 29], we addressed the binary classification problem and applied min-max normalization to scale the values of individual criteria to the range $[0, 1]$. To demonstrate the capabilities of each method, we conducted a comparative analysis across three scenarios for each dataset, varying the distribution of alternatives between the A^R (reference set) and A^T (test set) subsets in the ratios of $20 - 80$, $50 - 50$, and $80 - 20$.

In the *supplementary material*, we illustrate using three algorithms (GEN, FSS, and PSO) to the ESL problem. We demonstrate how they minimize the loss function in a single run and discuss the obtained model parameter values.

5.2 Quality measures

To evaluate the algorithms, we used the following two performance measures:

- *Area Under Curve* [**auc**] is the percentage of accurately replicated comparisons among pairs of test alternatives $a_i, a_k \in A^T = A \backslash A^R$ that were anticipated to be categorized into distinct classes:

$$auc(M) = \frac{\sum_{a_i: I_{DM}(a_i)=1} \sum_{a_k: I_{DM}(a_k)=2} c_M(a_i, a_k)}{|a_i : I_{DM}(a_i) = 1| \cdot |a_k : I_{DM}(a_k) = 2|}, \; a_i, a_k \in A^T,$$
$$c_M(a_i, a_k) = \begin{cases} 1, & \text{if } Ch_M(a_i) < Ch_M(a_k), \\ 0, & \text{otherwise,} \end{cases}, \; a_i, a_k \in A^T. \tag{17}$$

- *Classification accuracy* [**acc**] reflect the model's average proficiency in correctly classifying test alternatives $a \in A^T$ into their respective classes. This metric is determined by the proportion of alternatives correctly assigned to the class specified by the DM out of the total number of alternatives evaluated:

$$acc(M) = \frac{|a \in A^T : I_M(a) = I_{DM}(a)|}{|A^T|}. \tag{18}$$

Table 2: Average and standard deviation of *acc* obtained by GEN, FSS, and PSO for three datasets, various p_R values, and 30 seconds of optimization.

Alg.	p_R	BCC	CPU	ESL
GEN	0.2	0.7234 ± 0.0249	0.9036 ± 0.0326	0.9197 ± 0.0112
	0.5	0.7278 ± 0.0287	0.9411 ± 0.0281	0.9228 ± 0.0139
	0.8	0.7355 ± 0.0540	0.9488 ± 0.0315	0.9271 ± 0.0246
FSS	0.2	0.7265 ± 0.0277	0.9053 ± 0.0305	0.9211 ± 0.0118
	0.5	0.7271 ± 0.0271	0.9214 ± 0.0243	0.9241 ± 0.0135
	0.8	0.7370 ± 0.0569	0.9257 ± 0.0362	0.9271 ± 0.0242
PSO	0.2	0.7271 ± 0.0223	0.9054 ± 0.0334	0.9209 ± 0.0126
	0.5	0.7297 ± 0.0272	0.9301 ± 0.0239	0.9261 ± 0.0141
	0.8	0.7336 ± 0.0589	0.9260 ± 0.0358	0.9286 ± 0.0253

5.3 Experimental setup

The experimental analysis was divided into two phases:

- *Hyperparameter Optimization – Preliminary Analysis*: The objective was to identify the optimal hyperparameter values for each combination (d, p_R, alg), where d is one of the five datasets, p_R denotes the proportion of the reference set A^R size relative to $|A|$, and alg signifies one of the eight algorithms. Table 1 shows the hyperparameters for each algorithm. To select the best setting, 10-fold Monte Carlo Cross Validation was used for each combination (d, p_R, alg). This approach creates 10 different problems by randomly dividing all alternatives into reference (A^R) and test (A^T) subsets, without replacement, per the assumed p_R. The assignments for A^R constituted preference information and enabled the evaluation of solutions at the optimization stage. In turn, the assignments for A^T were unknown to the algorithms, being used to estimate the quality of the obtained models in line with *acc* and *auc*. Subsequently, for each (d, p_R, alg), a set of hyperparameters was chosen that yielded the highest average value of $\frac{acc(m)+auc(m)}{2}$. Across all considered scenarios, the execution timeout was set to 10 seconds.
- *Comparative Analysis of Algorithms*: In this phase, we conducted a 100-fold Monte Carlo Cross-Validation for each (d, p_R, alg) combination using the chosen hyperparameters. We assumed a 30-second execution timeout for each algorithm.

5.4 Results

This section discusses the most interesting experimental results.

5.4.1 Impact of the reference set size (p_R) on model quality. Increasing $|A^R|$ enriches the model with more detailed preference information, enhancing its ability to align with the incoming data and reducing the likelihood of learning from biased information. However, this also elevates the complexity of the optimization task and extends the computational time required for model evaluation.

Table 2 displays the mean *acc* values for $p_r \in \{0.2, 0.5, 0.8\}$ across three algorithms and three datasets. For all pairs of scenarios except one, there are slight improvements in *acc* with an increase in $|A^R|$, suggesting that enriching the model with more preference information positively influences the quality measures. Notably, for BCC and ESL, the standard deviations also rise with an increase in $|A^R|$, hinting at enhanced result stability when $p_R = 0.2$. The observations for *auc* values align with these findings.

5.4.2 Comparison of algorithm performance. Tables 3 and 4 display the average *auc* and *acc* values achieved by all algorithms across various problems, with a reference set proportion $p_R = 0.8$, along with their respective rankings. We denote in bold the algorithms for which the differences to the best performer were not statistically significant. The standout performers were:

- **GEN**: Exhibiting the best average ranks for both *auc* ($r^{0.8}_{auc} = 1.2$) and *acc* ($r^{0.8}_{acc} = 2.6$) across all datasets. GEN secured the best average *auc* values for nearly all problems except one and maintained strong performance in *acc*, ranking in the upper half. However, its advantage is not statistically significant for some benchmarks, even though it shares the top place with various approaches. Its least favorable outcomes were observed for CPU and BCC, suggesting a potential preference for problems with a larger set of alternatives.
- **PSO** and **FSS**: Both algorithms demonstrated comparable performance, with average ranks of 3.8 for *auc* and 3.8 and 4.0 for *acc*, respectively, trailing only behind GEN. The Wilcoxon Signed-Rank Test, with $\alpha = 0.05$, revealed statistically significant superior *auc* scores for FSS over PSO on ESL (0.9839 vs. 0.9830, p-value = 0.029) and highlighted PSO's advantage in *auc* for BCC (0.7432 vs. 0.7393, p-value = 0.009) and LEV (0.8879 vs. 0.8853, p-value = 0.009). When considering $p_r \in \{0.2, 0.5\}$, the statistically significant advantage of PSO and FSS was confirmed for eight out of 20 other combinations (d, p_r, *acc*/*auc*); the inverse was true only for (LEV, 0.5, *auc*), hence slightly favoring PSO. The detailed results are presented in the supplementary material.

Examining the performance of other approaches reveals that their $r^{0.8}_{auc}$ and $r^{0.8}_{acc}$ ranks are significantly lower. The algorithms employing Mathematical Programming have excelled in specific quality metrics and datasets, such as MNR outperforming others in *auc* for CPU, and MMR leading in *acc* for both BCC and CPU. Despite these achievements, Mathematical Programming and Local Search strategies generally occupy the lowest average rank positions. Specifically, MMR and MNR show the weakest performance in *auc*, while GLS, SLS, and SAN fall behind in *acc*.

Within the Local Search algorithms, SLS emerged as the top performer based on $r^{0.8}_{auc}$. Delving into dataset-specific analyses, the Wilcoxon Test, with $\alpha = 0.05$, reveals that SLS's superiority is statistically validated only against SAN for CPU, achieving 0.9917 compared to SAN's 0.9833 (p-value = $6.82 \cdot 10^{-13}$). Conversely, for LEV, SAN significantly outperforms SLS, with scores of 0.8895 against 0.8860 (p-value = $1.98 \cdot 10^{-5}$). Regarding *acc*, GLS appears to lead in this group. However, its superiority over SAN is statistically significant only for CPU (0.9483 vs. 0.9279, p-value = $7.56 \cdot 10^{-7}$), and over SLS solely for LEV (0.8269 vs. 0.8218, p-value = 0.006). Notably, for LEV, SAN surpasses GLS (0.8301 vs. 0.8269, p-value = 0.033). Therefore, determining the most effective method among these three is challenging, with the choice heavily influenced by the specific dataset.

The analogous results attained by several state-of-the-art preference learning methods are reported in [23]. Our best-performing methods prove better than many variants of linear regression, rule-based algorithms, or outranking- and value-based methods incorporating mathematical programming. They are also competitive to Choquistic regression [29], outperforming it on many datasets in

Table 3: Average and standard deviation of *auc* for 80% train data and 20% test data for five datasets.

Algorithm	BCC		CPU		ESL		ERA		LEV		Avg. rank ($r^{0.8}_{auc}$)
MMR	**0.7430 ± 0.0669**	(3)	**0.9911 ± 0.0085**	(4)	0.9737 ± 0.0145	(8)	0.7258 ± 0.0370	(8)	0.8702 ± 0.0333	(8)	6.2
MNR	0.7301 ± 0.0705	(8)	**0.9927 ± 0.0071**	(1)	0.9750 ± 0.0125	(7)	0.7515 ± 0.0349	(7)	0.8802 ± 0.0277	(7)	6.0
GLS	0.7366 ± 0.0657	(5)	**0.9907 ± 0.0102**	(5)	0.9802 ± 0.0110	(6)	0.7562 ± 0.0320	(4)	0.8852 ± 0.0231	(6)	5.2
SLS	0.7366 ± 0.0662	(6)	**0.9917 ± 0.0091**	(3)	0.9806 ± 0.0095	(5)	0.7540 ± 0.0335	(5)	0.8860 ± 0.0226	(4)	4.6
SAN	0.7357 ± 0.0682	(7)	0.9833 ± 0.0147	(7)	0.9811 ± 0.0097	(4)	0.7532 ± 0.0318	(6)	**0.8895 ± 0.0237**	(2)	5.2
GEN	**0.7440 ± 0.0663**	(1)	**0.9924 ± 0.0076**	(2)	**0.9842 ± 0.0086**	(1)	**0.7634 ± 0.0300**	(1)	**0.8901 ± 0.0227**	(1)	1.2
FSS	0.7393 ± 0.0648	(4)	0.9840 ± 0.0129	(6)	**0.9839 ± 0.0082**	(2)	0.7591 ± 0.0286	(2)	0.8853 ± 0.0226	(5)	3.8
PSO	**0.7432 ± 0.0658**	(2)	0.9826 ± 0.0136	(8)	0.9830 ± 0.0093	(3)	0.7583 ± 0.0322	(3)	0.8879 ± 0.0229	(3)	3.8

Table 4: Average and standard deviation of *acc* for 80% train data and 20% test data for five datasets.

Algorithm	BCC		CPU		ESL		ERA		LEV		Avg. rank ($r^{0.8}_{acc}$)
MMR	**0.7411 ± 0.0572**	(1)	**0.9660 ± 0.0238**	(1)	0.9201 ± 0.0265	(7)	0.6846 ± 0.0388	(8)	0.8065 ± 0.0320	(8)	5.0
MNR	**0.7355 ± 0.0612**	(3)	0.9507 ± 0.0296	(2)	0.9169 ± 0.0262	(8)	0.7085 ± 0.0313	(5)	0.8221 ± 0.0277	(5)	4.6
GLS	0.7316 ± 0.0566	(6)	0.9483 ± 0.0332	(5)	0.9241 ± 0.0250	(4)	0.7053 ± 0.0288	(7)	0.8269 ± 0.0232	(3)	5.0
SLS	0.7304 ± 0.0522	(7)	0.9495 ± 0.0318	(3)	0.9213 ± 0.0253	(6)	0.7087 ± 0.0310	(4)	0.8218 ± 0.0249	(6)	5.2
SAN	0.7257 ± 0.0571	(8)	0.9279 ± 0.0395	(6)	0.9223 ± 0.0244	(5)	0.7074 ± 0.0292	(6)	**0.8301 ± 0.0236**	(2)	5.4
GEN	0.7355 ± 0.0540	(3)	0.9488 ± 0.0315	(4)	**0.9271 ± 0.0246**	(2)	**0.7162 ± 0.0296**	(3)	**0.8304 ± 0.0242**	(1)	2.6
FSS	**0.7370 ± 0.0569**	(2)	0.9257 ± 0.0362	(8)	**0.9271 ± 0.0242**	(2)	**0.7176 ± 0.0253**	(1)	0.8213 ± 0.0242	(7)	4.0
PSO	0.7336 ± 0.0589	(5)	0.9260 ± 0.0358	(7)	**0.9286 ± 0.0253**	(1)	**0.7166 ± 0.0293**	(2)	0.8254 ± 0.0252	(4)	3.8

auc. These observations confirm the high potential of the delineated research direction. However, the most advanced deep preference learning methods using complex preference models attain slightly better outcomes on all benchmark problems [23].

For the other reference set proportions (p_R) (see supplementary material for tables with the detailed results), the average *auc* rank values for $p_R = 0.5$ resemble those in Table 3, albeit with certain variations. Notably, SLS shows the most significant improvement, dropping from $r^{0.8}_{auc} = 4.6$ to $r^{0.5}_{auc} = 4.0$. In turn, FSS experiences the largest decline in its average rank (from 3.8 to 4.2). For $p_R = 0.2$, SLS sees a marked decrease to $r^{0.2}_{auc} = 5.2$. Also, PSO emerges as the highest-ranked algorithm, surpassing GEN with 1.8 versus 2.4 and achieving the best average *auc* across the BCC, ESL, and ERA datasets. In terms of *acc*, GEN retains the lowest average rank ($r^{0.5}_{acc} = 2.6$), but falls behind PSO at $r^{0.2}_{acc}$ with 3.2 versus 1.8. FSS consistently ranks third across different p_R values ($r^{0.5}_{acc} = 4.2$ and $r^{0.2}_{acc} = 3.6$). Identifying clear patterns for the other methods proves challenging, aside from SAN consistently recording the lowest average rank among all approaches. The supplementary material discusses the detailed outcomes of statistical tests.

6 CONCLUSIONS

We introduced the Choquet integral as a preference model for addressing sorting problems. It is adept at capturing interactions among the monotonic criteria. Recognizing the limitations of conventional optimization approaches, especially their inefficiency with a vast array of capacities and model constraints, we explored alternative methods for parameter estimation. We proposed eight optimization algorithms, encompassing two mathematical programming-based, three local search, and three nature-inspired techniques that emulate evolutionary processes or the collective behaviors observed in various animal species. Additionally, we detailed two post-optimization strategies to enhance the quality of the optimal solution yielded by each algorithm. An illustrative example was provided to elucidate the algorithms' functionality and the interpretation of model parameters.

We conducted comprehensive computational experiments on five benchmark datasets, varying in attributes and alternatives, to assess the model quality in terms of classification accuracy and preference reconstruction between alternative pairs. The experimental analysis was conducted in two phases: (a) identifying the optimal hyperparameters for each algorithm and problem and (b) evaluating the performance of models generated by different methods. The Genetic Algorithm emerged as the most effective, closely followed by Fish School Search and Particle Swarm Optimization.

Future research directions include refining the current algorithms with alternative strategies for managing constraint space violations, incorporating ensemble meta-algorithms like bagging or boosting to enhance performance, and expanding the algorithmic suite with additional nature-inspired metaheuristics suitable for constrained continuous optimization. Moreover, extending the experimental analysis to include more benchmarks, artificially generated datasets, a more exhaustive exploration of hyperparameter spaces, and evaluating algorithm performance under varied stopping conditions would be beneficial.

ACKNOWLEDGMENTS

Michał Wójcik was supported by the Polish National Science Center (DEC-2019/34/E/HS4/00045). Miłosz Kadziński was supported by the Polish Ministry of Science and Higher Education (0311/SBAD).

REFERENCES

[1] P. A. Alvarez, A. Ishizaka, and L. Martínez. 2021. Multiple-criteria decision-making sorting methods: A Survey. *Expert Systems with Applications* 183 (2021), 115368.

[2] S. Angilella, S. Greco, and B. Matarazzo. 2010. Non-additive robust ordinal regression: A multiple criteria decision model based on the Choquet integral. *European Journal of Operational Research* 201, 1 (2010), 277–288.

[3] S. Arcidiacono, S. Corrente, and S. Greco. 2021. Robust stochastic sorting with interacting criteria hierarchically structured. *European Journal of Operational Research* 292, 2 (2021), 735–754.

[4] C. Bastos Filho, F. B. de Lima Neto, A. Lins, A. Nascimento, and M. Lima. 2008. A novel search algorithm based on fish school behavior. In *2008 IEEE International Conference on Systems, Man and Cybernetics*. IEEE.

[5] J. Branke, S. Corrente, S. Greco, R. Słowiński, and P. Zielniewicz. 2016. Using Choquet integral as preference model in interactive evolutionary multiobjective optimization. *European Journal of Operational Research* 250, 3 (2016), 884–901.

[6] L. Breiman. 1996. Bagging predictors. *Machine Learning* 24, 2 (1996), 123–140.

[7] G. Choquet. 1954. Theory of capacities. *Annales de l'institut Fourier* 5 (1954), 131–295.

[8] K. Ciomek and M. Kadziński. 2021. Polyrun: A Java library for sampling from the bounded convex polytopes. *SoftwareX* 13 (2021), 100659.

[9] H.E. de Oliveira, L. Tomazeli Duarte, and J.M. Travassos Romano. 2022. Identification of the Choquet integral parameters in the interaction index domain by means of sparse modeling. *Expert Systems with Applications* 187 (2022), 115874.

[10] J. Fürnkranz and E. Hüllermeier. 2011. *Preference Learning*. Springer.

[11] M. Grabisch. 1996. The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research* 89, 3 (1996), 445–456.

[12] S. Greco, V. Mousseau, and R. Słowiński. 2010. Multiple criteria sorting with a set of additive value functions. *European Journal of Operational Research* 207, 3 (2010), 1455–1470.

[13] R. Hecht-Nielsen. 1989. Theory of the backpropagation neural network. In *International Joint Conference on Neural Networks*. IEEE.

[14] M. Herin, P. Perny, and N. Sokolovska. 2024. Learning preference representations based on Choquet integrals for multicriteria decision making. *Annals of Mathematics and Artificial Intelligence* (2024). https://doi.org/10.1007/s10472-024-09930-0

[15] E. Hüllermeier and R. Słowiński. 2024. Preference learning and multiple criteria decision aiding: differences, commonalities, and synergies—part II. *4OR* (2024). https://doi.org/10.1007/s10288-023-00561-5

[16] E. Hüllermeier and R. Słowiński. 2024. Preference learning and multiple criteria decision aiding: differences, commonalities, and synergies–part I. *4OR* (2024). https://doi.org/10.1007/s10288-023-00560-6

[17] J. Kennedy and R. Eberhart. 1995. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks (ICNN-95)*. IEEE. https://doi.org/10.1109/icnn.1995.488968

[18] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. 1983. Optimization by Simulated Annealing. *Science* 220, 4598 (1983), 671–680.

[19] A. Lambora, K. Gupta, and K. Chopra. 2019. Genetic Algorithm – A Literature Review. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. IEEE.

[20] Z. Liao, H. Liao, and X. Zhang. 2023. A contextual Choquet integral-based preference learning model considering both criteria interactions and the compromise effects of decision-makers. *Expert Systems with Applications* 213 (2023), 118977.

[21] A. Lipowski and D. Lipowska. 2012. Roulette-wheel selection via stochastic acceptance. *Physica A: Statistical Mechanics and its Applications* 391, 6 (2012), 2193–2196.

[22] J. Liu, M. Kadziński, X. Liao, and X. Mao. 2021. Data-Driven Preference Learning Methods for Value-Driven Multiple Criteria Sorting with Interacting Criteria. *INFORMS Journal on Computing* 33 (2021), 586–606. Issue 2.

[23] K. Martyn and M. Kadziński. 2023. Deep preference learning for multiple criteria decision analysis. *European Journal of Operational Research* 305, 2 (2023), 781–805.

[24] B. Miller and D. Goldberg. 1996. Genetic Algorithms, Selection Schemes, and the Varying Effects of Noise. *Evolutionary Computation* 4, 2 (1996), 113–131.

[25] R. Pelissari, A. Abackerli, and L. Tomazeli Duarte. 2022. Choquet capacity identification for multiple criteria sorting problems: A novel proposal based on Stochastic Acceptability Multicriteria Analysis. *Applied Soft Computing* 120 (2022), 108727.

[26] R. Smith. 1984. Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed Over Bounded Regions. *Operations Research* 32, 6 (1984), 1296–1308.

[27] O. Sobrie, V. Mousseau, and M. Pirlot. 2018. Learning monotone preferences using a majority rule sorting model. *International Transactions in Operational Research* 26, 5 (2018), 1786–1809.

[28] É. D. Taillard. 2023. *Design of Heuristic Algorithms for Hard Optimization: With Python Codes for the Travelling Salesman Problem*. Springer International Publishing.

[29] A. Tehrani, W. Cheng, K. Dembczyński, and E. Hüllermeier. 2012. Learning monotone nonlinear models using the Choquet integral. *Machine Learning* 89, 1 (2012), 183–211.

[30] A.F. Tehrani, W. Cheng, and E. Hüllermeier. 2012. Preference Learning Using the Choquet Integral: The Case of Multipartite Ranking. *IEEE Transactions on Fuzzy Systems* 20, 6 (2012), 1102–1113.

[31] H. Q. Vu, G. Beliakov, and G. Li. 2014. A Choquet Integral Toolbox and Its Application in Customer Preference Analysis. In *Data Mining Applications with R*, Yanchang Zhao and Yonghua Cen (Eds.). Academic Press, Boston, 247–272.

Supplementary material [P3]

Nature-inspired Preference Learning Algorithms Using the Choquet Integral – Appendix

Michał Wójcik*
michal.wojcik@cs.put.poznan.pl
Faculty of Computing and Telecommunication,
Poznań University of Technology
Poznań, Poland

Miłosz Kadziński
milosz.kadzinski@cs.put.poznan.pl
Faculty of Computing and Telecommunication,
Poznań University of Technology
Poznań, Poland

ACM Reference Format:
Michał Wójcik and Miłosz Kadziński. 2024. Nature-inspired Preference Learning Algorithms Using the Choquet Integral – Appendix. In *Genetic and Evolutionary Computation Conference (GECCO '24), July 14–18, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3638529.3654054

1 ILLUSTRATIVE STUDY

This section first illustrates how to calculate the value of the Choquet integral for an example alternative using a model with established capacities. Furthermore, it showcases examples of the optimization processes undertaken by individual algorithms, highlighting the progression of learning through visualizations. It delves into the interpretation of model parameters and includes a discussion on how the model classifies a non-reference alternative into a specific class.

1.1 Determining the Choquet integral value

To illustrate the use of the Choquet integral, we consider a three-criteria problem. Figure 1 shows the *capacities* of the Choquet integral for each subset of criteria. They indicate:

- a synergy (positive interaction) for a pair (g_1, g_2) since $\mu_M(\{g_1, g_2\}) > \mu_M(\{g_1\}) + \mu_M(\{g_2\})$;
- a redundancy (negative interaction) for a pair (g_2, g_3) since $\mu_M(\{g_2, g_3\}) < \mu_M(\{g_2\}) + \mu_M(\{g_3\})$;
- no interaction for a pair (g_1, g_3) since $\mu_M(\{g_1, g_3\}) = \mu_M(\{g_1\}) + \mu_M(\{g_3\})$.

Let us consider the following performances of alternative a: $g_1(a) = 0.5$, $g_2(a) = 0.9$, and $g_3(a) = 0.3$. The Choquet integral value is computed according to the following equation:

$$Ch_M(a) = \sum_{j=1}^{m} [g_{(j)}(a) - g_{(j-1)}(a)] \cdot \mu_M(G_{(j)}).$$

When considering the performances on various subsets of criteria, they are as strong as their weakest (least) element. This fact is reflected in Figure 2. It shows that a attains performance of 0.3 for all three criteria ($\{g_1, g_2, g_3\}$), 0.5 for a pair $\{g_1, g_2\}$, and 0.9 for $\{g_2\}$. Hence, we order the criteria indices in line with the non-decreasing performances $g_j(a)$, $j = 1, 2, 3$. Since $g_3(a) = 0.3 \leq g_1(a) = 0.5 \leq g_2(a) = 0.9$, we get $(\cdot) = (3, 1, 2)$. Moreover, according to the model's assumptions, we assume $g_{(0)}(a) = 0$. Then, the Choquet integral for alternative a can be computed as follows:

$$\begin{aligned} Ch_M(a) = {} & [g_3(a) - g_{(0)}(a)] \cdot \mu_M(\{g_1, g_2, g_3\}) \\ & + [g_1(a) - g_3(a)] \cdot \mu_M(\{g_1, g_2\}) \\ & + [g_2(a) - g_1(a)] \cdot \mu_M(\{g_2\}), \end{aligned}$$

which is equivalent to:

$$\begin{aligned} Ch_M(a) &= (0.3 - 0) \cdot 1 + (0.5 - 0.3) \cdot 0.8 + (0.9 - 0.5) \cdot 0.35 \\ &= 0.3 + 0.16 + 0.14 = 0.6. \end{aligned}$$

Each subsequent component of the above sum is calculated based on the minimum performance of the alternative for a given subset of criteria and the capacity associated with this subset.

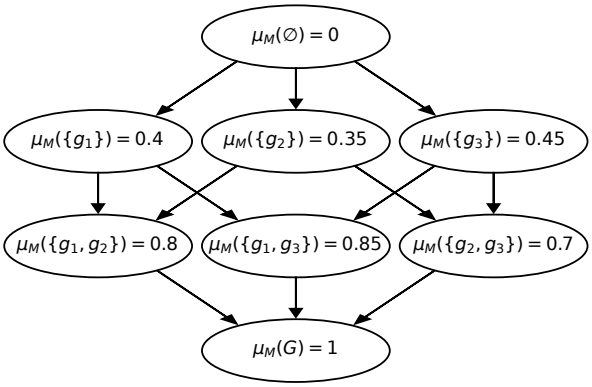


Figure 1: *Capacities* of the Choquet integral used in the example.

1.2 Learning progress for optimization algorithms

This section demonstrates the optimization process through the outcomes of three distinct approaches (GEN, FSS, PSO) applied to a two-class sorting problem derived from the Employee Selection [ESL] dataset. Preference information was provided as expected class assignments for all 488 alternatives.

The objective of executing these algorithms was to assess the quality of both intermediate and final models produced. A uniform stopping condition of 30 seconds was set for all algorithms. The

*Corresponding author.

GECCO '24, July 14–18, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0494-9/24/07…$15.00
https://doi.org/10.1145/3638529.3654054

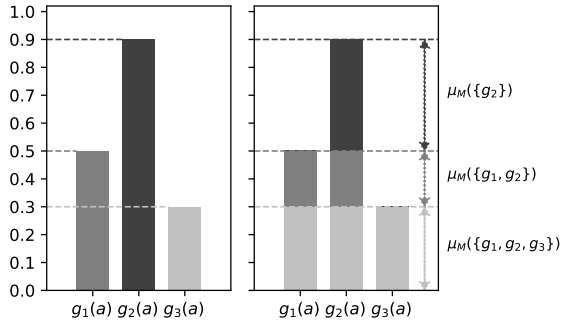


Figure 2: Performances of alternative a on the three criteria and associated model *capacities* used in the example.

primary goal for each algorithm was to minimize the loss function value, $L(m^*)$, of the best solution m^* identified up to that point. The progression of loss function improvements over processing time is depicted in Figure 3, with the initial assumption that $L(m^*)$ equals infinity from $t = 0$ seconds until the discovery of the first m^*.

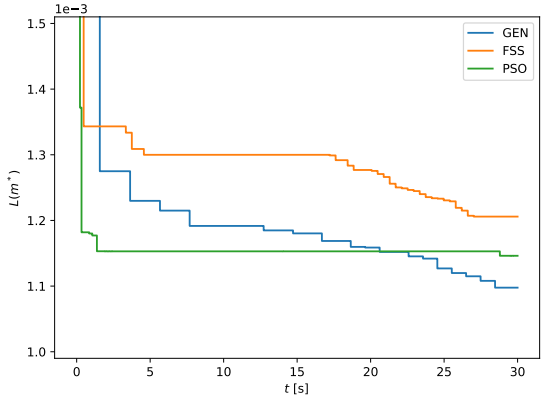


Figure 3: The change of loss function $L(m^*)$ value over time for three preference learning algorithms.

The GEN algorithm commenced its optimization by generating the initial m^* model at 1.57 seconds into the execution. This delay is attributed to the evaluation of a relatively large initial population size ($S = 250$) required to identify the first optimal individual. Notably, the loss function value, $L(m^*)$, for GEN exhibited a consistent decline throughout the optimization period. It achieved the second-best outcome at the 10-second mark ($L(m^*) = 1.192 \cdot 10^{-3}$) and the best result at the conclusion of the 30-second execution window ($L(m^*) = 1.098 \cdot 10^{-3}$). The intervals marking successive enhancements in $L(m^*)$ appeared to be quite uniform, suggesting a correlation with the computational effort allocated for generating successive generations.

The FSS algorithm produced its initial solution at 0.48 seconds, followed by a series of significant advancements that led to a loss function value of $L(m^*) = 1.3 \cdot 10^{-3}$ shortly after 4.5 seconds. However, it then experienced a period exceeding 12 seconds during which it could not enhance its optimal model m^*. After this hiatus, the algorithm gradually reduced the loss function value, culminating in $L(m^*) = 1.206 \cdot 10^{-3}$ by the end of the execution period. Despite these incremental improvements, FSS achieved the least favorable outcomes compared to the other algorithms.

PSO distinguished itself by identifying an initial optimal model m^* merely 0.23 seconds into the process, with a loss function value of $1.372 \cdot 10^{-3}$. The most significant reductions in $L(m^*)$ were achieved within the initial 1.38 seconds, leading to a solution characterized by $L(m^*) = 1.153 \cdot 10^{-3}$. Subsequently, the model's progress plateaued, registering only marginal improvements over the ensuing 27.4 seconds, which diminished the loss by a mere $1.31 \cdot 10^{-7}$. This phase likely involved some particles refining the vicinity of the previously identified optimal solution, while others continued to explore the broader solution space. This exploration phase yielded positive results at 28.79 seconds when the algorithm unveiled its optimal solution, marked by $L(m^*) = 1.146 \cdot 10^{-3}$. Following this discovery, the model experienced minimal enhancements for over a second before the algorithm concluded.

1.3 Model parameters

Table 1 presents $L(m^*)$ indicating the performance of the model, alongside fourteen model parameters $\mu(G')$, $G' \subset G \neq \emptyset$, and a single threshold t_1 for each method, presented in three variations:

(a) first ever obtained m^*;
(b) m^* attained after 10 seconds of algorithm execution;
(c) m^* obtained upon completion of the algorithm execution.

The remaining model parameters resulting from the constraints are common to all variants, i.e., $\mu(\emptyset) = 0$, $\mu(G) = 1$, $t_0 = 0$, and $t_2 = 1 + \varepsilon$.

The variation in parameter magnitudes is more pronounced between models (a) and (b). Specifically, in models derived using the GEN algorithm, the average deviation among $\mu(G'), G' \subset G \neq \emptyset$ is approximately 0.053, whereas the average discrepancy between parameters in models (b) and (c) is around 0.039. The disparity is even more significant in models generated by PSO, with differences amounting to 0.059 and 0.007, respectively. This trend aligns with the intuition that the most substantial changes in $L(m^*)$ occur in the initial moments of the algorithm's execution, diminishing over time as the incremental improvements in subsequent models become less significant. As the optimization process advances, identifying superior models becomes more and more challenging, and the magnitude of enhancements diminishes. This characteristic is commonly observed in optimization strategies.

Notably, criterion g_4 emerges as the most important factor in the GEN models. This is evidenced by $\mu(\{g_4\})$ achieving the highest values across all single-criterion subsets in all three model variants. Conversely, among subsets comprising three criteria, $\mu(\{g_1, g_2, g_3\})$ registers the lowest value, suggesting that the $Ch(a)$ is most negatively affected when $g_4(a)$ is low. Similarly, for subsets containing two criteria, the combinations $\mu(\{g_1, g_4\})$, $\mu(\{g_2, g_4\})$, and $\mu(\{g_3, g_4\})$ yield the highest values, with the sole exception being $\mu(\{g_3, g_4\}) = 0.3940$ in model (a). However, this value sees a significant increase in subsequent, better performing models (b) and (c), culminating in

Table 1: Model parameters m^*, t_1, and $L(m^*)$ for each algorithm and model variant.

Algorithm	Model	$L(m^*)$	$\mu(\{g_1\})$	$\mu(\{g_2\})$	$\mu(\{g_3\})$	$\mu(\{g_4\})$	$\mu(\{g_1,g_2\})$	$\mu(\{g_1,g_3\})$	$\mu(\{g_1,g_4\})$	$\mu(\{g_2,g_3\})$	$\mu(\{g_2,g_4\})$	$\mu(\{g_3,g_4\})$	$\mu(\{g_1,g_2,g_3\})$	$\mu(\{g_1,g_2,g_4\})$	$\mu(\{g_1,g_3,g_4\})$	$\mu(\{g_2,g_3,g_4\})$	t_1
GEN	(a)	$1.275 \cdot 10^{-3}$	0.0516	0.1907	0.1195	0.2375	0.2884	0.4586	0.5067	0.4978	0.5744	0.3940	0.5653	0.6248	0.7138	0.6397	0.5923
	(b)	$1.192 \cdot 10^{-3}$	0.1162	0.1760	0.1323	0.2982	0.3311	0.4219	0.5478	0.4001	0.5517	0.5022	0.6056	0.7098	0.7327	0.7330	0.6009
	(c)	$1.098 \cdot 10^{-3}$	0.1319	0.1758	0.1528	0.2706	0.3767	0.4864	0.6404	0.4391	0.5714	0.5636	0.6287	0.7312	0.7840	0.8012	0.6118
FSS	(a)	$1.343 \cdot 10^{-3}$	0.2716	0.0456	0.4007	0.4096	0.3242	0.5815	0.6614	0.4267	0.4959	0.5608	0.6640	0.7435	0.8000	0.7973	0.6163
	(b)	$1.300 \cdot 10^{-3}$	0.1796	0.1765	0.1491	0.2868	0.3350	0.3939	0.6055	0.4141	0.6096	0.5973	0.6388	0.8224	0.7542	0.8388	0.6120
	(c)	$1.206 \cdot 10^{-3}$	0.1419	0.1321	0.1810	0.1847	0.4663	0.5492	0.6258	0.5220	0.5330	0.5650	0.6106	0.8334	0.7706	0.7983	0.6157
PSO	(a)	$1.372 \cdot 10^{-3}$	0.2622	0.2748	0.3803	0.1404	0.4268	0.5106	0.7961	0.4768	0.5395	0.7716	0.5564	0.9282	0.9191	0.8094	0.6322
	(b)	$1.153 \cdot 10^{-3}$	0.2090	0.2154	0.1987	0.1404	0.4192	0.5442	0.7094	0.5539	0.5659	0.7327	0.6515	0.8550	0.8601	0.8414	0.6263
	(c)	$1.146 \cdot 10^{-3}$	0.2062	0.2312	0.2084	0.1453	0.4253	0.5551	0.7110	0.5625	0.5877	0.7331	0.6602	0.8646	0.8604	0.8443	0.6270

0.6287 in model (c). Similarly, g_1 is perceived as the least important criterion in this case.

1.3.1 Interactions between criteria. When using the Choquet integral, it is possible to determine both the nature (positive or negative) and the magnitude of interactions among subsets of criteria, thereby capturing complex relationships. Such nuanced interactions are beyond the scope of strictly additive models, such as the Additive Value Function (AVF), where the additive principle dictates that for any two disjoint subsets of criteria $G_1, G_2 \subseteq G$ (i.e., $G_1 \cap G_2 = \emptyset$), the value $\mu(G_1 \cup G_2)$ is always equal to the sum $\mu(G_1) + \mu(G_2)$. This principle does not hold in non-additive models like the one under consideration, enabling the evaluation of the difference $\mu(G_1 \cup G_2) - (\mu(G_1) + \mu(G_2))$. This difference provides insight into the strength and character of the interactions between the subsets.

For model (c) derived using GEN, the interaction for the subset $\{g_1, g_4\}$ is calculated as $\mu(\{g_1, g_4\}) - (\mu(\{g_1\}) + \mu(\{g_4\})) = 0.6404 - (0.1319 + 0.2706) = 0.2379$. This value represents the most significant interaction among the 2-criteria subsets within this model. Conversely, the minimal interaction value is observed for the subset comprising criteria g_1 and g_2, amounting to 0.0690. Despite being considerably lower, this value still signifies a positive interaction between these criteria.

In the case of the best model produced by FSS, all 2-criteria interactions exhibit positive values. This model attributes the highest and lowest interaction strengths to the same pairs of criteria as the GEN model (0.2993 and 0.1923, respectively). However, model (c) generated by PSO identifies pair $\{g_1, g_2\}$ as having the least interaction, with a value of $\mu(\{g_1, g_2\}) = -0.012$, suggesting a negative interaction between these features. Additionally, this model highlights a different pair, $\{g_3, g_4\}$, as having the maximum synergy, with an interaction strength of $\mu(\{g_3, g_4\}) = 0.3795$.

