Università degli Studi di Cagliari

Ph.D. DEGREE

ECONOMICS AND BUSINESS

Cycle XXXIV

TITLE OF THE Ph.D. THESIS

Advances On The Analysis Of Ordinal Data Expressed As Rankings And Paired Comparisons: Distance-based Approaches, New Probabilistic Models, And Tree-based Applications

Scientific Disciplinary Sector(s)

SECS-S/01 Economics and Statistics

Ph.D. Student: Alessio Baldassarre

Supervisor Claudio Conversano

Final exam. Academic Year 2020/2021
Thesis defence: April 2022 Session

Advances On The Analysis Of Ordinal Data Expressed As Rankings And Paired Comparisons: Distance-based Approaches, New Probabilistic Models, And Tree-based Applications

---

PhD candidate: Alessio Baldassarre

Supervisor: Claudio Conversano

PhD in Economics and Business

SECS-S/01-Economics and Statistics

XXXIV cycle

Università degli Studi di Cagliari

Facoltà di Scienze Economiche ed Aziendali

Cagliari, Italy

Thesis Defense: April 2022

Navigando sul mare color vino

verso uomini di altre lingue

Alla mia famiglia,

ai miei amici,

quelli storici e quelli che lo diventeranno

Abstract

This thesis represents an original contribution to knowledge on ordinal data, which constitutes the leitmotif of the entire research product. Specifically, a short review of preference data presents the leading distance and correlations measures with a focus on two weighted measures (i.e., distance and correlation). A simulation study investigates the effect of introducing positional weights when preferences are analyzed through distance-based approaches. The weighted correlation coefficient is then used as cost function within an evolutionary algorithm for finding the consensus ranking. Then, we will focus on analyzing ordinal data through probabilistic approaches, presenting a new tree-based model, the Bradley-Terry Regression Trunk model (BTRT). Again, a simulation study is conducted to evaluate the performance of the pruning procedure implemented in the new algorithm. This model is applied on two different datasets: the first is composed of self-reported data by students from the University of Cagliari; the second derives from well-known databases and contains financial information about tax revenues by central governments worldwide and their socio-economic characteristics. The BTRT model is applied to the first dataset to partition students based on their preference rankings about the attributes they expect from an ideal professor. For the second dataset, the goal is to apply the BTRT model for partitioning countries based on the size of their tax revenues and how their socio-economic characteristics influence these revenues. The BTRT model furnishes an easy-to-read partition in the form of a small regression tree, called trunk, able to capture the interactions between covariates that cause the most significant decrease in model deviance. Hence, it finds the best interactions between covariates by simultaneously considering their main effects. Finally, the last chapter shows an advance for the BTRT model by following the Mallows specification of the Bradley-Terry model. The Mallows specification works on rankings instead of paired comparisons. It assumes independence across the ranked objects so that the probability of observing a specific ranking is proportional to the product of the estimated worth parameters for each object. The BTRT model with the Mallows specification is applied to the financial dataset to discover the causal effect between government expenditure and tax revenues.

Acknowledgment

I would like to express my gratitude to my supervisor, Claudio Conversano, who guided me throughout my Ph.D. course. I would also like to thank professors Antonio D'Ambrosio from the Università Federico II di Napoli, Elise Dusseldorp, and Mark de Rooij from the University of Leiden. They all supported me and offered deep insight into the study. Furthermore, I would like to thank my colleagues at the Local Taxation Office at the Italian Ministry of Economy and Finance and my manager for allowing me to complete my studies as a visiting Ph.D. candidate at the University of Leiden.

Alessio Baldasarre

Contents

Chapter 1

Introduction

Analyzing ordinal data is ubiquitous in many scientific fields, such as social sciences, economics, political sciences, computer science, psychometrics, behavioral sciences, and many others. This thesis aims to resume the literature about a specific branch of ordinal data, the preference data, when expressed as rankings or paired comparisons. Specifically, we will analyze preference data through distance-based approaches and probabilistic models. The first is generally aimed at determining the "consensus ranking" through the minimization/maximization of distance/correlation measures. Hence, a short review will be presented on the most common distance and correlation measures, focusing on two position-weighted measures. The adoption of position weights allows calculating the accurate distance between two rankings, considering where the differences occur (e.g., the top-ranked objects or the last ones). The consensus ranking problem can be approached through a new differential evolutionary algorithm when position weights are taken into account. It aims at finding a solution for the search of the consensus ranking when different weights are assigned to the different positioning of a set of objects within a ranking. Finally, a study of simulations presents how the choice of position weights affects the consensus ranking problem.

As far as probabilistic approaches are concerned, an innovative tree-based model is presented for ordinal data expressed through paired comparisons. This approach is based on the Bradley-Terry model and uses the specification of the STIMA algorithm

(Dusseldorp et al., 2010) for the construction of a particular regression tree called regression trunk. The combination of these tools gives rise to the Bradley-Terry Regression Trunk model (BTRT), from which it is possible to obtain a breakdown of H individuals based on the preferences expressed by them on n_o objects. The heterogeneity of individuals is taken into account by considering their characteristics and how these interact with the individuals' choices. The main advantage of this model is finding interaction effects when no a priori information is known about them. The model performance is computed through a simulation study to choose the tuning parameter for the pruning procedure. Then, we will show an application of the BTRT model to two different datasets. The first one was collected at the Università degli Studi di Cagliari, where the first-year students were asked to order a set of objects based on their preferences; the second dataset is country-sectional and contains financial information on a set of countries worldwide. Four government tax revenues categories are ranked based on their size for each country. Given the high number of predictors, we apply a feature selection with the Bradley-Terry-Luce Lasso method for the second application. Finally, we present an extension of the BTRT model through the Mallows specification. This new model constitutes an advance of the BTRT model. In the end, we apply this model to the dataset containing financial data to investigate the causal relationship between government expenditure and government tax revenues.

This thesis presents the candidate's research products and is structured as follows: The introductory Chapter 2.1 defines the preference data and the different methodologies to analyze them. It focuses on the geometric representation of rankings and how to analyze them with a distance-based approach. Here, we present an insight into a weighted correlation coefficient for rankings and a simulation study on how position weights can affect the calculations. Subsequently, Chapter 2.6 focuses on evolutive algorithms for the consensus ranking problem, presenting the WDECOR algorithm for finding consensus ranking through the maximization of a weighted correlation coefficient. Then, in Chapter 3.1, we move on to probabilistic approaches, focusing on a new tree-based model, the Bradley-Terry Regression Trunk model. The subsections in this chapter show how the algorithm works and how it performs through a simulation study. The model is applied to the student's dataset to include the procedure and results in Chapter 4.1. Chapter 5.1

presents a study of tax data aimed at partitioning a set of countries worldwide based on the ordering of their main tax revenues and their socio-economic characteristic. Chapter 6.1 presents an advance on the BTRT model through the Mallows specification and its application to financial data. Finally, the conclusions are reported in Chapter 7.

Chapter 2

Preference data

2.1 Introduction to preference data

Preference data are ordinal data and they can be expressed in several ways. However, for the rest of this thesis, we will focus on preferences expressed as rankings, orderings, and paired comparisons. Preference data are generally expressed through numerical vectors or lists of objects, called rankings and orderings, respectively. While their meaning is different, these terms will be used interchangeably. Specifically, H judges can express their preferences relating to n_o objects by assigning values from 1 to n_o, where 1 indicates the best-ranked object and n_o the worst. Sometimes, instead of assigning a numeric score to each item, people can place in order the objects by forming a list in which the preferences are stated simply by looking at the order in which each object appears in the list. This list is called ordering (or order vector), and it can be transformed into a ranking (or rank vector) when, given any arbitrary order of the set of the objects, the rank of each of them is reported (Marden, 1996). For example, let four objects (A, B, C, D) and the order expressed by the h-judge be (B, C, A, D), then the ranking associated with this ordering is (3, 1, 2, 4). This expression indicates that object A falls into the third position, while B and D are the most and least preferred objects. The rankings indicate which object is preferred to another, but they do not offer any information about the nature and intensity of this preference. Ordinal data do not have

metric information. Although the response options might be numerically labelled as consecutive integers, the numerals only indicate order and do not indicate equal intervals between levels. In fact, in the previous example, a ranking of this type (40, 2, 30, 70) would keep the order expressed by the judge unchanged. When an individual assigns different values from 1 to n_o to all objects, we speak of full rankings. If, on the other hand, two or more objects are preferred in the same way, and they assume the same position, then a tied (or weak) ranking is obtained. Finally, we speak of partial rankings when judges express their preferences only on a subset of objects. The latter two cases can be found in numerous datasets, to the point of considering their presence a rule rather than an exception (D'Ambrosio et al., 2015). Sometimes objects are presented in pairs to judges, producing the so-called paired comparisons: this could be the natural experimental procedure when the objects to be ranked are similar, and the introduction of others may be confusing (David, 1969). Given a ranking of n_o objects, it is always possible to determine the relative $n_o \times (n_o - 1)/2$ pairwise preferences. On the other hand, a set of $n_o \times (n_o - 1)/2$ paired comparisons does not always correspond to a ranking because of the phenomenon of non-transitivity of the preferences. Such non-transitivity could be avoided by ensuring that 'individuals comparisons are independent or nearly' (David, 1969, p. 11). In analyzing rank data, the goal is often to find one ranking that best represents all the preferences stated by the individuals. When dealing with rank vectors, this goal is known as the consensus ranking problem, the Kemeny problem, or the rank aggregation problem (D'Ambrosio et al., 2019). When dealing with paired comparisons, the goal is to determine the probability that objects i is preferred to object j for all the possible pairs of them: the outcome is thus a probabilistic determination of the central ranking (Kendall and Smith, 1940; Bradley and Terry, 1952; Mallows, 1957). Finding the central ranking is a very important step when rank data are analyzed (Cook and Seiford, 1982; Emond and Mason, 2002; Meila et al., 2007; Amodio et al., 2016; Aledo et al., 2017) either as a final analysis tool, when homogeneity among people is assumed or as a part of a more complex analysis strategy when heterogeneity among judges is assumed.

The preferences can be visualized through a geometric representation within the permutation polytope (Thompson, 1993; W. Heiser, 2004), a convex figure containing

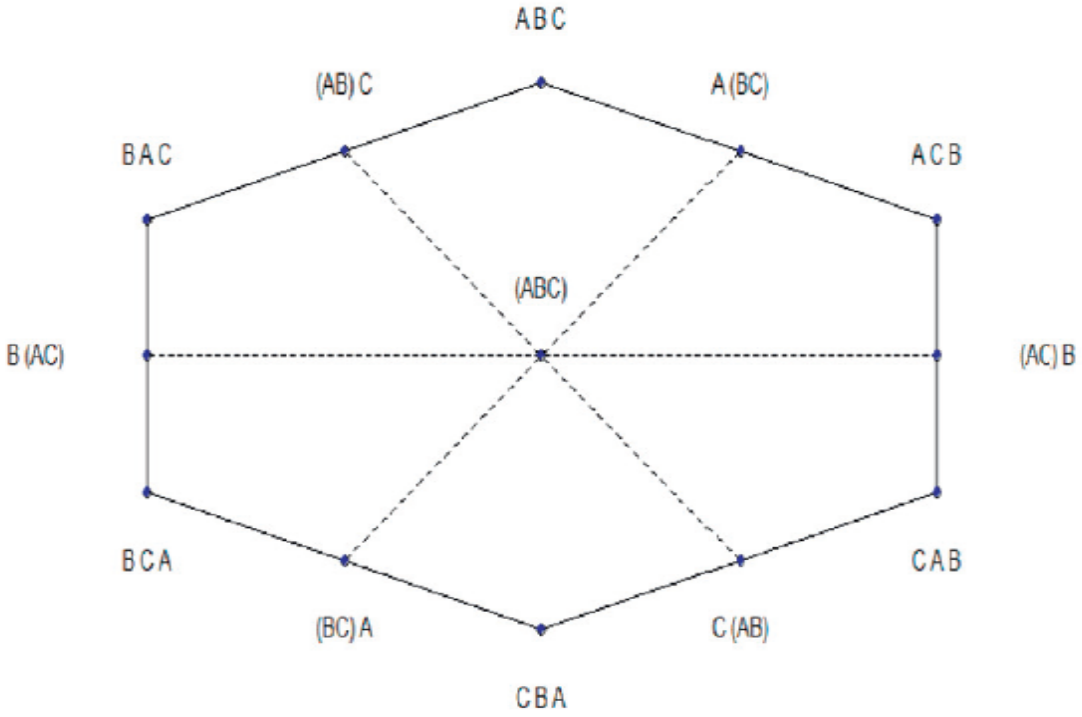


Figure 2.1: Generalized permutation polytope, full (and tied) rankings, three objects

all the permutations of the n_o objects to be classified. The polytope defined in $\mathbb{R}^{n_o}$ is discrete, symmetrical, and finite space, with sides of equal size, a number of vertices equal to $n_o!$ and dimensions equal to $n_o - 1$. For this reason, the rankings space can only be displayed when the objects are three or four. Its construction requires knowledge of the number of objects involved in the analysis, and it is not necessary to have either the preferences expressed by the judges or their frequency.

The permutations are arranged on the vertices so that in passing from one adjacent corner to another, only one exchange is made between pairs of objects. With a number of objects equal to three (A B C), the preference space is a two-dimensional hexagon, equating to opposite vertices with full inverse rankings. If ties are allowed, the latters will be arranged between one vertex and another, thus obtaining the generalized permutation polytope, where, in the center, there is an all-tied ranking, which is a vector that considers all objects in ties and equidistant from all other points of the hexagon. Figure 2.1 shows the generalized permutation polytope of full and tied rankings with three objects.

With a number of objects equal to four, (A B C D), we obtain a truncated octa-

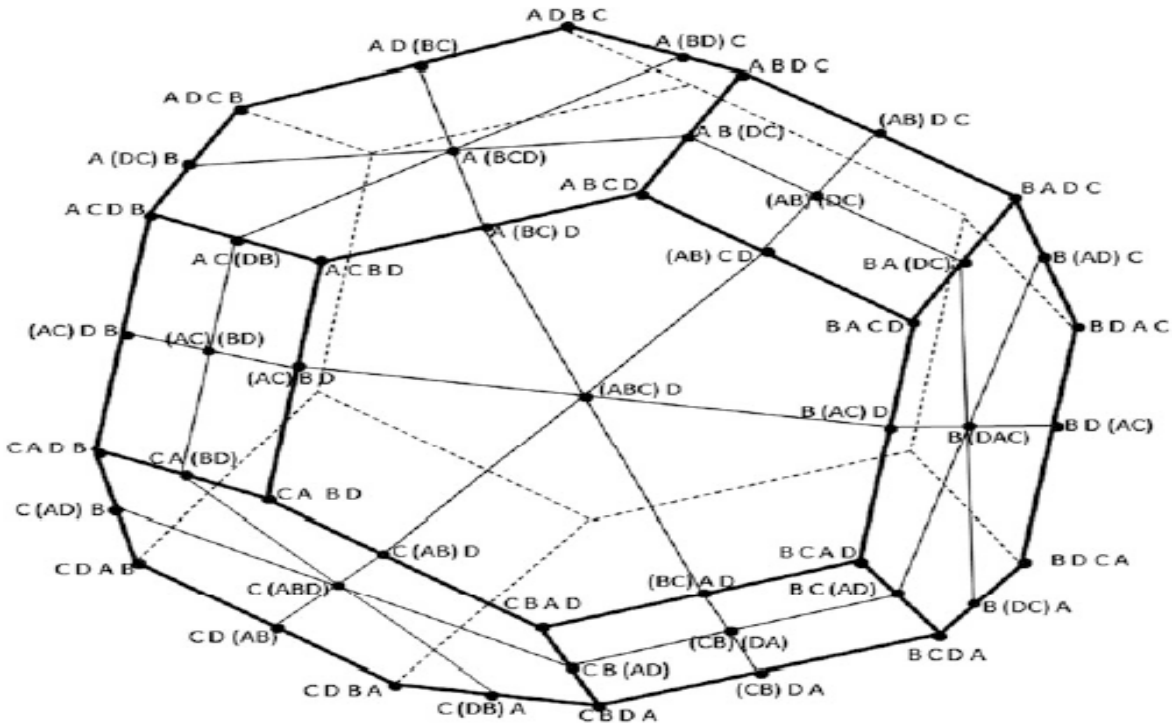


Figure 2.2: Generalized permutation polytope, full and tied (in brackets) rankings, 4 objects

hedron of three dimensions and 24 vertices (Figure 2.2), formed by six squares and eight hexagons, of which the first four always have the same object in the first position. The remaining four hexagons indicate the same object as the least preferred on all vertices. As for the squares, the vertices are associated with rankings that always show the same pair of objects as a favorite.

The vertices of the polytope can be interpreted as the centers of gravity of the objects. Therefore the truncated octahedron can be inscribed inside a pyramid whose vertices act as poles of attraction. The centers of the four hexagons with the same object in the first position are attracted to the vertex corresponding to the latter. Likewise, the vertices of the pyramid represent poles of repulsion for the faces of the hexagons that are in the opposite position. Figure 2.3 shows the truncated octahedron inscribed inside a pyramid.