1.4 Assigning alternatives to classes

This section illustrates the computation of $Ch(b)$ for a given model M and the subsequent classification of an alternative into a decision class. We will elucidate this process using models (a) and (c) obtained with GEN. Consider an alternative b characterized by the following performances: $g_1(b) = 0.7$, $g_2(b) = 0.5$, $g_3(b) = 0.4$, and $g_4(b) = 0.8$.

This section illustrates the classification of an example alternative into a decision class. We will elucidate this process using models (a) and (c) obtained with GEN. Consider an alternative b characterized by the following performances: $g_1(b) = 0.7$, $g_2(b) = 0.5$, $g_3(b) = 0.4$, and $g_4(b) = 0.8$.

Similar to the example presented in Section 1.1, it is first necessary to determine the order of criteria indices aligning with the non-decreasing order of performances $g_j(b)$, $j = 1, 2, 3, 4$. For the considered alternative b, it is $(\cdot) = (3, 2, 1, 4)$, and $g_{(0)}(b) = 0$. For alternative b, the value of the Choquet integral is computed as follows:

$$\begin{aligned} Ch_M(b) = {} & [g_3(b) - g_{(0)}(b)] \cdot \mu_M(G) \\ & + [g_2(b) - g_3(b)] \cdot \mu_M(\{g_1, g_2, g_4\}) \\ & + [g_1(b) - g_2(b)] \cdot \mu_M(\{g_1, g_4\}) \\ & + [g_4(b) - g_1(b)] \cdot \mu_M(\{g_4\}), \end{aligned}$$

which is equivalent to:

$$\begin{aligned} Ch_M(b) = {} & 0.4 \cdot \mu_M(G) + 0.1 \cdot \mu_M(\{g_1, g_2, g_4\}) \\ & + 0.2 \cdot \mu_M(\{g_1, g_4\}) + 0.1 \cdot \mu_M(\{g_4\}). \end{aligned}$$

For the two models mentioned above, these values are $Ch_{GEN_{(a)}}(b) = 0.5876$ and $Ch_{GEN_{(c)}}(b) = 0.6283$, respectively. For model (a), this leads to assigning b to class C_1, as $t_1 = 0.5923 > Ch_{GEN_{(a)}}(b)$. In turn, for model (c), $t_1 = 0.6118 < Ch_{GEN_{(c)}}(b)$, and therefore, b is assigned to C_2.

2 PRELIMINARY ANALYSIS OF THE USEFULNESS OF BACKPROPAGATION

To verify the usefulness of the backpropagation postoptimization technique, we performed a 10-fold Monte Carlo Cross Validation for all five datasets and eight algorithms, assuming $p_R = 0.8$. The execution timeout was set to 10 seconds, and the subject of the comparison were the results obtained by the algorithms in four variants – without backpropagation (marked as 0%) and using this technique for 5%, 10%, and 20% of the assumed execution time. The average values of *auc* and *acc* obtained by all algorithms in the four scenarios are presented in Table 6.

For both measures, the best average results were achieved by the approach in which 5% of the execution time was allocated to backpropagation optimization of the best-found model. However, the advantage over the variant with 10% timeshare devoted to backpropagation is negligible. The average values indicate the superiority of these two settings over the variant without post-optimization. Furthermore, the Wilcoxon signed-rank test shows a statistically significant advantage for 5% setting over 0% counterpart in the context of both quality measures (the p-values are 0.015 for *auc* and 0.035 for *acc*). Additionally, for 20%, the *auc* values are, on average, worse than for the variants in which less time was spent on backpropagation and comparable to the results without backpropagation. This may indicate a tendency for the results to deteriorate with increasing time spent on post-optimization. For this reason,

Table 2: Average and standard deviation of *auc* for 20% train data and 80% test data for five datasets.

Algorithm	BCC		CPU		ESL		ERA		LEV		Avg. rank ($r^{0.2}_{auc}$)
MMR	0.7291 ± 0.0179	(7)	0.9794 ± 0.0157	(2)	0.9707 ± 0.0081	(8)	0.7376 ± 0.0171	(8)	0.8659 ± 0.0156	(8)	6.6
MNR	0.7054 ± 0.0309	(8)	**0.9828 ± 0.0197**	(1)	0.9727 ± 0.0065	(7)	0.7469 ± 0.0207	(7)	0.8674 ± 0.0197	(7)	6.0
GLS	0.7341 ± 0.0179	(4)	0.9692 ± 0.0193	(8)	0.9778 ± 0.0052	(4)	0.7486 ± 0.0156	(5)	0.8792 ± 0.0089	(4)	5.0
SLS	0.7324 ± 0.0185	(6)	0.9713 ± 0.0169	(7)	0.9760 ± 0.0058	(5)	**0.7523 ± 0.0131**	(3)	0.8791 ± 0.0098	(5)	5.2
SAN	0.7329 ± 0.0201	(5)	0.9713 ± 0.0165	(6)	0.9755 ± 0.0056	(6)	0.7508 ± 0.0137	(4)	0.8794 ± 0.0097	(3)	4.8
GEN	0.7342 ± 0.0179	(3)	0.9736 ± 0.0143	(3)	0.9781 ± 0.0049	(3)	**0.7542 ± 0.0133**	(2)	**0.8837 ± 0.0078**	(1)	2.4
FSS	0.7346 ± 0.0194	(2)	0.9735 ± 0.0150	(5)	**0.9799 ± 0.0052**	(2)	0.7473 ± 0.0172	(6)	0.8751 ± 0.0106	(6)	4.2
PSO	**0.7378 ± 0.0189**	(1)	0.9735 ± 0.0146	(4)	**0.9805 ± 0.0051**	(1)	**0.7546 ± 0.0133**	(1)	**0.8830 ± 0.0087**	(2)	1.8

Table 3: Average and standard deviation of *acc* for 20% train data and 80% test data for five datasets.

Algorithm	BCC		CPU		ESL		ERA		LEV		Avg. rank ($r^{0.2}_{acc}$)
MMR	**0.7232 ± 0.0267**	(5)	**0.9352 ± 0.0337**	(1)	0.9167 ± 0.0121	(7)	0.6885 ± 0.0204	(8)	0.8001 ± 0.0200	(8)	5.8
MNR	0.7066 ± 0.0396	(8)	**0.9284 ± 0.0426**	(2)	0.9114 ± 0.0150	(8)	**0.7066 ± 0.0198**	(2)	0.8095 ± 0.0201	(7)	5.4
GLS	**0.7250 ± 0.0220**	(3)	0.8962 ± 0.0367	(8)	0.9194 ± 0.0119	(4)	0.7011 ± 0.0164	(6)	0.8150 ± 0.0148	(5)	5.2
SLS	0.7212 ± 0.0234	(7)	0.9026 ± 0.0350	(6)	0.9185 ± 0.0115	(5)	0.7052 ± 0.0158	(4)	0.8174 ± 0.0137	(3)	5.0
SAN	0.7216 ± 0.0226	(6)	0.9008 ± 0.0327	(7)	0.9179 ± 0.0121	(6)	0.6994 ± 0.0158	(7)	0.8152 ± 0.0139	(4)	6.0
GEN	**0.7234 ± 0.0249**	(4)	0.9036 ± 0.0326	(5)	0.9197 ± 0.0112	(3)	0.7060 ± 0.0151	(3)	**0.8197 ± 0.0120**	(1)	3.2
FSS	**0.7265 ± 0.0277**	(2)	0.9053 ± 0.0305	(4)	**0.9211 ± 0.0118**	(1)	0.7029 ± 0.0177	(5)	0.8124 ± 0.0139	(6)	3.6
PSO	**0.7271 ± 0.0223**	(1)	0.9054 ± 0.0334	(3)	**0.9209 ± 0.0126**	(2)	**0.7087 ± 0.0165**	(1)	**0.8179 ± 0.0137**	(2)	1.8

Table 4: Average and standard deviation of *auc* for 50% train data and 50% test data for five datasets.

Algorithm	BCC		CPU		ESL		ERA		LEV		Avg. rank ($r^{0.5}_{auc}$)
MMR	0.7334 ± 0.0307	(6)	0.9885 ± 0.0112	(3)	0.9731 ± 0.0081	(8)	0.7337 ± 0.0246	(8)	0.8683 ± 0.0166	(8)	6.6
MNR	0.7222 ± 0.0343	(8)	**0.9916 ± 0.0065**	(1)	0.9746 ± 0.0066	(7)	0.7500 ± 0.0257	(7)	0.8736 ± 0.0244	(7)	6.0
GLS	0.7349 ± 0.0334	(4)	0.9859 ± 0.0116	(6)	0.9788 ± 0.0065	(5)	0.7551 ± 0.0174	(6)	0.8856 ± 0.0119	(4)	5.0
SLS	0.7340 ± 0.0293	(5)	0.9875 ± 0.0096	(4)	0.9789 ± 0.0061	(4)	0.7559 ± 0.0165	(4)	0.8861 ± 0.0120	(3)	4.0
SAN	0.7320 ± 0.0315	(7)	0.9855 ± 0.0098	(7)	0.9784 ± 0.0059	(6)	0.7553 ± 0.0173	(5)	0.8863 ± 0.0115	(2)	5.4
GEN	**0.7388 ± 0.0296**	(1)	**0.9893 ± 0.0101**	(2)	0.9815 ± 0.0052	(2)	**0.7629 ± 0.0155**	(1)	**0.8899 ± 0.0106**	(1)	1.4
FSS	0.7355 ± 0.0315	(3)	0.9834 ± 0.0088	(8)	**0.9819 ± 0.0052**	(1)	0.7598 ± 0.0147	(3)	0.8843 ± 0.0115	(6)	4.2
PSO	**0.7372 ± 0.0302**	(2)	0.9860 ± 0.0085	(5)	0.9815 ± 0.0052	(3)	0.7611 ± 0.0142	(2)	0.8850 ± 0.0115	(5)	3.4

Table 5: Average and standard deviation of *acc* for 50% train data and 50% test data for five datasets.

Algorithm	BCC		CPU		ESL		ERA		LEV		Avg. rank ($r^{0.5}_{acc}$)
MMR	**0.7307 ± 0.0262**	(1)	**0.9585 ± 0.0226**	(1)	0.9198 ± 0.0137	(5)	0.6885 ± 0.0280	(8)	0.8021 ± 0.0211	(8)	4.6
MNR	0.7157 ± 0.0336	(8)	0.9514 ± 0.0237	(2)	0.9158 ± 0.0156	(8)	0.7105 ± 0.0222	(4)	0.8167 ± 0.0198	(7)	5.8
GLS	0.7247 ± 0.0331	(5)	0.9305 ± 0.0273	(5)	0.9192 ± 0.0161	(6)	0.7063 ± 0.0170	(7)	0.8206 ± 0.0165	(4)	5.4
SLS	0.7246 ± 0.0270	(6)	0.9366 ± 0.0305	(4)	0.9200 ± 0.0138	(4)	0.7089 ± 0.0172	(5)	0.8247 ± 0.0133	(2)	4.2
SAN	0.7246 ± 0.0291	(6)	0.9298 ± 0.0287	(7)	0.9185 ± 0.0137	(7)	0.7081 ± 0.0164	(6)	0.8230 ± 0.0132	(3)	5.8
GEN	**0.7278 ± 0.0287**	(3)	0.9411 ± 0.0281	(3)	0.9228 ± 0.0139	(3)	**0.7158 ± 0.0165**	(3)	**0.8276 ± 0.0125**	(1)	2.6
FSS	**0.7271 ± 0.0271**	(4)	0.9214 ± 0.0243	(8)	0.9241 ± 0.0135	(2)	**0.7163 ± 0.0168**	(2)	0.8206 ± 0.0144	(5)	4.2
PSO	**0.7297 ± 0.0272**	(2)	0.9301 ± 0.0239	(6)	**0.9261 ± 0.0141**	(1)	**0.7172 ± 0.0169**	(1)	0.8204 ± 0.0145	(6)	3.2

Table 6: Comparison of the average values of *auc* and *acc* for approaches without and with backpropagation post-optimization.

Measure	0%	5%	10%	20%
auc	0.8647	0.8667	0.8664	0.8646
acc	0.8095	0.8173	0.8172	0.8168

we decided to include this technique in the experimental analysis, devoting 5% of algorithm execution time to this approach.

3 COMPARISON OF ALGORITHMS' PERFORMANCE

This section displays the average *auc* and *acc* values achieved by all algorithms across various problems, with reference set proportions $p_R = 0.2$ (see Tables 2 and 3) and $p_R = 0.5$ (see Tables 4 and 5). Similar to the main paper, we denote in bold the algorithms for which the differences to the best performer were not statistically significant.

For all considered proportions p_r, the algorithms recorded the highest *auc* scores for the CPU and ESL datasets. Conversely, the

Table 7: The comparison of differences in *auc* for all pairs of algorithms using Wilcoxon signed rank test (p-value) for 20% train data and 80% test data for five datasets.

Alg. 1	Alg. 2	BCC	CPU	ESL	ERA	LEV
GEN	PSO	$\prec (0.002)$	?	$\prec (< 10^{-7})$	?	?
	FSS	?	?	$\prec (< 10^{-4})$	$\succ (< 10^{-4})$	$\succ (< 10^{-10})$
	SLS	?	?	$\succ (< 10^{-6})$	?	$\succ (< 10^{-10})$
	GLS	?	$\succ (0.003)$	?	$\succ (< 10^{-5})$	$\succ (< 10^{-11})$
	SAN	?	?	$\succ (< 10^{-7})$	$\succ (< 10^{-4})$	$\succ (< 10^{-7})$
	MNR	$\succ (< 10^{-14})$	$\prec (< 10^{-5})$	$\succ (< 10^{-11})$	$\succ (0.002)$	$\succ (< 10^{-11})$
	MMR	$\succ (< 10^{-3})$	$\prec (0.001)$	$\succ (< 10^{-15})$	$\succ (< 10^{-13})$	$\succ (< 10^{-17})$
PSO	FSS	$\succ (0.002)$	?	?	$\succ (< 10^{-5})$	$\succ (< 10^{-8})$
	SLS	$\succ (< 10^{-4})$	?	$\succ (< 10^{-12})$	?	$\succ (< 10^{-6})$
	GLS	$\succ (< 10^{-3})$	$\succ (0.002)$	$\succ (< 10^{-8})$	$\succ (< 10^{-5})$	$\succ (< 10^{-6})$
	SAN	$\succ (< 10^{-4})$	?	$\succ (< 10^{-11})$	$\succ (< 10^{-3})$	$\succ (< 10^{-5})$
	MNR	$\succ (< 10^{-15})$	$\prec (< 10^{-6})$	$\succ (< 10^{-15})$	$\succ (< 10^{-3})$	$\succ (< 10^{-10})$
	MMR	$\succ (< 10^{-9})$	$\prec (0.002)$	$\succ (< 10^{-17})$	$\succ (< 10^{-13})$	$\succ (< 10^{-16})$
FSS	SLS	?	?	$\succ (< 10^{-10})$	$\prec (0.006)$	$\prec (< 10^{-3})$
	GLS	?	$\succ (0.009)$	$\succ (< 10^{-5})$	?	$\prec (< 10^{-3})$
	SAN	$\succ (0.025)$	?	$\succ (< 10^{-10})$	?	$\prec (< 10^{-3})$
	MNR	$\succ (< 10^{-14})$	$\prec (< 10^{-5})$	$\succ (< 10^{-15})$	?	$\succ (0.001)$
	MMR	$\succ (< 10^{-3})$	$\prec (0.001)$	$\succ (< 10^{-16})$	$\succ (< 10^{-5})$	$\succ (< 10^{-7})$
SLS	GLS	?	?	$\prec (< 10^{-3})$	$\succ (0.002)$	?
	SAN	?	?	?	?	?
	MNR	$\succ (< 10^{-12})$	$\prec (< 10^{-6})$	$\succ (< 10^{-5})$	$\succ (0.021)$	$\succ (< 10^{-7})$
	MMR	$\succ (0.007)$	$\prec (< 10^{-3})$	$\succ (< 10^{-11})$	$\succ (< 10^{-11})$	$\succ (< 10^{-13})$
GLS	SAN	?	?	$\succ (< 10^{-5})$	$\prec (0.043)$	?
	MNR	$\succ (< 10^{-14})$	$\prec (< 10^{-7})$	$\succ (< 10^{-11})$	?	$\succ (< 10^{-7})$
	MMR	$\succ (< 10^{-3})$	$\prec (< 10^{-4})$	$\succ (< 10^{-13})$	$\succ (< 10^{-8})$	$\succ (< 10^{-14})$
SAN	MNR	$\succ (< 10^{-13})$	$\prec (< 10^{-6})$	$\succ (< 10^{-4})$	?	$\succ (< 10^{-7})$
	MMR	$\succ (0.002)$	$\prec (< 10^{-4})$	$\succ (< 10^{-9})$	$\succ (< 10^{-10})$	$\succ (< 10^{-14})$
MNR	MMR	$\prec (< 10^{-11})$	$\succ (0.005)$	$\succ (0.006)$	$\succ (< 10^{-4})$	?

Table 8: The comparison of differences in *acc* for all pairs of algorithms using Wilcoxon signed rank test (p-value) for 20% train data and 80% test data for five datasets.

Alg. 1	Alg. 2	BCC	CPU	ESL	ERA	LEV
GEN	PSO	?	?	?	$\prec (0.033)$	?
	FSS	?	?	$\prec (0.040)$	?	$\succ (< 10^{-5})$
	SLS	?	?	$\succ (0.043)$	?	$\succ (0.032)$
	GLS	?	$\succ (0.031)$	?	$\succ (< 10^{-3})$	$\succ (< 10^{-4})$
	SAN	?	?	$\succ (0.015)$	$\succ (< 10^{-5})$	$\succ (< 10^{-3})$
	MNR	$\succ (< 10^{-5})$	$\prec (< 10^{-6})$	$\succ (< 10^{-7})$	?	$\succ (< 10^{-5})$
	MMR	?	$\prec (< 10^{-9})$	$\succ (0.010)$	$\succ (< 10^{-10})$	$\succ (< 10^{-14})$
PSO	FSS	?	?	?	$\succ (0.005)$	$\succ (0.001)$
	SLS	$\succ (0.006)$	?	$\succ (0.017)$	$\succ (0.023)$	?
	GLS	?	$\succ (< 10^{-3})$	?	$\succ (< 10^{-5})$	$\succ (0.023)$
	SAN	$\succ (0.021)$	$\succ (0.044)$	$\succ (0.006)$	$\succ (< 10^{-7})$	$\succ (0.019)$
	MNR	$\succ (< 10^{-6})$	$\prec (< 10^{-6})$	$\succ (< 10^{-7})$	?	$\succ (< 10^{-3})$
	MMR	?	$\prec (< 10^{-9})$	$\succ (< 10^{-3})$	$\succ (< 10^{-11})$	$\succ (< 10^{-13})$
FSS	SLS	$\succ (0.010)$	?	$\succ (0.003)$	?	$\prec (< 10^{-3})$
	GLS	?	$\succ (0.006)$	$\succ (0.043)$	?	?
	SAN	$\succ (0.013)$	?	$\succ (0.001)$	$\succ (0.035)$	$\prec (0.029)$
	MNR	$\succ (< 10^{-5})$	$\prec (< 10^{-5})$	$\succ (< 10^{-8})$	?	?
	MMR	?	$\prec (< 10^{-9})$	$\succ (< 10^{-3})$	$\succ (< 10^{-7})$	$\succ (< 10^{-8})$
SLS	GLS	?	$\succ (0.044)$	?	$\succ (0.005)$	$\succ (0.007)$
	SAN	?	?	?	$\succ (< 10^{-3})$	$\succ (0.044)$
	MNR	$\succ (0.002)$	$\prec (< 10^{-6})$	$\succ (< 10^{-5})$	?	$\succ (< 10^{-3})$
	MMR	?	$\prec (< 10^{-9})$	$\succ (0.029)$	$\succ (< 10^{-9})$	$\succ (< 10^{-12})$
GLS	SAN	$\succ (0.039)$	?	?	?	?
	MNR	$\succ (< 10^{-5})$	$\prec (< 10^{-8})$	$\succ (< 10^{-6})$	$\prec (0.003)$	$\succ (0.007)$
	MMR	?	$\prec (< 10^{-11})$	$\succ (0.022)$	$\succ (< 10^{-5})$	$\succ (< 10^{-10})$
SAN	MNR	$\succ (< 10^{-4})$	$\prec (< 10^{-8})$	$\succ (< 10^{-5})$	$\prec (< 10^{-3})$	$\succ (0.012)$
	MMR	?	$\prec (< 10^{-11})$	?	$\succ (< 10^{-5})$	$\succ (< 10^{-10})$
MNR	MMR	$\prec (< 10^{-4})$	?	$\prec (< 10^{-3})$	$\succ (< 10^{-8})$	$\succ (< 10^{-4})$

ERA and BCC datasets emerged as the most challenging, exhibiting vastly lower mean *auc* values. A similar pattern is observed for *acc* scores, albeit with marginally lower absolute figures. An intriguing aspect to consider is the variation across datasets, reflected in the range of mean values achieved by the algorithms. For CPU and ESL, the disparity between the highest and lowest mean *auc* values was relatively narrow. In contrast, for LEV and ERA, the differences were more pronounced. This suggests that benchmark datasets with a larger set of alternatives may offer a better opportunity to highlight significant distinctions between algorithms in terms of *auc*. However, a parallel examination of *acc* values does not universally support these findings. Here, the greatest variances were observed for CPU and ERA, with the smallest differences noted for ESL and BCC.

To reflect a comprehensive picture enabling the comparison of algorithms in the considered scenarios, Tables 7–12 summarize the results verifying the statistical significance of differences for all pairs of algorithms, for both quality measures, taking into account the analyzed distributions of alternatives to the reference and test sets – $p_r \in \{0.2, 0.5, 0.8\}$. The symbols in the tables indicate whether and which algorithm in a given pair attained significantly better results according to the Wilcoxon Signed-Rank Test with $\alpha = 0.05$. The p-value of the test performed is provided in round brackets.

Among all problems with known assignments for $p_r = 20\%$ of the alternatives, PSO outperformed other approaches, achieving both $r^{0.2}_{auc}$ and $r^{0.2}_{acc}$ of 1.8. For all problems except CPU, the average values obtained by PSO were among the best two, and none of the other methods statistically significantly outperformed this algorithm. Considering CPU, it was second only to the mathematical programming approaches, which performed the worst in all other problems, except the *acc* for the ERA problem, where MNR achieved competitive results.

The second choice is GEN, with average rankings of 2.4 for *auc* and 3.2 for *acc*. This approach achieved the best results for LEV and was among the best methods for ERA on *auc* and BCC on *acc*. In other cases, it was among the best three algorithms, with one exception of *acc* for the CPU problem (0.9036), where it was significantly inferior to MMR (0.9352, p-value $< 10^{-9}$) and MNR (0.9284, p-value $< 10^{-6}$).

The following best-performing approach was FSS with $r^{0.2}_{auc} = 4.2$ and $r^{0.2}_{acc} = 3.6$. It obtained one of the best results for ESL, but in other cases, it was inferior to other approaches. For example, for LEV, it performed significantly worse performance than GEN, PSO, SAN, and SLS for both measures, as well as for GLS on *auc* (0.8751 vs. 0.8792, p-value $< 10^{-3}$).

Among the group of local search approaches, the results for *auc* are similar, and in the case of SAN and SLS, no significant differences were observed. In turn, *GLS* (0.9778) outperformed both of these methods on ESL (p-value $< 10^{-3}$ for both comparisons), and for ERA (0.7486), it performed significantly worse (p-value = 0.002 for SLS and 0.043 for SAN). In turn, for *acc*, SLS obtained the best average ranking of $r^{0.2}_{acc} = 5.0$, outperforming both methods on ERA and LEV, as well as GLS on the CPU problem (0.9026 vs. 0.8962, p-value = 0.044).

For problems in which half of the alternatives provided reference assignments ($p_r = 0.5$), the advantage of GEN over the other algorithms is evident. Taking into account *auc*, for two problems,

Table 9: The comparison of differences in *auc* for all pairs of algorithms using Wilcoxon signed rank test (p-value) for 50% train data and 50% test data for five datasets.

Alg. 1	Alg. 2	BCC	CPU	ESL	ERA	LEV
GEN	PSO	?	$\succ (< 10^{-6})$	?	$\succ (0.023)$	$\succ (< 10^{-9})$
	FSS	$\succ (0.004)$	$\succ (< 10^{-9})$	$\prec (0.012)$	$\succ (< 10^{-3})$	$\succ (< 10^{-10})$
	SLS	$\succ (< 10^{-3})$	$\succ (< 10^{-3})$	$\succ (< 10^{-10})$	$\succ (< 10^{-7})$	$\succ (< 10^{-8})$
	GLS	$\succ (0.008)$	$\succ (< 10^{-4})$	$\succ (< 10^{-12})$	$\succ (< 10^{-8})$	$\succ (< 10^{-8})$
	SAN	$\succ (< 10^{-4})$	$\succ (< 10^{-5})$	$\succ (< 10^{-13})$	$\succ (< 10^{-7})$	$\succ (< 10^{-6})$
	MNR	$\succ (< 10^{-9})$	?	$\succ (< 10^{-16})$	$\succ (< 10^{-7})$	$\succ (< 10^{-12})$
	MMR	$\succ (< 10^{-3})$	?	$\succ (< 10^{-17})$	$\succ (< 10^{-17})$	$\succ (< 10^{-17})$
PSO	FSS	?	$\succ (< 10^{-3})$	$\prec (0.028)$	?	?
	SLS	$\succ (0.021)$	$\prec (0.006)$	$\succ (< 10^{-7})$	$\succ (< 10^{-3})$	?
	GLS	$\succ (0.036)$	?	$\succ (< 10^{-8})$	$\succ (< 10^{-5})$	?
	SAN	$\succ (< 10^{-3})$	?	$\succ (< 10^{-10})$	$\succ (< 10^{-5})$	$\prec (0.026)$
	MNR	$\succ (< 10^{-9})$	$\prec (< 10^{-6})$	$\succ (< 10^{-17})$	$\succ (< 10^{-5})$	$\succ (< 10^{-5})$
	MMR	$\succ (0.012)$	$\prec (0.020)$	$\succ (< 10^{-16})$	$\succ (< 10^{-16})$	$\succ (< 10^{-15})$
FSS	SLS	?	$\prec (< 10^{-6})$	$\succ (< 10^{-10})$	$\succ (0.006)$	$\prec (0.018)$
	GLS	?	$\prec (< 10^{-4})$	$\succ (< 10^{-11})$	$\succ (< 10^{-3})$	?
	SAN	$\succ (0.006)$	$\prec (0.006)$	$\succ (< 10^{-13})$	$\succ (< 10^{-3})$	$\prec (0.006)$
	MNR	$\succ (< 10^{-7})$	$\prec (< 10^{-11})$	$\succ (< 10^{-16})$	$\succ (< 10^{-4})$	$\succ (< 10^{-3})$
	MMR	?	$\prec (< 10^{-5})$	$\succ (< 10^{-17})$	$\succ (< 10^{-16})$	$\succ (< 10^{-15})$
SLS	GLS	?	?	?	?	?
	SAN	?	$\succ (0.006)$	?	?	?
	MNR	$\succ (< 10^{-5})$	$\prec (< 10^{-5})$	$\succ (< 10^{-11})$	?	$\succ (< 10^{-6})$
	MMR	?	?	$\succ (< 10^{-13})$	$\succ (< 10^{-14})$	$\succ (< 10^{-17})$
GLS	SAN	?	?	?	?	?
	MNR	$\succ (< 10^{-6})$	$\prec (< 10^{-6})$	$\succ (< 10^{-9})$	?	$\succ (< 10^{-6})$
	MMR	?	$\prec (0.016)$	$\succ (< 10^{-14})$	$\succ (< 10^{-13})$	$\succ (< 10^{-16})$
SAN	MNR	$\succ (< 10^{-5})$	$\prec (< 10^{-7})$	$\succ (< 10^{-9})$	?	$\succ (< 10^{-6})$
	MMR	?	$\prec (0.007)$	$\succ (< 10^{-12})$	$\succ (< 10^{-14})$	$\succ (< 10^{-16})$
MNR	MMR	$\prec (< 10^{-5})$	$\succ (0.020)$	$\succ (0.028)$	$\succ (< 10^{-7})$	$\succ (< 10^{-3})$

Table 10: The comparison of differences in *acc* for all pairs of algorithms using Wilcoxon signed rank test (p-value) for 50% train data and 50% test data for five datasets.

Alg. 1	Alg. 2	BCC	CPU	ESL	ERA	LEV
GEN	PSO	?	$\succ (< 10^{-4})$	$\prec (< 10^{-3})$	?	$\succ (< 10^{-7})$
	FSS	?	$\succ (< 10^{-8})$	$\prec (0.042)$	?	$\succ (< 10^{-6})$
	SLS	?	$\succ (0.041)$	$\succ (0.002)$	$\succ (< 10^{-3})$	$\succ (0.002)$
	GLS	?	$\succ (< 10^{-4})$	$\succ (< 10^{-3})$	$\succ (< 10^{-7})$	$\succ (< 10^{-6})$
	SAN	?	$\succ (< 10^{-4})$	$\succ (< 10^{-5})$	$\succ (< 10^{-5})$	$\succ (< 10^{-4})$
	MNR	$\succ (0.001)$	$\prec (< 10^{-3})$	$\succ (< 10^{-6})$	$\succ (0.024)$	$\succ (< 10^{-6})$
	MMR	?	$\prec (< 10^{-6})$	$\succ (0.009)$	$\succ (< 10^{-13})$	$\succ (< 10^{-16})$
PSO	FSS	?	$\succ (0.001)$	$\succ (0.029)$	?	?
	SLS	$\succ (0.007)$	$\prec (0.006)$	$\succ (< 10^{-5})$	$\succ (< 10^{-5})$	$\prec (< 10^{-3})$
	GLS	$\succ (0.013)$	?	$\succ (< 10^{-6})$	$\succ (< 10^{-7})$	?
	SAN	$\succ (0.013)$	?	$\succ (< 10^{-7})$	$\succ (< 10^{-5})$	$\prec (0.032)$
	MNR	$\succ (< 10^{-4})$	$\prec (< 10^{-8})$	$\succ (< 10^{-10})$	$\succ (0.009)$	?
	MMR	?	$\prec (< 10^{-11})$	$\succ (< 10^{-6})$	$\succ (< 10^{-13})$	$\succ (< 10^{-11})$
FSS	SLS	$\succ (0.048)$	$\prec (< 10^{-6})$	$\succ (< 10^{-3})$	$\succ (< 10^{-4})$	$\prec (< 10^{-3})$
	GLS	?	$\prec (< 10^{-3})$	$\succ (< 10^{-4})$	$\succ (< 10^{-6})$	?
	SAN	?	$\prec (0.004)$	$\succ (< 10^{-5})$	$\succ (< 10^{-4})$	?
	MNR	$\succ (< 10^{-3})$	$\prec (< 10^{-13})$	$\succ (< 10^{-6})$	$\succ (0.014)$	$\succ (0.032)$
	MMR	?	$\prec (< 10^{-14})$	$\succ (< 10^{-4})$	$\succ (< 10^{-13})$	$\succ (< 10^{-12})$
SLS	GLS	?	$\succ (0.018)$	?	$\succ (0.033)$	$\succ (< 10^{-3})$
	SAN	?	$\succ (0.020)$	?	?	$\succ (0.042)$
	MNR	$\succ (0.018)$	$\prec (< 10^{-5})$	$\succ (< 10^{-3})$	?	$\succ (< 10^{-3})$
	MMR	$\prec (0.010)$	$\prec (< 10^{-8})$	?	$\succ (< 10^{-8})$	$\succ (< 10^{-15})$
GLS	SAN	?	?	?	?	?
	MNR	?	$\prec (< 10^{-7})$	$\succ (0.006)$	$\prec (0.025)$	?
	MMR	$\prec (0.018)$	$\prec (< 10^{-10})$	?	$\succ (< 10^{-8})$	$\succ (< 10^{-11})$
SAN	MNR	$\succ (0.026)$	$\prec (< 10^{-7})$	$\succ (0.018)$	$\prec (0.023)$	$\succ (0.005)$
	MMR	$\prec (0.035)$	$\prec (< 10^{-10})$	?	$\succ (< 10^{-9})$	$\succ (< 10^{-14})$
MNR	MMR	$\prec (< 10^{-4})$	$\prec (0.003)$	$\prec (0.002)$	$\succ (< 10^{-8})$	$\succ (< 10^{-8})$

it was significantly better than all the others (p-values ≤ 0.023 for ERA and $< 10^{-6}$ for LEV), and for the next two (BCC and CPU), it ranked among the best methods. The only exception was the ESL problem, for which GEN was second only to FSS (0.9815 vs. 0.9819, p-value = 0.012). Similarly, for *acc*, GEN scored the best $r^{0.5}_{acc} = 2.6$, and it was only surpassed by MNR and MMR for CPU (p-values $< 10^{-3}$) and FSS and PSO for ESL (p-values ≤ 0.042).

PSO ($r^{0.5}_{auc} = 3.4$) achieved slightly worse results than GEN. However, in the case of *auc*, it was among the best algorithms – when considering the statistical significance of differences – only for the BCC problem. For the remaining four datasets, GEN turned out to be better on all of them, and additionally, FSS for the ESL dataset. In the case of *acc*, for BCC and ERA problems, it significantly outperformed local search and mathematical programming approaches (except MMR in the case of BCC, where no significant differences were noticed). Moreover, considering the ESL problem, PSO has gained a significant advantage over all other algorithms. However, for the remaining two problems, it performed worse, scoring the sixth average value among all approaches. Similar phenomena can also be observed for FSS, which was usually slightly inferior to PSO on both quality measures, except for *auc* for ESL, where it achieved a significantly better result (0.9819 vs. 0.9815, p-value = 0.028) and *acc* for LEV, where the difference was not significant.

Among local search approaches, SLS shows the best ranking, with $r^{0.5}_{auc} = 4.0$ and $r^{0.5}_{acc} = 4.2$, while for the remaining methods, the average ranking ranged from 5.0 to 5.8. It is also worth noting that none of these methods is among the leading algorithms for any problem. When considering *auc*, the results did not differ significantly from each other; the only exception was the values for CPU, where SLS confirmed its advantage over SAN (0.9875 vs. 0.9855, p-value = 0.006). However, for *acc*, the advantage of SLS over the others becomes more visible, as it outperforms GLS and SAN on the CPU and LEV problems, as well as GLS on the ERA problem (0.7089 vs. 0.7063, p-value = 0.033).

When comparing MNR and MMR, the former gains a significant advantage over the latter on *auc* on four out of five datasets and scores the best on CPU among all algorithms. The exception is BCC, where MMR is better (0.7334 vs. 0.7222, p-value $< 10^{-5}$). On the other hand, for *acc*, MMR reveals a better average ranking of $r^{0.5}_{acc} = 4.6$ compared to MNR and $r^{0.5}_{acc} = 5.8$ and obtaining the best results among all approaches for BCC and CPU datasets. Apart from these two problems, it is statistically significantly better than MNR on ESL. For ERA and LEV, the situation is reversed, and MNR achieves significantly better results. Nevertheless, apart from the CPU problem, one of these methods always scores the worst among all analyzed approaches.

Table 11: The comparison of differences in *auc* for all pairs of algorithms using Wilcoxon signed rank test (p-value) for 80% train data and 20% test data for five datasets.