The distance between the vertices can be interpreted as the minimum number of transpositions of adjacent objects necessary to transform one ranking into another. If ties are allowed, the measure that best fits the permutation polytope is the Kemeny distance,

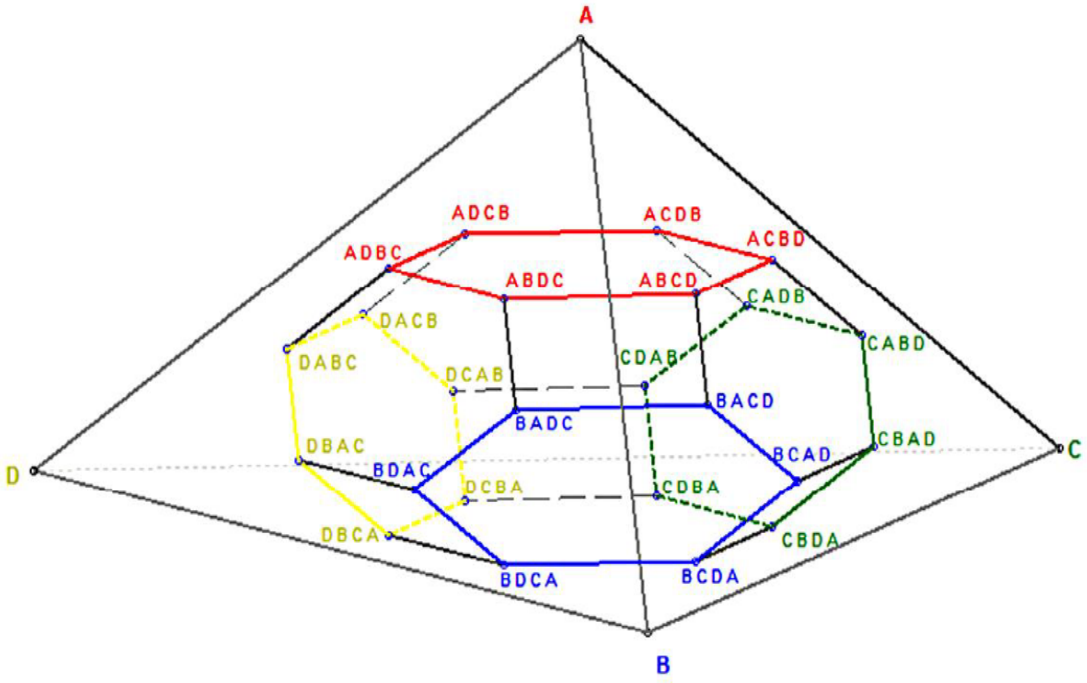


Figure 2.3: Permutation polytope of 4 objects inscribed in a pyramid

which calculates the number of exchanges between pairs of objects necessary to transform one (partial) ranking into another (Kemeny, 1959 ,W. J. Heiser and D'Ambrosio, 2013). All these geometric figures represent a valuable tool for various analyses, since they allow the use of two measures commonly used to study full rankings, namely the Kendall coefficient τ (see Section 2.2) and Spearman's ρ. These indices provide a natural geometric interpretation of the permutation polytope.

The Spearman index is proportional to the linear distance between the vertices of the permutation polytope. Also, if the sides are of length equal to $\sqrt{2}$, then Spearman's ρ coincides with the Euclidean distance between two vertices. The Spearman distance between two rankings R_1 and R_2 of n_o objects can be calculated as follows

$$d(R_1, R_2) = \sum_{i=1}^{n_o} (R_{1,i} - R_{2,i})^2 \tag{2.1}$$

From which the Spearman correlation coefficient ρ derives

$$\rho = 1 - \frac{6 \sum_{i=1}^{n_o} d(R_{1,i}, R_{2,i})}{n_o^3 - n_o} \tag{2.2}$$

According to the approach of Deza and Deza, 2009, a distance metrics satisfies three properties whose meaning is not always attributable to that of metrics. For values $A, B, C \in R_h$, with $R_h \neq \emptyset$ and judges $h = 1, \ldots, H$:

1. $d(R_1, R_2) \geq 0$ (non negativity)
2. $d(R_1, R_2) = d(R_2, R_1)$ (symmetry)
3. $d(R_1, R_1) = 0$ (reflexivity)

If d satisfies the following conditions, then it is both distance and metric:

1. $d(R_1, R_2) = 0$ if and only if $R_1 = R_2$ (identity)
2. $d(R_1, R_2) \leq d(R_1, R_3) + d(R_2, R_3)$ (triangular inequality)

Distance and correlation represent two equivalent approaches calculations that can give relevant information about the degree of agreement between rankings. It is, in fact, demonstrable that the distance can be converted into a correlation coefficient through a linear transformation. The correlation varies between ± 1: a value of -1 indicates that the judges are in total disagreement, while the index is equal to +1 when the rankings are equal. It is possible to convert any correlation coefficient to a distance through the linear transformation $d = 1 - c$. When the correlation is maximum, it is evident that the distance assumes a value equal to zero. Similarly, starting from the distance value, the correlation coefficient associated with it can be calculated (Edmond and Mason, 2002) as follows

$$c = 1 - \frac{2d}{D_{max}} \tag{2.3}$$

2.2 Kendall distance and correlation indexes

In 1938 Kendall provided his first contribution to the study of rankings which, despite the work previously carried out by other authors, marked the starting point for the first wave of contributions to the analysis of rankings, intended as a new branch of statistics. He defined a distance measure that naturally fits the structure of the permutation polytope (Kendall, 1938). Given two different rankings, R_1 and R_2, consisting of n_o objects, the Kendall distance equals the minimum number of interchanges between adjacent objects necessary to transform R_1 into R_2. For example, looking at Figure 2.1, the Kendall distance between (A B C) and (B C A) is equal to two, since two exchanges between pairs of adjacent objects are required to transform the first ordering into the second.

Given two rankings R_1 and R_2 a pair of objects i, j is defined as discordant if the two judges have opposite relative preferences. In this case, Kendall's distance will be

$$d(R_1, R_2) = \sum \sum_{1 \leq i \leq j \leq n_o} I[(R_{1,i} - R_{1,j})(R_{2,i} - R_{2,j}) < 0], \tag{2.4}$$

where I is the indicator function. If two judges express the following preferences R_1 = (A B) and R_2 = (B A), then the associated rankings will be R_1= (1 2) and R_2 = (2 1). In this case $(R_{1,i} - R_{1,j})$ indicates the difference between A and B in the first ranking. Since A is preferred to lower than B, then it turns out that this difference equals to 1. On the contrary, the second judge prefers B to A so that $(R_{2,i} - R_{2,j})$= 1. The product of these two differences is a negative value, and the indicator function is applied to this result. We obtain that the distance between the two rankings is equal to 1. The Kendall distance is widespread within numerous models, such as, for example, in the Mallows-ϕ model (Mallows, 1957) discussed in Chapter 3.4 as a criterion for generating data samples for a simulations study. Kendall introduced the concept of a ranking matrix (i.e., score matrix) to calculate a correlation coefficient. Each rankings R_i of n_o objects is associated with a matrix $n_o \times n_o$ whose elements α_{ij} are obtained as follows

$$\alpha_{ij} = \begin{cases} +1 & \text{if i is preferred to j} \\ -1 & \text{if j is preferred to i} \\ 0 & \text{if i and j are in tie or i=j} \end{cases}$$

At this point the correlation between R_1 and R_2 with respective score matrices α_{ij} and β_{ij} is defined as

$$\tau_b(R_1, R_2) = \frac{\sum_{i=1}^{n_o} \sum_{j=1}^{n_o} \alpha_{ij}\beta_{ij}}{\sqrt{\sum_{i=1}^{n_o} \sum_{j=1}^{n_o} \alpha_{ij}^2 \beta_{ij}^2}} \tag{2.5}$$

Kendall defined τ_b as a "measure of the proximity of the agreement between two given rankings, in the sense that it measures how accurate one of the two rankings would be if the other were objective" (Kendall and Smith, 1940). Before reaching this conclusion, Kendall experimented with another index, called τ_a. It differs from τ_b for the value expressed in the denominator, which corresponds to the maximum distance, that is $n_o \times (n_o - 1)$. However, τ_a was little used because it does not satisfy the identity property. In fact, in the presence of weak orderings, the correlation between a ranking and itself is less than one (Emond and Mason, 2002). He showed that τ_b distributes like a normal. However, it presents problems when ties are allowed. The correlation between an all-tied ranking and another one is equal to the indefinite form 0/0. This result could be treated as if it corresponded to a value equal to zero, but this would imply that the correlation between the ranking mentioned above and any other, including itself, must be zero. Furthermore, τ_b does not respect the triangular distance property and gives illogical results even in simple cases. For example, for the following preferences (A B), (A B), and (B A), for the criterion of maximum correlation, the consensus ranking is (A B). However, when a third item C is entered, which the three judges rank last, the consensus ranking becomes (A-B C), although expected to be (A B C). The introduction of the third irrelevant object has produced a result such that A and B are in tie and, therefore, an illogical solution. In conclusion, the Kendall distance correlation coefficient calculates the disagreement between full rankings using the minimum number of interchanges between adjacent objects necessary to transform one ranking into another. This index works

satisfactorily when ties are not allowed.