Alg. 1	Alg. 2	BCC	CPU	ESL	ERA	LEV
GEN	PSO	?	$\succ (< 10^{-13})$	$\succ (0.004)$	$\succ (< 10^{-3})$	$\succ (0.002)$
	FSS	$\succ (0.016)$	$\succ (< 10^{-12})$	?	$\succ (< 10^{-3})$	$\succ (< 10^{-6})$
	SLS	$\succ (0.002)$	?	$\succ (< 10^{-11})$	$\succ (< 10^{-7})$	$\succ (< 10^{-6})$
	GLS	$\succ (0.006)$	?	$\succ (< 10^{-10})$	$\succ (< 10^{-5})$	$\succ (< 10^{-7})$
	SAN	$\succ (0.007)$	$\succ (< 10^{-12})$	$\succ (< 10^{-11})$	$\succ (< 10^{-7})$	?
	MNR	$\succ (0.001)$	?	$\succ (< 10^{-16})$	$\succ (< 10^{-6})$	$\succ (< 10^{-7})$
	MMR	?	?	$\succ (< 10^{-16})$	$\succ (< 10^{-16})$	$\succ (< 10^{-13})$
PSO	FSS	$\succ (0.009)$	?	$\prec (0.029)$	?	$\succ (0.009)$
	SLS	$\succ (< 10^{-3})$	$\prec (< 10^{-12})$	$\succ (< 10^{-5})$	$\succ (0.013)$	$\succ (0.015)$
	GLS	$\succ (0.002)$	$\prec (< 10^{-10})$	$\succ (< 10^{-5})$	?	$\succ (< 10^{-3})$
	SAN	$\succ (< 10^{-3})$	?	$\succ (< 10^{-5})$	$\succ (< 10^{-3})$	$\prec (0.026)$
	MNR	$\succ (0.003)$	$\prec (< 10^{-8})$	$\succ (< 10^{-15})$	$\succ (0.009)$	$\succ (< 10^{-4})$
	MMR	?	$\prec (< 10^{-5})$	$\succ (< 10^{-14})$	$\succ (< 10^{-14})$	$\succ (< 10^{-10})$
FSS	SLS	?	$\prec (< 10^{-11})$	$\succ (< 10^{-10})$	$\succ (0.002)$	?
	GLS	?	$\prec (< 10^{-10})$	$\succ (< 10^{-9})$	$\succ (0.023)$	?
	SAN	?	?	$\succ (< 10^{-8})$	$\succ (< 10^{-3})$	$\prec (< 10^{-4})$
	MNR	?	$\prec (< 10^{-7})$	$\succ (< 10^{-16})$	$\succ (0.001)$	?
	MMR	$\prec (0.036)$	$\prec (< 10^{-5})$	$\succ (< 10^{-16})$	$\succ (< 10^{-15})$	$\succ (< 10^{-6})$
SLS	GLS	?	?	?	?	?
	SAN	?	$\succ (< 10^{-12})$	?	?	$\prec (< 10^{-4})$
	MNR	?	?	$\succ (< 10^{-10})$	?	?
	MMR	$\prec (0.002)$	?	$\succ (< 10^{-11})$	$\succ (< 10^{-13})$	$\succ (< 10^{-9})$
GLS	SAN	?	$\succ (< 10^{-8})$	?	?	$\prec (< 10^{-6})$
	MNR	?	?	$\succ (< 10^{-10})$	?	?
	MMR	$\prec (0.005)$	?	$\succ (< 10^{-13})$	$\succ (< 10^{-14})$	$\succ (< 10^{-6})$
SAN	MNR	?	$\prec (< 10^{-7})$	$\succ (< 10^{-12})$	?	$\succ (< 10^{-5})$
	MMR	$\prec (0.003)$	$\prec (< 10^{-5})$	$\succ (< 10^{-13})$	$\succ (< 10^{-11})$	$\succ (< 10^{-12})$
MNR	MMR	$\prec (0.007)$	?	?	$\succ (< 10^{-11})$	$\succ (< 10^{-3})$

Table 12: The comparison of differences in *acc* for all pairs of algorithms using Wilcoxon signed rank test (p-value) for 80% train data and 20% test data for five datasets.

Alg. 1	Alg. 2	BCC	CPU	ESL	ERA	LEV
GEN	PSO	?	$\succ (< 10^{-8})$	?	?	$\succ (0.007)$
	FSS	?	$\succ (< 10^{-8})$	?	?	$\succ (< 10^{-5})$
	SLS	?	?	$\succ (< 10^{-3})$	$\succ (0.005)$	$\succ (< 10^{-5})$
	GLS	?	?	$\succ (0.014)$	$\succ (< 10^{-5})$	$\succ (0.018)$
	SAN	$\succ (0.006)$	$\succ (< 10^{-6})$	$\succ (< 10^{-3})$	$\succ (< 10^{-3})$	?
	MNR	?	?	$\succ (< 10^{-7})$	$\succ (0.005)$	$\succ (< 10^{-3})$
	MMR	$\prec (0.042)$	$\prec (< 10^{-5})$	$\succ (< 10^{-3})$	$\succ (< 10^{-10})$	$\succ (< 10^{-9})$
PSO	FSS	?	?	?	?	?
	SLS	?	$\prec (< 10^{-9})$	$\succ (< 10^{-3})$	$\succ (0.001)$	?
	GLS	?	$\prec (< 10^{-7})$	$\succ (0.006)$	$\succ (< 10^{-4})$	?
	SAN	$\succ (0.018)$	?	$\succ (< 10^{-3})$	$\succ (< 10^{-3})$	$\prec (0.013)$
	MNR	?	$\prec (< 10^{-6})$	$\succ (< 10^{-6})$	$\succ (0.006)$	?
	MMR	$\prec (0.017)$	$\prec (< 10^{-12})$	$\succ (< 10^{-4})$	$\succ (< 10^{-10})$	$\succ (< 10^{-6})$
FSS	SLS	$\succ (0.031)$	$\prec (< 10^{-8})$	$\succ (< 10^{-3})$	$\succ (< 10^{-3})$	?
	GLS	$\succ (0.047)$	$\prec (< 10^{-8})$	$\succ (0.013)$	$\succ (< 10^{-5})$	$\prec (0.012)$
	SAN	$\succ (0.001)$	?	$\succ (< 10^{-4})$	$\succ (< 10^{-3})$	$\prec (< 10^{-4})$
	MNR	?	$\prec (< 10^{-6})$	$\succ (< 10^{-6})$	$\succ (< 10^{-3})$	?
	MMR	?	$\prec (< 10^{-13})$	$\succ (< 10^{-3})$	$\succ (< 10^{-11})$	$\succ (< 10^{-4})$
SLS	GLS	?	?	?	?	$\prec (0.006)$
	SAN	?	$\succ (< 10^{-7})$	?	?	$\prec (< 10^{-5})$
	MNR	?	?	$\succ (0.004)$	?	?
	MMR	$\prec (0.003)$	$\prec (< 10^{-5})$	?	$\succ (< 10^{-7})$	$\succ (< 10^{-4})$
GLS	SAN	?	$\succ (< 10^{-6})$	?	?	$\prec (0.033)$
	MNR	?	?	$\succ (< 10^{-4})$	?	?
	MMR	$\prec (0.005)$	$\prec (< 10^{-4})$	$\succ (0.042)$	$\succ (< 10^{-5})$	$\succ (< 10^{-7})$
SAN	MNR	$\prec (0.014)$	$\prec (< 10^{-5})$	$\succ (< 10^{-3})$	?	$\succ (< 10^{-3})$
	MMR	$\prec (< 10^{-6})$	$\prec (< 10^{-11})$	?	$\succ (< 10^{-6})$	$\succ (< 10^{-9})$
MNR	MMR	?	$\prec (< 10^{-4})$	?	$\succ (< 10^{-6})$	$\succ (< 10^{-4})$

Publication [P4]

M. Kadziński, M. Wójcik, and M. Ghaderi, "From investigation of expressiveness and robustness to a comprehensive value-based framework for multiple criteria sorting problems", *Omega*, 2024. Status: accepted for publication

From investigation of expressiveness and robustness to a comprehensive value-based framework for multiple criteria sorting problems

Miłosz Kadziński[a], Michał Wójcik[a], Mohammad Ghaderi[b,c,d,*]

[a] *Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznań, Poland*
[b] *Department of Economics and Business, Pompeu Fabra University, 08005 Barcelona, Spain*
[c] *Barcelona School of Economics, Barcelona, Spain*
[d] *Barcelona School of Management, Barcelona, Spain*

Abstract

We adopt an experiment-oriented perspective to investigate two essential characteristics – expressiveness and robustness – of multiple criteria sorting methods. We focus on the approaches from the family of UTADIS, learning the parameters of a value-driven threshold-based model from the Decision Maker's assignment examples. Even if the considered properties are crucial for the methods' reliability and usefulness in real-world scenarios, their verification through explicit numerical tests has been so far neglected. On the one hand, expressiveness captures the models' flexibility to reproduce different preferences, including simple and complex ones, meaningfully and accurately. On the other hand, robustness reflects the ability to deliver valid recommendations and ensure proper conclusiveness given the multiplicity of compatible preference model instances. We consider different variants of UTADIS, from assuming monotonic and preferentially independent criteria to more advanced settings that relax the monotonicity constraints or represent interactions. The experimental results capture the trade-off between the considered quality dimensions, indicating that richer models are characterized by greater expressiveness and lesser robustness. We also formulate a comprehensive framework indicating when some variant should be used, given the nature of supplied preferences or problem characteristics. These findings aid decision analysts in making robust recommendations in different contexts and help refine preference modeling assumptions. The framework's practical use is illustrated in a case study involving sorting mobile phone models into pre-defined preference-ordered classes.

Keywords: Multiple criteria decision aiding, Sorting, Model expressiveness, Recommendation robustness, Interactions, Non-monotonicity

1. Introduction

Multiple Criteria Decision Aiding (MCDA) aims at developing approaches that support solving complex decision problems [18]. This complexity derives from the multiplicity of alternative ways to attain a decision goal and pertinent factors relevant to their assessment [25]. The essence of MCDA tools consists of processing an objective problem's description and the stakeholders' subjective preferences to make recommendations.

Over the last decades, hundreds of MCDA methods have been proposed [6, 22, 51]. They all have been conceived with specific intentions on when they might be helpful and how they should perform. When it comes to the former aspect, the relevant characteristics can be divided into four major categories. First, the methods tackle problems with distinct types and structures. Second, they apply various preference models to faithfully represent the stakeholders' judgments and aggregate the performances on multiple criteria. Third, they vary in type and modality of preferences and frequency of the elicitation process. Fourth, they apply different strategies for exploiting the preference relation induced by the model to compute the recommendation. Such objective features are often

*Corresponding author: Department of Economics and Business, Pompeu Fabra University, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. Tel. +34 93 542 2682.

Email addresses: milosz.kadzinski@cs.put.poznan.pl (Miłosz Kadziński), michal.wojcik@cs.put.poznan.pl (Michał Wójcik), mohammad.ghaderi@upf.edu (Mohammad Ghaderi)

used to select an appropriate method for a given decision problem. However, the aspects related to the performance of MCDA methods are often neglected. Nonetheless, checking experimentally whether these tools conform to what was expected from them is worthwhile.

This paper focuses on multiple criteria sorting, i.e., assigning alternatives to pre-defined, preference-ordered classes [1]. In particular, we consider a preference disaggregation setting where the parameters of an assumed model are induced from the assignment examples provided by the Decision Makers (DMs) [13]. Such holistic judgments specify desired classifications for a subset of reference alternatives, representing the DMs' decision policy and value system. Even if preference disaggregation approaches are considered more user-friendly due to reducing cognitive effort, their use implies two significant problems. On the one hand, the indirect preferences may be incompatible with an assumed model, leading to an empty space of feasible parameters [38]. On the other hand, when the method can represent the supplied information, typically multiple consistent model instances or feasible parameter sets exist and may lead to various recommendations on the set of non-reference alternatives [19, 48].

In the context of preference disaggregation sorting methods, only a few studies verified some of their desirable properties through explicit numerical tests. In particular, The and Mousseau [46] considered the inference procedures for ELECTRE TRI-B. They analyzed the amount of indirect information needed to infer the parameters reliably, the method's ability to detect inconsistencies, and the results' stability given various objective functions. Additionally, Vetschera et al. [47] investigated the properties of two methods, case-based distance sorting and simple additive weighting, to capture the impact of various problem dimensions on three characteristics: a) compatibility reflecting the size of the space of feasible parameters, b) robustness capturing the tendency of alternatives to be assigned to the same class for all feasible parameters, and c) validity interpreted as the probability of alternatives being sorted in the right class. Further, Doumpos et al. [14] considered five value-based sorting procedures to examine their predictive abilities and relation with the robust recommendations that can be formulated based on the DM's reference judgments. The latter study has been extended in [48] regarding the number of accounted procedures and investigated measures.

We aim to adopt an experiment-oriented perspective to the methods from the family of UTADIS [8, 42, 49]. They apply a value-driven threshold-based sorting procedure in which the comprehensive quality of each alternative is quantified using an additive value function, and the value ranges for all classes are delimited by the lower and upper thresholds [19, 44]. Such a model adequately represents how individuals make classification decisions for different options. Consequently, it is appreciated in the MCDA community for the intuitiveness and high interpretability of the delivered results. Therefore, the UTADIS methods have already been applied to solve real problems such as supplier classification [36], credit risk assessment [50], classification of securities [9], and subcontractor assessment [39].

The basic variant of UTADIS assumes the monotonicity of per-criterion preferences and the condition of preferential independence [8, 42]. The former implies that all criteria must be gain- or cost-type, and therefore, preferences are represented by non-decreasing or non-increasing marginal value functions. The latter conjectures that the impact of one attribute on an alternative's comprehensive score should not be influenced by the alternative's performance on other attributes. However, various extensions of value-based methods have been proposed to relax the limitations mentioned above. On the one hand, some procedures are oriented toward admitting non-monotonicity of marginal functions [16, 30]. On the other hand, the interactions between criteria can be incorporated into an additive value model using the bonuses and penalties related to observing specific combinations of performances on a subset of criteria [21]. These modifications can influence the vital properties of the underlying sorting approaches, deciding upon their suitability for being used in real-world problems with incomplete preference information. In particular, we expect richer models to be applicable in a greater variety of decision settings but lead to less stable results. However, these hypotheses need to be verified experimentally.

Our main contribution consists of performing an extensive computational study investigating the fundamental properties of six variants of UTADIS (some of them newly proposed). These characteristics are crucial for the methods' reliability and usefulness in real-world scenarios [29]. To capture a trade-off between the flexibility

of value-based preference models and their ability to reproduce the DM's indirect preferences, we will assess the expressiveness. This feature refers to the model's ability to capture the actual preferences of individuals in a meaningful and accurate way. A practically helpful preference model should be flexible enough to accommodate different preferences, including simple and complex ones. This means it should not impose overly restrictive assumptions limiting its applicability. Also, it should minimize errors and discrepancies between actual choices and preferences it intends to represent or predict.

While expressiveness is valuable, overly complex models may deliver recommendations that lack conclusiveness. Therefore, striking a balance between expressiveness and robustness is essential. Thus, we will also verify the methods' robustness, understood as the ability to ensure that the representations, predictions, or recommendations they deliver are valid and accurate under different conditions. Specifically, we will investigate the stability of outcomes computed with various methods based on the same preferences given the respective set of compatible model instances. The consistency of results produced with a more robust model enhances the trust of the DMs, making the respective recommendations more likely to be implemented in real-world decision-making [2, 30].

The conducted experiment involves a broad range of problems characterized by various numbers of classes, criteria, characteristic points of marginal functions, and reference alternatives per class, and algorithms used to simulate performances of non-dominated alternatives. We consider seven measures to quantify expressiveness and robustness. Regarding the former, we focus on the proportion of scenarios for which the indirect preferences are fully consistent with an assumed model [29] and the misclassification error [19]. As for the latter, we build five metrics referring to the precision of possible assignments for all alternatives [19] and the variability of class acceptability indices [31]. The possible classifications are confirmed by at least one compatible model instance. In turn, the stochastic acceptabilities represent the shares of feasible instances suggesting specific assignments, serving as the base for the entropy-inspired measures. Analyzing the expressiveness and robustness provides insights into the usefulness of UTADIS variants in different decision-aiding contexts and the amount of preference information needed from the DMs to restore their views faithfully. These insights can be used by the decision analysts, who are responsible for interacting with the DM as well as operating and selecting the methods when facing a decision problem.

Our experimental results enable the formulation of guidelines for selecting the appropriate model based on the nature of supplied preferences. The framework outlines the necessary steps to improve recommendations' robustness in different contexts and to revise the preference modeling assumptions, particularly concerning non-monotonicity and interactions. Its use is illustrated in the problem of sorting mobile phone models based on the preferences of three DMs. Nevertheless, we also formulate some taxonomy-based guidelines on selecting an appropriate variant of UTADIS based on the characteristics of the tackled decision problem. They refer to the features regarding problem formulation, preference model, and preference information.

The paper's remainder is organized in the following way. Section 2 reminds the primary UTADIS method and defines modified variants of this approach. In Section 3, we discuss the concepts of expressiveness and robustness along with proposed quality measures, experimental settings, and analysis of the obtained results. In Section 4, we discuss frameworks to support the choice of an appropriate method. Section 5 illustrates the use of the framework. The last section concludes the paper and outlines promising future research directions.

2. Reminder on the UTADIS method and its extensions

In this section, we present variants of the UTADIS method. We start from the basic approach and then demonstrate how to relax the constraints on monotonicity or preferential independence. Also, we discuss the robustness analysis methods whose results will subsequently serve to define the measures relevant to the experimental verification. We use the following notation:

- $A = \{a_1, a_2, \ldots, a_i, \ldots, a_n\}$ – a finite set of n alternatives, each evaluated in terms of m criteria;

- $A^R = \{a_1^*, a_2^*, \ldots, a_r^*\}$ – a finite set of r reference alternatives; $A^R \subseteq A$; for each reference alternative, the DM provides a desired assignment;
- $G = \{g_1, g_2, \ldots, g_j, \ldots, g_m\}$ – a finite set of m criteria, $g_j : A \to \mathbb{R}$ for all $j \in J = \{1, \ldots, m\}$; without loss of generality, for now, we assume that all criteria in G are of gain type;
- $X_j = \{g_j(a_i), a_i \in A\}$ – a finite set of performances of all alternatives in A on criterion g_j;
- $x_j^1, x_j^2, \ldots, x_j^{n_j(A)}$ – the ordered values of X_j, $x_j^{k-1} < x_j^k, k = 2, \ldots, n_j(A)$, where $n_j(A) = |X_j|$ and $n_j(A) \leq n$; thus, $X = \prod_{j=1}^m X_j$ is the performance space; note that X_j can also be enriched with the extreme values of the performance scale that are not attained by any alternative;
- $C_1, C_2, \ldots, C_p$ – p pre-defined and preference-ordered classes so that C_l is preferred to C_{l-1} for $l = 2, \ldots, p$.

2.1. Basic model

The first considered preference disaggregation method is **UTADIS**, proposed in [8]. For each alternative $a \in A$, this approach quantifies a comprehensive quality using an Additive Value Function (AVF) [33]:

$$U(a) = \sum_{j=1}^{m} u_j(g_j(a)), \tag{1}$$

where u_j, $j = 1, \ldots, m$, are Marginal Value Functions (MVFs). These are piecewise linear, monotonic functions, defined on the set of γ_j pre-defined and equally distributed characteristic points $\beta_j^1, \beta_j^2, \ldots, \beta^{\gamma_j}$, such that:

$$\beta_j^s = x_j^1 + (x_j^{n_j(A)} - x_j^1)\frac{s-1}{\gamma_j - 1}, j = 1, \ldots, m, s = 1, \ldots, \gamma_j. \tag{2}$$

To ensure that comprehensive values $U(a), \forall a \in A$, are normalized to the $[0, 1]$ range, the following constraints are incorporated: $u_j(\beta_j^1) = 0$, for $j = 1, \ldots, m$, and $\sum_{j=1}^{m} u_j(\beta_j^{\gamma_j}) = 1$. For gain-type criteria, MVFs are assumed to be non-decreasing, which is modeled as follows:

$$u_j(\beta_j^s) - u_j(\beta_j^{s-1}) \geq 0,\ j = 1, \ldots, m, s = 2, \ldots, \gamma_j. \tag{3}$$

To determine the marginal values for $x_j^k \in [\beta_j^s, \beta_j^{s+1}]$, linear interpolation is used:

$$u_j(x_j^k) = u_j(\beta_j^s) + (u_j(\beta_j^{s+1}) - u_j(\beta_j^s))\frac{x_j^k - \beta_j^s}{\beta_j^{s+1} - \beta_j^s},\ j = 1, \ldots, m,\ k = 1, \ldots, n_j(A). \tag{4}$$

To assign alternatives to pre-defined, preference-ordered classes, UTADIS applies a threshold-based value-driven procedure. In this approach, each class C_l is delimited by the lower t_{l-1} and upper t_l thresholds (see Figure 1). For simplicity, we omit the lower limit of the least-preferred class C_1 and the upper limit of the most-preferred class C_p. Hence, the model includes $p-1$ thresholds $t = [t_1, \ldots, t_{l-1}, t_l, \ldots, t_{p-1}]$. To ensure the minimum width of the range of values for each class, arbitrarily small positive value ε is introduced into the constraints ensuring adequate relations between thresholds: $t_1 \geq \varepsilon$, $t_l - t_{l-1} \geq \varepsilon$, for $l = 2, \ldots, p-1$, and $t_{p-1} + \varepsilon \leq 1$.

We assume the DM provides the desired class assignment $a^* \to C_l$ for each reference alternative $a^* \in A^R$. Function I indicates to which class a^* is assigned:

$$\forall a^* \in A^R,\ a^* \to C_l \iff I(a^*) = l. \tag{5}$$

To construct a model defined above that would be compatible with the DM's indirect preferences, we need to ensure the comprehensive value of reference alternative $a^* \in A^R$ is within the range $[t_{l-1}, t_l)$, corresponding to the desired

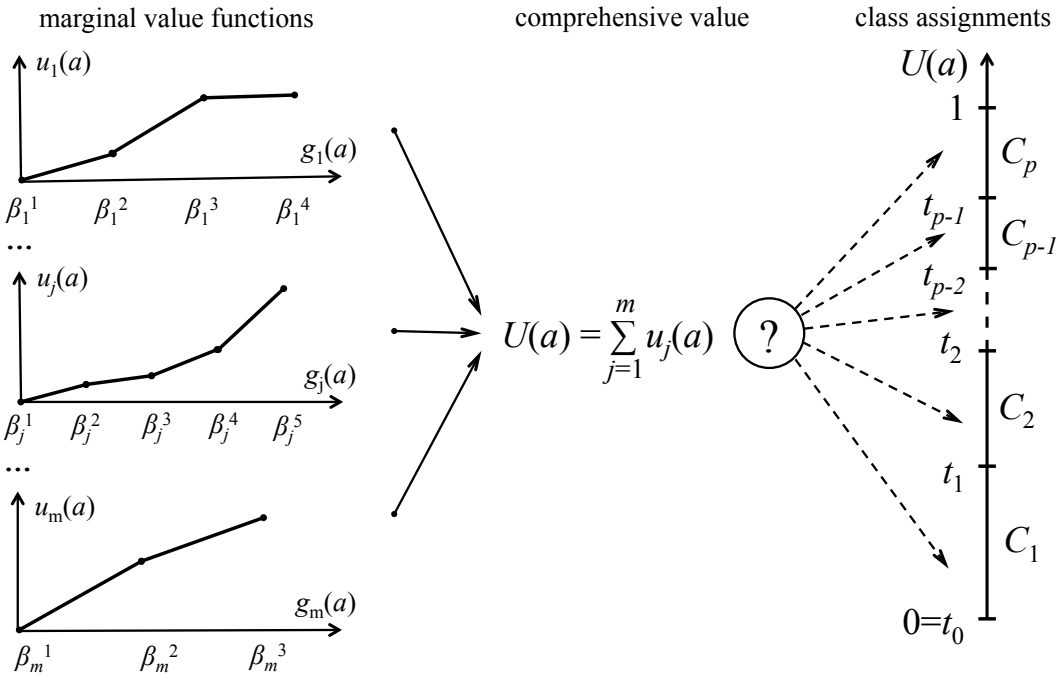


Figure 1: Threshold-based value-driven sorting procedure involving multiple criteria.

class C_l:

$$\forall a^* \in A^R:\ I(a^*) = l \in \{1,\ldots,p-1\} \quad \Longrightarrow \quad t_l - U(a^*) \geq \delta + \varepsilon, \tag{6}$$

$$\forall a^* \in A^R:\ I(a^*) = l \in \{2,\ldots,p\} \quad \Longrightarrow \quad U(a^*) - t_{l-1} \geq \delta, \tag{7}$$

where δ allows for controlling the distance of the alternatives' comprehensive values from the thresholds limiting the class to which they are assigned. Overall, a set of linear constraints E^{A^R}, defining a set $\mathcal{U}^R$ of all compatible AVFs and class thresholds, can be formulated as follows:

$$\left.\begin{array}{l}
\left.\begin{array}{l} u_j(\beta_j^1) = 0, j = 1,\ldots,m, \\ \sum_{j=1}^{m} u_j(\beta_j^{\gamma_j}) = 1, \end{array}\right\}(E^N) \\
\left.\begin{array}{l} u_j(\beta_j^s) - u_j(\beta_j^{s-1}) \geq 0,\ j = 1,\ldots,m, s = 2,\ldots,\gamma_j, \end{array}\right\}(E^M) \\
\left.\begin{array}{l} t_1 \geq \varepsilon, \\ t_l - t_{l-1} \geq \varepsilon,\ l = 2,\ldots,p-1, \\ 1 - t_{p-1} \geq \varepsilon, \end{array}\right\}(E^T) \\
\left.\begin{array}{l} \forall a^* \in A^R:\ I(a^*) = l \in \{1,\ldots,p-1\} \implies t_l - U(a^*) \geq \delta + \varepsilon, \\ \forall a^* \in A^R:\ I(a^*) = l \in \{2,\ldots,p\} \implies U(a^*) - t_{l-1} \geq \delta, \end{array}\right\}(E^{DM})
\end{array}\right\}(E^{A^R}) \tag{8}$$

where ε is an arbitrarily small positive constant and $\delta \geq 0$. Note that the role of constant ε is to transform strict inequalities into their weak counterparts. Moreover, all sorting models for which δ is non-negative are compatible with the DM's preferences as they ensure that for all $a^* \in A^R$ such that $a^* \to C_l$, $U(a^*) \in [t_{l-1}, t_l)$. When E^{A^R} is feasible, $\mathcal{U}^R$ contains at least one and possibly infinitely many instances compatible with the model's assumptions and the DM's preferences. As proven in [19], set $\mathcal{U}^R$ is convex.

To deliver a precise sorting recommendation that is compatible with the supplied assignment examples, one such instance must be selected arbitrarily. This can be conducted in various ways [12, 14, 48]. We decided to choose the *most discriminant* AVF, which can be obtained by maximizing δ, representing the minimum difference between the

comprehensive values of reference alternatives and the corresponding class thresholds:

$$\text{Maximize } \delta, \text{ subject to } E^{A^R}. \tag{9}$$

The variables in the model are as follows:

- $u_j(\beta_j^s),\ j = 1, \ldots, m,\ s = 1, \ldots, \gamma_j$ – value of the marginal function for criterion g_j at characteristic point β_j^s;
- $t_l,\ l = 1, \ldots, p-1$ – thresholds separating intervals of comprehensive values associated with each class;
- δ – the minimum difference between the comprehensive values of reference alternatives and the thresholds associated with the corresponding class.

The objective function defined above is related to increasing confidence in the model's ability to reflect DM's preferences when alternatives are further distant from the respective class thresholds. This way, we capture a solution in which the differences between the comprehensive values of the reference alternatives are as far as possible from the thresholds of the corresponding classes, hence representing DM's preferences in the most robust manner. To ensure comparability of results, unless otherwise explicitly stated, we will use the same objective function in the remaining UTADIS variants.

2.2. Modeling non-monotonic marginal value functions

The basic variant of UTADIS assumes that all criteria are associated with pre-defined preference directions, and hence the pre-criterion preferences are represented with monotonic MVFs. However, this assumption can be relaxed to let the method construct possibly non-monotonic marginal functions. We denoted the variants in this stream as the **UTADIS-NM (NM)** group. They are useful in scenarios where the knowledge of the preference for the performances on each criterion is missing, and needs to be discovered from the DM's indirect preferences.

The first variant, called **UTADIS-NM-1**, adopts the proposal formulated in [16] in the context of ranking problems. It removes the monotonicity constraints, introducing the lower and upper bounds for all characteristic points of MVFs:

$$\left.\begin{array}{l} u_j(\beta_j^k) \geq 0, \\ u_j(\beta_j^k) \leq 1. \end{array}\right\} \forall j \in \{1, \ldots, m\}, \forall k \in \{1, \ldots, \gamma_j\} \left.\right\} (E^{A^R}_{bound}) \tag{10}$$

In this way, we avoid unbounded solution space and allow freedom regarding the shape of MVFs. The normalization of AVF and threshold values is performed after optimization and obtaining a consistent solution. To implement this model, it is necessary to calculate the *slope change* ϕ_j^k of each MVF's segment. It is defined as the difference between the marginal values for two consecutive characteristic points divided by the distance between these points:

$$\left.\begin{array}{l} \frac{u_j(\beta_j^k) - u_j(\beta_j^{k-1})}{\beta_j^k - \beta_j^{k-1}} - \frac{u_j(\beta_j^{k-1}) - u_j(\beta_j^{k-2})}{\beta_j^{k-1} - \beta_j^{k-2}} \leq \phi_j^k, \\ \frac{u_j(\beta_j^{k-1}) - u_j(\beta_j^{k-2})}{\beta_j^{k-1} - \beta_j^{k-2}} - \frac{u_j(\beta_j^k) - u_j(\beta_j^{k-1})}{\beta_j^k - \beta_j^{k-1}} \leq \phi_j^k. \end{array}\right\} \forall j \in \{1, \ldots, m\}, \forall k \in \{3, \ldots, \gamma_j\} \left.\right\} (E^{A^R}_{slope}) \tag{11}$$

The understanding of the notation used when modeling the slope change is supported by the example marginal value function illustrated in Figure 2. A set of linear constraints defining all consistent solutions and the objective function for finding a precise recommendation by NM-1 are formulated as follows:

$$\text{Minimize } \frac{\sum_{j=1}^{m} \sum_{k=3}^{\gamma_j} \phi_j^k}{\delta}, \text{ subject to } \left. E^T, E^{DM}, E^{A^R}_{bound}, E^{A^R}_{slope} \right\} (E^{A^R}_{\text{NM-1}}) \tag{12}$$

This way, we prefer the most discriminant model (maximization of δ in the denominator) for which the sum of slope changes is as small as possible. Such an objective allows some degree of freedom in representing the per-criterion preferences using non-monotonic functions. Therefore, it is less constrained and more flexible than the primary

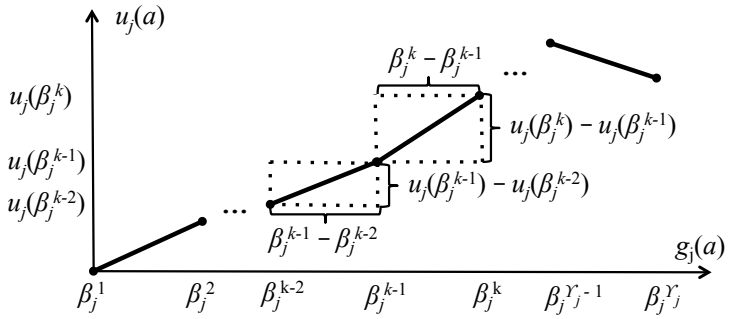


Figure 2: Illustration of the slope change for a marginal value function.

variant of UTADIS. Apart from the same variables used in UTADIS, the above model incorporates the following additional ones:

- $\phi_j^k,\ j = 1, \ldots, m,\ k = 3, \ldots, \gamma_j$ – absolute value of the slope change of the marginal value function u_j for criterion g_j, based on the analysis of neighboring ranges $(\beta_j^{k-2}, \beta_j^{k-1})$ and $(\beta_j^{k-1}, \beta_j^k)$ delimited by three consecutive characteristic points β_j^{k-2}, β_j^{k-1}, and β_j^k.

Minimizing the variations in the slope of MVFs aligns with searching for the parsimonious explanation of indirect judgments, being most likely the correct way [4, 16]. In this case, the simplest additive value model is assumed to incorporate, if possible, monotonic MVFs, which are the most linear ones, i.e., minimizing the deviation from the linearity. The goal is to avoid abrupt changes in MVFs, leading to unrealistic preference models (e.g., zigzag functions changing direction in each characteristic point). Controlling variation in slope has been used before in [11] to determine the minimum number of criteria sub-intervals and in [20] to obtain parsimonious preference models. In the considered method, the complexity of the preference model is optimized against its discriminatory power.

Even if the objective function combining the above aims is non-linear, leading to a Linear-fractional Programming (LFP) problem, it can be easily transformed to the Linear Programming (LP) problem [5]. The above approach cannot be used with linear MVFs that involve only two characteristic points, as in this case, no slope change can be represented. Therefore, for such scenarios, we only maximize δ.

A more complex objective function considered by NM-1 implies potential difficulties in comparing its outcomes with the results produced by other variants of UTADIS. Therefore, we also propose a modification of the concept presented in [16], called UTADIS-NM-2. It aims to infer the most discriminant sorting model, referring to the same variables as in UTADIS:

$$\text{Maximize } \delta, \ \text{ subject to } \quad \left. E^T, E^{DM}, E^{A^R}_{bound} \right\} (E^{A^R}_{\text{NM-2}}) \tag{13}$$

Compared to NM-1, $E^{A^R}_{slope}$ has been omitted because ϕ_j^k is not optimized, and slope changes are not required to define the model. Apart from ϕ_j^k, the remaining parameters in both models are subject to the same constraints. However, due to the different objective functions, both approaches may lead to different solutions and, hence, various recommendations on the set of non-reference alternatives.

To present the process of normalizing the model obtained with NM-1 or NM-2, let us denote it by U'. It consists of MVFs u'_j, class threshold values t'_l, and the minimum difference between the comprehensive values of the reference alternatives and corresponding thresholds δ'. In the first step, the minimum value in each MVF is subtracted from the values assigned to all of its characteristic points:

$$\forall j \in \{1, \ldots, m\} \ \ u''_j(x) = u'_j(x) - \min_{k \in \{1, \ldots, \gamma_j\}} u'_j(\beta_j^k). \tag{14}$$

Then, the minimum value of each modified MVF equals zero. Decreasing the value of each MVF reduces the comprehensive value of each alternative by exactly the same value. Thus, to keep the solution feasible, it is

necessary to reduce the threshold values in the same way, i.e., by subtracting the sum of the minimum values of each u'_j:

$$\forall l \in \{1, \ldots, p-1\} \;\; t''_l = t'_l - \sum_{j=1}^{m} \min_{k \in \{1,\ldots,\gamma_j\}} u'_j(\beta_j^k). \tag{15}$$

The value of δ does not change because the minimum distance of the thresholds and comprehensive values of the reference alternatives remains unchanged, i.e., $\delta'' = \delta'$.

The second step focuses on normalizing the values so that the maximum comprehensive value equals one. For this purpose, all MVFs are divided by the sum of their maximum values, denoted as ρ. Then, we proceed analogously with the values of thresholds and δ:

$$\rho = \sum_{j=1}^{m} \max_{k \in \{1,\ldots,\gamma_j\}} u''_j(\beta_j^k), \tag{16}$$

$$\forall j \in \{1, \ldots, m\} \quad u_j(x) = \tfrac{u''_j(x)}{\rho}, \quad \forall l \in \{1, \ldots, p-1\}\; t_l = \tfrac{t''_l}{\rho}, \quad \delta = \tfrac{\delta''}{\rho}. \tag{17}$$

The latter is needed because reducing the comprehensive values of reference alternatives and thresholds will also reduce the minimum distance between them. After the above transformations, the model is normalized to the $[0, 1]$ range, and hence δ has the same interpretation as for the primary variant of UTADIS.

We propose another variant admitting non-monotonicity, called UTADIS-NM-3, which is inspired by the ideas presented in [30]. It represents preferences on each potentially non-monotonic criterion g_j using a sum of two monotonic functions, one non-decreasing u_j^{ND} and another non-increasing u_j^{NI}, each adhering to standard weak monotonicity constraints:

$$\left.\begin{array}{l}
\left.\begin{array}{l} u_j^{\mathrm{ND}}(\beta_j^1) = 0,\; u_j^{\mathrm{ND}}(\beta_j^{\gamma_j}) \le 1,\; j = 1, \ldots, m, \\ u_j^{\mathrm{NI}}(\beta_j^{\gamma_j}) = 0,\; u_j^{\mathrm{NI}}(\beta_j^1) \le 1,\; j = 1, \ldots, m, \end{array}\right\} (E^B_{\text{NM-3}}) \\
\left.\begin{array}{l} u_j^{ND}(\beta_j^s) - u_j^{\mathrm{ND}}(\beta_j^{s-1}) \ge 0,\; j = 1, \ldots, m, s = 2, \ldots, \gamma_j, \\ u_j^{NI}(\beta_j^s) - u_j^{\mathrm{NI}}(\beta_j^{s-1}) \le 0,\; j = 1, \ldots, m, s = 2, \ldots, \gamma_j, \end{array}\right\} (E^M_{\text{NM-3}}) \\
\left.\begin{array}{l} u_j(\beta_j^s) = u_j^{ND}(\beta_j^s) + u_j^{NI}(\beta_j^s),\; j = 1, \ldots, m, s = 1, \ldots, \gamma_j, \\ 0 \le u_j(\beta_j^s) \le 1,\; j = 1, \ldots, m, s = 1, \ldots, \gamma_j. \end{array}\right\} (E^{\text{MVF}}_{\text{NM-3}})
\end{array}\right\} (E_{\text{NM-3}}) \tag{18}$$

The composition of a pair of monotonic functions with opposing preference directions results in the potentially non-monotonic MVF as shown by the example marginal value functions in Figure 3. Again, we select the most discriminant model as the representative solution:

$$\text{Maximize } \delta, \text{ subject to } \; E^T, E^{DM}, E_{\text{NM-3}} \;\Big\} (E^{A^R}_{\text{NM-3}}) \tag{19}$$

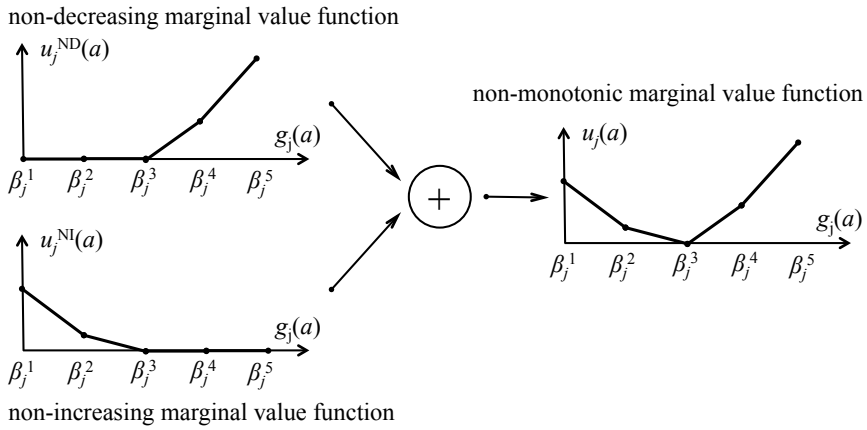


Figure 3: Example non-decreasing and non-increasing components resulting in a non-monotonic marginal value function.