The Kendall correlation index τ can be considered as a "disorder coefficient" because it is linked to the Kendall distance as shown in equation 2.6.

$$\tau = 1 - \frac{2d}{n_o(n_o - 1)/2} \tag{2.6}$$

This equivalence between a correlation and a distance measure highlights their connection with the permutation polytope and demonstrates their fundamental importance for the study of rankings: the natural distance measure defined on the permutation polytope is the Kendall distance. Given two different rankings on the permutation polytope of n_o objects, the Kendall distance counts the total number of steps to migrate from R_1 and R_2 by reversing adjacent pairs of objects (W. Heiser, 2004). The value in the denominator is equal to the maximum computable distance on $\frac{n_o(n_o-1)}{2}$ pairs of objects.

2.3 Kemeny distance and the score matrix concept

In 1962 Kemeny & Snell proposed a set of four axioms applicable to any distance index $d(R_1, R_2)$ between two weak orderings R_1 and R_2 of n_o objects Kemeny and Snell, 1962

1. $d(R_1, R_2) \geq 0$ and assumes a value equal to zero if and only if $R_1 = R_2$;
2. $d(R_1, R_2) = d(R_2, R_1)$;
3. $d(R_1, R_2) + d(R_2, R_3) \geq d(R_1, R_3)$;
4. If R'_1 and R'_2 result from the same permutation of objects applied to both R_1 and R_2, then $d(R_1, R_2) = d(R'_1, R'_2)$;
5. If R_1 and R_2 coincide except for a set S of k objects, then $d(R_1, R_2)$ could be computed as if these k objects were the only ones classified;
6. The minimum positive distance is equal to 1.

The first three axioms, already shown before, concern the properties that each distance must respect to be considered as a metric. The fourth axiom guarantees that no permutation of objects from their initial position is decisive for calculating distance (invariance to random permutations). Axiom 5 requires consistency in measurement when the number of objects varies, so if two rankings of n_o objects indicate the same preferences, except for a subset of k objects placed in the middle of the ranking, then only the latter will be considered for the distance calculation. This condition ensures that the inclusion of irrelevant alternatives does not affect consensus ranking. Finally, the last axiom requires that the distance is also a unit of measurement. Kemeny & Snell, referring to Kendall's studies, proved that only one metric satisfies these axioms. They followed the procedure illustrated for the construction of the score matrix α_{ij} and arrived at the following formulation

$$d_{Kem}(R_1, R_2) = \frac{1}{2} \sum_{i=1}^{n_o} \sum_{j=1}^{n_o} |\alpha_{ij} - \beta_{ij}| \tag{2.7}$$

This formula allows to calculate the distance between two rankings R_1 and R_2, even in the presence of ties and partial orderings, in terms of the interchanges required between pairs of adjacent objects to transform one (partial) ranking into another. If ties are allowed, it is preferable to use Kemeny's distance; otherwise, it leads to the same results as Kendall's distance. The procedure for constructing the score matrices associated with each ranking is identical to that described above for Kendall's distance. For example, if $R_1 = (A, C, D, B)$ and $R_2 = (C, B, D, A)$ then the score matrixes associated with the two rankings $R_1 = (1\ 4\ 2\ 3)$ and $R_2 = (4\ 2\ 1\ 3)$ are shown in Figure 2.4.

For the Kemeny distance, the point-by-point differences of these matrices must be added in absolute value. Since the result must be divided by two, this is equivalent to making the following calculation

$$d_{Kem}(R_1, R_2) = 8 \tag{2.8}$$

Note that, for the formula $n_o \times (n_o - 1)$, the maximum distance that can be

α	A	B	C	D
A	0	1	1	1
B	-1	0	-1	-1
C	-1	1	0	1
D	-1	1	-1	0

β	A	B	C	D
A	0	-1	-1	-1
B	1	0	-1	1
C	1	1	0	1
D	1	-1	-1	0

Figure 2.4: Score matrices α and β associated to rankings R_1 and R_2, respectively, where R_1 = (1 4 2 3) and R_2 = (4 2 1 3)

calculated between rankings of four objects is equal to 12. In the case shown, the distance between R_1 and R_2 is equal to 8, the two judges are in disagreement regarding only two preferences (both prefer the object C to D and B). The distance between two rankings R_1 and R_2 can be calculated with an analogous formula in which the sign function is present. This variant generates the same results as the one seen previously and is set out for the sake of completeness

$$d_{Kem}(R_1, R_2) = \sum_{i<j=1}^{n_o} |sgn(R_{1,i} - R_{1,j}) - sgn(R_{2,i} - R_{2,j})|, \tag{2.9}$$

where $i < j$. This formula does not require the construction of score matrices: the distance between the two rankings is calculated by applying the sign function to the difference between objects i and j. The sign function returns different values depending on the results obtained. The differences and the presence of the absolute value make this procedure similar to that which requires the construction of the score matrix. Indicating with Δ the resultant of the difference $(R_{1,i} - R_{1,j})$, the score matrices are composed by the following values

$$sgn(\Delta) = \begin{cases} +1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \\ 0 & \text{if } a = 0 \end{cases}$$

Kemeny and Snell applied d_{Kem} to the search for the median ranking, defined as that point in the rankings space Z^{n_o} more in agreement with the preferences expressed by the judges. When ties are allowed, the number of all possible rankings of n_o objects is equal to

$$\frac{1}{2}\left(\frac{1}{ln(2)}\right)^{n_o+1} n_o! \tag{2.10}$$

The larger n_o is, the larger the space the search is carried out becomes (Gross, 1962). More formally, given a set of H rankings $R_{i=1}^{H}$, the median ranking S is that point (or those points) for which the sum of the distances between it and all the others is minimal.

$$\sum_{i=1}^{H} d(R_i, S) = min, \tag{2.11}$$

with $S \in Z^n$

Similarly, the median ranking can be searched through the correlation coefficient obtainable from the distance measure using the linear transformation seen previously. Therefore, the median ranking can be considered as that point in the space of rankings such that the sum of the correlations between it and all the others is maximum.

$$\sum_{i=1}^{H} \tau(R_i, S) = max \tag{2.12}$$

2.4 Edmond and Mason extended correlation coefficient

Due to the problems related to the Kendall correlation coefficient for the search for consensus ranking, Edmond & Mason,(2002), proposed an extension of this index and reviewed the concept of "interchange" treated by Kendall. They applied the two-step method of half-flip for rankings without ties. Two adjacent objects (A B) can be interchanged with a first half-flipping in such a way as to form a tie (A-B), to then be divided and swapped positions in a second step (B A). The distance between rankings admits ties and is closely related to the new correlation coefficient τ_x proposed by Edmond & Mason. They proved that this index satisfies the Kemeny-Snell axioms as well as the distance calculated with the half-flip. Note that the two passes of the half-flip form a single interchange in Kendall's distance. The extended correlation coefficient τ_x differs from Kendall's one in the way ties are treated. A weak ordering R_1 of n_o objects can be represented by a new score matrix of size $n_o \times n_o$, whose elements α_{ij}^* take on the following values

$$\alpha_{ij}^* = \begin{cases} +1 & \text{if } i \text{ is preferred to } j \text{ or if they are in tie} \\ -1 & \text{if } j \text{ is preferred to } i \\ 0 & \text{if } i = j \end{cases}$$

The correlation between R_1 and R_2 is given by the point-by-point product of their associated matrices α^* and β^*

$$\tau_x(R_1, R_2) = \frac{\sum_{i=1}^{n_o} \sum_{j=1}^{n_o} \alpha_{ij}^* \beta_{ij}^*}{n_o(n_o - 1)} \tag{2.13}$$