Since the inferred parameter values may be outside the range between 0 and 1, a normalization process needs to be conducted, similarly as for the NM-1 and NM-2 procedures (see Eqs. (14)–(17)). Apart from the same variables used in UTADIS, UTADIS-NM-3 incorporates the following additional ones:

- $u_j^{\mathrm{ND}}(\beta_j^s),\ j=1,\ldots,m,\ s=1,\ldots,\gamma_j$ – the value of the non-decreasing component of the marginal function u_j for criterion g_j at characteristic point β_j^s,
- $u_j^{\mathrm{NI}}(\beta_j^s),\ j=1,\ldots,m,\ s=1,\ldots,\gamma_j$ – the value of the non-increasing component of the marginal function u_j for criterion g_j at characteristic point β_j^s.

2.3. Modeling interactions between criteria

Using AVF and, thus, the basic variant of UTADIS requires the fulfillment of preferential independence. It means that the DM's preferences over any subset of attributes are independent of its complement. However, in some decision scenarios, one needs to represent interactions between criteria and reflect a non-additive nature of preferences. This option has been successfully implemented in [21] in the context of ranking problems. We adapt it to the multiple criteria sorting, giving rise to the **UTADIS-INT** methods.

In this approach, the positive and negative interactions between all criteria pairs: $g_q, g_r \in G : q < r$ are modeled using bonuses $\mathrm{syn}_{q,r}^{+}$ and penalties $\mathrm{syn}_{q,r}^{-}$. Then, the comprehensive value of alternative $a \in A$ is expressed as follows:

$$U(a) = \sum_{j=1}^{m} u_j(g_j(a)) + \sum_{q=1}^{m-1}\sum_{r=q+1}^{m} \mathrm{syn}_{q,r}^{+}(g_q(a), g_r(a)) - \sum_{q=1}^{m-1}\sum_{r=q+1}^{m} \mathrm{syn}_{q,r}^{-}(g_q(a), g_r(a)). \tag{20}$$

Functions syn^{+} and syn^{-} need to satisfy the following normalization (E_{INT}^{N}) and monotonicity (E_{INT}^{M}) conditions:

$$\left.\begin{array}{l}
\left.\begin{array}{l}
\mathrm{syn}_{q,r}^{+}(\beta_q^1, \beta_r^1) = 0,\ \forall q, r \in \{1,\ldots,m\} : q < r, \\
\mathrm{syn}_{q,r}^{-}(\beta_q^1, \beta_r^1) = 0,\ \forall q, r \in \{1,\ldots,m\} : q < r, \\
\mathrm{syn}_{q,r}^{+}(\beta_q^{\gamma_q}, \beta_r^{\gamma_r}) \le \mu\lambda_{q,r},\ \forall q, r \in \{1,\ldots,m\} : q < r, \\
\mathrm{syn}_{q,r}^{-}(\beta_q^{\gamma_q}, \beta_r^{\gamma_r}) \le \mu\lambda_{q,r},\ \forall q, r \in \{1,\ldots,m\} : q < r, \\
\tau = \sum\limits_{j=1}^{m} u_j(\beta_j^{\gamma_j}), \\
\tau^{+} = \sum\limits_{q=1}^{m-1}\sum\limits_{r=q+1}^{m} \mathrm{syn}_{q,r}^{+}(\beta_q^{\gamma_q}, \beta_r^{\gamma_r}), \\
\tau^{-} = \sum\limits_{q=1}^{m-1}\sum\limits_{r=q+1}^{m} \mathrm{syn}_{q,r}^{-}(\beta_q^{\gamma_q}, \beta_r^{\gamma_r}), \\
\tau + \tau^{+} - \tau^{-} = 1, \\
\sum\limits_{r=1}^{q-1} \lambda_{r,q} + \sum\limits_{r=q+1}^{m} \lambda_{q,r} \le \sigma,\ \forall q \in \{1,\ldots,m\}, \\
\lambda_{q,r} \in \{0,1\},\ \forall q, r \in \{1,\ldots,m\} : q < r,
\end{array}\right\} (E_{\mathrm{INT}}^{N}) \\
\forall q, r \in \{1,\ldots,m\} : q < r \text{ and } \forall s, u \in \{1,\ldots,\gamma_q\} \ :\ s \ge u \text{ and } \forall t, v \in \{1,\ldots,\gamma_r\} \ :\ t \ge v : \\
\left.\begin{array}{l}
\mathrm{syn}_{q,r}^{+}(\beta_q^s, \beta_r^t) \ge \mathrm{syn}_{q,r}^{+}(\beta_q^u, \beta_r^v), \\
\mathrm{syn}_{q,r}^{-}(\beta_q^s, \beta_r^t) \ge \mathrm{syn}_{q,r}^{-}(\beta_q^u, \beta_r^v), \\
u_q(\beta_q^s) + u_r(\beta_r^t) + (\mathrm{syn}_{q,r}^{+}(\beta_q^s, \beta_r^t) - \mathrm{syn}_{q,r}^{-}(\beta_q^s, \beta_r^t)) \\
\quad \ge u_q(\beta_q^u) + u_r(\beta_r^v) + (\mathrm{syn}_{q,r}^{+}(\beta_q^u, \beta_r^v) - \mathrm{syn}_{q,r}^{-}(\beta_q^u, \beta_r^v)).
\end{array}\right\} (E_{\mathrm{INT}}^{M})
\end{array}\right\} (E_{\mathrm{INT}}) \tag{21}$$

Hence, both syn^{+} and syn^{-} attain zero when parameterized with the least preferred performances on the two criteria, whereas their maximal value is constrained by constant μ. Following [21], we use $\mu = 1$. Binary variable $\lambda_{q,r}$ indicates if an interaction is active for a given pair of criteria, and σ is the maximum number of active interactions for each criterion. A comprehensive additive value function enriched with bonuses and penalties cannot take values greater than one. Further, functions syn^{+} and syn^{-} are monotonic in both of their arguments (i.e., performances on the respective criteria). That is, the bonuses for positively interacting criteria or penalties for negatively interacting criteria cannot decrease with the increase of any performance. Moreover, the interaction coefficients cannot change

the relation between marginal values corresponding to more preferred performances on any pair of attributes. Note that syn^+ and syn^- are defined only for pairs of characteristic points. Since this is a two-dimensional function, it is necessary to use bilinear interpolation to determine the value of these functions for any pair of performances. The Mixed-Integer Linear Programming (MILP) problem that needs to be solved to select a representative model instance can be formulated as follows:

$$\text{Maximize } \delta - \sum_{\forall q,r \in \{1,\ldots,m\}: q<r} \lambda_{q,r}, \quad \text{subject to} \quad \left. E^{A^R}, E_{\text{INT}} \right\} (E^{A^R}_{\text{INT}}) \tag{22}$$

The objective function formulated above represents a lexicographic objective, in which the primary aim is to minimize the number of active interactions between all criteria pairs, and the secondary aim is to maximize the value of δ. The former is consistent with the original postulate formulated in [21] to explain the DM's assignment example using a compatible value function with as few interactions as possible. We intend to use interactions in a parsimonious way. In particular, when the DM's preference statements can be represented just by a simple additive value function, then no interaction is considered. Further, if just a limited number of interaction components is needed to fit the supplied preference information, in line with Ockham's razor principle, entities should not be multiplied beyond necessity, i.e., above a minimum required number of interacting pairs of criteria. Therefore, similarly to NM-1, this approach infers a model that is as simple as possible and capable of correctly representing the DM's policy. As the secondary target, **UTADIS-INT** maximizes δ, hence searching for the most discriminant solution among all compatible models with the lowest possible number of active interactions. Let us emphasize that by *active interactions*, we mean the interacting (either positively or negatively) criteria pairs.
Overall, UTADIS-INT uses the same variables as in UTADIS plus the following ones:

- $\text{syn}^+_{q,r}(\beta^s_q, \beta^t_r),\ q,r \in \{1,\ldots,m\} : q < r,\ s = 1,\ldots,\gamma_q,\ t = 1\ldots,\gamma_r$ – the value of the positive interaction function between criteria g_q and g_r, for a pair of performances β^s_q, β^t_r,
- $\text{syn}^-_{q,r}(\beta^s_q, \beta^t_r),\ q,r \in \{1,\ldots,m\} : q < r,\ s = 1,\ldots,\gamma_q,\ t = 1\ldots,\gamma_r$ – the value of the negative interaction function between criteria g_q and g_r, for a pair of performances β^s_q, β^t_r,
- $\lambda_{q,r},\ q,r \in \{1,\ldots,m\} : q < r$ – a binary value indicating the activation of the positive or negative interaction function for criteria pair g_q and g_r.

Additionally, for clarity, symbols τ, τ^+, and τ^- are introduced to represent the sums of the individual components in the normalization constraints. The remaining symbols (μ, ε, and σ) have pre-defined, constant values, so they are not considered decision variables.

In what follows, we will consider two variants of **UTADIS-INT**, differing in terms of the maximal number of criteria with which each attribute can interact. Specifically, **UTADIS-INT-1** follows the assumption made in [21], letting each criterion interact with at most one other attribute ($\sigma = 1$). This ensures significant interpretability of the inferred model. In turn, **UTADIS-INT-∞** postulates that the number of interacting criteria pairs is unlimited ($\sigma = \infty$). Effectively, this means that each criterion can interact with at most $m - 1$ other criteria.

Another popular model handling the interactions between criteria is the Choquet integral [17]. The advantages of the considered value-based tool when compared with the non-additive integral derive from a) not requiring the evaluations on all criteria to be expressed on the same scale to ensure full commensurability [3], b) the ability to represent adequately more advanced interactions between couples of criteria that the Choquet integral cannot handle [21], c) generalizing the 2-additive Choquet integral, which is a particular case of the applied model [21], d) offering clear justification of the recommended decision (e.g., preference or assignment) in terms of values of model parameters that are more interpretable than non-additive weights (capacity) [21], e) more significant potential for increased predictive accuracy as proven by the extensive computational experiments in [35].

2.4. Robustness analysis

The primary issue related to the practical use of incomplete preferences in UTADIS derives from multiple or even infinitely many instances of the sorting model compatible with the DM's indirect statements. The presentation of all variants of UTADIS involved the selection of the most discriminant model among them. This was attained by maximizing δ, possibly coupled with other objectives, ensuring the parsimony of the selected model, e.g., in terms of the shape of MVFs or the number of interacting pairs of criteria. Such a selection of a single representative preference model instance consistutes an important stream in ordinal regression (see Figure 4).

However, a single δ-maximizer model is not necessarily the only possible solution compatible with the DM's classification examples. Specifically, all models with δ greater or equal to zero are feasible, being consistent with the DM's decision policy represented by the assignments of reference alternatives. The application of such models on the set of non-reference alternatives potentially leads to ambiguous recommendations. In this perspective, it may be relevant to conduct robustness analysis. As noted in [14], even when the compatible solution is unique or the feasible space of models is empty due to inconsistencies or restrictive assumptions imposed by the model, the robustness concern remains relevant, the same as in the highly noisy context of statistical learning theory.

Let us emphasize that the notion of *robustness* still remains vague in the entire Operations Research and Decision Analysis domain [41]. Multiple meanings accorded to the term robust include flexibility, stability, sensitivity, and even equity. Following the directions indicated by Roy [41], our treatment of robustness is closely tied to a capacity for withstanding "zones of ignorance" arising from the disparity between the model and real-life decision context. To this aim, we account for various sensible versions of the problem formulation. Each version represents a reality that should be considered, reflecting a combination of the options related to the model's frailty points. Specifically, we are interested in investigating the robustness of the provided conclusions, i.e., whether they are valid for all or for the most plausible sets of compatible preference model instances.

Given their multiplicity in the context of the UTADIS variants, it is relevant to verify the robustness of sorting results [12]. The need for carefully exploiting the set of multiple compatible models was first emphasized in [26]. The suggested approach was based on a heuristic post-optimality procedure seeking to identify some characteristic alternative models corresponding to corner points of the feasible polyhedron. However, such techniques provide only a limited view of the complete set of models compatible with the DM's preferences. Therefore, the two prevailing streams for robustness analysis in this context are Robust Ordinal Regression (ROR) and Stochastic Ordinal Regression (SOR) (see Figure 4).

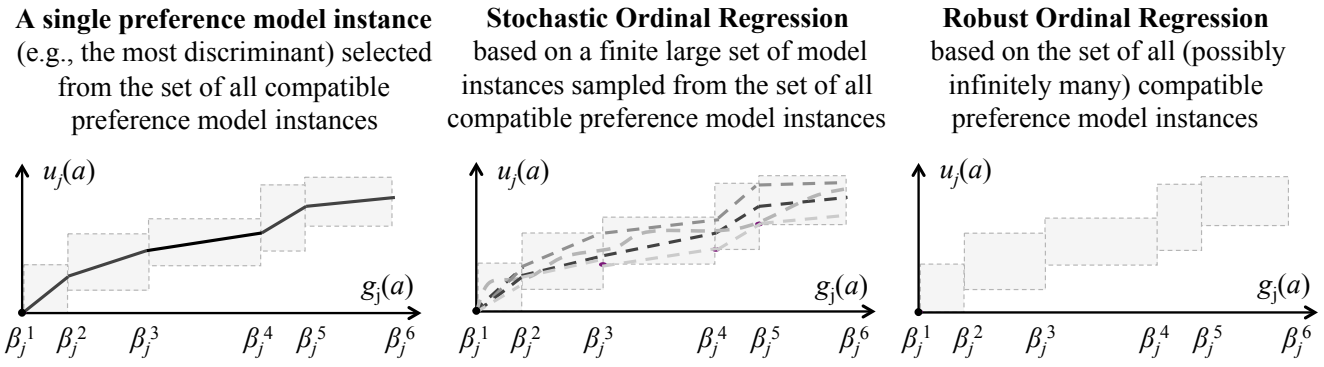


Figure 4: Three main methodological streams in ordinal regression.

ROR computes the possible and necessary assignments for each alternative based on the analysis of all compatible models [19, 34]. The former reflects the classifications attained for at least one feasible model $U \in \mathcal{U}^R$ [19], i.e.:

$$\forall a \in A, \forall l \in \{1, \ldots, p\} : \; a \rightarrow^P C_l \iff \exists U \in \mathcal{U}^R : a \rightarrow_U C_l, \tag{23}$$

where $a \rightarrow_U C_l$ denotes that model U assigns alternative a to class C_l, i.e., $I_U(a) = l$. A set of all possible

assignments for $a \in A$ is denoted by:

$$PCA_{\mathcal{U}^R}(a) = \{C_l : a \to^P C_l\}. \tag{24}$$

In turn, the necessary assignment needs to be confirmed by all feasible models $U \in \mathcal{U}^R$ [19], i.e.:

$$\forall a \in A, \forall l \in \{1, \ldots, p\} : \; a \to^N C_l \iff \forall U \in \mathcal{U}^R : a \to_U C_l. \tag{25}$$

In particular, when $\mathcal{U}^R \neq \emptyset$, all reference alternatives $a^* \in A^R$ are necessarily assigned to the classes specified by the DM.

Another approach, called SOR, provides quantitative information, estimating how often a given assignment occurs in the set of all compatible model instances [31]. To exploit $\mathcal{U}^R$, SOR uses the Monte Carlo simulation to sample a sufficiently large subset of uniformly distributed sorting models $S \subseteq \mathcal{U}^R$ ($|S| \ll |\mathcal{U}^R|$). For this purpose, we employ the Hit-And-Run (HAR) algorithm [45] implemented in [7]. Then, Class Acceptability Index (CAI) quantifies the share of all compatible model instances assigning each alternative to a given class. Its approximation (CAI') based on the simulation results is computed as follows [31]:

$$\forall a \in A \; \forall l \in \{1, \ldots, p\} : \; CAI'(a, C_l) = \frac{|\{U \in S : I_U(a) = l\}|}{|S|}. \tag{26}$$

CAI' takes values in the range between 0 and 1. However, it can also be interpreted as the percentage of the feasible models, from 0% to 100%. Let us emphasize that the ideas underlying ROR and SOR can be adjusted to all variants of UTADIS, as long as the respective space of feasible models is defined using the linear constraints.

3. Experimental analysis

This section is devoted to computational experiments verifying the characteristics of the six variants of UTADIS. First, we define quality measures considered in the comparative analysis. Second, we describe how the experiment was conducted. Third, we discuss the results given the expressiveness and robustness dimensions.

3.1. Quality measures

This section defines quality measures that capture the expressiveness of the preference model and the robustness of the recommendation delivered by a specific method. They are adjusted to the scope of multiple criteria sorting preference disaggregation approaches. The examples supporting the understanding of all measures are provided in Section 5, devoted to a case study.

3.1.1. Expressiveness of the preference model

The model's expressiveness reflects its ability to reconstruct the DM's indirectly expressed preferences. We use the following two measures to compare different approaches in this regard.

Preference recoverability (**PR**) is measured as a ratio of scenarios for which DM's assignment examples are entirely consistent with an assumed preference model. Then, at least one compatible value function and a set of class thresholds exist with $\delta^* \geq 0$ [29]. Higher PR values indicate the method's ability to infer model parameters consistent with a broader spectrum of sorting policies. This emphasizes greater flexibility of the model to fit the DM's indirect preferences.

Maximum delta ($\boldsymbol{\delta^*}$) indicates the minimum difference between the comprehensive value of the reference alternatives and the lower and upper thresholds of the class to which it was assigned by the DM [29]. It reflects the ability of the threshold-based value-driven procedure to discriminate between alternatives with various desired classifications. The higher the value of δ^*, the greater the model's ability in robustly reproducing the DM's preferences. For each approach, δ^* is derived from the optimal problem solution that implicitly or explicitly maximizes its value in the objective function. Normalizing each method's results to the range between 0 and 1 ensures the comparability of this parameter's values between different approaches.

3.1.2. Robustness of the sorting recommendation

Robustness captures the stability and credibility level of the recommendation suggested by a particular method, given the multiplicity of model instances that can reproduce the DM's assignment examples. In this aspect, we distinguish five measures based on the analysis of results for non-reference (holdout) alternatives to prevent overfitting concerns.

Average possible class assignment (**APCA**) is based on the average number of classes to which it is possible to assign each non-reference alternative from $A^T = A \setminus A^R$ [27]. To make the results for various problem sizes comparable, this measure is normalized as follows:

$$APCA(\mathcal{U}^{\mathcal{R}}) = 1 - \frac{1}{|A^T|} \sum_{a \in A^T} \frac{|PCA_{\mathcal{U}^{\mathcal{R}}}(a)| - 1}{p - 1}, \tag{27}$$

where p is the number of classes and $PCA_{\mathcal{U}^{\mathcal{R}}}$ is defined in Eq. (24). The measure reaches a maximum value of 1 when all non-reference alternatives have non-empty and precise necessary assignments. In this case, the model's recommendations are unambiguous and most reliable. A minimum value of 0 indicates that each $a \in A^T$ can be assigned to any class. This means the variability of recommendations obtained in the set of compatible sorting model instances is enormous.

Certain assignments ratio (**CAR**) reflects the share of non-reference alternatives assigned precisely to some class by all compatible sorting model instances [27]:

$$CAR(\mathcal{U}^{\mathcal{R}}) = \frac{|\{a \in A^T : |PCA_{\mathcal{U}^{\mathcal{R}}}(a)| = 1\}|}{|A^T|}. \tag{28}$$

The maximum value of 1 indicates the complete model's confidence regarding the assignments of all non-reference alternatives. In turn, 0 denotes a hesitation in the recommended classification for all alternatives.

Entropy class acceptability index (**ECAI**) is based on Shannon's concept of entropy [43]. It is calculated based on CAIs for each $a \in A^T$ [27]:

$$ECAI_{alt}(a) = -\sum_{l=1}^{p} CAI'(a, C_l) \log_2 CAI'(a, C_l). \tag{29}$$

Note that $ECAI_{alt}(a) = 0$ if and only if there is class C_l such that $CAI'(a, C_l) = 1$, indicating the agreement in the suggested recommendation for all sorting model instances. Conversely, the maximum possible value of $ECAI_{alt}(a)$ is $\log_2(p)$. It is obtained if $CAI'(a, C_l) = \frac{1}{p}$ for each $l \in \{1, \ldots, p\}$, suggesting the same support given to all classes in the set of all compatible model instances. To aggregate the outcomes for all non-reference alternatives and normalize the measure to the $[0, 1]$ interval, we define **ECAI** as follows:

$$ECAI(\mathcal{U}^{\mathcal{R}}) = 1 - \frac{1}{\log_2(p) \cdot |A^T|} \sum_{a \in A^T} ECAI_{alt}(a). \tag{30}$$

Mean class acceptability index (**MCAI**) is the average value of CAIs for all non-reference alternatives and the classes they were univocally assigned to by a given method [14]:

$$MCAI(U) = \frac{1}{|A^T|} \sum_{a \in A^T : a \to_U C_l} CAI'(a, C_l). \tag{31}$$

A higher value of $MCAI$ indicates greater support given to the method's recommendation by all feasible model instances. This measure captures how representative is the instance selected by a given approach for the entire space of instances when considering the variety of assignments observed in this space.

Confirmed class assignment (**CCA**) builds on the ambiguity in setting the thresholds separating the classes by a specific method. These thresholds are determined to encompass the comprehensive value of reference alternatives

in a given class. However, this still leaves some freedom in placing them between the highest value of some reference alternative in a less preferred class and the lowest value of some reference alternative in a more preferred class. Non-reference alternatives with comprehensive values in the above range would change their assignments if the thresholds were set differently, even for the same additive value model. Hence, we want to quantify the share of non-reference alternatives with *confirmed assignments*, i.e., implied by scores between the extreme values associated with reference alternatives in the assigned class. For this, we first determine the *class boundaries*, i.e., the lowest (C_l^{LB}) and highest (C_l^{UB}) comprehensive values for reference alternatives in each class:

$$C_l^{LB} = \min_{a \in A^R:\, a \to C_l} U(a) \quad \text{and} \quad C_l^{UB} = \max_{a \in A^R:\, a \to C_l} U(a). \tag{32}$$

Note that we set C_1^{LB} to 0 and C_p^{UB} to 1 because the values of these extreme thresholds do not influence the assignments of non-reference alternatives. Then, we compute the share of non-reference alternatives with comprehensive values that guarantee their assignments would not change if the class thresholds were set differently:

$$CCA(U) = \frac{|a \in A^T :\, a \to_U C_l \wedge C_l^{LB} \leq U(a) \leq C_l^{UB}|}{|A^T|}. \tag{33}$$

If the measure reaches a maximum value of 1, then the assignments of all non-reference alternatives are *confirmed* by reference ones. If CCA is equal to 0, the classification of each non-reference alternative could change if the thresholds were set differently in the admissible range.

Overall, the first two measures describing the expressiveness of the preference models focus on their abilities to reproduce the DM's preferences (PR) and to highlight differences between comprehensive values of reference alternatives from different classes (δ^*). The former aggregates the binary indication of the complete consistency between the preference information and the model's assumptions from all considered scenarios, whereas the latter captures the quantitative information expressed on the conjoint interval scale only for settings for which all DM's assignment examples can be reproduced. The following five measures focus on the recommendations' robustness. Some build on the results of exact ($APCA$ and CAR) or ($ECAI$) robustness analysis as conducted in ROR and SOR, respectively, capturing the level of compliance between sorting results given the entire space of consistent sorting models. However, they represent complementary perspectives, reflecting if some outcomes are ever possible ($APCA$ and CAR), to what degree they are possible ($ECAI$), if they are precise ($APCA$), and how imprecise they are (CAR and $ECAI$). Other measures focus on the certainty of precise assignments recommended for non-reference alternatives by some method (in our case, the most discriminant preference model instance) given other feasible sorting results. This aspect is reflected by the support offered to these classifications in the set of all compatible models ($MCAI$) and the share of alternatives for which the assignments would not change if other, though still compatible, thresholds were set (CCA). This way, we verify the robustness of the recommendation delivered by the specific method while changing the parameter values within the feasible space and assuming that the performances, preferences, and model assumptions (e.g., characteristic points) are fixed. Overall, our robustness metrics build on three important methodological streams in ordinal regression, i.e., selection of a representative model instance, ROR, and SOR.

Such a broad spectrum of measures aims to provide different interpretations of expressiveness and robustness, referring to various scenarios, information scales, and types of results. It also increases the reliability of subsequent experimental analysis, which is not biased by an arbitrary selection of one interpretation. The fact that the results for measures representing even the same dimension do not need to align helps address the nuances of applying various methods and models and formulating more robust conclusions.

3.2. Simulation design

In the experimental comparison, we considered decision problems with the following dimensions:

- the number of classes – $p \in \{2, 3, 4, 5\}$;

- the number of criteria – $m \in \{2, 3, 4, 5\}$;
- the number of characteristic points for the marginal value function u_j on each criterion g_j – $\gamma_j \in \{2, 3, 4, 5\}$;
- the number of reference alternatives assigned by the DM to each of p classes – $r \in \{1, 2, 3, 4, 5\}$;
- the algorithm used to generate performances of a non-dominated set of alternatives – $c \in \{sphere, random\}$.

In this way, we covered problems with different complexities, starting from simple problems with binary classification and two conflicting criteria and ending up with more complex ones involving five classes and attributes. We also allowed various flexibility of MVFs, from linear to piecewise linear ones with five characteristic points. To check how well the methods and models cope with different amounts of DM's preference information, we assumed various numbers of assignment examples for each class.

To generate a non-dominated set of alternatives, we employed two procedures. The one called ***sphere*** randomly selects points with all non-negative coordinates from the unit m-sphere and then assigns its values as some alternative's performances. This procedure also makes it easier to notice preference dependencies. The other procedure, called ***random***, generates the performances for each alternative independently form a uniform distribution in the range $[0, 1)$. In this case, the newly generated alternative is added to set A if and only if it neither dominates nor is dominated by any member of A. This process continues until the pre-defined size of A is reached. Figure 5 shows example alternatives' performances generated by both procedures for a bi-criteria problem. Using the two approaches allows for considering problems with more diverse characteristics. Note that the lack of dominance in set A makes solving these problems more challenging.

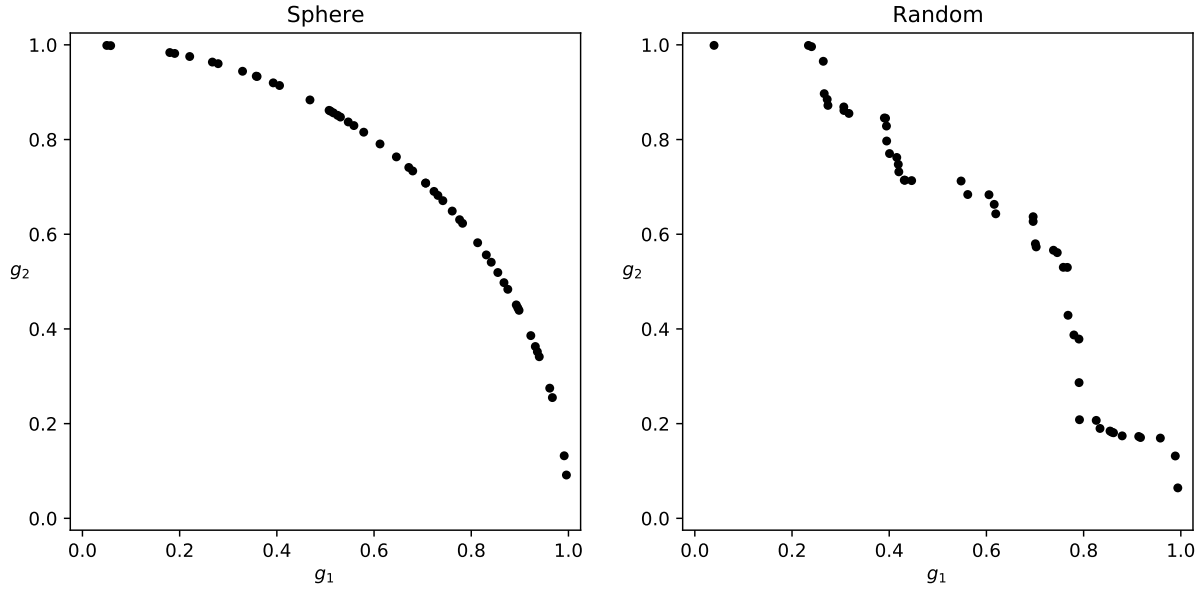


Figure 5: Example performances of fifty alternatives generated by the *sphere* and *random* procedures for a bi-criteria problem.

Moreover, for each considered problem, we randomly generate $|A^T| = 25$ non-reference alternatives, i.e., hold-outs. This way, we mimic realistic scenarios where the DM typically provides preference information for less than half of the considered alternatives ($|A^R| \leq |A^T|$). The results obtained for non-reference alternatives are used to compute the quality measures, so their large number gives greater credibility to the obtained values. For each combination of problem dimensions, we generated 100 instances. Overall, this gives $4 \cdot 4 \cdot 4 \cdot 5 \cdot 2 \cdot 100 = 64,000$ repetitions. Such a large number of analyzed problems allows for concluding the characteristics, similarities, and differences between the considered variants of UTADIS. To generate each instance, we followed four steps:

- Using procedure $c \in \{sphere, random\}$, a set A of $n = |A^R| + |A^T|$ (where $|A^R| = p \cdot r$) non-dominated alternatives was generated, each with m performances from the range between 0 and 1.

- A subset of $|A^R|$ alternatives was randomly selected from A to create a reference set.
- To generate reference assignments and simulate the DM's decision policy, alternatives from A^R were randomly distributed into p subsets $A^R_{C_1}, A^R_{C_2}, \ldots, A^R_{C_p}$, each containing r alternatives. Each alternative in $A^R_{C_l}$, $l = 1, \ldots, p$, was assigned to class C_l. Please note that the procedure for generating the performances of alternatives guarantees that the considered assignments are based on the weakest possible assumption, i.e., they do not violate the dominance relation.
- For each criterion g_j, the values of γ_j characteristic points were determined so that $\beta_j^1 = 0$, $\beta_j^{\gamma_j} = 1$, and the remaining points were set at equal intervals between them.

3.3. Results

This section describes the obtained results. We compare different variants of UTADIS in terms of expressiveness and robustness, and discuss the impact of various problem dimensions. For each setting, the statistical significance of the observed differences was verified using the Wilcoxon signed-rank test [24] for paired samples with a p-value of 0.05.

3.3.1. Preference recoverability

Reproducibility of DM's preferences is one of the most intuitive factors reflecting the expressiveness of a model. The higher the ratio of scenarios with fully reproduced sets of assignment examples, the richer the variety of problems and respective decision policies a given procedure applies to. Tables 1 and 2 show the PR values, both comprehensive and broken down into analyzed problem dimensions. Figure 6 reveals the proportion of problems that different subsets of UTADIS variants have solved. Since all three *non-monotonic* approaches solved the same subset of problems, their results are grouped under the NM name. In addition, when both approaches modeling interactions between criteria could find a consistent solution for the same problem instances, they are labeled as the INT group in Figure 6.

Table 1: Preference recoverability ratio for all problem instances and sub-groups with different numbers of classes and criteria.

PR		Number of classes				Number of criteria			
Procedure	**All settings**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
UTADIS	0.485	0.784	0.536	0.366	0.254	0.300	0.465	0.556	0.619
NM	0.621	0.899	0.693	0.515	0.376	0.407	0.596	0.704	0.775
INT-1	0.597	0.851	0.641	0.500	0.395	0.396	0.588	0.679	0.725
INT-∞	0.757	0.898	0.780	0.703	0.647	0.396	0.819	0.885	0.928

Table 2: Preference recoverability ratio for sub-groups of problem instances with different numbers of characteristic points, reference assignments per class, and performance generation algorithms.

PR	Number of ch. points				Number of reference assignments					Generation algorithm	
Procedure	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**sphere**	**random**
UTADIS	0.271	0.450	0.568	0.651	0.898	0.620	0.420	0.283	0.203	0.565	0.404
NM	0.361	0.586	0.723	0.812	0.969	0.777	0.590	0.440	0.326	0.616	0.625
INT-1	0.349	0.592	0.690	0.757	0.941	0.727	0.544	0.430	0.342	0.715	0.479
INT-∞	0.556	0.794	0.825	0.852	0.966	0.839	0.730	0.653	0.597	0.798	0.716

Let us focus first on the comprehensive results. More than 75% of all problem instances were solved by INT-∞. This model offers the greatest flexibility among all considered variants, and its advantage in preference recoverability is significant. The NM group delivered a solution for 62.1% of simulated scenarios and INT-1 – for 59.7%. UTADIS performed the worst, solving less than half of the considered problems (48.5%). This is due to more restrictive assumptions of the primary model. As confirmed by Figure 6, all problem instances solved by UTADIS were also successfully solved by the remaining approaches. Adding the possibility of representing non-monotonic per-criteria preferences or at least one interaction for each criterion leads to a consistent solution for an additional several percent of instances. Note that when comparing the performance of NM and INT-∞ approaches, NM was able to

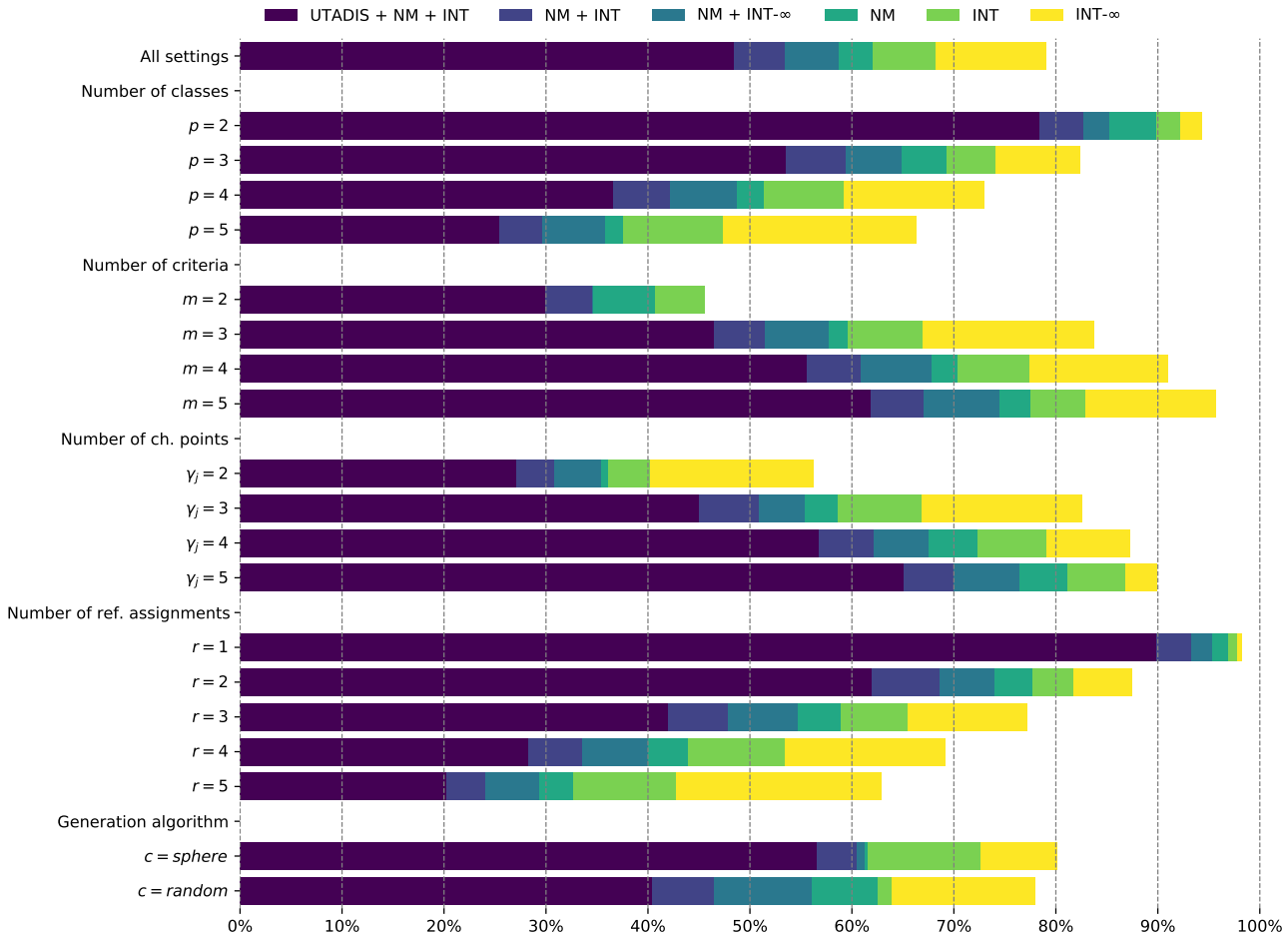


Figure 6: Preference recoverability ratio attained by different subsets of methods for all problem instances and sub-groups with various characteristics of the considered problems.

solve around 3.4% of instances that INT-∞ could not handle (see Figure 6). In comparison, the inverse observation holds for about 19% of considered problems. No methodological variant was capable of reproducing all assignment examples in about 21% simulated scenarios.