This index is identical to Kendall's τ_a when ties are not allowed. Otherwise, the Edmond & Mason coefficient differs since ties assume a value equal to 1 rather than 0. The denominator indicates the maximum distance calculated on the entire score matrix so that the coefficient can assume a maximum value equal to ad 1. The matrix of scores

associated with a ranking has extra-diagonal values equal to $\pm$ 1. However, if partial orderings are allowed, there may be cells whose value equals 0. This means that there is no information about the preference for that pair of objects. It should also be specified that the method proposed by Edmond & Mason is based on a different interpretation of the concept of a tie: the assignment of the same value to multiple objects by a judge does not imply indifference. Indeed, this means that that judge would be fully satisfied if the final choice fell on all the objects that he likes equally. Therefore, equality can be interpreted as a positive declaration of agreement between two or more objects. Given H weak rankings $R_{h=1}^{H}$, each of these can be associated with a weight w_i related to the importance of the preference expressed by a judge over the others. At this point, the S ranking is sought that maximizes the weighted average correlation between it and all the other H rankings

$$\frac{\sum_{h=1}^{H} w_h \tau_x(S, R_h)}{\sum_{h=1}^{H} w_h} \tag{2.14}$$

By replacing the coefficient with the formula seen above and indicating with s_{ij} the score matrix associated with the consensus ranking, the condition to be checked is the following

$$\frac{\sum_{h=1}^{H} w_h (\sum_{i=1}^{n_o} \sum_{j=1}^{n_o} s_{ij} \alpha_{ij}^{h})}{n_o(n_o-1) \sum_{h=1}^{H} w_h} = max, \tag{2.15}$$

where α_{ij}^{h} is the score matrix, as defined by Edmond & Mason, associated with the ranking R_h. For the maximization problem, the denominator can be ignored. By moving the sum of the weights within the parenthesis to the numerator, the final expression to be maximized is obtained

$$\sum_{i=1}^{n_o} \sum_{j=1}^{n_o} s_{ij} c_{ij}, \tag{2.16}$$

where $c_{ij} = \sum_{h=1}^{H} w_h \alpha_{ij}^{h}$. This matrix is given by the sum of all the scores matrices and is called a combined input matrix. It contains all the information about the starting

data set, so the search for the consensus ranking becomes faster since it will be enough to calculate the product point by point between the score matrices of the candidate rankings and the combined input matrix. The ranking that maximizes this product will solve the research problem of consensus ranking (Emond and Mason, 2002, p. 23). The combined input matrix satisfies the following properties

1. If $c_{ij} = 0 \ \forall i, j$, then each ranking constitutes a solution;
2. If $sgn(c_{ij}) = 1 \ \forall i, j$, then the solution is an all tied ranking;
3. If the combined input matrix is a valid score matrix, the only solution is that ranking represented by that same matrix.

The latter case rarely occurs. Unfortunately, the space of possible consensus candidate rankings expands as the number of objects n_o increases; in fact, three objects correspond to 13 weak orderings, four objects correspond to 75, and so on. To solve this issue, algorithms have been implemented, such as branch-and-bound (Emond and Mason, 2002), which in the presence of weak orderings of about 20 objects provide a solution (or more solutions) in a reasonable amount of time.

2.5 Position weights for distance and correlation measures

Garcıa-Lapresta and Pérez-Román, 2010, introduce in the distance calculation the possibility of weighting the discrepancies between weak orderings to take into account where such disagreements occur. They start from the assumption that in some decision problems, it may be helpful to understand whether the judges' choices differ about the objects classified in the first positions rather than in the last ones. The introduction of different weights allows guesses where these discrepancies occur. Bosch, 2006, introduced the concept of consensus measure as an index that assigns a number from 0 to 1 to any linear order and satisfies three properties:

1. Unanimity: for each subgroup of agents, the highest degree of consensus is reached only in the case in which all individuals express the same preferences;

2. Anonymity: the degree of consent does not undergo variations as a result of permutations of agents;

3. Neutrality: the degree of consent does not vary as alternatives are exchanged.

Garcìa-Lapresta and Pérez-Romàn extended these properties to weak orders, in which ties between objects are allowed, and considered two other properties that consensus measures should satisfy:

1. Maximum dissent: in each subset of two agents, the minimum consent occurs whenever linear orders represent the agents' preferences, and each of these is the inverse of the others;

2. Reciprocity: if all the orders are reversed, the consensus does not change. They pointed out that Kemeny distance does not consider the position in which there is disagreement among the judges.

For example, given a set of objects (A B C D) and three rankings R_1, R_2, R_3 a situation like this can occur:
R_1 = (A B C D);
R_2 = (A B D C);
R_3 = (B A C D).

The first ranking R_1 and the second R_2 differ for the objects ranked in the last position. By calculating the Kemeny distance, we obtain that $d_{Kem}(R_1, R_2) = 2$. Conversely, R_1 and R_3 differ for the choices in the first position. The third judge prefers object B to A. Similar to the previous case, we obtain that $d_{Kem}(R_1, R_3) = 2$. Despite this, it seems reasonable that the first ranking is more similar to the second than the third since it is only in the last positions that there is a discrepancy between R_1 and R_2. For this reason, Garcìa-Lapresta and Pérez-Romàn introduce the weighted Kemeny distance. Given a vector of weights $w = (w_1, ..., w_{n_o-1}) \in [0,1]^{n_o-1}$ such that $\sum w_i = 1$ and

$w_1 \geq ... \geq w_{n_o-1}$, the weighted Kemeny distance between two weak orderings A and B is defined as follows

$$\begin{aligned} d_{K,w}(R_1, R_2) =& \frac{1}{2}\left[\sum_{i<j=1}^{n_o} w_i |sgn(R_{1,i}^{\sigma_1} - R_{1,j}^{\sigma_1}) - sgn(R_{2,i}^{\sigma_1} - R_{2,j}^{\sigma_1})| + \right. \\ &\left. + \sum_{i<j=1}^{n_o} w_i |sgn(R_{2,i}^{\sigma_2} - R_{2,j}^{\sigma_2}) - sgn(R_{1,i}^{\sigma_2} - R_{1,j}^{\sigma_2})|, \right. \end{aligned} \tag{2.17}$$

where:

- σ_1, σ_2 long to the set of permutations S_{n_o} that can be performed at n_o objects;
- σ_1, σ_2 are such that $R_1^{\sigma_1} = R_2^{\sigma_2} \equiv (1, 2, ..., n_o)$;
- $(R_{1,1}, .., R_{1,n_o}) \equiv R_1, (R_{2,1}, .., R_{2,n_o}) \equiv R_2$

By applying this formula to the previous example, we obtain that $d_{K,w}(R_1, R_2) = 1/3$ and $d_{K,w}(R_1, R_3) = 1$. The introduction of weights confirmed what seemed logical and had not been possible to deduce with the use of unweighted distance measures. This weighted distance has the following properties:

1. $d_{K,w}$ is a neutral distance in the set of weak orderings;
2. $d_{K,w}$ does not always verify the triangular inequality;
3. $d_{K,w}$ checks the identity property if and only if $w_{n-1} > 0$;
4. The maximum distance is $max(d_{K,w}) = 2\sum_{i=1}^{n_o-1}(n_o - 1)w_i$

The exact distance can be calculated using the construction of the score matrices α and β with the Edmond and Mason method. In this case the formula becomes:

$$d_{K,w}(R_1, R_2) = \frac{1}{4}\sum_{i=1}^{n_o}\sum_{j=1}^{n_o}\left[w_{ij}\left(\left|\alpha_{ij}^{\sigma_1} - \beta_{ij}^{\sigma_1}\right| + \left|\beta_{ij}^{\sigma_2} - \alpha_{ij}^{\sigma_2}\right|\right)\right] \tag{2.18}$$

	a	b	...	n
a	0	w_a	w_a	w_a
b	$-w_a$	0	w_b	w_b
...	$-w_a$	$-w_b$	0	$w_{...}$
n	$-w_a$	$-w_b$	$-w_{...}$	0

Figure 2.5: Score matrices of weights applied to each ranking of n_o objects. Note that the vector of weights require $n_o - 1$ values. It is like we assign a weight equal to zero to the last object, even if, given the score matrix structure, we don't see this value.

Note that w_{ij} is the asymmetric matrix of weights of size $n_o \times n_o$ associated to the vector of weights $w = (w_1, ..., w_{n_o-1})$. Given a set of objects $R = (a, b, ..., n_o)$, the score matrix of weights is shown in Figure 2.5.

The attribution of different weights to each object allows calculating the distance between rankings giving appropriate relevance to the alternatives presented to the judges. The final aim is to find that ranking is more in agreement with the others, and, as seen in the previous chapters, this problem can be solved by using an appropriate correlation coefficient (Plaia and Sciandra, 2019). As with weighted distance, it is possible to adopt a correlation index that attributes the right weight to each object to be classified. This index respects the Kemeny axioms and can be obtained through a linear transformation operated on the weighted distance. Starting from the weighted Kemeny distance and using the formula that transforms a distance into a correlation index, Plaia et al., 2019, presented an extension of the correlation coefficient proposed by Edmond & Mason. Given two rankings, R_1 and R_2, of n_o objects, the weighted correlation between the two is equal to

$$\tau_x^w(R_1, R_2) = \frac{\sum_{i<j=1}^{n} a_{ij}^{\sigma_1} b_{ij}^{\sigma_1} w_i + \sum_{i<j=1}^{n} b_{ij}^{\sigma_2} a_{ij}^{\sigma_2} w_i}{2\sum_{i=1}^{n-1}(n-i)w_i}, \tag{2.19}$$

where:

- α_{ij}, β_{ij} represent the score matrices built on rankings R_1 and R_2 by following the Edmond and Mason procedure;
- σ_1, σ_2 belong to the set of permutations S applicable to n_o objects;
- σ_1, σ_2 are such that $\alpha^{\sigma 1} = \beta^{\sigma_2} \equiv (1, 2, ..., n_o)$;
- w_i is the weight vector $w = (w_1, ..., w_{n_o-1})$.