Let us now pass to considering the impact of various problem dimensions. Table 1 shows that increasing the number of classes p decreased the fraction of problems that could be handled by each method. The most significant deterioration is visible for UTADIS, which was able to solve over 78% of 2-class problems and about 25% of 5-class problems. The performance of INT-∞ was the least sensitive to changing p – from nearly 90% for binary classification problems to less than 65% for problems with five classes. Despite the advantage of NM over INT-1 for 2-, 3- and 4-class problems, the latter could reproduce almost 2% more 5-class problems.

The inverse trend can be observed when changing the number of criteria m. There is a 2-fold increase in recoverable problem instances between two and five criteria. Note that when $m = 2$, it is possible to introduce only one interaction. Then, the results of INT-1 and INT-∞ were the same and their adaptability to DM's indirect preferences was limited (PR = 39.6%). In turn, the NM approaches were marginally better, providing a consistent solution for 40.7% instances. For a greater number of attributes, the advantage of the NM group over INT-1 is still visible. However, INT-∞ outperforms all remaining variants of UTADIS, taking advantage of the greater flexibility in representing interactions between all criteria pairs.

The change in the shares of problem instances handled by different methods with the increase in the number of characteristic points is best visible in Figure 6. There is a significant decrease in the fraction of problems solved exclusively by INT-∞, from 16.2% for $\gamma_j = 2$ to 3.2% for $\gamma_j = 5$. For the NM group, the trend is inverse – these approaches exclusively solved 0.7% problems with linear MVFs and 4.7% for functions involving five characteristic

points. This tendency is also confirmed by Table 6, where the difference between these two approaches for 2-point MVFs is almost 20% (55.6% for INT-∞ and 36.1% for NM), while for 5-point problems it drops to 4% (85.2% and 81.2%, respectively). Thus, increasing the flexibility of MVFs has a more significant impact on the performance of non-monotonic approaches than on INT-∞. Moreover, additional constraints on the monotonicity of the interaction-oriented functions *syn* (see E^M_{INT}) limit the improvement of recoverability for the INT methods.

A greater number of assignment examples for each class (r) implies a decrease in preference recoverability for each method (see Table 2). Figure 6 shows that the number of problem instances unsolvable by any approach increased from 1.7% for problems with one assignment per class to 37% when $r = 5$. This is due to additional constraints introduced by more assignment examples, reducing the space of potential solutions. For problems with poor knowledge of DM's preferences, the differences between the methods are relatively small. In this case, even the basic UTADIS model could deliver a solution for almost 90% instances. The NM group coped best with such problems, providing a consistent solution for 96.9% of problems. INT-∞ fared slightly worse, reaching 96.6%, while INT-1 scored 94.1%. The increase in indirect preference information had the greatest impact on the UTADIS and NM methods. They were able to correctly reproduce, respectively, 20.3% and 32.6% of the problems with five reference assignments per class. The impact was slightly less for INT-1, which solved 34.2% instances, while INT-∞ could handle 59.7% of problems. Thus, if the DM provides more assignment examples, in some cases, INT-∞ may be the only possible choice.

Table 2 shows that the UTADIS and INT methods were better at dealing with problems where performances were generated by *sphere* rather than *random* sampling. In both cases, INT-∞ performed best, solving almost 80% of problems generated with m-sphere approach and over 70% instances with randomly drawn performances. INT-1 outperformed the NM approaches when using *sphere*, scoring 71.5%. However, it performed worse for *random*, scoring 47.9%. In turn, NM reached 61.6% and 62.5%, respectively. The primary UTADIS method performed the worst, providing a consistent solution for 56.5% and 40.4% instances, respectively.

Overall, INT-∞ proved the best at reproducing diverse DM preferences. Nevertheless, the NM methods are slightly better in the case of simple problems with two classes, two criteria, and poor knowledge of DM's preferences. In some cases, the NM approaches perform slightly worse than INT-1, e.g., when the number of classes or reference assignments is the greatest. UTADIS is inferior in preference reproducibility, delivering a consistent model for significantly fewer instances for any combination of problem dimensions.

3.3.2. Analysis for problem instances handled by all methods

The first group of analyzed problem instances are those for which UTADIS found a feasible solution. Then, all other methods also managed to deliver results compatible with all DM's assignment examples. Such a selection of a subset of problems makes it possible to compare all methods in terms of quality measures other than preference recoverability. This selection criterion determines the distortion of the proportions between the different variants of problems. Overall, 48.5% of all problems were selected, and their characteristics can be seen in Figure 7. The advantage of problems with a small number of reference classes and assignments and a larger number of criteria and characteristic points can be noticed. Moreover, the *sphere* algorithm generated more problems solved by the primary variant of UTADIS.

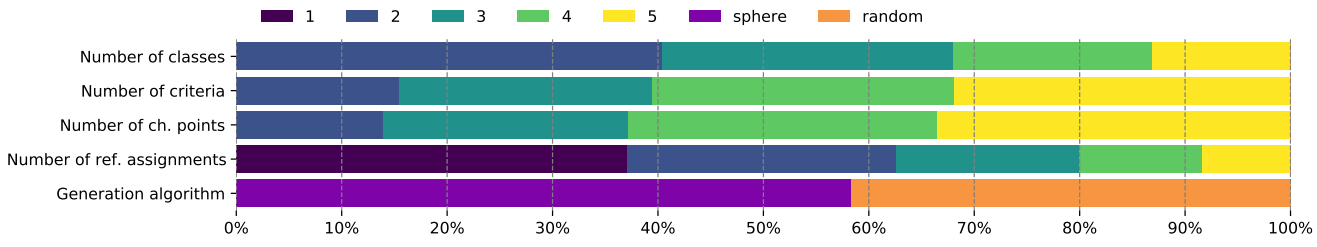


Figure 7: Characteristics of problem instances for which all methods delivered a feasible solution.

The analysis of δ^* complements the conclusions formulated based on PR. The average results for the entire subset of problem instances and the sub-groups with common characteristics are presented in Tables 3 and 4. Since these instances were recoverable by UTADIS and the primary goal for both methods from the INT group is to minimize the number of interactions between criteria in the objective function, the results for UTADIS and INT are the same for all quality measures.

As for δ^*, NM-3 obtained the best results, regardless of the problem size. NM-2 follows it. Both methods optimize the value of δ but differ in their assumptions about the MVF shapes and the post-normalization steps. The only scenarios for which NM-2 led to statistically comparable results involved MVFs with five characteristic points. UTADIS and INT methods fared worse than NM-2, except for simple problems with linear MVFs and one reference assignment per class. According to the Wilcoxon signed-rank test, the differences were not statistically significant in these cases. NM-1 performed the worst, as it is the only one considering slope changes in the objective function, which significantly impacts the δ^* values. The trends observed for various problem dimensions are similar as for the expressiveness expressed with PR. The value of δ^* decreased with more classes and assignments and increased with more complex shapes of MVF. The only difference is in the influence of the number of criteria. For UTADIS + INT and NM-1, δ^* increased slightly, and for NM-2 and NM-3, it decreased. However, these trends are less significant than for other dimensions.

Table 3: Average δ^* values for all problem instances reproducible by all methods and their sub-groups with different numbers of classes and criteria.

δ^*		Number of classes				Number of criteria			
Procedure	**All settings**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
UTADIS + INT	0.081	0.142	0.055	0.030	0.019	0.078	0.078	0.081	0.084
NM-1	0.030	0.050	0.022	0.013	0.009	0.026	0.030	0.031	0.031
NM-2	0.090	0.144	0.070	0.044	0.032	0.097	0.092	0.089	0.086
NM-3	0.099	0.162	0.074	0.046	0.033	0.102	0.103	0.099	0.094

Table 4: Average δ^* values for problem instances reproducible by all methods with different numbers of characteristic points, reference assignments per class, and performance generation algorithms.

δ^*	Number of ch. points				Number of reference assignments					Generation algorithm	
Procedure	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**sphere**	**random**
UTADIS + INT	0.073	0.075	0.081	0.088	0.134	0.071	0.045	0.030	0.024	0.068	0.098
NM-1	0.014	0.032	0.032	0.033	0.044	0.026	0.020	0.017	0.016	0.023	0.040
NM-2	0.067	0.079	0.093	0.105	0.125	0.085	0.066	0.055	0.049	0.075	0.111
NM-3	0.093	0.093	0.099	0.105	0.138	0.096	0.071	0.058	0.051	0.080	0.126

A similar analysis was performed for the APCA measure. The respective results are presented in Tables 5 and 6. In this case, the models delivered by the NM methods implied the same set of possible assignments and thus have equal APCA values. In all cases, the models for UTADIS and INT gave significantly more unambiguous recommendations. Hence, the APCA values are several times higher than for the NM methods, regardless of the problem characteristics. The most significant changes can be observed when increasing the number of characteristic points – APCA for UTADIS + INT is more than twice larger for linear MVFs (0.335 vs. 0.158) and nine times greater for MVFs with five characteristic points (0.081 vs. 0.009). A significantly greater value for the *random* generation of performances obtained by UTADIS + INT is also noteworthy. For these methods, the average recommendation robustness captured by APCA is almost twice as high as that for the *sphere* algorithm. The observations regarding CAR and ECAI are analogous. Thus they are included in the eAppendix (supplementary material available online). Generally, the trends for different problem characteristics are opposite to those observed for the expressiveness measures. Their values increase with more classes or assignments and fewer criteria or characteristic points. Also, since both APCA and CAR focus on the unambiguity of delivered recommendations, in the eAppendix, we discuss the relation between their values in all simulation runs.

The MCAI values – built on the robustness of classifications suggested by the selected model instances – are presented in Tables 7 and 8. Again, they confirm the advantage of UTADIS and INT over all non-monotonic

Table 5: Average values of APCA for all problem instances reproducible by all methods and their sub-groups with different numbers of classes and criteria.

APCA		Number of classes				Number of criteria			
Procedure	**All settings**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
UTADIS + INT	0.146	0.125	0.150	0.165	0.171	0.279	0.162	0.118	0.094
NM	0.039	0.036	0.038	0.045	0.043	0.114	0.045	0.024	0.012

Table 6: Average values of APCA for problem instances reproducible by all methods with different numbers of characteristic points, reference assignments per class, and performances generation algorithms.

APCA	Number of ch. points				Number of reference assignments					Generation algorithm	
Procedure	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**sphere**	**random**
UTADIS + INT	0.335	0.165	0.115	0.081	0.106	0.155	0.172	0.181	0.185	0.106	0.201
NM	0.158	0.038	0.018	0.009	0.025	0.042	0.049	0.054	0.054	0.042	0.035

approaches. The exceptions are problems with two criteria, where NM-2 scores slightly better (0.851 vs. 0.828). NM-2 is the leader among non-monotonic methods, performing distinctly better than NM-1. We can conclude that maximizing δ produces more robust results than optimizing it while limiting slope changes. This is also confirmed by the advantage of NM-3 over NM-1. The exceptions to the last observation are instances with linear MVF. This is understandable as NM-1 optimizes only δ in this setting because there are no slope changes.

The trends for different problem dimensions are similar to those observed previously for APCA. The exception is a decrease in MCAI with the number of classes. The more classes, the lower the values of CAI can be. For example, for 2-class problems, the highest CAI value for a given alternative must be at least 0.5, while for 5-class problems, it is 0.2. This can cause a decrease in MCAI even if the methods still suggest the assignments supported by the most significant number of feasible sorting instances. As for the previously considered robustness measures, UTADIS and INT perform better for the *random* performance generation algorithm. In turn, the NM methods attain greater APCA values for the *sphere* algorithm. Due to the high similarity of conclusions regarding CCA, the results are discussed in eAppendix.

Table 7: Average values of MCAI for all problem instances reproducible by all methods and their sub-groups with different numbers of classes and criteria.

MCAI		Number of classes				Number of criteria			
Procedure	**All settings**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
UTADIS + INT	0.754	0.818	0.746	0.697	0.658	0.828	0.766	0.743	0.720
NM-1	0.651	0.740	0.634	0.575	0.519	0.783	0.671	0.623	0.595
NM-2	0.713	0.771	0.700	0.663	0.636	0.851	0.741	0.686	0.651
NM-3	0.674	0.752	0.659	0.604	0.565	0.801	0.695	0.649	0.619

The analysis indicates that for problem instances reproducible by all methods, UTADIS achieves the most robust recommendations while still being characterized by the lowest preference recoverability. This observation applies to quality measures based on both stochastic analysis and the support given to the results implied by the selected preference model instance.

3.3.3. Analysis for problem instances handled by the NM and INT approaches

To compare the models admitting the use of non-monotonic MVFs or the interactions between criteria, we analyzed the results for problems for which both groups of methods delivered a feasible solution. This makes it possible to unleash the potential of INT methods, which, in this case, consider interactions for at least one pair of criteria. Overall, 10.2% of the initially considered problem instances were included in the analysis. These can be divided into problems solved by all NM and INT methods (5%) and those solved by the NM approaches and INT-∞ (5.2%). Tables 9–14 contain the average values of the quality measures consistent with this division. The upper part of each table corresponds to the subset of problems solved by all NM and INT methods. In contrast, the lower part exhibits the results for the subset of problems reproducible by the NM approaches and INT-∞.

The proportions between sub-problems with particular feature values for both subsets are shown in Figures 8

Table 8: Average values of MCAI for the problem instances reproducible by all methods with different numbers of characteristic points, reference assignments per class, and performance generation algorithms.

MCAI	Number of ch. points				Number of reference assignments					Generation algorithm	
Procedure	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**sphere**	**random**
UTADIS + INT	0.828	0.767	0.747	0.721	0.693	0.760	0.796	0.817	0.834	0.742	0.772
NM-1	0.809	0.666	0.624	0.597	0.594	0.641	0.688	0.725	0.750	0.676	0.616
NM-2	0.823	0.739	0.696	0.665	0.639	0.716	0.766	0.794	0.816	0.739	0.678
NM-3	0.788	0.697	0.657	0.625	0.597	0.677	0.723	0.758	0.783	0.694	0.646

and 9. In the first case, the shares of problem instances are balanced for all numbers of classes (from 21.3% for 5-class to 29.2% for 3-class problems) and criteria (from 23.2% for 2-criteria to 26% for 5-criteria problems). There is a smaller representation of problems with linear MVFs (18.8%) and extreme numbers of preference information pieces (13.6% for one assignment and 15% for five assignments per class). The other subset shows an increased representation of higher numbers of classes, characteristic points, and assignments per class. In this case, there are no problems involving only two criteria. This is because, for such problems, INT-1 and INT-∞ work the same, having the possibility of establishing only one interaction. In both subsets, most instances had performances generated using the *random* algorithm. This suggests that such problems are, on average, more challenging.

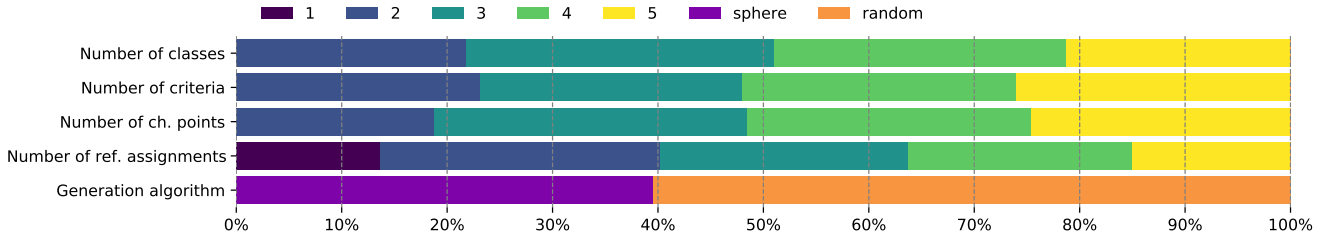


Figure 8: Characteristics of problem instances for which all NM and INT methods delivered a feasible solution.

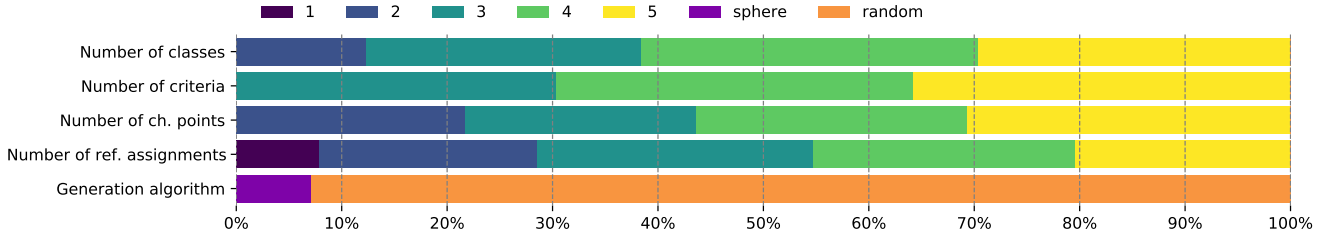


Figure 9: Characteristics of problem instances for which all NM methods and INT-∞ delivered a feasible solution.

Regarding the obtained δ^* values, for the first subset of problem instances, the INT methods had a weaker expressiveness than the NM methods. On the contrary, when considering only problems solved by INT-∞, this method performed better than the NM approaches for most problem characteristics. The only exception is the subset of problems with performances generated by the *sphere* algorithm, where the solutions obtained by INT-∞ are not statistically significantly better than those obtained by NM-2 and NM-3. These results confirm that increasing the number of interactions in the INT approaches positively impacted expressiveness. The detailed results – considering various problem characteristics – are available in Tables 9 and 10.

For the first group of problem instances, INT-∞ achieved better results than INT-1. For 2- and 3-criteria problems, they got the same values because both methods find the same solution involving a single interaction for one pair of criteria. For 4- and 5-criteria problems, both methods represented at most two additional interactions. However, since INT-1 cannot involve the same criterion in multiple interactions, while INT-∞ does not have such a limitation, the average value of δ^* was significantly higher for INT-∞. Among the NM methods, similar to instances considered in Section 3.3.2, NM-3 achieved the best results.

Table 9: Average δ^* values for all problem instances reproducible by the NM and INT methods and their sub-groups with different numbers of classes and criteria.

δ^*		Number of classes				Number of criteria			
Procedure	**All settings**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
NM-1	0.012	0.012	0.013	0.011	0.010	0.004	0.012	0.014	0.016
NM-2	0.022	0.029	0.023	0.018	0.016	0.011	0.023	0.025	0.027
NM-3	0.024	0.034	0.025	0.020	0.018	0.011	0.025	0.028	0.030
INT-1	0.006	0.010	0.006	0.006	0.005	0.007	0.007	0.006	0.006
INT-∞	0.009	0.011	0.008	0.008	0.009	0.007	0.007	0.010	0.012
NM-1	0.013	0.015	0.016	0.013	0.010		0.013	0.013	0.014
NM-2	0.022	0.038	0.025	0.019	0.016		0.022	0.022	0.022
NM-3	0.024	0.043	0.027	0.020	0.018		0.023	0.024	0.025
INT-∞	0.032	0.047	0.034	0.028	0.027		0.030	0.032	0.033

Table 10: Average δ^* values for the problem instances reproducible by the NM and INT methods with different numbers of characteristic points, reference assignments per class, and performances generation algorithms.

δ^*	Number of ch. points				Number of reference assignments					Generation algorithm	
Procedure	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**sphere**	**random**
NM-1	0.006	0.012	0.013	0.015	0.003	0.011	0.013	0.013	0.014	0.003	0.017
NM-2	0.023	0.019	0.020	0.025	0.018	0.023	0.023	0.022	0.021	0.008	0.030
NM-3	0.028	0.022	0.021	0.026	0.021	0.026	0.024	0.024	0.022	0.009	0.034
INT-1	0.008	0.007	0.006	0.005	0.007	0.007	0.006	0.005	0.006	0.006	0.006
INT-∞	0.010	0.009	0.008	0.009	0.007	0.010	0.009	0.010	0.007	0.007	0.011
NM-1	0.007	0.014	0.016	0.016	0.006	0.012	0.014	0.015	0.015	0.003	0.014
NM-2	0.026	0.021	0.020	0.022	0.028	0.025	0.020	0.021	0.020	0.010	0.023
NM-3	0.029	0.023	0.022	0.023	0.033	0.027	0.022	0.023	0.022	0.011	0.025
INT-∞	0.032	0.034	0.029	0.032	0.038	0.036	0.031	0.029	0.030	0.011	0.033

The APCA values for both considered groups of problem instances are presented in Tables 11 and 12. They confirm the higher robustness of recommendations delivered by models incorporating interactions rather than non-monotonicity. All NM methods attained the same average results, while INT-1 was at least as good as INT-∞, regardless of the problem characteristics. Also, we observe a significant decrease in the average value of APCA for INT-∞ between the first and second group of problems (0.401 and 0.206). This means that a greater number of active interactions increases the variability of possible results, hence decreasing their robustness. For the NM methods, this loss is not so evident (0.196 vs. 0.170), suggesting a greater stability of results obtained with the non-monotonic approaches with the increasing problem complexity.

When analyzing the trends for the average value of APCA for different numbers of criteria, we observe that increasing the number of criteria (m) leads to decreasing APCA for the NM approaches. This is due to the greater flexibility of the underlying preference models. In turn, for the INT approaches, we observe an increase in the mean value of APCA with more attributes. This suggests that the number of active interactions did not grow substantially for problems involving more criteria, hence not increasing the space of feasible solutions vastly and the variability of possible results. Even if the INT methods produce, in general, more robust results, the NM approaches are superior in this regard for problems with two criteria and interactions that are representable with INT (0.353 vs. 0.248). Analogous conclusions can be drawn based on the analysis of CAR and ECAI. For a detailed discussion on these measures, see eAppendix.

Table 11: Average APCA values for all problem instances reproducible by the NM and INT methods and their sub-groups with different numbers of classes and criteria.

APCA		Number of classes				Number of criteria			
Procedure	**All settings**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
NM	0.196	0.245	0.207	0.186	0.146	0.353	0.222	0.137	0.091
INT-1	0.401	0.435	0.403	0.382	0.387	0.248	0.388	0.460	0.491
INT-∞	0.373	0.419	0.386	0.354	0.332	0.248	0.388	0.425	0.418
NM	0.170	0.267	0.175	0.148	0.150		0.214	0.177	0.126
INT-∞	0.206	0.404	0.225	0.164	0.152		0.192	0.211	0.213

Tables 13 and 14 show the average MCAI values for both subsets of considered problem instances. In the

Table 12: Average APCA values for the problem instances reproducible by the NM and INT methods with different numbers of characteristic points, reference assignments per class, and performances generation algorithms.

APCA	Number of ch. points				Number of reference assignments					Generation algorithm	
Procedure	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**sphere**	**random**
NM	0.392	0.233	0.139	0.065	0.296	0.195	0.158	0.190	0.177	0.331	0.108
INT-1	0.613	0.405	0.338	0.304	0.418	0.391	0.390	0.418	0.396	0.249	0.500
INT-∞	0.588	0.379	0.313	0.267	0.416	0.366	0.353	0.375	0.375	0.242	0.459
NM	0.385	0.207	0.103	0.048	0.195	0.192	0.173	0.153	0.155	0.460	0.148
INT-∞	0.500	0.198	0.123	0.072	0.330	0.234	0.181	0.187	0.185	0.451	0.187

first case, the precise recommendations suggested by the INT approaches achieved, on average, greater support among the consistent model instances than the results offered by the NM methods. Once again, INT-1 with more restrictive constraints led to slightly more robust recommendations than INT-∞. We also observe similar trends to those noted for APCA for a different number of criteria.

Among the NM methods, for both subsets of problems, the most robust results were obtained by NM-2, followed by NM-3. For the more challenging subset of problem instances reproducible by INT-∞, NM-2 performed significantly better than INT-∞. This is due to a noticeable decrease in the mean value between the two subsets of problems for INT-∞ (0.825 to 0.757) and a less intense decrease for NM-2 (0.816 to 0.798). This observation strengthens the hypothesis about the negative impact of the number of interactions on the robustness of recommended assignments. The detailed results for CCA are available in eAppendix.

Table 13: Average MCAI values for all problem instances reproducible by the NM and INT methods and their sub-groups with different numbers of classes and criteria.

MCAI		Number of classes				Number of criteria			
Procedure	**All settings**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
NM-1	0.717	0.832	0.742	0.681	0.610	0.788	0.742	0.688	0.657
NM-2	0.816	0.896	0.837	0.789	0.739	0.890	0.844	0.793	0.745
NM-3	0.759	0.866	0.786	0.722	0.661	0.826	0.779	0.737	0.703
INT-1	0.836	0.894	0.856	0.814	0.776	0.799	0.810	0.854	0.874
INT-∞	0.825	0.891	0.852	0.803	0.748	0.799	0.810	0.844	0.842
NM-1	0.715	0.869	0.749	0.689	0.648		0.753	0.717	0.681
NM-2	0.798	0.888	0.821	0.785	0.755		0.839	0.800	0.761
NM-3	0.741	0.871	0.776	0.718	0.680		0.772	0.744	0.711
INT-∞	0.757	0.897	0.800	0.732	0.686		0.748	0.755	0.765

Table 14: Average MCAI values for the problem instances reproducible by the NM and INT methods with different numbers of characteristic points, reference assignments per class, and performances generation algorithms.

MCAI	Number of ch. points				Number of reference assignments					Generation algorithm	
Procedure	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**sphere**	**random**
NM-1	0.838	0.737	0.684	0.636	0.685	0.677	0.703	0.751	0.787	0.799	0.662
NM-2	0.885	0.847	0.797	0.746	0.791	0.791	0.807	0.839	0.861	0.898	0.762
NM-3	0.844	0.793	0.737	0.678	0.722	0.726	0.747	0.796	0.818	0.837	0.708
INT-1	0.892	0.845	0.817	0.800	0.776	0.817	0.844	0.862	0.871	0.804	0.856
INT-∞	0.887	0.837	0.806	0.782	0.775	0.807	0.829	0.845	0.864	0.803	0.839
NM-1	0.841	0.746	0.684	0.628	0.660	0.662	0.721	0.734	0.757	0.838	0.705
NM-2	0.871	0.826	0.782	0.740	0.739	0.760	0.805	0.815	0.831	0.887	0.791
NM-3	0.840	0.774	0.717	0.667	0.677	0.697	0.745	0.757	0.785	0.841	0.733
INT-∞	0.872	0.766	0.736	0.685	0.741	0.728	0.737	0.777	0.792	0.846	0.750

In general, the INT approaches generated more robust recommendations than NM for problems that are too complex for the primary UTADIS approach. Moreover, INT-1 slightly outperformed INT-∞, regardless of the quality measure analyzed. This confirms the validity of the strategy of activating the least possible number of interactions per each criterion. Among the non-monotonic algorithms, the best performer was NM-2. It obtained better results for measures based on acceptability indices. The premises indicating the dependence of robustness on the number of interactions in the INT methods raise doubts. Hence, this topic will be considered in the next section.

3.3.4. Robustness and expressiveness in the context of the number of active interactions in the INT method

We aim to test the impact of the number of active interactions between pairs of criteria on the recommendation robustness. For this purpose, we consider all problems reproducible simultaneously by the NM methods and INT-∞. Undoubtedly, a higher number of active interactions increases the model's expressiveness because it gives more freedom in adjusting the impact of marginal functions and interactions on the comprehensive values of alternatives. At the same time, due to the minimization of the number of active interactions in the objective function, it can be presumed that solutions with a higher number of interactions apply to more demanding problems. This, in turn, means that a direct comparison of quality measure values between solutions obtained using INT-∞ without considering the number of active interactions may not be trustworthy. Hence, one needs to analyze each such subset of problems separately.

Still, comparing these values to the results attained by other methods from the NM group is possible. Figure 10 shows the structure of the set of considered problems, and Table 15 exhibits the average values of quality measures while dividing the set of problem instances based on the number of active interactions in the solution obtained by INT-∞. There is a clear advantage in the frequency of problems involving sets of alternatives whose performances were generated by the *random* algorithm. Moreover, problems with greater complexity (i.e., with higher numbers of classes, criteria, characteristic points, and reference assignments) occur more often.

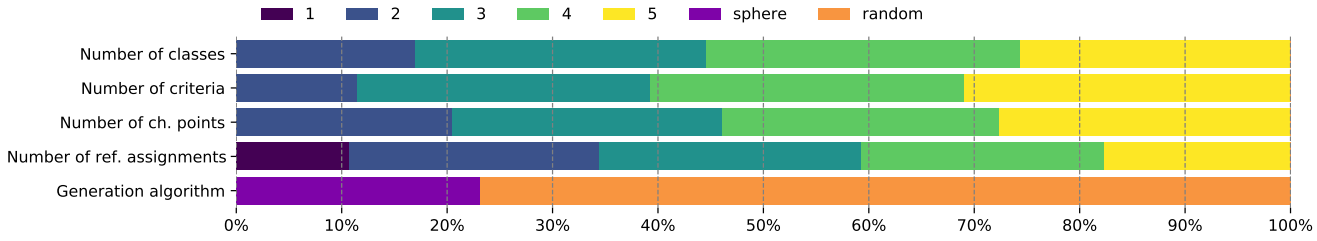


Figure 10: Characteristics of problem instances for which NM and INT-∞ with at least one active interaction delivered a feasible solution.

Table 15 suggests that the subsets of problem instances for which the model delivered by INT-∞ has one or two active interactions are of similar size. They jointly constitute about 93.5% of all considered instances. Due to the small number of problems with more than two active interactions, the remaining solutions are grouped and marked as 3+. Considering the δ^* values, with the increasing number of interactions, the expressiveness of INT-∞ increases too, whereas, for the NM group, it increases slightly for two interactions and then decreases rapidly for three or more interactions. When comparing different methods, the INT model with only one active interaction is, on average, less expressive than all non-monotonic approaches. Conversely, with two or more interactions needed, it is significantly more flexible than the remaining methods.

Considering the quality measures that quantify the robustness by exploiting set $\mathcal{U}^R$, i.e., APCA, CAR, and ECAI, in all three cases, INT-∞ generated more robust recommendations than the NM group when it needed at most two active interactions. Both in terms of the possible class assignments and consensus between the compatible sorting model instances, significant increases in robustness are seen for the NM methods between problems requiring solutions with two and three interactions. This increase is much smaller for INT-∞ in the case of APCA and CAR, and for ECAI, there is even a slight decrease.

The analysis of the remaining two measures leads to the same conclusions. The exception is observed for MCAI in the case of two interactions. Then, NM-2 achieved an average value of 0.785, better than 0.758 attained by INT-∞. The observations above suggest that the minimum number of active interactions outputted by INT-∞ may be essential for the robustness of delivered recommendations. If the number of interactions does not exceed two, then the model with interactions leads to more robust recommendations. However, if there are more interactions, then using one of the non-monotonic approaches is more beneficial. Still, the DM's indications of non-monotonicity of preferences and/or interactions between criteria should be critical in the method selection process. Choosing

Table 15: Shares of problem instances and average values of six measures for different methods and different number of active interactions in the solutions obtained with the INT-∞ method.

Number of interactions	**1**	**2**	**3+**
% of considered problems	**45.24%**	**48.26%**	**6.51%**
δ^*			
NM-1	0.011	0.014	0.008
NM-2	0.021	0.023	0.016
NM-3	0.024	0.025	0.017
INT-∞	0.007	0.032	0.034
APCA			
NM	0.204	0.135	0.410
INT-∞	0.387	0.201	0.224
CAR			
NM	0.159	0.104	0.315
INT-∞	0.317	0.153	0.203
ECAI			
NM-1, NM-2	0.711	0.671	0.795
NM-3	0.660	0.604	0.758
INT-∞	0.814	0.766	0.699
MCAI			
NM-1	0.721	0.694	0.846
NM-2	0.821	0.785	0.876
NM-3	0.764	0.725	0.840
INT-∞	0.830	0.758	0.736
CCA			
NM-1	0.740	0.726	0.805
NM-2	0.704	0.689	0.782
NM-3	0.705	0.680	0.780
INT-∞	0.812	0.738	0.689

a model based on the number of active interactions should only be considered when the DM does not opt for using either approach because of the apparent characteristics of relevant attributes.

3.3.5. Robustness and expressiveness within the NM and INT methods

The shares of problem instances solved by either all NM approaches or both INT approaches are 3.37% and 6.19%, respectively. The structure of problems reproducible by the NM group is represented in Figure 11. Once again, there is a considerable predominance of problems with performances generated by the *random* algorithm. Problems with fewer classes and criteria, more complex MVFs, and more reference assignments are also more common.

Table 16 shows the average measure values for the considered problem instances. They confirm previous observations. NM-3 can be considered the most flexible approach based on the analysis of δ^*. Yet, the values of PR, APCA, and CAR for all three NM methods are the same. However, the models used by NM-1 and NM-2 lead to more robust results in terms of ECAI. Moreover, among these two approaches, NM-2 typically suggests a more robust recommendation because it attains significantly higher MCAI values than NM-1. In turn, NM-1 is the best for CCA, maximizing the number of non-reference alternatives with *confirmed* assignments. This fact correlates with the low value of δ^* for NM-1, leading to narrower value ranges in which assignments are uncertain than for other methods.

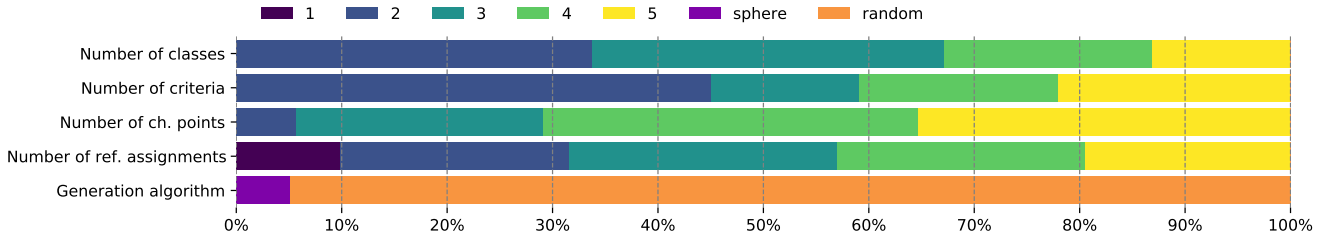


Figure 11: The characteristics of problem instances reproducible by the NM methods.

Table 16: Average values of six measures for problem instances reproducible by the NM methods.

Measure	NM-1	NM-2	NM-3
δ^*	0.015	0.034	0.036
APCA	0.187	0.187	0.187
CAR	0.148	0.148	0.148
ECAI	0.641	0.641	0.588
MCAI	0.715	0.797	0.747
CCA	0.750	0.666	0.657

A similar analysis was performed for problem instances for which the solutions were generated exclusively by the INT methods. This subset involves more instances with performances generated by the *sphere* algorithm, higher numbers of classes and assignments, and three or four criteria or characteristic points (see Figure 12). The quality measures shown in Table 17 are consistent with the previous observations. The δ^* value indicates a higher expressiveness of INT-∞, whereas all other measures confirm a statistically significant advantage of INT-1 in ensuring higher recommendation robustness. For both groups, NM and INT, more detailed data showing the impact of various model parameters on the values of quality measures are also consistent with previous observations. The relevant discussion is included in eAppendix.

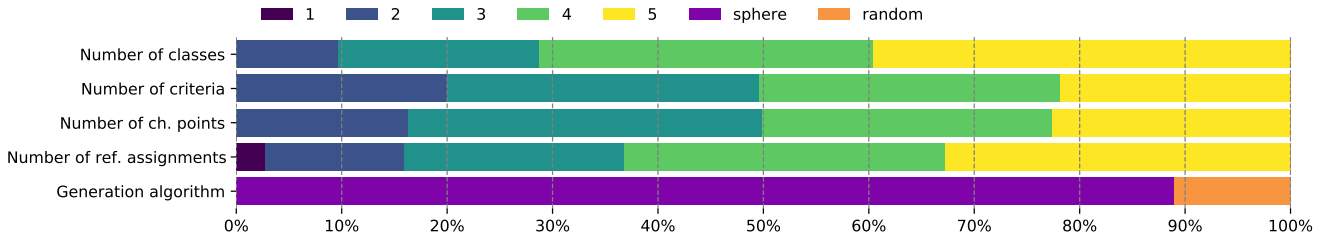


Figure 12: The characteristics of problem instances reproducible by the INT methods.

Table 17: Average values of six measures for problem instances reproducible by the INT methods.

Measure	INT-1	INT-∞
δ^*	2.26E-03	2.64E-03
APCA	0.323	0.306
CAR	0.248	0.232
ECAI	0.810	0.802
MCAI	0.813	0.806
CCA	0.820	0.810

The above analyses confirm that NM-2 is the most advantageous method among non-monotonic approaches. It combines increased flexibility and the greatest robustness, suggesting recommendations highly consistent with the results produced by a set of all compatible models. Due to the significantly higher robustness of the delivered solutions, the recommended approach among the INT methods is INT-1. INT-∞ should be used for more demanding problems when more interactions between criteria are needed to reproduce the DM's preferences.

4. Which is the most suitable UTADIS variant that should be used for a given problem?

This section presents two frameworks for recommending the appropriate variant of UTADIS for a particular problem. One is based on the experimental results, hence referring to the concepts of expressiveness and robustness. The other is taxonomy-based, taking into account the problem's characteristics and the DM's requirements.

4.1. Framework for recommending the adequate variant of UTADIS based on the experimental results

The conclusions from the experimental analysis led us to formulate a framework that supports selecting an adequate model for a given problem. We assume that the ability to reproduce the DMs' preferences is crucial to providing

a solution consistent with their value system and that recommendations should be trustworthy in terms of their robustness. Figure 13 shows a flowchart underlying the selection procedure.

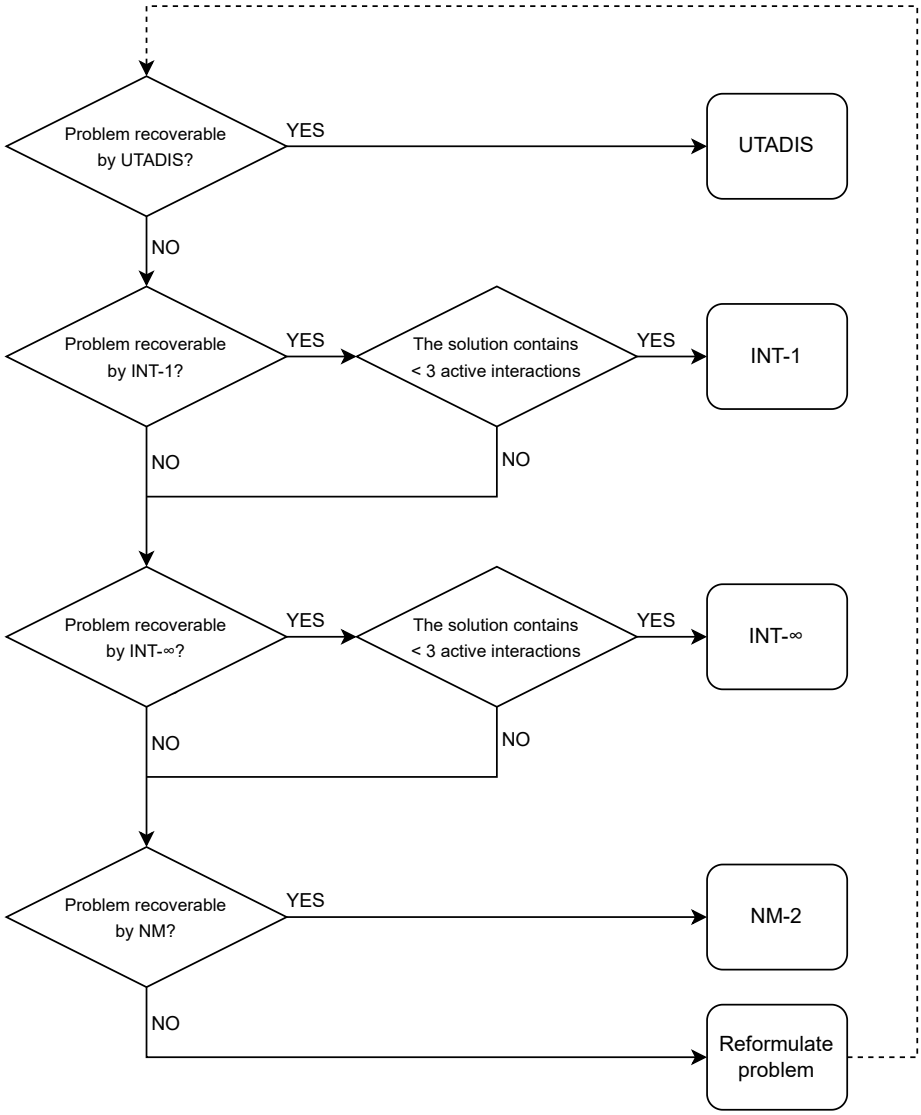


Figure 13: A framework for recommending an adequate variant of UTADIS for a given decision problem.

Its underlying idea is to recommend an approach that, in most cases, would deliver results that are as robust as possible, given that all constraints resulting from the expressed preferences are satisfied. The first condition verifies whether the assumed model is expressive enough to reproduce DM's preferences. According to the experimental analysis results, UTADIS was characterized by significantly better robustness than other methods. Hence, it should be considered the first choice under the consistency setting. Its results should be the most conclusive and easy to interpret for the user. If UTADIS could not reproduce the DM's assignment examples, then INT-1 or INT-∞ (in this order) should be employed because their recommendations' robustness was higher than for the non-monotonic approaches. However, this advantage stands true when the number of interactions does not exceed two. When the interaction-oriented approaches cannot reproduce the preference information, one should check if the NM group of methods can do so. In the case of a positive answer, the recommended method should be NM-2, as it provides more robust recommendations than its counterparts. Let us emphasize that the framework's application is justified if the choice of the specific model is consulted between the analyst and the DM. Simply, the acceptance of modifying the assumptions regarding preferential independence or monotonicity needs to make sense, given the problem's characteristics. Finally, suppose it is impossible to reproduce the preferences using any of the methods

considered. In this case, one needs to reformulate the problem and/or the DM's preferences using algorithmic support or interacting with the DM [19, 38]. After modifying the problem's input or assumptions (e.g., revising some judgments to restore the consistency or accepting some level of inconsistency), the framework can be reapplied.

To remain concise, the formulated guidelines refer to sufficient expressiveness and the number of active interactions, neglecting, e.g., the number of classes or criteria. However, the experimental results discussed in the previous subsections supplement the framework, providing more detailed hints on the method selection based on the parameters of the considered problem and quality measures exhibiting different interpretations of robustness, provided that adequate reproducibility is guaranteed. For example, when considering problems reproducible by NM and INT-∞, Tables 13 and 14 indicate that NM-2 outperformed INT-∞ in terms of $MCAI$, especially for a large number of classes, reference alternatives, and characteristic points. Therefore, when the DM's assignment examples need a complex model to be reproduced and one cares about the support given to the delivered assignments by all compatible sorting models, we can opt for using NM-2 over INT-∞.

The proposed framework is valid when there are no solid reasons or DM's preference that would directly indicate the need to introduce non-monotonic MVFs or interactions between criteria. Therefore, the framework's workflow should be perceived as a set of guidelines when no other arguments for using a specific method can be expressed. Otherwise, an appropriate variant of UTADIS should be selected irrespective of the results attained in the comparative study based on the agreement of the methods' features with the problem characteristics and DM's requirements. To support dialogue in this scenario, in Section 4.2, we formulate a set of questions to enable the selection of an appropriate UTADIS method.

4.2. Taxonomy-based framework for recommending the adequate variant of UTADIS based on the problem characteristics

The variant of UTADIS appropriate for a given decision problem can also be selected based on the problem's characteristics and the DM's requirements. A comprehensive framework for performing such a selection has been proposed in [6]. In Table 18, we report the questions and answers that lead to selecting all UTADIS variants. They refer to the features regarding problem formulation, preference model, and preference information. In particular, the accounted UTADIS variants support sorting problems with completely ordered classes without cardinality constraints, flat criteria structures, and deterministic performances on a complete family of criteria while applying a cardinal scale to lead the assignments. Regarding the model, the performances are used quantitatively, compared by the DM with respect to non-graded preference intensity, and aggregated while admitting full compensation. As for the preferences, all variants accept indirect assignment examples.

Table 18: Questions and answers leading to the recommendation of all considered methods from the family of UTADIS.

Symbol	**Question**	**Answer**
	Problem typology	
Q-PT-I	What type of decision recommendation is requested?	Sorting
Q-PT-II	What order of classes is requested?	Complete
Q-PT-III	What scale leading the recommendation is requested?	Cardinal
Q-PT-IV	What cardinality of classes is required?	Without constraints
Q-PT-V	What is the structure of the criteria used for the assessment?	Flat
Q-PT-VI	What is the type of performance of the criteria?	Deterministic
Q-PT-VII	What is the completeness status of the criteria set?	Complete
	Preference model	
Q-PM-I	How should the input information/performance data be used by the method(s)?	Quantitatively
Q-PM-II	What type of method that considers the quantitative information from the criteria performances should be selected?	Performance-based
Q-PM-III	How should the comparison of the performances on the criteria be performed?	Performances are compared by the DM with respect to non-graded intensity of preference
Q-PM-IV	How much can the good performance on a criterion compensate for the bad performance on another criterion?	Fully
	Preference information	
Q-PI-I	What type of preference information is provided?	Indirect
Q-PI-II	What type of indirect preferences would you like to account for?	Assignments of reference alternatives to classes

Table 19 presents the questions that lead to particular variants of UTADIS. Specifically, the traditional UTADIS should be used under the assumptions of preferential independence between the criteria, for which the preference for the performances is known and monotonic. The INT methods are recommended when interactions between criteria should be considered. Then, the contribution of the performance on some criterion into the alternative's comprehensive evaluation may be affected by the performances on the remaining criteria. If a criterion can interact with at most one other attribute, then INT-1 should be prioritized over INT-∞. This assumption makes constructing a sorting model more manageable and interpretation more straightforward. These two variants also require that the order of preference for the performances on all criteria is known, meaning there is a clear, pre-defined correspondence between attributes and class assignments. When the DM does not know a priori if such a monotonic dependency is present and admits that it can be non-monotonic, the NM variants should be employed. To discriminate between them, one needs to indicate whether a) sudden changes in the functions' directions should be prevented, hence minimizing slope changes and opting for the most parsimonious model (NM-1), b) the most discriminant model should be prioritized, maximizing the difference between comprehensive values of alternatives from various classes (NM-2), or c) it is desired to distinguish gain and cost components for the potentially non-monotonic criteria (NM-3).

Table 19: Questions and answers leading to the recommendation of different methods from the family of UTADIS.

Symbol	Question	Answers (Methods)
	Interactions between criteria	
Q-INT-I	Should interactions between criteria be considered to reflect a non-additive nature of preferences?	Yes (INT-1, INT-∞) No (UTADIS, NM)
Q-INT-II	Can each criterion interact with more than one other attribute?	Yes (INT-∞), No (INT-1)
	Potential non-monotonicity of preference directions	
Q-NM-I	What is the knowledge of the preference for the values of each criterion?	Known, monotonic (UTADIS, INT) To discover, potentially non-monotonic (NM)
Q-NM-II	When handling potential non-monotonicity, should sudden changes in the functions' directions be prevented?	Yes (NM-1, NM-2) No (NM-3)
Q-NM-III	When handling potential non-monotonicity, should obtaining the most discriminant model be prioritized?	Yes (NM-1, NM-3) No (NM-2)
Q-NM-IV	Is it desired to distinguish pair of components with monotonic potentially positive and negative relationships for the non-monotonic criteria?	Yes (NM-3) No (NM-1, NM-2)

5. A case study

To illustrate the applicability of the presented UTADIS methods and quality measures, we consider the problem of sorting 30 mobile phone models into three classes: C_1, C_2 and C_3, where C_1 is the least preferred and C_3 is the most preferred one. The alternatives are evaluated in terms of four criteria: g_1 – display size (inches), g_2 – storage (gigabytes), g_3 – battery capacity (mAh), and g_4 – price (Ukrainian Hryvnia – UAH). Their performances are given in Table 20. The data comes from [40].

For each criterion, we selected the following four characteristic points: g_1 – 5.1, 6.1, 7.1, 8.1; g_2 – 64, 128, 256, 512; g_3 – 2000, 3000, 4000, 5000; g_4 – 15000, 30000, 45000, 60000. The extreme observed performances on each criterion define the range of acceptable evaluations. In addition, the breakpoints for criteria g_1, g_3, and g_4 were selected according to Equal Width Binning [10]. Since there are only four possible performances on g_2, they were all selected as characteristic points. In the basic definition of the problem, we considered the first three criteria to be of a gain type and the last criterion – a cost type. We also admitted non-monotonicity in the case of display size (g_1). Finally, we considered three artificial DMs – DM_1, DM_2, and DM_3, simulating a dialogue with each. Their preferences are generated so that to illustrate the use of different variants of UTADIS and the proposed frameworks for method selection while referring to the least possible number of DMs and interactions.

In the next step, nine reference alternatives were selected. They were then precisely classified by each DM. All three DMs agreed to assign alternatives a_8, a_{13} and a_{20} to class C_1, a_2 to C_2, and a_{12} to C_3. In the case of a_3, a_5, a_{17}, and a_{19}, they had conflicting preferences, but each of these alternatives was assigned to either C_2 or C_3. The

Table 20: Performance table for the problem of sorting mobile phone models.

Alternative	g_1 (display size)	g_2 (storage)	g_3 (battery capacity)	g_4 (price)
a_1	8.00	512	4500	56082
a_2	7.30	256	4380	55338
a_3	6.10	512	2815	46503
a_4	8.10	256	3577	44232
a_5	6.90	512	4500	39524
a_6	6.80	128	5000	37630
a_7	5.80	512	3190	36188
a_8	6.10	128	2815	35507
a_9	6.50	256	4000	34165
a_{10}	6.70	512	4260	32530
a_{11}	6.50	64	3969	32583
a_{12}	6.58	512	4200	31656
a_{13}	5.40	256	2227	31369
a_{14}	6.10	256	4000	29927
a_{15}	5.20	64	4500	25649
a_{16}	6.20	128	2510	22760
a_{17}	6.78	256	4510	22587
a_{18}	6.00	128	4080	21879
a_{19}	6.67	128	5000	21435
a_{20}	6.10	128	3140	17475
a_{21}	7.60	256	4500	51460
a_{22}	6.80	256	5000	39620
a_{23}	6.50	512	3174	27190
a_{24}	6.70	256	4200	29999
a_{25}	6.67	256	4500	27999
a_{26}	6.67	256	5000	27417
a_{27}	5.99	64	4100	15048
a_{28}	5.10	64	2510	15000
a_{29}	6.20	128	2000	28543
a_{30}	7.00	512	4260	60000
min	5.10	64	2000	15000
max	8.10	512	5000	60000

upper part of Table 21 summarizes the initially collected preference information. The remaining 21 alternatives not listed in this section of the table were not evaluated by the DMs in the first step.

For each DM, we simulated two iterations so that the analysis of the results delivered by the primary UTADIS model after the first iteration stimulated the provision of additional assignment examples, which are presented in the lower part of Table 21. The class thresholds, alternatives' comprehensive values, and assignments for all discussed models are provided in Table 22. Also, for each iteration, we sampled $100,000$ uniformly distributed feasible model instances in $\mathcal{U}^R$.

For illustrative purposes, we also determined the values of six quality measures capturing the models' expressiveness and robustness of the delivered results (see Table 23). This way, the readers could better understand their meaning while referring to an example sorting problem.

Analysis for DM_1. Let us assume DM_1, supported by the decision analyst, opted for using a standard UTADIS. Hence, he agreed that the impact of all criteria in a comprehensive score does not depend on any other attribute. Moreover, he claimed that g_1, g_2, and g_3 are gain criteria, with greater performance being more favorable, whereas g_4 is a cost criterion. The model compatible with the nine reference assignments of DM_1 is denoted as UTADIS_1 (see Table 22). The comprehensive value ranges for the three classes are as follows: $[0, 0.647)$, $[0.647, 0.842)$, and $[0.842, 1]$. There are 7, 11, and 12 alternatives in C_1, C_2, and C_3, respectively. However, the DM judged this recommendation unsatisfactory because a_{23} was assigned to C_2. Considering its low price (27190 UAH) and large capacity (512 GB), the DM opted for assigning a_{23} to C_3 (see Table 21). The new solution, denoted by UTADIS_2, was still delivered by the primary UTADIS model while respecting the pre-defined monotonicity constraints and not violating preferential independence.

When comparing the MVFs obtained for DM_1 in the two iterations (see Figure 14), there is a significant increase in the maximal value for u_2 (for 512 GB) and higher appreciation of prices in the 15000-30000 UAH range compared

Table 21: Reference assignments provided by three DMs.

Alternative	$\mathbf{DM_1}$	$\mathbf{DM_2}$	$\mathbf{DM_3}$
First iteration			
a_2	C_2	C_2	C_2
a_3	C_2	C_3	C_2
a_5	C_3	C_3	C_2
a_8	C_1	C_1	C_1
a_{12}	C_3	C_3	C_3
a_{13}	C_1	C_1	C_1
a_{17}	C_3	C_2	C_3
a_{19}	C_2	C_2	C_3
a_{20}	C_1	C_1	C_1
Second iteration			
a_1		C_1	
a_4		C_1	
a_{15}			C_3
a_{23}	C_3		

to the 30000-45000 UAH range. The comprehensive value for the new reference alternative a_{23} increased slightly (from 0.812 to 0.820). Since the class thresholds were assigned lesser values, a_{23} is now assigned to the most preferred class C_3. The two non-reference alternatives whose classifications were affected by the model change are a_1 and a_4 assigned to C_3 and C_2, respectively.

When comparing the two models obtained for DM_1, the value of δ^* decreased after enriching the constraint set with the one implied by the desired assignment of a_{23} (see Table 23). The number of non-reference alternatives decreased from 21 to 20. Among them, only two alternatives had a non-empty necessary assignment ($a_{10} \rightarrow^N C_3$, $a_{30} \rightarrow^N C_3$). Thus, the value of CAR is equal to $\frac{2}{21}$ and $\frac{2}{20}$ in the two iterations. In addition, there is a slight increase in the values of ECAI (from 0.618 to 0.630) and MCAI (from 0.621 to 0.665), confirming the positive impact of additional references assignments on the robustness of delivered results.

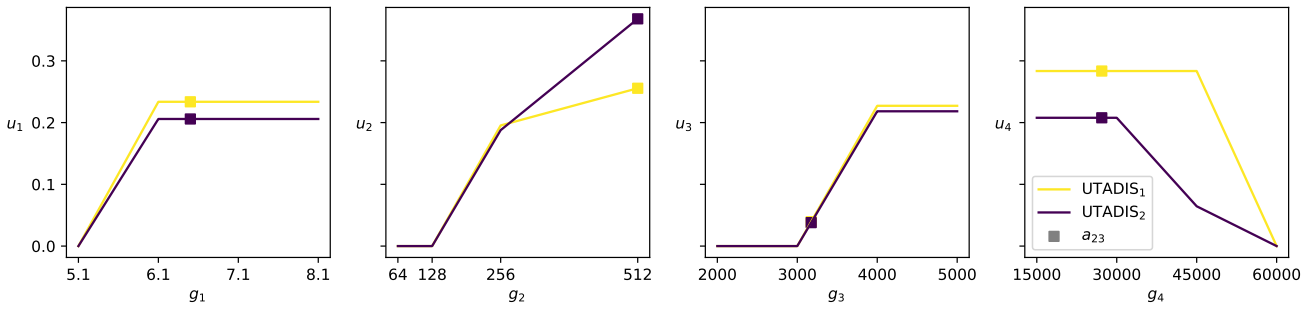


Figure 14: Marginal value functions obtained in the two iterations for DM_1.

Analysis for DM_2. The analysis for DM_2 starts while tolerating the preferential independence and pre-defined preference directions for the four criteria. The model selected by the primary UTADIS method is presented in Figure 15 and Table 22. The greatest share in the comprehensive value is associated with g_2, while g_1 and g_4 have a negligible impact on the alternatives' scores. Also, the class thresholds were vastly different than for DM_1, with the upper limits equal to 0.187 for C_1 and 0.493 for C_2. This affected the cardinalities of alternatives assigned to each class: C_1 – 6, C_2 – 16, and C_3 – 8.

Table 24 presents the possible class assignments for all non-reference alternatives. Only a_{10}, a_{23}, and a_{30} have non-empty necessary assignments; for ten alternatives – there are two possible classes, whereas for the remaining eight options – all three classes are observed for at least one feasible sorting model. As a result, $APCA(\mathcal{U}^R) = 1 - \frac{8 \cdot 2 + 10 \cdot 1 + 3 \cdot 0}{21 \cdot 2} = 0.381$. In Table 24, the bold values of $CAI'(a, C_l)$ are associated with the class to which the selected model assigns a given alternative. For 17 alternatives, this assignment is confirmed by most samples, and only for alternatives a_{11}, a_{15}, a_{18}, and a_{27} – CAI' is higher for some other class. Still, their average is relatively high ($MCAI(U) = 0.783$), confirming the high robustness of the suggested classification. Further, the analysis of

Table 22: Threshold values, alternatives assignments, and comprehensive values obtained in each iteration for the three DMs.

DM	**DM$_1$**		**DM$_2$**				**DM$_3$**	
Method	**UTADIS$_1$**	**UTADIS$_2$**	**UTADIS**	**NM-1**	**NM-2**	**NM-3**	**UTADIS**	**INT-1**
Threshold								
t_1	0.647	0.538	0.187	0.547	0.571	0.478	0.442	0.176
t_2	0.842	0.726	0.493	0.636	0.810	0.674	0.714	0.682
Alternative								
a_1	0.791 (C_2)	0.809 (C_3)	**0.961(C_3)**	**0.543(C_1)**	**0.508(C_1)**	**0.406(C_1)**	0.611 (C_2)	0.388 (C_2)
a_2	0.744 (C_2)	0.632 (C_2)	0.340 (C_2)	0.551 (C_2)	0.635 (C_2)	0.551 (C_2)	0.459 (C_2)	0.247 (C_2)
a_3	0.744 (C_2)	0.632 (C_2)	0.646 (C_3)	0.640 (C_3)	0.873 (C_3)	0.746 (C_3)	0.459 (C_2)	0.247 (C_2)
a_4	0.844 (C_3)	0.591 (C_2)	**0.207(C_2)**	**0.353(C_1)**	**0.205(C_1)**	**0.023(C_1)**	0.355 (C_1)	0.205 (C_2)
a_5	1.000 (C_3)	0.909 (C_3)	0.955 (C_3)	0.716 (C_3)	0.937 (C_3)	0.949 (C_3)	0.697 (C_2)	0.612 (C_2)
a_6	0.744 (C_2)	0.559 (C_2)	0.344 (C_2)	0.578 (C_2)	0.635 (C_2)	0.584 (C_2)	0.610 (C_2)	0.442 (C_2)
a_7	0.746 (C_2)	0.702 (C_2)	0.693 (C_3)	0.753 (C_3)	0.764 (C_2)	0.691 (C_3)	0.541 (C_2)	0.342 (C_2)
a_8	0.517 (C_1)	0.361 (C_1)	0.000 (C_1)	0.486 (C_1)	0.508 (C_1)	0.406 (C_1)	0.339 (C_1)	0.000 (C_1)
a_9	0.940 (C_3)	0.780 (C_3)	0.292 (C_2)	0.677 (C_3)	0.682 (C_2)	0.530 (C_2)	0.514 (C_2)	0.689 (C_3)
a_{10}	1.000 (C_3)	0.976 (C_3)	0.930 (C_3)	0.748 (C_3)	0.906 (C_3)	0.898 (C_3)	0.735 (C_3)	0.897 (C_3)
a_{11}	0.737 (C_2)	0.600 (C_2)	0.250 (C_2)	0.594 (C_2)	0.477 (C_1)	0.508 (C_2)	0.385 (C_1)	0.628 (C_2)
a_{12}	1.000 (C_3)	0.984 (C_3)	0.922 (C_3)	0.767 (C_3)	0.899 (C_3)	0.868 (C_3)	0.730 (C_3)	0.932 (C_3)
a_{13}	0.549 (C_1)	0.444 (C_1)	0.034 (C_1)	0.543 (C_1)	0.426 (C_1)	0.301 (C_1)	0.426 (C_1)	0.105 (C_1)
a_{14}	0.940 (C_3)	0.820 (C_3)	0.280 (C_2)	0.740 (C_3)	0.682 (C_2)	0.429 (C_1)	0.580 (C_2)	0.859 (C_3)
a_{15}	0.534 (C_1)	0.447 (C_1)	0.284 (C_2)	0.805 (C_3)	0.212 (C_1)	0.242 (C_1)	**0.408(C_1)**	**0.753(C_3)**
a_{16}	0.517 (C_1)	0.414 (C_1)	0.003 (C_1)	0.416 (C_1)	0.508 (C_1)	0.431 (C_1)	0.426 (C_1)	0.000 (C_1)
a_{17}	0.940 (C_3)	0.820 (C_3)	0.340 (C_2)	0.633 (C_2)	0.746 (C_2)	0.602 (C_2)	0.735 (C_3)	0.859 (C_3)
a_{18}	0.721 (C_2)	0.611 (C_2)	0.251 (C_2)	0.704 (C_3)	0.481 (C_1)	0.388 (C_1)	0.431 (C_1)	0.753 (C_3)
a_{19}	0.744 (C_2)	0.632 (C_2)	0.340 (C_2)	0.599 (C_2)	0.635 (C_2)	0.551 (C_2)	0.730 (C_3)	0.753 (C_3)
a_{20}	0.549 (C_1)	0.444 (C_1)	0.034 (C_1)	0.543 (C_1)	0.508 (C_1)	0.406 (C_1)	0.426 (C_1)	0.105 (C_1)
a_{21}	0.818 (C_2)	0.649 (C_2)	0.350 (C_2)	0.503 (C_1)	0.507 (C_1)	0.353 (C_1)	0.495 (C_2)	0.247 (C_2)
a_{22}	0.940 (C_3)	0.728 (C_3)	0.379 (C_2)	0.629 (C_2)	0.808 (C_2)	0.607 (C_2)	0.732 (C_3)	0.466 (C_2)
a_{23}	**0.812(C_2)**	**0.820(C_3)**	0.701 (C_3)	0.640 (C_3)	0.873 (C_3)	0.848 (C_3)	0.696 (C_2)	0.378 (C_2)
a_{24}	0.940 (C_3)	0.820 (C_3)	0.314 (C_2)	0.645 (C_3)	0.707 (C_2)	0.581 (C_2)	0.641 (C_2)	0.859 (C_3)
a_{25}	0.940 (C_3)	0.820 (C_3)	0.336 (C_2)	0.650 (C_3)	0.745 (C_2)	0.574 (C_2)	0.732 (C_3)	0.859 (C_3)
a_{26}	0.940 (C_3)	0.820 (C_3)	0.375 (C_2)	0.650 (C_3)	0.808 (C_2)	0.574 (C_2)	0.884 (C_3)	0.859 (C_3)
a_{27}	0.719 (C_2)	0.609 (C_2)	0.253 (C_2)	0.680 (C_3)	0.449 (C_1)	0.386 (C_1)	0.436 (C_1)	0.753 (C_3)
a_{28}	0.284 (C_1)	0.208 (C_1)	0.000 (C_1)	0.564 (C_2)	0.112 (C_1)	0.223 (C_1)	0.237 (C_1)	0.000 (C_1)
a_{29}	0.517 (C_1)	0.414 (C_1)	0.003 (C_1)	0.325 (C_1)	0.508 (C_1)	0.431 (C_1)	0.426 (C_1)	0.000 (C_1)
a_{30}	1.000 (C_3)	1.000 (C_3)	0.940 (C_3)	0.701 (C_3)	0.906 (C_3)	0.975 (C_3)	0.775 (C_3)	1.000 (C_3)

*CAI*s leads to the following entropy: $\sum_{a\in A^T} ECAI_{alt}(a) = 6.9999$. After normalizing it, we obtained the value of the entropy-oriented quality measure: $ECAI(\mathcal{U}^R) = 1 - \frac{6.9999}{\log_2(3)\cdot 21} = 0.790$. It should be perceived as relatively high, suggesting that the variability of sorting recommendations in the set of feasible sorting model instances is rather low.

Nevertheless, the solution proposed by UTADIS was not approved by DM_2 due to overestimating a_1 and a_4. Hence, the DM stated that these two alternatives should be assigned to C_1 (see Table 21) mainly because he judged a display size of at least eight inches too large. Note that this was not possible when using any model compatible with the assumptions of the primary UTADIS method (see Table 24).

After including the two additional assignment examples, the UTADIS method could not find any feasible solution. Hence, the DM – supported by the decision analyst – assumed that the non-monotonic shape of u_1 can be accepted to increase the model's flexibility while suitably representing his preferences. Indeed, he confirmed that it is acceptable that the preference should be the least for small or large display sizes. In contrast, the most preferred screens have intermediate sizes, ensuring a proper balance between usability, comfort, and conveniently storing the phone in a pocket. For the remaining three criteria, MVFs were still required to respect the pre-defined preference directions. The DM was offered three solutions, each obtained using a different approach from the NM group (see Table 22 and Figure 15). He assessed that the recommendations obtained using NM-2 best reflected his preferences.

Figure 15 confirms the non-monotonic character of u_1. In the case of NM-2, it assigns the highest scores to phones with intermediate display sizes between 6.1 and 7.1 inches. In addition, compared to UTADIS, there is an apparent decrease in the value for alternatives with the highest storage (g_2) and battery (g_3) capacities. The impact of price (g_4) was found negligible for all four models obtained based on the preferences of DM_2. When comparing

Table 23: Values of six quality measures for the sorting model instances selected in each iteration for the three DMs.

DM	**DM$_1$**		**DM$_2$**				**DM$_3$**	
Method	**UTADIS$_1$**	**UTADIS$_2$**	**UTADIS**	**NM-1**	**NM-2**	**NM-3**	**UTADIS**	**INT-1**
δ^*	0.098	0.094	0.153	0.004	0.063	0.072	0.016	0.071
APCA	0.238	0.225	0.381	0.158	0.158	0.158	0.476	0.150
CAR	0.095	0.100	0.143	0	0	0	0.143	0.050
ECAI	0.618	0.630	0.790	0.781	0.781	0.775	0.848	0.802
MCAI	0.621	0.665	0.783	0.449	0.894	0.796	0.898	0.862
CCA	0.619	0.450	0.381	0.947	0.842	0.632	0.905	0.850

Table 24: The possible assignments PCA and class acceptability indices CAI' for DM$_2$ when using the primary UTADIS model for deriving the recommendation.

DM	**DM$_2$**				
Method	**UTADIS**				
Alternative	**PCA$_{\mathcal{U}^{\mathcal{R}}}$(a)**	**CAI$'$(a, C$_1$)**	**CAI$'$(a, C$_2$)**	**CAI$'$(a, C$_3$)**	**ECAI$_{\text{alt}}$(a)**
a_1	$\{C_2, C_3\}$	0	4.00E-05	**1**	5.84E-04
a_4	$\{C_1, C_2, C_3\}$	0.035	**0.791**	0.174	0.876
a_6	$\{C_1, C_2, C_3\}$	0.118	**0.881**	2.40E-04	0.528
a_7	$\{C_1, C_2, C_3\}$	6.00E-05	0.030	**0.970**	0.195
a_9	$\{C_1, C_2\}$	0.167	**0.833**	0	0.651
a_{10}	$\{C_3\}$	0	0	**1**	0.000
a_{11}	$\{C_1, C_2\}$	0.890	**0.110**	0	0.500
a_{14}	$\{C_1, C_2\}$	0.306	**0.694**	0	0.889
a_{15}	$\{C_1, C_2\}$	0.937	**0.063**	0	0.339
a_{16}	$\{C_1, C_2\}$	**1**	1.80E-04	0	0.002
a_{18}	$\{C_1, C_2\}$	0.730	**0.270**	0	0.841
a_{21}	$\{C_2, C_3\}$	0	**0.905**	0.095	0.453
a_{22}	$\{C_1, C_2, C_3\}$	0.002	**0.933**	0.064	0.365
a_{23}	$\{C_3\}$	0	0	**1**	0.000
a_{24}	$\{C_1, C_2\}$	0.034	**0.966**	0	0.214
a_{25}	$\{C_1, C_2\}$	0.005	**0.995**	0	0.045
a_{26}	$\{C_1, C_2, C_3\}$	5.70E-04	**0.898**	0.101	0.480
a_{27}	$\{C_1, C_2, C_3\}$	0.858	**0.142**	2.00E-05	0.590
a_{28}	$\{C_1, C_2, C_3\}$	**1**	5.00E-05	3.00E-05	0.001
a_{29}	$\{C_1, C_2, C_3\}$	**0.997**	0.003	6.00E-05	0.030
a_{30}	$\{C_3\}$	0	0	**1**	0.000

the results obtained using UTADIS and NM-2, the latter proposed a less preferred class for more alternatives. Specifically, apart from the change for reference alternatives a_1 and a_4, five additional alternatives (a_{11}, a_{15}, a_{18}, a_{21}, a_{27}) were assigned to C_1, whereas another one option (a_7) was placed in C_2.

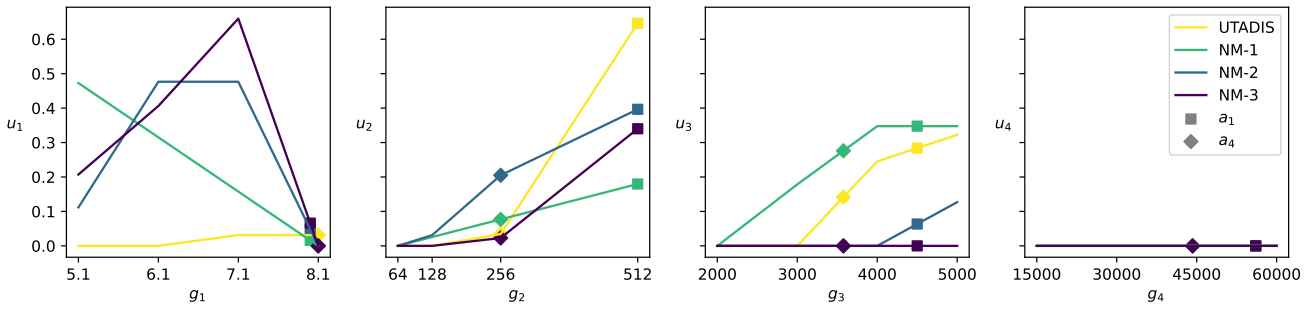


Figure 15: Marginal value functions obtained in the two iterations for DM_2.

Analysis for DM_3. When reproducing the nine reference assignments desired by DM_3, UTADIS delivered the model presented in Table 22 and Figure 16. It indicates that the classes are well-separated ($t_1 = 0.442$ and $t_2 = 0.714$), and each criterion significantly impacts the alternatives' comprehensive scores. Nevertheless, the DM did not accept this solution as he felt that the preference for phones with a low price and a high battery capacity was too low. Therefore, he indicated that a_{15} should be assigned to C_3 motivated by combining one of the highest performances on g_3 and one of the lowest on g_4. In this case, UTADIS was unable to find a satisfying solution.

In line with the experimental-based framework proposed in Section 4.1, we attempted to apply INT-1. It suggested a model that respected the pre-defined monotonicity constraints but incorporated a positive interaction between g_3 and g_4. The solution was approved by DM_3 as it suitably represented the desired impact of battery capacity and price on the phone's comprehensive quality.

In general, the interaction component for pairs of criteria allows the introduction of additional dependencies into the model that cannot be expressed in the standard UTADIS. In particular, one can increase or decrease the preference for combinations of performances on two criteria using bonuses or penalties. In this case, DM_3 wanted to emphasize the positive perception of phones with both large battery capacity and low price, as the first model did not properly reflect these preferences. In DM's opinion, better performance on only one of the two criteria (g_3 or g_4) was less important than a favorable combination of both values of these attributes, hence the introduction of synergy for these criteria. In a way, the DM desired to assign a bonus to models with large battery capacity and low prices. The analysis of various phones confirmed that such combinations were possible, and hence, these alternatives should be promoted by increasing their comprehensive values and ranks. The DM did not want to pay more for a better battery, as he could have a decent one while spending less.

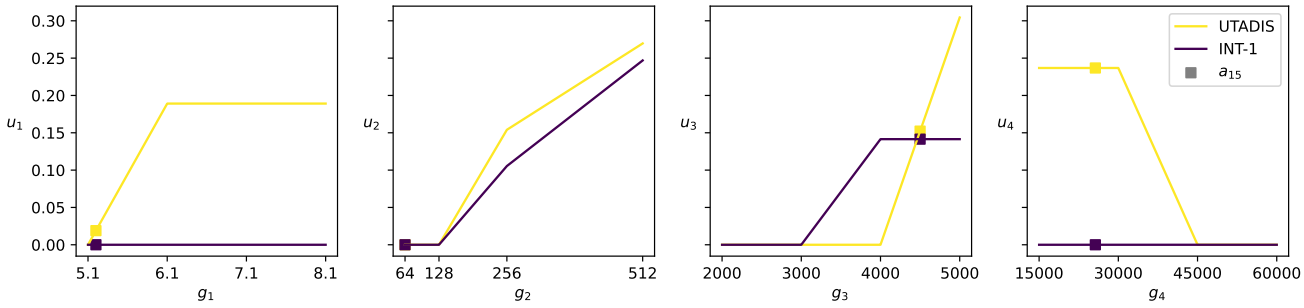


Figure 16: Marginal value functions obtained in the two iterations for DM_3.