This index satisfies the equality that links the weighted distance $d_{K,w}$ to the coefficient $\tau_{x,w}$ through the following linear transformation

$$\tau_{x,w}(R_1, R_2) = 1 - \frac{2d(R_1, R_2)}{d_{max}} \tag{2.20}$$

The correlation coefficient is an index that allows faster and more natural interpretations than the distance. For this reason, it may be appropriate to prefer a correlation measure for finding and interpreting the median ranking (or social choice). Due to its complexity, the determination of social choice may require the use of computational methods such as, for example, genetic algorithms belonging to the broadest class of evolutionary algorithms. The following paragraph will show how they work, focusing on a specific algorithm capable of identifying the ranking that best represents the set of preferences when a set of weights occurs. Subsequently, we will proceed with an exposition of some simulations and applications designed to verify how introducing a set of weights can affect the detection of the final consensus ranking.

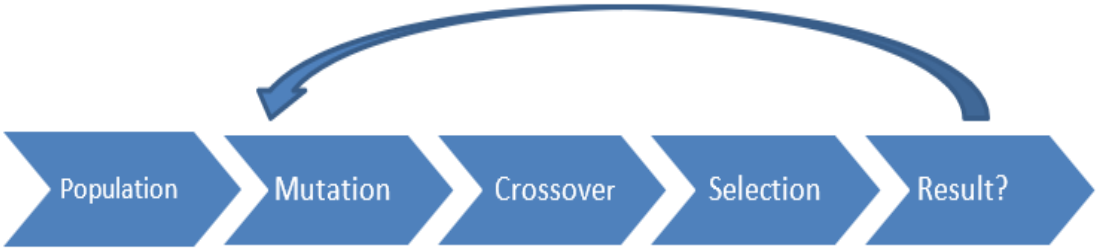


Figure 2.6: A generic process of differential evolution

2.6 The Weighted Differential Evolutionary algorithm for finding Consensus Ranking (WDECoR)

Evolutionary algorithms constitute a set of meta-heuristic methods used successfully in problems of great complexity. They are inspired by the process of evolution and the Darwinian theory of natural selection. Specifically, given a population of individuals, the data is perturbed to generate a selection that improves results. Usually, an initial population is randomly generated, made up of potential solutions, and then adopts a function that acts as an adaptation measure. The initial population is mutated based on a mechanism by which the best mutation persists and takes the place of the previous one. Through an iterative process, the population evolves generation by generation, leading to continuous optimization. Among these algorithms, the most famous are the genetic algorithms. Differential evolution solves optimization problems through alterations and selection operators. The method uses a fixed number P of vectors x, which constitute the initial population in each generation G. This set is chosen randomly. The method used to generate new vectors plays a key role, as the mutation and crossover operations depend on it. If the new vector fits the results better than the previous one, then the new one takes the place of the old one in the next generation. For each generation, the best individual is memorized, that is, the vector $x_{best,G}$, which minimizes the cost function.

For the research of the consensus ranking, the method called "Differential Evolutionary for Consensus Ranking" (DECoR) was proposed and implemented in *ConsRank* *R* package (D'Ambrosio et al., 2019). The algorithm requires as cost function to be min-

imized the sum of the Kemeny distances between the consensus S and all the rankings represented by the combined input matrix c_{ij}.

$$cost(S) = \sum_{h=1}^{H} w_h \frac{n_o(n_o-1)}{2}[1 - \tau_x(c_{ij}, s_{ij})] \tag{2.21}$$

Note that $tau_x(c_{ij}, s_{ij})$ is the coefficient that measures the correlation between the combined input matrix c_{ij} and the consensus candidate ranking s_{ij}. Furthermore, w_k indicates the frequency of the k-th ranking. It acts as weight within the cost function but should not be confused with the weights mentioned in the previous paragraph, which do not represent frequencies, but the weight attributed to each object within the classification. $cost(S)$ is none other than Kemeny distance multiplied by the frequency of each ranking. From the linear transformation, it is possible to resort to the following formulation

$$[1 - \tau_x(c_{ij}, s_{ij}] = \frac{2d(c_{ij}, s_{ij})}{n_o(n_o-1)} \tag{2.22}$$

By replacing this value, the cost function to be minimized is obtained

$$cost(S) = \sum_{h=1}^{H} w_h \frac{n_o(n_o-1)}{2} \frac{2d(c_{ij}, s_{ij})}{n_o(n_o-1)} \tag{2.23}$$

Once the appropriate simplifications have been made, it is clear that the cost function to be minimized is the distance, weighted by the frequencies, between the combined input matrix and the consensus candidate ranking

$$cost(S) = \sum_{h=1}^{H} w_h d(c_{ij}, s_{ij}) \tag{2.24}$$

Note that $d(c_{ij}, s_{ij})$ is identically equal to the sum of the distances between all rankings contained in the combined input matrix and the candidate for consensus ranking. Therefore, if we had three rankings A, B, C, whose respective score matrices are represented by α, β, γ, we obtain the following equivalence, valid for both distance and correlation

$$d(c_{ij}, s_{ij}) = d(a_{ij}, s_{ij}) + d(\beta_{ij}, s_{ij}) + d(\gamma_{ij}, s_{ij}) \tag{2.25}$$

$$\tau_x(c_{ij}, s_{ij}) = \tau_x(a_{ij} + \beta_{ij} + \gamma_{ij}, s_{ij}). \tag{2.26}$$

When different weights are attributed to objects classified by the judges, it is necessary to resort to this equality since it is impossible to substitute the formula of distance or weighted correlation in Equation 2.15. In fact, such weighted calculations would require the construction of a weighted combined input matrix and permutations, which would lead to meaningless results. For this reason, when calculating the distance/correlation between multiple rankings with the adoption of weights, the sum of the weighted distances/ correlations between each ranking of the starting dataset and the consensus candidate is used. This procedure increases the computational cost of the consensus ranking search. The combined input matrix constitutes a synthesis of all the data available, so its use would speed up the search for consensus ranking. Future studies and insights are directed in this direction to solve this problem and be able to use more efficient evolutionary algorithms, which quickly return solutions even in the presence of weighted objects. The operation of the DECoR algorithm is summarized below as a pseudo-code:

- **input**: population size NP and number of generations L in which no improvements appear (arbitrary stop criterion)
- population initialization;
- the best ranking of the first generation is stored;
- **while** repeat until improvements are made (number of improvements ¡L);
- **for** i loop = 1:NP;
- Evolution = population mutation;
- Evolution = population crossover;
- Evolution = population discretization;

- **If** cost of evolution $\leq$ cost of the h-th ranking of the population:
- the h-th ranking of the population = evolution;
- **End of cycle for**;
- the best of this generation is stored;
- **If** the cost of the best of this generation is equal to the cost associated with the best of the previous generation, then the number of missed improvements L increases by one, otherwise $L = 0$;
- **End while loop**
- **output**: all the best solutions obtained in the last generation.

For a complete description of DECoR refer to D'Ambrosio et al. (2016). Note that the DECoR algorithm includes the input parameters F and CR, corresponding to the scale factor and the crossover ratio. These two parameters are used to change the initial population through the following procedure: three different individuals are chosen at random, and a new vector is generated, then the crossover is applied, which consists in generating random numbers between 0 and 1 for each element vector. If these random numbers are more significant than the crossover ratio, the values are accepted and represent a solution; otherwise, the current index values remain. The R function $DECORcore$ descends from the DECoR algorithm, which recalls the primary function of the more generic version. Several simulations show that the algorithm is accurate, although it often provides optimal local and not global solutions. DECoR is more robust and faster than other algorithms and has the same accuracy as branch-and-bound and other heuristic algorithms. When the number of objects is less than 50, it is preferable to use the QUICK algorithm (Amodio et al., 2016). If, on the other hand, the number of objects is greater than 100, DECoR provides a solution much faster than the QUICK algorithm. A weakness of the DECoR, and all the heuristic algorithms, is the phase of choosing the parameters F and CR.