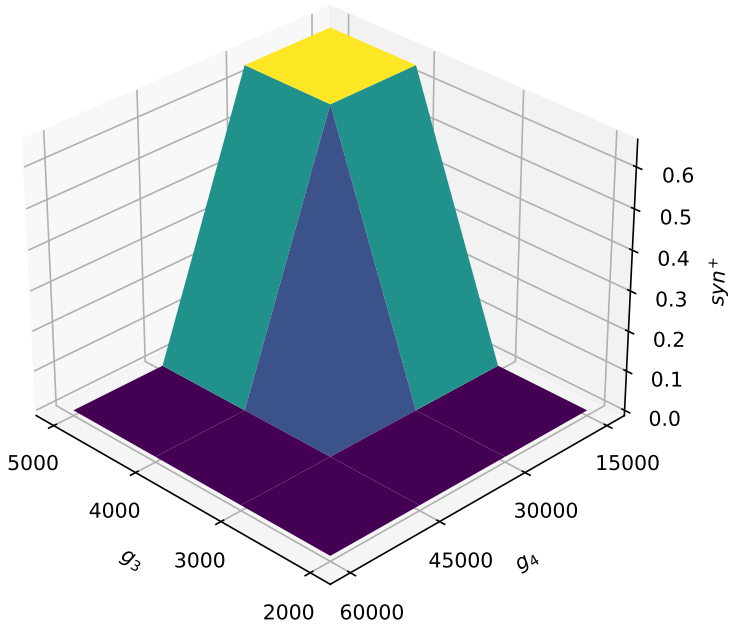


Figure 17: Positive interaction function $syn^+_{g_3,g_4}$ between criteria g_3 and g_4 for results obtained by the INT-1 method for DM_3.

The respective MVFs are given in Figure 16, and the three-dimensional plot of the $syn^+_{g_3,g_4}$ interaction function is shown in Figure 17. The latter has the greatest maximal impact on the alternatives' comprehensive scores. Such maximum scores of $syn^+_{g_3,g_4}$ are assigned to all alternatives with $g_3(a) \geq 4000$ and $g_4(a) \leq 30000$, which coincides with the DM's preferences. The introduction of synergy brought the desired effect and better reflected the DM's

value system. At the same, the importance of the individual criteria was reduced (e.g., the impact of u_1 and u_4 considered separately was found to be negligible).

Table 23 reveals an increase of δ^*. This results from using a more expressive model than the primary UTADIS method. At the same time, we observed a decrease in the values of all robustness measures. For example, the possible assignments are now more diverse, as confirmed by the deterioration of $APCA$ from 0.476 to 0.150, and the number of necessary assignments decreased from three to one among all non-reference alternatives (CAR decreased from 0.143 to 0.050). When it comes to CCA, for UTADIS – $CCA(U) = \frac{19}{21} = 0.905$ and for INT-1 – $CCA(U) = \frac{17}{20} = 0.850$. For the former, only the assignments of a_{18} and a_{27} are unconfirmed by the reference alternatives. This is because their comprehensive values of 0.431 and 0.436 are between $C_1^{UB} = 0.426$ (attained by reference alternatives a_{13} and a_{20}) and $C_2^{LB} = 0.459$ (attained by reference alternatives a_2 and a_3). For INT-1, the unconfirmed assignments are associated with three non-reference alternatives (a_4 with the $U = 0.205$ between $C_1^{UB} = 0.105$ and $C_2^{LB} = 0.247$, and a_{11} and a_9 with comprehensive values of 0.628 and 0.689 between $C_2^{UB} = 0.612$ and $C_3^{LB} = 0.753$).

The above examples showed the impact of DM's indirect preferences on the attained sorting results. They also emphasized the importance of collecting reliable assignment examples representing the DM's value system. Also, we illustrated how important it is to involve the DMs, possibly supported by an analyst, in the process of selecting a model that would reflect their decision policy in the best way.

6. Conclusions

While many MCDA methods have been proposed over the years, the focus primarily has been on their objective features, with less attention to their performance in practice. This paper aimed to address this gap by focusing on the performance aspects of MCDA methods. We accounted for preference disaggregation approaches in the context of multiple criteria sorting. Specifically, we considered the family of UTADIS methods, inferring a value-based aggregation model and class thresholds from the DM's reference assignments. These approaches are known for their intuitiveness, interpretability, and convenience to exploit their outputs.

We discussed a basic variant of UTADIS, where an alternative's score is computed using an additive value function. Then, we extended it by suitably adapting proposals existing in the context of ranking problems. On the one hand, we discussed how to incorporate the dependencies between criteria while accepting various assumptions on the number of active interactions for each attribute. On the other hand, we presented how to discover the preference directions for various criteria while tolerating that marginal functions may be non-monotonic. Overall, we considered six variants of UTADIS that differed in their assumptions, influencing their performance.

We introduced the concepts reflecting the performance of multiple criteria sorting methods in real-world decision-making – the model's expressiveness and the robustness of the delivered recommendations. Expressiveness refers to the ability of a method to accurately represent the preferences of DMs, while robustness stands for the stability and validity of the recommendations across different conditions. We proposed seven measures capturing the performance in these two dimensions. They were used to quantify the outputs of an extensive computational experiment. This way, we proposed the sorting-based counterpart of the framework proposed in [29] in the context of ranking and choice problems.

The best performance in terms of expressiveness was attained by the INT-∞ method, which does not pose any limits on the number of interacting criteria pairs. It was followed by the NM methods, admitting non-monotonicity. The least expressive was the basic variant of UTADIS. Regarding robustness, the latter approach delivered the most stable recommendations for the scenarios handled by all approaches. In the remaining cases, the best performer depended on the number of interactions needed to ensure consistency with the DM's preferences. When it was not higher than two, it was better to use the interaction-oriented methods. Otherwise, the NM approaches led to more robust results.

We used the above observations to support decision analysts in selecting the appropriate MCDA model. On the one hand, the guidelines are based on the nature of supplied preferences for a specific decision problem. We

aimed to attain the most robust recommendation derived from the model whose complexity is adjusted to the DM's assignment examples. On the other hand, the recommendations should always be confronted against the DM's requirements and problem characteristics. To support such a confrontation, we formulated a set of questions and answers leading to the selection of various UTADIS variants. The essential ones refer to the features regarding problem formulation and preference model.

We confirmed that expressiveness and robustness are conflicting. Moreover, the challenge of comprehensive reproduction of DM's indirect preference increased with more classes and reference assignments, fewer criteria and characteristic points of marginal functions, and randomly generated performance of non-dominated alternatives. In turn, the robustness of the recommendation is positively affected by richer preference information and negatively impacted by a greater number of model parameters, which depends on the number of criteria and characteristic points.

In our experimental study, the non-dominated alternatives were randomly assigned to the ordered classes. Consequently, the model's goodness of fit might be negatively influenced as the number of alternatives increases. The relationship between the fit and the number of alternatives while controlling for the above factor, e.g., through a systematic assignment mechanism respecting the relation between potential dominance and desired classes, is worth exploring.

We envisage the following other directions for future research. First, it is possible to extend the analysis to other value-based methods that handle interactions between criteria [35] or non-monotonicity of per-criteria preferences [23]. Second, it would be interesting to account for the methods that tolerate inconsistency with the DM's preference information instead of being required to reproduce all assignment examples [28]. This way, we could capture the trade-off between accepting some positive misclassification error and increasing the preference model's complexity. Third, in situations where preference information from multiple DMs is available, such as our case study, it would be interesting to see how population-level insights could be exploited to improve individual-level results [15]. Fourth, it is desired to consider outranking-based multiple criteria methods such as ELECTRE TRI-B or TRI-C. However, modeling non-monotonicity with their use is still in its infancy [37]. Another challenge consists of elaborating a value-driven method that simultaneously considers interactions and non-monotonicity. In this case, the mathematical constraints would be non-linear, and hence one would need a heuristic optimization method. Finally, it would be interesting to extend the experiments concerning the predictive accuracy presented in [14, 48] to more advanced UTADIS variants that admit interactions and non-monotonicity. This would require knowing or generating the ground truth (i.e., the actual assignments of hold-out alternatives) and investigating different ways for selecting a single representative model (e.g., the most discriminant, average, central, parsimonious, or robust) [32, 48]. The latter is essential as even in the context of the basic UTADIS variant, the reported differences in the classification accuracy between the best and the worst-performing procedures were as significant as several percent.

Acknowledgments

We thank three anonymous referees for their constructive comments. Miłosz Kadziński was supported by the Polish Ministry of Science and Higher Education (grant no. 0311/SBAD). Michał Wójcik acknowledges financial support from the Polish National Science Center under the SONATA BIS project (grant no. DEC-2019/34/E/HS4/00045). Mohammad Ghaderi acknowledges financial support from the Spanish Agencia Estatal de Investigación (AEI) through the RYC2021-034981-I Grant.

References

[1] Alvarez, P. A., Ishizaka, A., and Martínez, L. (2021). Multiple-criteria decision-making sorting methods: A survey. *Expert Systems with Applications*, 183:115368.

[2] Angilella, S., Bottero, M., Corrente, S., Ferretti, V., Greco, S., and Lami, I. M. (2016). Non additive robust ordinal regression for urban and territorial planning: an application for siting an urban waste landfill. *Annals of Operations Research*, 245(1):427–456.

[3] Angilella, S., Corrente, S., and Greco, S. (2015). Stochastic multiobjective acceptability analysis for the Choquet integral preference model and the scale construction problem. *European Journal of Operational Research*, 240(1):172–182.

[4] Branke, J., Greco, S., Słowiński, R., and Zielniewicz, P. (2015). Learning value functions in interactive evolutionary multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 19(1):88–102.

[5] Charnes, A. and Cooper, W. W. (1962). Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9(3-4):181–186.

[6] 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. *European Journal of Operational Research*, 302(2):633–651.

[7] Ciomek, K. and Kadziński, M. (2021). Polyrun: A Java library for sampling from the bounded convex polytopes. *SoftwareX*, 13:100659.

[8] Devaud, J., Groussaud, G., and Jacquet-Lagrèze, E. (1980). UTADIS: Une méthode de construction de fonctions d'utilité additives rendant compte de jugements globaux. In *EURO Working Group on MCDA, Bochum, Germany*.

[9] Dimitras, A. (2002). Evaluation of Greek Construction Companies' Securities Using UTADIS Method. *European Research Studies Journal*, 5(1-2):95–107.

[10] Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and Unsupervised Discretization of Continuous Features. In Prieditis, A. and Russell, S., editors, *Machine Learning Proceedings 1995*, pages 194–202. Morgan Kaufmann, San Francisco (CA).

[11] Doumpos, M. and Zopounidis, C. (2002). *Multicriteria Decision Aid Classification Methods*. Kluwer Academic Publishers, Dordrecht.

[12] Doumpos, M. and Zopounidis, C. (2014). *The Robustness Concern in Preference Disaggregation Approaches for Decision Aiding: An Overview*, pages 157–177. Springer, New York, NY.

[13] Doumpos, M. and Zopounidis, C. (2018). Disaggregation Approaches for Multicriteria Classification: An Overview. In Matsatsinis, N. and Grigoroudis, E., editors, *Preference Disaggregation in Multiple Criteria Decision Analysis: Essays in Honor of Yannis Siskos*, pages 77–94. Springer, Cham.

[14] Doumpos, M., Zopounidis, C., and Galariotis, E. (2014). Inferring robust decision models in multicriteria classification problems: An experimental analysis. *European Journal of Operational Research*, 236(2):601–611.

[15] Ghaderi, M. and Kadziński, M. (2021). Incorporating uncovered structural patterns in value functions construction. *Omega*, 99:102203.

[16] Ghaderi, M., Ruiz, F., and Agell, N. (2017). A linear programming approach for learning non-monotonic additive value functions in multiple criteria decision aiding. *European Journal of Operational Research*, 259(3):1073 – 1084.

[17] Grabisch, M. (1996). The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, 89(3):445–456.

[18] Greco, S., Ehrgott, M., and Figueira, J. (2016). *Multiple Criteria Decision Analysis – State of the Art Surveys*. International Series in Operations Research & Management Science. Springer, New York.

[19] Greco, S., Mousseau, V., and Słowiński, R. (2010). Multiple criteria sorting with a set of additive value functions. *European Journal of Operational Research*, 207(3):1455 – 1470.

[20] Greco, S., Mousseau, V., and Słowiński, R. (2011). Parsimonious preference models for robust ordinal regression. In *EURO Working Group on MCDA, Yverdon, Switzerland*.

[21] Greco, S., Mousseau, V., and Słowiński, R. (2014). Robust ordinal regression for value functions handling interacting criteria. *European Journal of Operational Research*, 239(3):711 – 730.

[22] Greco, S., Słowiński, R., and Wallenius, J. (2024). Fifty years of multiple criteria decision analysis: From classical methods to robust ordinal regression. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2024.07.038.

[23] Guo, M., Liao, X., and Liu, J. (2019). A progressive sorting approach for multiple criteria decision aiding in the presence of non-monotonic preferences. *Expert Systems with Applications*, 123:1–17.

[24] Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric Statistical Methods*, volume 751. John Wiley & Sons, Hoboken, New Jersey.

[25] Ishizaka, A. and Nemery, P. (2013). *Multi-Criteria Decision Analysis: Methods and Software*. International Series in Operations Research & Management Science. John Wiley & Sons, Ltd, West Sussex, United Kingdom.

[26] Jacquet-Lagrèze, E. and 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.

[27] Kadziński, M. and Ciomek, K. (2021). Active learning strategies for interactive elicitation of assignment examples for threshold-based multiple criteria sorting. *European Journal of Operational Research*, 293(2):658–680.

[28] Kadziński, M., Ghaderi, M., and Dabrowski, M. (2020). Contingent preference disaggregation model for multiple criteria sorting problem. *European Journal of Operational Research*, 281(2):369–387.

[29] Kadziński, M., Ghaderi, M., Wasikowski, J., and Agell, N. (2017). Expressiveness and robustness measures for the evaluation of an additive value function in multiple criteria preference disaggregation methods: An experimental analysis. *Computers & Operations Research*, 87:146–164.

[30] Kadziński, M., Martyn, K., Cinelli, M., Słowiński, R., Corrente, S., and Greco, S. (2021). Preference disaggregation method for value-based multi-decision sorting problems with a real-world application in nanotechnology. *Knowledge-Based Systems*, 218:106879.

[31] Kadziński, M. and Tervonen, T. (2013). Stochastic ordinal regression for multiple criteria sorting problems. *Decision Support Systems*, 55(1):55–66.

[32] Kadziński, M., Wójcik, M., and Ciomek, K. (2022). Review and experimental comparison of ranking and choice procedures for constructing a univocal recommendation in a preference disaggregation setting. *Omega*, 113:102715.

[33] Keeney, R. and Raiffa, H. (1993). *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press.

[34] Köksalan, M. and Bilgin Özpeynirci, S. (2009). An interactive sorting method for additive utility functions. *Computers & Operations Research*, 36(9):2565–2572.

[35] Liu, J., Kadziński, M., Liao, X., and Mao, X. (2021). Data-Driven Preference Learning Methods for Value-Driven Multiple Criteria Sorting with Interacting Criteria. *INFORMS Journal on Computing*, 33(2):586–606.

[36] Manshadi, E. D., Mehregan, M. R., and Safari, H. (2015). Supplier Classification Using UTADIS Method Based on Performance Criteria. *International Journal of Academic Research in Business and Social Sciences*, 5(2):31–45.

[37] Minoungou, P., Mousseau, V., Ouerdane, W., and Scotton, P. (2021). Learning MR-Sort Models from Non-Monotone Data. `https://doi.org/10.48550/arXiv.2107.09668`.

[38] Mousseau, V., Dias, L. C., and Figueira, J. (2006). Dealing with inconsistent judgments in multiple criteria sorting models. *4OR*, 4(2):145–158.

[39] Palha, R. P., de Almeida, A. T., and Alencar, L. H. (2016). A Model for Sorting Activities to Be Outsourced in Civil Construction Based on ROR-UTADIS. *Mathematical Problems in Engineering*, 2016:9236414.

[40] Pozdniakov, A. (2021). Mobile Phones Data. `https://www.kaggle.com/datasets/artempozdniakov/ukrainian-market-mobile-phones-data`. Accessed: 2023-10-10.

[41] Roy, B. (2010). Robustness in operational research and decision aiding: A multi-faceted issue. *European Journal of Operational Research*, 200(3):629 – 638.

[42] Ru, Z., Liu, J., Kadziński, M., and Liao, X. (2023). Probabilistic ordinal regression methods for multiple criteria sorting admitting certain and uncertain preferences. *European Journal of Operational Research*, 311(2):596–616.

[43] Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

[44] Siskos, Y., Grigoroudis, E., and Matsatsinis, N. (2005). UTA Methods. In Figueira, J., Greco, S., and Ehrgott, M., editors, *Multiple Criteria Decision Analysis: State of the Art Surveys*, pages 297–344. Springer Verlag, Boston.

[45] Smith, R. L. (1984). Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed Over Bounded Regions. *Operations Research*, 32(6):1296–1308.

[46] The, A. N. and Mousseau, V. (2002). Using assignment examples to infer category limits for the ELECTRE TRI method. *Journal of Multi-Criteria Decision Analysis*, 11(1):29–43.

[47] Vetschera, R., Chen, Y., Hipel, K. W., and Kilgour, M. (2010). Robustness and information levels in case-based multiple criteria sorting. *European Journal of Operational Research*, 202(3):841–852.

[48] Wójcik, M., Kadziński, M., and Ciomek, K. (2023). 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.

[49] Zopounidis, C. and Doumpos, M. (2000). PREFDIS: a multicriteria decision support system for sorting decision problems. *Computers & Operations Research*, 27(7):779–797.

[50] Zopounidis, C. and Doumpos, M. (2001). A preference disaggregation decision support system for financial classification problems. *European Journal of Operational Research*, 130(2):402–413.

[51] Zopounidis, C. and Doumpos, M. (2020). Multicriteria sorting methods. In Pardalos, P. M. and Prokopyev, O. A., editors, *Encyclopedia of Optimization*, pages 1–20. Springer International Publishing, Cham.

Supplementary material [P4]

From investigation of expressiveness and robustness to a comprehensive value-based framework for multiple criteria sorting problems – eAppendix

Miłosz Kadziński[a,*], Michał Wójcik[a], Mohammad Ghaderi[b]

[a] *Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznań, Poland*
[b] *Department of Economics and Business, Pompeu Fabra University, 08005 Barcelona, Spain*

1. Analysis of robustness for the problem instances handled by all methods

Tables 1 and 2 show the CAR, ECAI, and CCA values for the subset of problem instances handled by all methods. The observations for CAR are analogous to those reported for APCA in the main paper. The only exception is the opposite trend regarding the dependence of the measure on the increasing number of classes when CAR decreases slightly. This difference is probably due to the increased difficulty of obtaining unambiguous recommendations confirmed by all compatible models as the number of classes increases. CAR is also more restrictive because it rewards alternatives that can be assigned to only one class. In turn, APCA offers greater granularity, rewarding models in which, e.g., an alternative can be assigned to two out of five classes.

ECAI also confirms the advantage of UTADIS and INT over the NM approaches. The only noticeable difference is that, unlike before, the results for NM-3 are worse than those obtained by NM-1 and NM-2. This is due to the different distribution of all models in the space of consistent solutions $\mathcal{U}^R$ that lowers consistency between the recommendations they suggest.

Observations for CCA values also indicate an advantage of the primary UTADIS method over non-monotonic approaches in the analyzed subset of problem instances. Among the NM methods, NM-1 achieves the best results. This is related to the attained δ^* values. Low δ^* means that the comprehensive values of the reference alternatives cover a large portion of all possible comprehensive values, leaving only small value ranges for the *uncertain* assignments.

Table 1: Average CAR, ECAI, and CCA values for the problem instances handled by all methods and different numbers of classes and criteria.

		Number of classes				Number of criteria			
Measure	**All settings**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
CAR									
UTADIS + INT	0.115	0.125	0.116	0.107	0.094	0.219	0.129	0.093	0.074
NM	0.030	0.036	0.027	0.026	0.021	0.087	0.034	0.018	0.009
ECAI									
UTADIS + INT	0.633	0.590	0.646	0.670	0.686	0.773	0.653	0.604	0.577
NM-1, NM-2	0.464	0.408	0.478	0.515	0.537	0.699	0.501	0.416	0.367
NM-3	0.416	0.375	0.427	0.454	0.469	0.635	0.446	0.372	0.328
CCA									
UTADIS + INT	0.669	0.733	0.660	0.611	0.570	0.717	0.696	0.659	0.633
NM-1	0.635	0.685	0.631	0.589	0.560	0.677	0.665	0.632	0.596
NM-2	0.493	0.527	0.488	0.463	0.441	0.630	0.532	0.465	0.422
NM-3	0.484	0.520	0.477	0.454	0.430	0.606	0.521	0.463	0.416

2. Analysis of robustness for the problem instances handled by the NM and INT methods

The average values of CAR, ECAI, and CCA for the problem instances handled by all methods except UTADIS are shown in Tables 3 and 4. Similarly to the observations made for APCA in the main paper, CAR and ECAI exhibit different

*Corresponding author: Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland. Tel. +48-61 665 3022.

Email addresses: milosz.kadzinski@cs.put.poznan.pl (Miłosz Kadziński), michal.wojcik@cs.put.poznan.pl (Michał Wójcik), mohammad.ghaderi@upf.edu (Mohammad Ghaderi)

Table 2: Average CAR, ECAI, and CCA values for the problems instances handled by all methods and different numbers of characteristic points, reference assignments per class, and performance generation algorithms.

	Number of ch. points				Number of reference assignments					Generation algorithm	
Measure	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**sphere**	**random**
CAR											
UTADIS + INT	0.276	0.129	0.089	0.061	0.068	0.121	0.147	0.162	0.171	0.082	0.161
NM	0.125	0.028	0.012	0.006	0.014	0.031	0.041	0.047	0.049	0.032	0.027
ECAI											
UTADIS + INT	0.685	0.640	0.634	0.606	0.548	0.650	0.693	0.714	0.724	0.610	0.665
NM-1, NM-2	0.624	0.494	0.439	0.400	0.376	0.471	0.527	0.555	0.578	0.517	0.391
NM-3	0.569	0.443	0.394	0.353	0.328	0.423	0.477	0.509	0.533	0.464	0.349
CCA											
UTADIS + INT	0.650	0.669	0.677	0.669	0.557	0.683	0.741	0.785	0.805	0.661	0.680
NM-1	0.569	0.641	0.648	0.648	0.460	0.682	0.751	0.791	0.812	0.656	0.606
NM-2	0.595	0.527	0.481	0.438	0.323	0.506	0.609	0.672	0.714	0.541	0.425
NM-3	0.580	0.518	0.474	0.429	0.313	0.502	0.598	0.662	0.704	0.528	0.423

trends for the NM and INT methods for various numbers of criteria. Also, for bi-criteria problems, the NM methods lead to statistically significantly better values. Moreover, non-monotonic approaches have an advantage for the problem instances with performances generated by the *sphere* algorithm.

The latter scenario is also the only one for which all NM approaches perform better in the context of CCA. In all other cases, there is a slight advantage of INT-1 over INT-∞, or both approaches achieve the same results. This can be observed for 2- and 3-criteria problems, where both methods provide the same results. Moreover, in all other cases, the INT methods produce better results than the NM approaches.

Table 3: Average CAR, ECAI, and CCA values for the problem instances handled by the NM and INT methods, and different numbers of classes and criteria.

		Number of classes				Number of criteria			
Measure	**All settings**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
CAR									
NM	0.154	0.245	0.167	0.124	0.081	0.272	0.174	0.106	0.077
INT-1	0.328	0.435	0.345	0.283	0.256	0.194	0.316	0.379	0.409
INT-∞	0.305	0.419	0.328	0.261	0.214	0.194	0.316	0.350	0.349
ECAI									
NM-1, NM-2	0.705	0.733	0.712	0.701	0.671	0.829	0.739	0.667	0.599
NM-3	0.652	0.691	0.665	0.640	0.612	0.794	0.684	0.604	0.545
INT-1	0.823	0.814	0.833	0.823	0.817	0.757	0.804	0.850	0.872
INT-∞	0.810	0.808	0.827	0.810	0.791	0.757	0.804	0.837	0.838
CCA									
NM-1	0.739	0.847	0.771	0.702	0.631	0.741	0.766	0.735	0.713
NM-2	0.701	0.811	0.739	0.661	0.588	0.744	0.733	0.692	0.642
NM-3	0.701	0.813	0.730	0.663	0.597	0.745	0.730	0.691	0.646
INT-1	0.821	0.888	0.857	0.777	0.758	0.747	0.805	0.849	0.873
INT-∞	0.803	0.883	0.848	0.756	0.721	0.747	0.805	0.831	0.824

Tables 5 and 6 exhibit results for the problem instances for which INT-1 was unable to reproduce the DM's preferences. In the case of a larger number of active interactions, the advantage of INT-∞ over the NM methods in terms of CAR and ECAI is not as strong as in the previous case. Similarly to APCA, there is an apparent decrease in the average values of CAR for INT-∞. Nevertheless, the advantage over non-monotonic approaches is still observable, regardless of the analyzed problem dimension, except for 3-criteria problems. In this case, CAR and ECAI for INT-∞ are 0.154 and 0.737, while for NM-1 and NM-2, they are 0.166 and 0.742, respectively. However, the Wilcoxon signed-rank test with p-value $= 0.05$ indicates that these differences are insufficient to conclude the advantage of NM approaches in this aspect.

Regarding the CCA values, the results obtained by INT-∞ and NM-1 should be considered statistically equivalent. The results for these two methods differ in the strength of their trends depending on the problem dimension. For NM-1, the decrease in the CCA value is more noticeable as the number of criteria increases and the number of reference assignments decreases. For INT-∞, these differences are more significant when the number of classes and characteristic points change. NM-2 and NM-3 are again clearly inferior, regardless of the analyzed dimension.

Table 4: Average CAR, ECAI, and CCA values for the problem instances handled by the NM and INT methods and different numbers of characteristic points, reference assignments per class, and performance generation algorithms.

	Number of ch. points				Number of reference assignments					Generation algorithm	
Measure	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**sphere**	**random**
CAR											
NM	0.320	0.181	0.107	0.045	0.191	0.144	0.128	0.166	0.161	0.257	0.086
INT-1	0.532	0.335	0.264	0.235	0.302	0.299	0.322	0.368	0.358	0.206	0.408
INT-∞	0.507	0.314	0.244	0.207	0.299	0.278	0.291	0.333	0.340	0.198	0.375
ECAI											
NM-1, NM-2	0.781	0.744	0.685	0.621	0.729	0.688	0.690	0.712	0.725	0.835	0.620
NM-3	0.737	0.700	0.627	0.558	0.683	0.632	0.628	0.664	0.682	0.803	0.554
INT-1	0.861	0.821	0.805	0.815	0.800	0.818	0.823	0.838	0.829	0.750	0.870
INT-∞	0.852	0.811	0.794	0.796	0.799	0.806	0.808	0.819	0.821	0.747	0.852
CCA											
NM-1	0.726	0.771	0.732	0.716	0.579	0.713	0.759	0.797	0.814	0.780	0.712
NM-2	0.724	0.728	0.699	0.653	0.556	0.657	0.711	0.767	0.803	0.787	0.645
NM-3	0.734	0.730	0.699	0.645	0.573	0.661	0.710	0.763	0.788	0.796	0.640
INT-1	0.804	0.821	0.817	0.837	0.677	0.794	0.844	0.877	0.882	0.763	0.859
INT-∞	0.794	0.806	0.801	0.809	0.674	0.777	0.822	0.850	0.870	0.759	0.832

Table 5: Average CAR, ECAI, and CCA values for the problem instances handled by the NM methods and INT-∞ and different numbers of classes and criteria.

		Number of classes				Number of criteria			
Measure	**All settings**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
CAR									
NM	0.130	0.267	0.143	0.108	0.087		0.166	0.137	0.094
INT-∞	0.161	0.404	0.188	0.111	0.089		0.154	0.164	0.164
ECAI									
NM-1, NM-2	0.688	0.702	0.685	0.686	0.685		0.742	0.689	0.639
NM-3	0.625	0.665	0.628	0.614	0.617		0.675	0.628	0.580
INT-∞	0.759	0.820	0.775	0.749	0.730		0.737	0.765	0.771
CCA									
NM-1	0.737	0.859	0.766	0.725	0.672		0.761	0.735	0.718
NM-2	0.701	0.805	0.732	0.692	0.642		0.740	0.705	0.665
NM-3	0.694	0.799	0.726	0.682	0.634		0.732	0.697	0.658
INT-∞	0.736	0.876	0.785	0.716	0.656		0.744	0.738	0.728

3. Robustness and expressiveness within the NM and INT groups of methods

3.1. Analysis for the problem instances handled by the NM methods

The average values of quality measures for the problem instances handled only by the NM methods are shown in Tables 7 and 8. The results confirm previous observations, i.e., the greatest flexibility of NM-3, associated with the highest δ^* values, and the best robustness of recommendations obtained by NM-2, which is preceded only by NM-1 in the case of CCA. This fact is related to the lower values of δ^* for NM-1, which causes the *uncertainty* intervals between extreme reference assignments to neighboring classes to be much narrower, leading to greater CCA values than for NM-2. The trends and relationships between individual methods are preserved regardless of the analyzed problem dimension. The only exception is the CCA value for instances with performance generated by the *sphere* algorithm. Then, NM-1 obtains significantly worse results than the other two approaches.

3.2. Analysis for the problem instances handled by the INT methods

Tables 9 and 10 show the average values of quality measures for the problem instances handled by the INT methods. The δ^* values obtained by both approaches confirm the observations made in the preference recoverability analysis. That is, INT-∞ has higher expressiveness than INT-1. Conversely, the remaining quality measures confirm the higher robustness of INT-1. Except for instances with low complexity (e.g., problems with two criteria or one reference assignment per class) where both methods return the same results, INT-1 performs slightly better than INT-∞. Hence, INT-1 should be preferred over INT-∞ if its use is possible.

Table 6: Average CAR, ECAI, and CCA values for the problem instances handled by the NM methods and INT-∞ and different numbers of characteristic points, reference assignments per class, and performance generation algorithms.

	Number of ch. points				Number of reference assignments					Generation algorithm	
Measure	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**sphere**	**random**
CAR											
NM	0.313	0.150	0.075	0.033	0.123	0.129	0.136	0.128	0.131	0.345	0.114
INT-∞	0.433	0.148	0.082	0.042	0.242	0.175	0.137	0.154	0.154	0.384	0.144
ECAI											
NM-1, NM-2	0.757	0.721	0.673	0.626	0.636	0.652	0.701	0.700	0.711	0.805	0.679
NM-3	0.718	0.663	0.603	0.550	0.565	0.591	0.637	0.636	0.653	0.767	0.614
INT-∞	0.844	0.745	0.741	0.722	0.761	0.746	0.743	0.773	0.773	0.825	0.754
CCA											
NM-1	0.733	0.783	0.739	0.704	0.517	0.694	0.765	0.768	0.790	0.752	0.736
NM-2	0.735	0.738	0.700	0.653	0.530	0.633	0.723	0.736	0.767	0.736	0.699
NM-3	0.733	0.729	0.689	0.645	0.527	0.625	0.714	0.730	0.758	0.737	0.691
INT-∞	0.810	0.741	0.724	0.690	0.660	0.695	0.724	0.770	0.781	0.809	0.731

Table 7: Average values of six performance measures for the problem instances handled by the NM methods and different numbers of classes and criteria.

		Number of classes				Number of criteria			
Measure	**All settings**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
$\boldsymbol{\delta^*}$									
NM-1	0.015	0.019	0.016	0.013	0.010	0.007	0.020	0.023	0.023
NM-2	0.034	0.050	0.031	0.022	0.016	0.015	0.039	0.052	0.052
NM-3	0.036	0.054	0.032	0.023	0.017	0.015	0.044	0.056	0.055
APCA									
NM	0.187	0.186	0.178	0.182	0.222	0.344	0.154	0.043	0.012
CAR									
NM	0.148	0.186	0.139	0.108	0.132	0.270	0.131	0.032	0.008
ECAI									
NM-1, NM-2	0.641	0.611	0.639	0.661	0.691	0.802	0.654	0.504	0.419
NM-3	0.588	0.561	0.587	0.606	0.630	0.746	0.599	0.455	0.371
MCAI									
NM-1	0.715	0.783	0.703	0.672	0.637	0.800	0.735	0.653	0.584
NM-2	0.797	0.848	0.789	0.753	0.747	0.881	0.821	0.735	0.662
NM-3	0.747	0.818	0.738	0.699	0.657	0.817	0.778	0.694	0.629
CCA									
NM-1	0.750	0.843	0.750	0.665	0.637	0.761	0.781	0.746	0.709
NM-2	0.666	0.742	0.668	0.599	0.567	0.755	0.710	0.590	0.522
NM-3	0.657	0.732	0.654	0.599	0.560	0.742	0.692	0.599	0.511

4. The correlation between APCA and CAR

APCA and CAR are closely related, focussing on the unambiguity of delivered recommendations. For binary classification problems, their values are equal. In general, APCA can be viewed as an upper bound on CAR. This is because the value of APCA increases with each alternative whose $|PCA_{\mathcal{U}^R}(a)|$ is less than the number of classes p, while the value of CAR increases only when $|PCA_{\mathcal{U}^R}(a)|$ equals 1.

Considering all simulation experiments, 68.36% runs were associated with the same APCA and CAR values (for 8.78% scenarios, these values were equal to 0, and only for 0.13% cases, they were equal to 1). For the remaining 31.64% runs, APCA was greater than CAR. The average value of APCA across all runs was 0.126, and for CAR, it was 0.099. The Pearson correlation coefficient – defined on a scale -1 and 1 – for both quality measures is 0.824. The values of these measures obtained in all simulation runs are visible in Figure 1.

Table 8: Average values of six performance measures for the problem instances handled by the NM methods and different numbers of characteristic points, reference assignments per class, and performance generation algorithms.

	Number of ch. points				Number of reference assignments					Generation algorithm	
Measure	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**sphere**	**random**
δ^*											
NM-1	0.003	0.015	0.016	0.017	0.008	0.012	0.016	0.018	0.019	0.004	0.016
NM-2	0.015	0.029	0.035	0.038	0.024	0.028	0.035	0.038	0.037	0.008	0.035
NM-3	0.015	0.034	0.037	0.039	0.026	0.030	0.039	0.039	0.039	0.009	0.037
APCA											
NM	0.670	0.268	0.134	0.110	0.276	0.232	0.176	0.167	0.132	0.247	0.184
CAR											
NM	0.572	0.215	0.104	0.080	0.160	0.172	0.140	0.150	0.123	0.190	0.146
ECAI											
NM-1, NM-2	0.845	0.700	0.622	0.588	0.714	0.657	0.640	0.616	0.617	0.812	0.632
NM-3	0.822	0.653	0.564	0.531	0.646	0.600	0.590	0.568	0.566	0.761	0.578
MCAI											
NM-1	0.894	0.778	0.697	0.663	0.689	0.688	0.709	0.730	0.749	0.782	0.712
NM-2	0.921	0.842	0.790	0.753	0.782	0.771	0.796	0.805	0.823	0.875	0.792
NM-3	0.904	0.797	0.740	0.695	0.694	0.714	0.751	0.763	0.785	0.829	0.742
CCA											
NM-1	0.737	0.769	0.762	0.727	0.548	0.729	0.764	0.791	0.806	0.775	0.748
NM-2	0.732	0.700	0.669	0.629	0.506	0.632	0.675	0.694	0.739	0.792	0.659
NM-3	0.733	0.693	0.656	0.622	0.500	0.614	0.665	0.698	0.725	0.821	0.648

Table 9: Average values of six performance measures for the problem instances handled by the INT methods and different numbers of classes and criteria.

		Number of classes				Number of criteria			
Measure	**All settings**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
δ^*									
INT-1	0.002	0.005	0.003	0.002	0.002	0.002	0.003	0.002	0.002
INT-∞	0.003	0.005	0.003	0.002	0.002	0.002	0.003	0.003	0.003
APCA									
INT-1	0.323	0.478	0.409	0.302	0.260	0.323	0.297	0.322	0.359
INT-∞	0.306	0.474	0.393	0.282	0.241	0.323	0.298	0.293	0.317
CAR									
INT-1	0.248	0.478	0.350	0.219	0.167	0.236	0.232	0.252	0.278
INT-∞	0.232	0.474	0.333	0.201	0.150	0.236	0.232	0.227	0.238
ECAI									
INT-1	0.810	0.823	0.836	0.808	0.796	0.818	0.794	0.807	0.829
INT-∞	0.802	0.821	0.831	0.799	0.785	0.818	0.793	0.792	0.811
MCAI									
INT-1	0.813	0.902	0.861	0.804	0.775	0.803	0.794	0.820	0.836
INT-∞	0.806	0.902	0.858	0.796	0.765	0.803	0.794	0.810	0.819
CCA									
INT-1	0.820	0.891	0.863	0.811	0.789	0.769	0.824	0.833	0.845
INT-∞	0.810	0.889	0.855	0.802	0.775	0.769	0.824	0.815	0.822

Table 10: Average values of six performance measures for the problem instances handled by the INT methods and different numbers of characteristic points, reference assignments per class, and performance generation algorithms.

	Number of ch. points				Number of reference assignments					Generation algorithm	
Measure	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**sphere**	**random**
δ^*											
INT-1	0.004	0.002	0.002	0.002	0.003	0.003	0.002	0.002	0.002	0.002	0.005
INT-∞	0.005	0.003	0.002	0.002	0.003	0.003	0.003	0.003	0.002	0.002	0.006
APCA											
INT-1	0.678	0.384	0.220	0.102	0.581	0.408	0.340	0.305	0.273	0.281	0.663
INT-∞	0.656	0.349	0.212	0.101	0.581	0.401	0.320	0.285	0.254	0.265	0.629
CAR											
INT-1	0.568	0.294	0.153	0.067	0.313	0.300	0.260	0.245	0.218	0.210	0.556
INT-∞	0.546	0.263	0.147	0.066	0.313	0.294	0.241	0.226	0.201	0.196	0.523
ECAI											
INT-1	0.867	0.837	0.797	0.746	0.814	0.820	0.817	0.812	0.799	0.800	0.891
INT-∞	0.858	0.820	0.793	0.744	0.814	0.815	0.808	0.801	0.791	0.792	0.880
MCAI											
INT-1	0.895	0.836	0.793	0.743	0.727	0.799	0.816	0.819	0.818	0.804	0.885
INT-∞	0.889	0.823	0.790	0.740	0.727	0.796	0.808	0.812	0.810	0.797	0.875
CCA											
INT-1	0.819	0.842	0.820	0.789	0.519	0.767	0.817	0.844	0.848	0.817	0.843
INT-∞	0.806	0.823	0.815	0.788	0.519	0.759	0.806	0.829	0.839	0.808	0.825

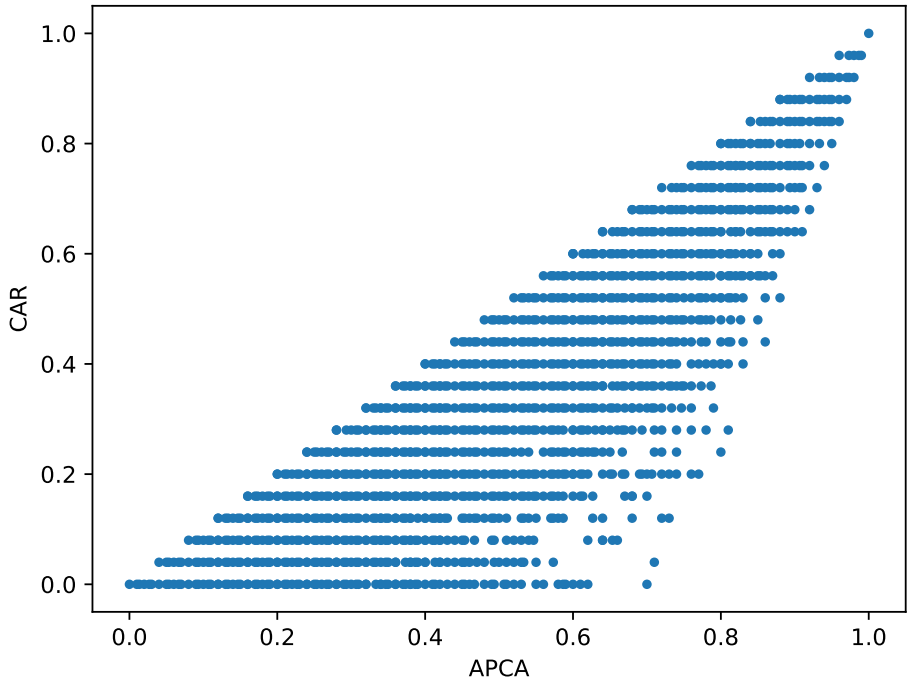


Figure 1: The relation between values of the APCA and CAR measures in all simulation runs.

Extended abstract in Polish

Eksperymentalna analiza własności modeli i metod wspomagania decyzji w kontekście wykorzystania holistycznych preferencji

Wprowadzenie

Jednym z fundamentalnych wyzwań, towarzyszących ludzkości od początku jej istnienia, jest rozwiązywanie różnorodnych problemów decyzyjnych. Mogą mieć one charakter indywidualny i być rozważane przez jednostki albo grupowy, kiedy pożądane jest wypracowanie kompromisowego rozwiązania, satysfakcjonującego wielu interesariuszy. Istnieją różnorodne sposoby radzenia sobie z takimi dylematami; zdolność do logicznego myślenia, intuicja, wiedza ekspercka czy czynniki losowe mogą wywierać istotny wpływ na podejmowane wybory. Zdarza się, że pomimo poświęconego czasu i wysiłku na dokładną analizę dostępnych opcji i konsekwencji ich wyboru, podejmowane są błędne decyzje, prowadzące do niezadowalających rezultatów. Bezpośrednią przyczyną takich decyzji może być np. brak zrozumienia istoty problemu, błędne postrzeganie rozważanych rozwiązań i priorytetów przez decydenta czy nieprawidłowy dobór kryteriów oceny. Jednym ze sposobów na zminimalizowanie ryzyka pomyłek podczas rozwiązywania rzeczywistych i istotnych dylematów, jest zastosowanie metod i praktyk wypracowanych w ramach rozwoju Wielokryterialnego Wspomagania Decyzji (WWD). Jest to dziedzina zorientowana na badanie, rozwój i systematyzowanie informacji, które pozwalają na skuteczne rozwiązywanie problemów decyzyjnych, w których preferencje decydenta odzwierciedlają jego stosunek do wielu, często sprzecznych ze sobą, kryteriów oceny.

Standardowe podejście do problemu decyzyjnego, które jest powszechnie wykorzystywane w WWD, zakłada istnienie zbioru wariantów decyzyjnych, nazywanych również alternatywami, opcjami lub akcjami. Każda z alternatyw jest oceniona przy pomocy co najmniej dwóch, często wzajemnie sprzecznych kryteriów. Oprócz tego, konieczne jest określenie rodzaju rozważanego problemu, związanego z oczekiwaną formą rekomendacji. Wyróżnia się cztery najpopularniejsze rodzaje – problem wyboru, którego rozwiązanie powinno wskazywać jedną lub więcej najlepszych albo wyróżniających się alternatyw; problem rankingu, dla którego konieczne jest uporządkowanie alternatyw od namniej do najbardziej pożądanej; problem sortowania, dla którego decydent oczekuje przypisania każdej z rozważanych alternatyw do jednej z wielu predefiniowanych klas, które są uporządkowane względem preferencji; problem opisu, który dostarcza informacji na temat konsekwencji podjęcia określonych decyzji.

Rozwój dziedziny spowodował, że na przestrzeni ostatnich pięćdziesięciu lat, powstało wiele opracowań podejmujących różne zagadnienia związane z procesami decyzyjnymi. Oferują one różnorodne narzędzia i procedury, których głównym celem jest ułatwienie podejmowania decyzji, a ponadto zagwarantowanie, że otrzymane odpowiedzi będą łatwe w interpretacji oraz spójne z oczekiwaniami decydenta. Opracowane modele i metody wspomagania decyzji tworzą kompleksowe rozwiązania, które umożliwiają systematyczne podejście do przeprowadzenia kolejnych kroków procesu decyzyjnego, takich jak określenie rodzaju rozważanego problemu i oczekiwanej formy utworzonych rekomendacji, zdefiniowanie zbioru możliwych alternatyw i istotnych kryteriów ich ewaluacji oraz określenie sposobu wyrażania preferencji i podejścia do ich reprezentacji.

Ze względu na sposób uzyskiwania informacji na temat przekonań decydenta, możemy wyróżnić podejścia, które mają bezpośrednie przełożenie na wartości parametrów i kształt określonego modelu. Taka koncepcja przekazywania informacji preferencyjnej wymaga od decydenta zrozumienia sposobu działania określonej procedury decyzyjnej i znaczenia poszczególnych parametrów, co sprawia, że skuteczne zastosowanie tego podejścia jest wymagające i nie gwarantuje dostarczenia wysokiej jakości rekomendacji. Z tego powodu, można zaobserwować rosnącą popularność metod i procedur, które bazują na pośredniej informacji preferencyjnej, wyrażającej przekonania decydenta na temat oczekiwanych relacji zachodzących wśród rozważanych alternatyw, które powinny zostać odzwierciedlone w otrzymanych rekomendacjach. Podejścia dostosowane do takiej formy przekazywania informacji o przekonaniach decydenta, wpisują się w paradygmat dezagregacji preferencji, który zakłada, że model reprezentujący preferencje może zostać wywiedziony na podstawie przykładowych decyzji, odnoszących się do niekompletnego podzbioru rozważanych alternatyw.

Wyrażone w ten sposób holistyczne preferencje umożliwiają decydentowi wyrażenie swoich przekonań w sposób intuicyjny i niewymagający specjalistycznej wiedzy dziedzinowej. Ponadto, istnienie dużej liczby dostępnych modeli i metod wspomagania decyzji, umożliwiających przetwarzanie tak sformułowanych preferencji dowodzi, że taka forma ekspresji decydenta jest uniwersalna, a ponadto umożliwia wiarygodne porównanie potencjału predykcyjnego wykorzystujących ją podejść. Niestety, to istotne zagadnienie badawcze jest bardzo często pomijane w literaturze naukowej. Publikacje prezentujące nowe procedury decyzyjne zwykle wskazują jedynie na praktyczne zastosowania, w jakich dana metoda została wykorzystana, abstrahując od przedstawienia porównania z istniejącymi, konkurencyjnymi podejściami, zdolynmi do rozwiązywania takich samych problemów.

Uzupełnienie tej istotnej luki badawczej stanowiło jedną z motywacji niniejszej pracy doktorskiej. Aby to umożliwić, przeprowadzono szereg badań, skoncentrowanych na przedstawieniu analizowanych modeli i procedur wspomagania decyzji, a ponadto ukazaniu ich różnorodności, przede wszystkim ze względu na jakość dostarczanych rekomendacji, rodzaj rozwiązywanych problemów i sposób wykorzystywania informacji preferencyjnej. Istniejące metody wraz z nowymi propozycjami i adaptacjami niektórych modeli i procedur decyzyjnych, zostały przeanalizowane pod kątem właściwości takich jak trafność rekomendacji, odporność oferowanego wyniku, ekspresywność założonego modelu wiedzy i zdolność do zastosowania w problemach wykorzystujących uczenie preferencji. Zaproponowano również szereg miar jakości, które zostały wykorzystane do przeprowadzenia eksperymentalnej analizy porównawczej. Dodatkowo, praca zawiera szczegółowy opis przeprowadzonych eksperymentów wraz z omówieniem uzyskanych rezultatów. Na ich podstawie, opracowane zostały wytyczne dla analityków decyzyjnych, ułatwiające ich pracę i wybór adekwatnej procedury decyzyjnej do rozważanego problemu.

Jakość i odporność procedur dezagregujących preferencje dla problemów wielokryterialnego rankingu i wyboru

Jednym z najpopularniejszych podejść opartych na paradygmacie dezagregacji preferencji i rozwiązującym problemy rankingu i wyboru jest metoda UTA. Wykorzystuje ona model addytywnej funkcji wartości (użyteczności) do reprezentacji preferencji decydenta. Procedura ta zyskała popularność ze względu na akceptację intuicyjnych stwierdzeń określających preferencje decydenta, w postaci porównań dla par alternatyw (np. alternatywa a jest preferowana względem b; alternatywa c jest co najmniej tak dobra jak d) oraz przejrzystą i zrozumiałą formę reprezentacji preferencji. Założenia metody dotyczące normalizacji użyteczności alternatyw i monotoniczności funkcji cząstkowych, wraz z dostarczonymi przez decydenta porównaniami, mogą być bowiem w prosty sposób reprezentowane jako ograniczenia w problemie programowania liniowego.

Ze względu na zakładaną niekompletność informacji preferencyjnej, która obejmuje wyłącznie ranking zupełny lub częściowy dla podzbioru wariatów referencyjnych, model ten może wyznaczyć nieskończenie wiele rozwiązań, które w pełni odzwierciedlają preferencje decydenta, a jednocześnie dostarczają różnych rekomendacji. Narzędziem do analizy tak reprezentowanych rozwiązań jest odporna regresja porządkowa, dostarczająca wniosków na temat koniecznych i możliwych relacji zachodzących dla poszczególnych alternatyw. Niestety, interpretacja tak przedstawionych rezultatów może być trudna do zrozumienia i tym samym nieakceptowalna przez decydenta oczekującego jasnych i wyjaśnialnych wskazań co do rekomendowanych wyborów.

Alternatywnym i powszechnie stosowanym podejściem do tego zagadnienia jest wzbogacenie procedury decyzyjnej o dodatkowy krok polegający na utworzeniu jednoznacznych rekomendacji w oparciu o dostarczony zbiór wszystkich kompatybilnych rozwiązań problemu. Literatura naukowa prezentuje wiele różnych podejść do tego zagadnienia, wprowadzając między innymi metody wyboru reprezentatywnej funkcji wartości. Metody te uzyskują rekomendacje poprzez wykorzystanie pojedynczego modelu spójnego z preferencjami decydenta. W zależności od metody, może to być model oferujący rekomendacje podkreślające ich centralny, średni, odporny albo najbardziej dyskryminujący charakter. Kolejna grupa metod stosuje reguły decyzyjne do zbudowania adekwatnych rekomendacji. Reguły te dostarczają rankingi w oparciu o między innymi porównanie eksteremalnych wartości użyteczności uzyskiwanych przez poszczególne alternatywy albo w oparciu o oczekiwaną pozycję w rankingu, wyznaczoną na podstawie przeprowadzonej analizy stochastycznej wszystkich spójnych rozwiązań.

Istnieją także procedury punktowania alternatyw, na przykład w związku z licznymi relacjami przewyższania innych wariantów decyzyjnych, które są podkreślane przez większość spójnych rozwiązań problemu. Ostatnia grupa metod dostarcza możliwie najbardziej odpornych rekomendacji, zbudowanych ńa podstawie wartości indeksów akceptowalności przypisania alternatywy do określonej pozycji w rankingu oraz wskaźników ukazujących jak często w kompatybilnych rozwiązaniach zachodzą określone relacje dla poszczególnych par alternatyw. Pomimo wielu propozycji metodologicznych dla tego samego problemu, publikacje w dziedzinie WWD nie rozważały dotychczas obszernej analizy porównwaczej zaproponowanych podejść, która mogłaby dostarczyć przesłanek do wyboru procedur posiadających największy potencjał do uzyskiwania wartościowych i trafnych rekomendacji. Z tego powodu, w ramach niniejszej pracy badawczej, zaproponowano eksperymentalną analizę jakości i odporności tych procedur.

Przeprowadzone badania obejmowały analizę porównawczą łącznie trzydziestu pięciu procedur dostarczających jednoznacznych rekomendacji, rozwiązujących problemy wielokryterialnego rankingu i wyboru w oparciu o rezultaty uzyskane przez metodę UTA. Zaproponowano łącznie siedem miar

jakości, spośród których cztery dostarczały informacji na temat poprawności odwzorowania preferencji decydenta, a trzy odzwierciedlały odporność dostarczanych rekomendacji. W każdej z dwóch grup, jedna z miar jakości odnosiła się do problemu wyboru, podczas gdy pozostałe wskazywały na jakość procedury w kontekście problemu tworzenia rankingu alternatyw. Omówiono także schemat wielowymiarowej analizy eksperymentalnej, skupionej na zbadaniu jakości generowanych rozwiązań, w zależności od parametrów rozważanego problemu, takich jak liczba alternatyw (od 6 do 14), liczba kryteriów (od 3 do 5), liczba punktów charakterystycznych dla funkcji marginalnych (od 2 do 4) i liczba dostarczonych porównań dla par wariantów decyzyjnych (od 4 do 10). Zaprezentowano również szczegółowy opis praktycznego zastosowania porównywanych procedur na przykładzie rzeczywistego problemu rankingu, którego celem było dostarczenie preferencyjnie uporządkowanej kolekcji sześciu modeli samochodów, w oparciu o ich ocenę dla pięciu różnych kryteriów decyzyjnych.

W kontekście problemu wyboru, analiza potwierdziła, że najlepsze średnie wyniki uzyskała procedura decyzyjna wykorzystująca rezultaty analizy stochastycznej rozwiązania, wskazująca na alternatywę, która była najczęściej wybieranym wariantem decyzyjnym wśród wszystkich akceptowalnych rozwiązań problemu. Z kolei dla problemu rankingu, wiodącymi metodami okazały się podejścia skupione na dostarczeniu rozwiązania, które w najlepszy sposób odzwierciedlało najpopularniejsze zależności (porównania par alternatyw, przypisania alternatyw do określonych pozycji w rankingu) zachodzące w całej przestrzeni kompatybilnych rozwiązań. Ponadto, istotność uzyskanych konkluzji została potwierdzona testem Wilcoxona dla par obserwacji i były one prawdziwe bez względu na rozważane parametry problemu, takie jak rozmiar, kształt funkcji marginalnych i bogactwo informacji preferencyjnej. Jednoznaczne wnioski pozwoliły na sformułowanie wskazówek dotyczących wyboru procedur uzyskujących średnio najlepsze rezultaty dla problemów rankingu i wyboru, w zależności od rodzaju problemu i rozważanego aspektu jakości dostarczanych rekomendacji.

Jakość i odporność procedur dezagregujących preferencje dla problemów wielokryterialnego sortowania

Podobnie do omówionej wcześniej analizy porównawczej procedur dostarczających jednoznacznych rekomendacji dla problemów rankingu i wyboru, istnieją analogiczne podejścia dla rozwiązania problemu wielokryterialnego sortowania. Wiele z nich działa w oparciu o rezultaty uzyskane przez metodę UTADIS, adaptującą podejście UTA poprzez wykorzystanie pośredniej informacji preferencyjnej, wyrażonej jako przykładowe przydziały alternatyw do klas oraz wzbogacenie modelu addytywnej funkcji użyteczności o wartości progowe separujące przedziały użyteczności alternatyw wraz z funkcją wykorzystującą owe wartości do jednoznacznego przyporządkowania warian-

tów do klas decyzyjnych. Literatura badawcza nie przedstawiła dotychczas jednoznacznych wskazań co do jakości rezultatów uzyskiwanych przez niniejsze procedury. W związku z tym, również to zagadnienie zostało rozważone w niniejszej rozprawie.

Eksperymentalna analiza własności obejmowała czternaście procedur pozwalających na rozwiązanie problemów wielokryterialnego sortowania poprzez jednoznaczne przypisanie rozważanych wariantów do klas decyzyjnych. Wszystkie rozpatrywane metody umożliwiają wyzaczenie reprezentatywnej instancji założonego modelu preferencji. Różnią się jednak co do sposobu jego wyboru, podążając choćby za ideą rozwiązania najbardziej dyskryminującego, centralnego, średniego, skapego czy odpornego. Trzy z rozważanych podejść zostały zaproponowane po raz pierwszy i stanowią dodatkowy wkład pracy do literatury przedmiotu. Zakładają one wybranie modelu, którego rekomendacje posiadają największe wsparacie w zbiorze wszystkich modeli spójnych z preferencjami decydenta. Zaprezentowano również opis poszczególnych metod, a także przypadek użycia ilustrujący praktyczne zastosowanie wszystkich procedur do rozwiązania rzeczywistego problemu sortowania trzydziestu miast europejskich pod kątem wdrażania polityki proekologicznej.

Ponownie, schemat zaproponowanych eksperymentów zakładał zmierzenie średnich wartości miar odnoszących się do istotnych aspektów jakościowych dostarczanych rekomendacji, takich jak trafność klasyfikacji, odporność uzyskanych rezultatów i ocenę podobieństwa wywiedzionego modelu do jego odpowiednika, który był wykorzystywany do zbudowania referencyjnych preferencji decydenta. Rezultaty przeprowadzonych eksperymentów zostały następnie poddane wielowymiarowej analizie, która oprócz identyfikacji najlepszych procedur, oceniła również wpływ parametrów rozważanego problemu na jakość uzyskanych rekomendacji. Analiza obejmowała problemy o różnej złożoności, utworzone w oparciu o kombinację następujących parametrów: liczba klas decyzyjnych (od 2 do 5), liczba kryteriów (od 3 do 9), liczba punktów charakterystycznych dla funkcji marginalnych (od 2 do 6) i liczba referencyjnych przypisań wariantów do klas (od 3 do 10).

Procedura wyznaczająca reprezentatywną funkcję wartości poprzez znalezienie analitycznego centrum hiperwielościanu, reprezentującego przestrzeń wszystkich kompatybilnych rozwiązań, dostarczała rozwiązań dla problemów wielokryterialnego sortowania, które charakteryzowały się najwyższą średnią trafnością klasyfikacji. Z drugiej strony, trzy nowe zaproponowane procedury, poszukujące rozwiązania, które najpełniej odzwierciedla najpopularniejsze przypisania do klas w całym zbiorze spójnych rekomendacji, uzyskiwały średnio najbardziej odporne rekomendacje. Przeprowadzona analiza statystyczna potwierdziła istotność zaobserwowanych zależności, niezależnie od parametrów charakteryzujących rozważane problemy decyzyjne. Zaprezentowane rezultaty umożliwiły sformułowanie kolejnych wskazówek dla analityków i tym samym dostarczyły przesłanek, potwierdzających hipotezę badawczą.

Odporność rekomendacji i ekspresywność modeli w podejściach rozwiązujących problemy wielokryterialnego sortowania

Oprócz modyfikacji procedur decyzyjnych, pozwalających na uzyskanie jednoznacznych rekomendacji w oparciu o ten sam model reprezentacji preferencji, istnieje również grupa procedur, które zmieniają założenia wykorzystywanego modelu. Model addytywnej funkcji wartości, wykorzystywany w metodach UTA i UTADIS, zakłada monotoniczność preferencji względem poszczególnych kryteriów, a ponadto traktuje niezależnie poszczególne kryteria oceny alternatyw i tym samym nie odzwierciedla zależności preferencji decydenta od interakcji międzykryterialnych. Istnieją jednak praktyczne problemy decyzyjne, dla których konieczne jest odzwierciedlenie wspomnianego niemonotonicznego charakteru preferencji lub zachodzących interakcji. Model addytywnej funkcji wartości nie ma możliwości reprezentacji takich preferencji, co powoduje, że te metody są niezdolne do dostarczenia jakościowych rekomendacji dla rozważanych problemów.

Przeprowadzona analiza umożliwiła identyfikacje jednego podejścia, które wzbogacało model addytywnej funkcji użyteczności o dodatkowe funkcje, reprezentujące interakcje dla par kryteriów oraz ujawniła dwa podejścia, które wprowadzają możliwość reprezentacji niemonotoniczności preferencji poprzez modyfikację podstawowych założeń modelu. Jednak wszystkie wymienione wyżej modele były zorientowane na rozwiązywanie problemów rankingu. Z tego powodu, w ramach przeprowadzonych prac badawczych, zaproponowano adaptację opisanych w literaturze metod, dostosowując ich działanie do rozwiązywania problemów wielokryterialnego sortowania.

Analiza eksperymentalna obejmowała porównanie podstawowej procedury UTADIS wraz z pięcioma zaproponowanymi adaptacjami wyżej wymienionych metod, spośród których dwie dotyczyły reprezentacji interakcji, a trzy umożliwiały uwzględnienie niemonotonicznych preferencji. Jednym z badanych aspektów była ekspresywność modeli, interpretowaną jako zdolność do odzwierciedlania preferencji decydenta, niezależnie od ich spójności i kompletności. Ponadto, zbadano również odporność dostarczanych rekomendacji i modeli, poprzez zmierzenie ich stabilności i zgodności z wszystkimi kompatybilnymi rozwiązaniami problemu. Oba te zagadnienia są często postrzegane jako sprzeczne, ze względu na to, że metody charakteryzujące się wyższą ekspresywnością bazują na bardziej wyrafinowanych i skomplikowanych modelach reprezentacji preferencji. To z kolei powoduje, że znacznie częściej mogą one dostarczać niejednoznacznych rekomendacji, co może powodować obniżenie stabilności i odporności proponowanych rozwiązań.

Wyniki przeprowadzonych analiz potwierdziły konkurencyjny charakter tych dwóch zagadnień. Metody budujące rekomendacje w oparciu o bardziej ekspresywne modele, uzyskują niższe rezultaty w kontekście miar jakości,

które odnoszą się do ich odporności i vice versa. Z tego powodu zaproponowano, aby do rozwiązywania problemów wielokryterialnego sortowania wykorzystywać metody dostarczające możliwie najodporniejszych rekomendacji, stopniowo przechodząc do bardziej ekspresywnych metod tylko w przypadku, gdy prostsze procedury nie są w stanie skutecznie odzwierciedlać przekonań decydenta. W przypadku rozważanych metod, w pierwszej kolejności jest to metoda UTADIS, dalej podejścia uwzględniające interakcje dla par kryteriów, a na końcu procedury odzwierciedlające niemonotoniczny charkater preferencji dla poszczególnych cech rozważanych alternatyw.

Przeprowadzona analiza ujawniła jednak, że odporność metod uwzględniających interakcję jest wyższa od tej uzyskiwanej przez niemonotoniczne procedury tylko wówczas, gdy liczba aktywnych interakcji nie przekracza dwóch. W przeciwnym przypadku, to metody zakładające niemonotoniczny kształt funkcji marginalnych dostarczają bardziej odpornych rekomendacji. Ponadto, jeśli żadna z metod nie jest w stanie w pełni odzwierciedlić preferencji decydenta, wówczas konieczna jest elicytacja informacji preferencyjnej. Należy również odnotować, że proponowany zbiór reguł prowadzący do wyboru najodporniejszej procedury, proponuje właściwy sposób postępowania tylko w sytuacji, gdy nie istnieją przesłanki świadczące o niemonotonicznym lub interakcyjnym charkaterze preferencji decydenta. Natomiast jeśli takowe przesłanki istnieją, należy zastosować podejście spójne z dodatkowymi założeniami dotyczącymi preferencji. Mimo to, zaproponowane reguły również stanowią użyteczne wskazówki, które mogą być skutecznie wykorzystywane przez analityków współpracujących z decydentami. Dodatkowo, aby ułatwić odbiór zaproponowanych reguł postępowania, zilustrowano ich zastosowanie na przykładzie problemu sortowania trzydziestu modeli telefonów komórkowych do trzech klas decyzyjnych, w kontekście preferencji dostarczonych przez trzech różnych decydentów.

Algorytmy wspomagające uczenie preferencji modelu całki Choquet inspirowane naturą

Kolejnym istotnym kierunkiem rozwoju metod wspomagania decyzji są podejścia do problemu uczenia preferencji, których głównym celem jest uzyskanie modelu skutecznie klasyfikującego alternatywy do preferencyjnie uporządkowanych klas. Wynikiem takich metod powinny być rozwiązania, które z jednej strony są w stanie skutecznie reprezentować preferencje wyrażone w oparciu o duże zbiory danych, mogące zawierać niespójne informacje preferencyjne, a z drugiej strony – dostarczają rekomendacji, które są łatwe do interpretacji i uzasadnienia.

Jedną z metod wspomagania decyzji rozważanych w literaturze przedmiotu w tym kontekście to model całki Choquet. Opiera on swoje działanie o parametry reprezentujące istotność ocen uzyskanych dla określonych

podziorów kryteriów. Tak sformułowane założenia powodują, że metoda ta jest zorientowana na odzwierciedlanie negatywnych i pozytywnych interakcji międzykryterialnych. Dzięki temu, ekspresywność całki Choquet jest znacznie wyższa niż dla mniej rozbudowanego modelu reprezentacji preferencji w metodzie UTADIS. Jednak wyznaczenie optymalnych wartości parametrów modelu stanowi wyzwanie, a zastosowanie klasycznych podejść opartych o programowanie liniowe jest często nieefektywne. Rozmiar sformułowanego problemu i złożoność ograniczeń liniowych, wynikających z dużej liczby dostarczonych informacji prefeerencyjnych, powodują, że czas potrzebny do optymalizacji rozwiązania ulega wydłużeniu. To z kolei sprawia, że dla wielu rzeczywistych zastosowań i rozważanych w nich problemów decyzyjnych, takie rozwiązanie jest nieakceptowalne.

Przeprowadzony przegląd literatury ujawnił istnienie różnych podejść, które umożliwiają optymalizację parametrów całki Choquet. Co prawda, nie gwarantują one uzyskania optymalnego rozwązania, ale w odróżnieniu od problemów programowania liniowego, mogą dostarczyć satysfakcjonujące rezultaty w bardzo krótkim czasie. Do przeprowadzenia eksperymentalnej analizy porównawczej, zaproponowano łącznie osiem podejść zorientowanych na optymalizację parametrów wspomnianego modelu. Dwa spośród nich wprowadzały pewne usprawnienia do sformułowanego problemu programowania liniowego, pozwalające na uzyskanie wartościowych rekomendacji w krótkim czasie. Kolejne trzy metody implementowały różne strategie oparte na koncepcji lokalnego przeszukiwania przestrzeni rozwiązań, a ostatnie trzy podejścia dostosowywały metaheurystyki inspirowane zjawiskami zachodzącymi w naturze, w szczególności zachowaniami stadnymi zwierząt i zjawiskiem selekcji naturalnej, do rozważanego problemu.

Prace badawcze obejmowały zaprezentowanie modelu całki Choquet oraz omówionych wyżej podejść do rozwiązania problemu. Zaprezentowano także przykład ilustrujący ewaluację wartości całki dla przykładowego wariantu decyzyjnego, a ponadto sposób działania rozważanych algorytmów. Z kolei przeprowadzona analiza eksperymentalna skupiała się w głównej mierze na trafności rezultatów binarnej klasyfikacji i poprawnym odzwierciedleniu relacji preferencji dla par alternatyw pochodzących z różnych klas. Obejmowała ona ewaluację wszystkich ośmiu podejść z wykorzystaniem pięciu referencyjnych zestawów danych, w podziale na trzy scenariusze o różnej liczbie informacji preferencyjnych wyrażonych przez decydenta. Analiza potwierdziła, że najlepsze rezultaty osiągają metaheurystyki inspirowane naturą, które niezależnie od rozważanej liczby dostarczonych preferencji, zajmowały średnio najwyższe pozycje w rankingu rozważanych metod. Dzięki temu, po raz kolejny możliwe było zdefiniowanie wskazówek rekomendujących zastosowanie algorytmów, które najdokładniej odzwierciedlały preferencje decydenta.

Podsumowanie

Rosnące zainteresowanie metodami WWD przyczyniło się do powstania wielu podejść spójnych z paradygmatem dezagregacji preferencji, który pozwala na uzyskanie interpretowalnych rozwiązań na podstawie prostych, holistycznych informacji preferencyjnych. Wzrost liczby opisanych w literaturze metod nie dał jednak odpowiedzi na pytanie: która z dostępnych procedur dostarcza jakościowych i odpornych rekomendacji w rozważanym kontekście decyzyjnym? Brak badań, które pomogłyby określić użyteczność poszczególnych metod wspomagania decyzji stanowi istotną lukę w literaturze naukowej, której wypełnienie stało się jednym z celów niniejszej dysertacji.

W ramach przeprowadzonych badań, zrealizowano łącznie cztery eksperymentalne analizy własności modeli i procedur wielokryterialnego wspomagania decyzji. Dotyczyły one dostarczenia jednoznacznych rekomendacji, odporności rozwiązań, ekspresywności modeli, a ponadto ich zdolności do uczenia się preferencji. Opracowania rezultatów przeprowadzonych badań obejmowały szczegółowy opis porównywanych metod wraz z prezentacją przypadków użycia, ilustrujących praktyczne zasotosowanie rozważanych modeli. Wśród analizowanych podejść do rozwiązywania problemów, obok istniejących metod wspomagania decyzji, znajdują się również takie, które prezentują nowe propozycje albo adaptacje istniejących procedur, które dostosowują je do rozważanego konteksktu decyzyjnego.

W celu uzyskania wartościowych wniosków i rekomendacji, zaproponowano szereg adekwatnych miar jakości, odzwierciedlających pożądane cechy generowanych rozwiązań, takie jak trafność rekomendacji, odporność preferencji i ekspresywność modelu. Wykonane badania potwierdziły również użyteczność nowo proponowanych metod dostarczania jednoznacznych preferencji, udowodniły przeciwstawny charakter odporności rekomendacji i ekspresywności modeli, a ponadto wykazały, że algorytmy optymalizacyjne mogą być skutecznym narzędziem do rozwązywania problemów uczenia preferencji.

Co jednak najistotniejsze, przeprowadzone analizy i zaproponowane eksperymentalne podejście do badania własności modeli i metod, mogą stanowić istotne uzupełnienie publikacji wprowadzających nowe procedury wspomagania decyzji, wzbogacając je o wnioski płynące z porównania rozwiązań uzyskiwanych przez rozważane procedury. Same zaś wyniki analiz pozwoliły na uzyskanie przesłanek, które usprawniają dobór adekwatnych metod do rozważanego problemu. Tym samym, ich skuteczne sformułowanie potwierdza hipotezę badawczą zawartą w niniejszej dysertacji.

Przyszłe kierunki badań powinny obejmować przeprowadzenie kolejnych eksperymentalnych analiz porównawczych, dostarczających przesłanek dedykowanych dla analityków decyzji, wskazujących na użyteczność określonych metod w odniesieniu do innych zagadnień, które są przedmiotem analiz prowadzonych w ramach rozwoju obszaru WWD. Po drugie, należy rozważyć

sformułowanie uniwersalnych kryteriów i reguł oceny nowo proponowanych podejść, co umożliwiłoby wskazanie zalet proponowanych rozwiązań i dostarczenie dowodów na ich wysoką użyteczność w rozważanym kontekście decyzyjnym. Zasady te powinny oferować kompleksowe podejście do analizy i ewaluacji metod, na przykład poprzez opracowanie kompletnego zbioru miar jakości oraz reprezentatywnej kolekcji referencyjnych zbiorów danych. Wykorzystanie miar jakości do ewaluacji dostarczanych rozwiązań dla predefiniowanego zbioru problemów, pozwoliłoby na ustandaryzowanie procesu ewaluacji, co ułatwiałoby porównywanie nowo wprowadzanych metod z dotychczas opisanymi w literaturze. Pożądanym z perspektywy rozwoju dziedziny byłoby także opracowanie meta-procedur wspomagania decyzji, które zgodnie z paradygmatem dezagregacji preferencji, poprzez holistyczną analizę informacji na temat rozważanego problemu decyzyjnego, rekomendowałyby wykorzystanie określonych metod wspomagania decyzji, umożliwiających skuteczne zaadresowanie danego problemu.

Declarations

September 20, 2024

DECLARATION

I hereby declare the following contribution as an author of the following papers:

M. Kadziński, M. Wójcik, K. Ciomek, Review and experimental comparison of ranking and choice procedures for constructing a univocal recommendation in a preference disaggregation setting, Omega, 2022, 113, 102715.

M. Wójcik, M. Kadziński, 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, 2023, 278, 110871.

- Co-authorship of the concept of comparing various methods from the UTA or UTADIS family within an experimental study
- Implementation of a subset of methods from the UTA or UTADIS family that were used in the experimental study
- Consultation on the measures used for quantifying the robustness concerns and predictive performance
- Conducting initial experiments that were not included in the papers, but allowed gaining insights into the performance of various methods
- Editing and revising the text of the publications

Krzysztof Ciomek

Krzysztof Ciomek

September 20, 2024

Universitat Pompeu Fabra Barcelona
Department of Economics and Business

Department of Economics and Business
Pompeu Fabra University (UPF)

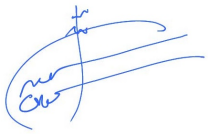
Ramon Trias Fargas 25-27
08005 Barcelona, Spain
mohammad.ghaderi@upf.edu

DECLARATION

I hereby declare the following contribution as an author of the following paper:

M. Kadziński, M. Wójcik, M. Ghaderi, *From investigation of expressiveness and robustness to a comprehensive value-based framework for multiple criteria sorting problems*, Omega, 2024, after 2nd revision.

- Co-authorship of the concept of considering the expressiveness and robustness of various UTADIS variants in an experimental study
- Consultation of the correctness of novel variants of UTADIS proposed in the paper
- Consultation of the experimental study design
- Close collaboration with other authors on the way experimental results are presented and the frameworks for method's selection are outlined
- Co-authorship of the text of the publication



Mohammad Ghaderi

Poznan University of Technology
Faculty of Computing and Telecommunications
Institute of Computing Science
Typeset using LaTeX in Computer Modern.

BibTeX:

```
@phdthesis{ Wojcik2024,
    author = "Michał Wójcik",
    title = "{Experimental analysis of the properties of models and decision support
methods in the context of the use of holistic preferences}",
    school = "Poznan University of Technology",
    address = "Pozna{\'n}, Poland",
    year = "2024",
}
```