As explained several times, we often resort to the use of weights in order to attribute greater or lesser importance to the objects of the classification. In this case,

it is not possible to use the DECoR algorithm to search for the consensus ranking, as it requires the combined input matrix as input. Calculating the weighted Kemeny distance and its correlation coefficient requires the performance of permutations on the rankings and the objects' weighing procedure, operations that should be applied on the combined input matrix and require more in-depth studies. The $WDECoR$ differential evolutionary algorithm is introduced to solve this problem, which is a weighted version of the code shown above. It follows the same scheme illustrated by the pseudo-code, with the difference that the cost function to be minimized is the weighted Kemeny distance. At the same time, the solution is represented by that ranking that maximizes the sum of the weighted average correlations between it and all the rankings present in the starting data set.

The Weighted Differential Evolutionary for Consensus Ranking (WDECoR) is an algorithm that searches for solutions efficiently and reasonably quickly. The higher the L index selected as an input, the more expensive the search for consensus ranking will be, but at the same time, it will provide solutions more similar to the global solution. If the default parameters are left unchanged[1], the algorithm returns solutions in about thirty seconds for a data set consisting of ten rankings and ten objects. The research time increases as the number of classified objects increases. When there are more than ten, it is recommended to reduce the value of L. However, this solution is not optimal in terms of results, as the algorithm will return local solutions that can be very different from the actual consensus ranking. A crucial point is how fast solutions are found when the number of rankings is high and only a few objects are ranked. For example, if the starting data set contains twenty rankings and five objects with unchanged input parameters, the output is returned in about ten seconds. Several simulations show that the time to search for consensus rankings increases more than proportional to the increase in the number of objects, but the increase will be less than proportional when the number of judges increases.

In the previous paragraphs we presented weights as values belonging to the vector of the type $w = (w_1, ..., w_{n_o-1}) \in [0,1]^{n_o-1}$ such that $\sum w_i = 1$, with $w_1 \geq \cdots \geq w_{n_o-1}$.

[1]population size NP=15, generations limit L=50, scaling rate for mutation FF=0.4, crossover range CR=0.9, search in the space of all possible permutations FULL=FALSE

However, it is necessary to investigate the nature of this vector and its influence on the rank aggregation problem. The choice of the weights to assign to the items affects the solutions obtained at the end of the research phase of the consensus ranking. Assigning zero weight to more than one object means focusing on the remaining alternatives. In order to investigate if the choice of the weights affects the determination of the consensus ranking, numerous simulations were carried out on data sets generated randomly. Specifically, the data were simulated through the Mallows-ϕ model, in which it is assumed that, given a consensus σ, a distance index d and a real parameter λ, the density function of a Uniform random variable is the following

$$f_\lambda(R;\sigma) = exp(\lambda d(\sigma, R) - \psi(\lambda)), \tag{2.27}$$

where R is a ranking and ψ is a normalization factor. The closest ranking to the median ranking has a higher probability of being extracted and this is due to the parameter λ, which quantifies the concentration of the distribution around the central value. In general: if λ tends to minus infinity, then the sum of the distances between all the rankings and the consensus is equal to zero (the rankings are all equals); if λ tends to infinity, then the sum of the distances is maximum; if λ is equal to zero, then the probability of extraction is equal for each ranking. When Kendall distance is used, the model is called Mallows-ϕ model. With Spearman distance, it is referred to as Mallows-θ model. Some authors define Equation 2.27 by setting the parameter λ with a negative sign. In this case, the density function becomes

$$f_\lambda(R;\sigma) = \frac{exp(-\lambda d(\sigma, R))}{\psi} \tag{2.28}$$

Hence a different interpretation of the concentration parameter: the higher the value of λ, the stronger the consensus around the ranking σ. Conversely, when λ is very low, a non-consensus situation is obtained. Finally, the interpretation of the parameter does not vary if it is zero. When using the Kendall distance, the distribution of the distances is known, so it is possible to estimate the parameter λ through the maximum likelihood method. The maximum likelihood estimate of λ can be obtained, for example,

with the central limit theorem or with the Newton-Raphson algorithm.

The simulations were carried out following the approach such that λ is a negative value using the *R* software and the *ConsRank* and *PerMallows* (Irurozki et al., 2016) packages . In the latter, the *rmm* function generates samples of H permutations from the Mallows model. In order to obtain generalizable results, ten samples composed of 50 and 100 rankings were generated, each of them composed of 5, 10, and 15 objects. Furthermore, three different absolute values of λ were chosen, namely 0, 0.4, and 0.8. The verification of these cases led to the generation of 180 data sets on which the consensus ranking was calculated using unweighted indices and, specifically, using the *QuickCons* function belonging to the *ConsRank* package. Subsequently, two weight vectors were introduced: w_1 that assigns a weight to each position as specified by Garcìa-Lapresta and Pérez-Romàn; w_2 that assigns a weight equal to 0.5 to the first two positions and 0 to all the others. At this point, the search for consensus ranking was carried out on all data sets through the weighted distance and correlation indices using the $WDECORcore$ function. Finally, the consensus ranking obtained without weights was compared with those obtained through the first and second weight vectors for each sample. The design factors are summarized as follows:

- observations $H = 50, 100$;
- objects $n_o = 5, 10, 15$;
- $\lambda = 0, 0.4, 0.8$;
- w_1 = decreasing weights;
- $w_2 = 0.5$ for the first two objects, 0 to the others.

At the end of the search, the code returned a list of ten result matrices. We compare the results obtained in the three procedures for each of them. For example, the first matrix of results calculated on a dataset of 100 rankings and five objects, for which a value of λ of zero has been chosen, is shown in Table 2.1.

Matrix #1	A	B	C	D	τ_x
Cons	4	3	2	1	0.062
Cons^{w_1}	4	3	2	1	0.055
Cons^{w_2}	3	2	2	1	0.049

Table 2.1: First matrix of results calculated on a dataset of $H = 100$ rankings, $n_o = 5$ objects, and $\lambda = 0$. The first row is the consensus ranking when no set of weights is introduced. The second row represents the consensus ranking when applying decreasing weights as specified by Garcia-Lapresta and Perez-Roman. The last row is the consensus ranking obtained when only the first two positions are weighted.

The first row corresponds to the consensus ranking (with the relative correlation coefficient in the last column) obtained with the calculation of the Kemeny distance; the second row is obtained by weighing all the objects; the last row represents the one for which only the first two objects are weighed. The first and second methods return the same solution, albeit with a different correlation index. On the other hand, by assigning values equal to 0.5 to the first two positions, a different solution is obtained, in which A and B are respectively in the third and second position. This is just one of the many observable cases for which it is necessary to resort to an approach that makes the results of the analysis generalizable. In order to verify that the consensus ranking calculated with the Kemeny distance and the one calculated with the first weight system is the same for most of the datasets, it may be convenient to calculate the correlation τ_x between these two. In the previous case in Table 2.1 it translates into calculating the correlation between the first and second row of the observed matrix. Repeating this calculation for the remaining nine matrices belonging to the list of results obtained from the datasets of size 100×5 with λ equal to zero, it is clear that the two methods generate results whose correlation is close to one.

The results indicate that the two methods return in six result consensus rankings with a correlation equal to one. The average of the column values is equal to 0.91. It can therefore be said that for the ten samples extracted from a data set consisting of 100 rankings, 5 objects and with a λ equal to zero, the method that assigns a weight to each object leads to results very similar to those obtained without the adoption of the weights

Matrix	$\tau_x(Cons, Cons^{w_1})$
#1	1.0
#2	1.0
#3	0.9
#4	1.0
#5	1.0
#6	1.0
#7	0.6
#8	0.8
#9	0.8
#10	1.0

Table 2.2: The Table shows in the second column the correlation coefficient τ_x between the first two rows of the 10 matrices. For instance, the first value 1 is the correlation between $Cons$ and $Cons^{w_1}$ in Table 2.1

themselves.

Carrying out the same procedure for all the result matrices, as λ increases, an increase in the average value of the coefficients shown in the column is noted. Specifically, on samples of 100 rankings and 5 objects, with $\lambda = 0.4$ the average of the column values is 0.93, while with $\lambda = 0.8$ the average rises to a value equal to 0.94. The same comparison method cannot be extended to the case in which only the first two positions are weighed. When a weight of zero is assigned to a position, the latter will assume a null value within the calculation of the weighted correlation index. The only evidence from the simulation is that this weighting method generates different solutions from those obtained in the other two cases. The matrices that compare the results of the three methods can be the subject of numerous other analyzes, including the two-factor analysis of variance (ANOVA). The simulation shows that by varying the weights adopted, more or less different consensus rankings can be obtained. Further research into weighting methods could bring out valuable results for those who must make decisions based on the preferences of other individuals. One of the main limitations of weighted approaches is that different judges can give different relevance to ranking positions. Giving greater weight to the objects in

the first positions, what weight to assign to the latter, and the differences in terms of results are the further questions to be asked following what has been demonstrated in this paragraph.

Here, we conducted further investigations to better understand how the choice of weights affects the final results in terms of consensus ranking. We use a dataset collected in 2019 at the Università degli Studi di Cagliari, where 100 first-year students were asked to rank five different objects (1) from the most interesting to the least one, (2) from the most difficult to the easiest one, and (3) from the most time consuming to the least one. The objects are the sequent: A = business, B = mathematics, C = law, D = microeconomics, E = statistics. The survey was conducted at the end of the first academic year so that all the participants had enough information to create a ranking of the five subjects. The simulations purpose is to investigate the correlation between the dispersion parameter λ presented in Equation 2.28 and the extended correlation coefficient τ_x in Equation 2.13. Recall that λ is a spread parameter of the Mallows model that quantifies the concentration around the consensus ranking. So, the larger the lambda, the stronger the consensus. If lambda is equal to zero, then each ranking in the universe is equally likely. Three different weight vectors were chosen to check for any changes in the solution obtained:

- $w_1 = (0.4, 0.3, 0.2, 0.1, 0)$;
- $w_2 = (0.5, 0.5, 0, 0, 0)$;
- $w_3 = (0.8, 0.2, 0, 0, 0)$.

The first weights vector w_1 assigns decreasing values as suggested by 2.17. The second, w_2, assigns the same value to the first two positions of the rank. Finally, the last weighs vector w_3 assigns values to the first two positions only. For each of the three parts of the dataset (i.e., (1) interest, (2) difficulty, and (3) time), we calculate the estimated $\hat{\lambda}$ with the function *lmm.theta* within the PerMallows *R* package. As result, the three parts of our dataset present three different and increasing values of concentration. In the specific, the concentration around the consensus ranking is greater when the students

were asked to rank the five subjects based on the time of preparation for the exam. Starting from the estimated values $\hat{\lambda}$, we generated ten datasets for interest, difficulty, and time, by using the *rmm* function within the PerMallows package. The input are the number of rows 100, the estimated $\hat{\lambda}$, and the consensus ranking. Table 2.3 summarizes data by showing, for each part of the dataset, the estimated $\hat{\lambda}$, and the consensus rankings when using no weights, w_1, w_2, and w_3.

Table 2.3: Simulation results: for each part of the dataset (i.e., interest, difficulty, and time), we show the estimated dispersion parameter $\hat{\lambda}$ (as specified in Eq. 2.28), the consensus ranking without weights *Cons*, the consensus rankings with the weights w_1, w_2, and w_3. The extended correlation coeffient τ_x (Eq. 2.13) is reported in parenthesis. For each part of the dataset, the objects are the same: A = business, B = mathematics, C = law, D = microeconomics, E = statistics.

	(1) Interest	(2) Difficulty	(3) Time
	$\hat{\lambda} = 0.35$	$\hat{\lambda} = 0.63$	$\hat{\lambda} = 0.8$
$Cons$	ADECB (0.28)	ECADB (0.46)	CEADB (0.56)
$Cons^{w_1}$	ADECB (0.31)	ECADB (0.51)	CEADB (0.57)
$Cons^{w_2}$	DACEB (0.28)	EACDB (0.52)	CEADB (0.61)
$Cons^{w_3}$	ADECB (0.34)	ECABD (0.61)	ECADB (0.57)

As expected, assigning decreasing weights to each object does not cause output changes. But, if we assign a weight of 0.5 only to the first two positions, then we get different results in the first two datasets. In fact, the weighted consensus rankings show different preferences in the top two positions. Finally, in the last case, the consensus ranking for dataset time is the only one that changes from the unweighted case. Looking at the last column, the solutions are very similar to each other regardless of the method chosen. Several simulations showed that as the concentration parameter of the Mallows model grows and the number of objects decreases, the weights allocation has a lesser impact on the results of the analysis. So, it is possible to say that the assignment of weights can generate different solutions, but these differences depend on the weight vector w chosen, the concentration λ of judges' preferences around the consensus ranking, the correlation coefficient τ_x associated with the consensus ranking, and the number of objects n_o.

Chapter 3

A new probabilistic approach: the Bradley-Terry regression trunk

3.1 Probabilistic approach

Preference rankings, and generally ordinal data, can be analyzed with several statistical models and methodologies, both supervised and unsupervised. Among these, there are methods based on the goodness-of-fit adaptation and probabilistic methods (W. J. Heiser and D'Ambrosio, 2013; Marden, 1996). The first category includes methods such as Principal Component Analysis (Carroll, 1972), Unfolding (Busing et al., 2005; Busing et al., 2010; Coombs, 1950, 1964; Van Deun et al., 2007), Multidimensional Scaling (W. J. Heiser and De Leeuw, 1981; Hooley, 1993) and Categorical Principal Component Analysis (Meulman et al., 2004). These methods are intended to describe the structure of rank data. On the other hand, the probabilistic methods can assume a homogeneous or heterogeneous distribution of judges. In the first case, they focus on the ranking process assuming solid homogeneity among the judges' preferences. In the second one, the methods are aimed at modeling the population of judges assuming substantial heterogeneity in their preferences. When homogeneity is assumed, probabilistic methods are based on the so-called Thurstonian models (Thurstone, 1927), distance-based and multistage models (Bradley and Terry, 1952; Luce, 1959; Mallows, 1957; Thurstone, 1927), mixtures

of Bradley-Terry-Luce models, mixtures of distance-based models (Croon, 1989; Gormley and Murphy, 2008a; Murphy and Martin, 2003), and probabilistic-distance methods (D'Ambrosio et al., 2019). The probabilistic methods that assume heterogeneity are based on a reasonable concept: different groups of subjects with specific characteristics may show different preference rankings (Strobl et al., 2011). Such heterogeneity can be accounted for by introducing subject-specific covariates, from which mixtures of known sub-populations can be estimated. In most cases, the methods that consider covariates are based either on generalized linear models (Böckenholt, 2001; Chapman and Staelin, 1982; Dittrich et al., 2000; Francis et al., 2002; Gormley and Murphy, 2008b; Skrondal and Rabe-Hesketh, 2003) or recursive partitioning methods (i.e., tree-based) (D'Ambrosio and Heiser, 2016; Lee and Yu, 2010; Plaia and Sciandra, 2019; Strobl et al., 2011).
There is relatively little work about tree-based models for rankings in the literature. Dittrich et al., 2000, proposed a parametric model for the analysis of rank-ordered preference through the Bradley-Terry (BT) type models when categorical subject-specific covariates are observed. Their idea was to transform the (complete) rankings data into paired comparisons and apply a log-linear model for a corresponding contingency table. The authors proposed a procedure for researching the interaction effects between covariates by applying a forward selection and backward elimination procedure. This approach is well suited for hypothesis-based modeling. However, this model requires an adequate selection of the covariates and a distinct choice of the functional form in which these covariates are added to the model (Strobl et al., 2011). For this reason, when no a priori hypotheses are known, it requires the arbitrary introduction of higher-order interactions.

Strobl et al., 2011, proposed a tree-based classifier, where the paired comparisons are treated as response variables in Bradley-Terry models. They found a way to discover interactions when no a priori hypothesis is known, suggesting a model-based recursive partitioning where splits are selected with a semi-parametric approach by looking for instability of the basic Bradley-Terry model object parameters. The final result provides the preference scales in each partition group that derives from the order of object-related parameters, but it does not offer information about how the subject-specific covariates affect the judges' preferences. Therefore, this semi-parametric model returns beta coefficients neither for the main effects nor for the interaction effects between the covariates.

Recently, Wiedermann et al., 2021, extended the Strobl model by combining the log-linear Bradley-Terry (LLBT) model with the model-based recursive partition (MOB) for detecting treatment effect heterogeneity. They proposed a semi-parametric model which distinguishes between focal independent variables and covariates for recursive partition. A score-based procedure, the M-fluctuation test (Zeileis and Hornik, 2007; Zeileis et al., 2008), is used to assess the stability of model parameters, and the pruning procedure is conducted using the AIC.

To overcome the drawbacks characterizing the works of Dittrich et al., 2000 and Strobl et al., 2011, we propose an utterly parametric approach that fits a generalized linear model with a Poisson distribution by combining its main effects with a parsimonious number of interaction effects. Our proposal is framed within the Simultaneous Threshold Interaction Modeling Algorithm (STIMA) proposed by Dusseldorp et al., 2010 and Conversano and Dusseldorp, 2017 that, in the case of a numerical response, is based on the Regression Trunk Approach Dusseldorp and Meulman, 2004. The differences with the Wiedermann model are due to the different split search procedures based on the MOB model. As pointed out by the authors, the testing procedure for the split search can be very challenging (Wiedermann et al., 2021). They use the M-fluctuation test to research the best split covariates, while our method is based on the easy-to-compute decrease in deviance by following the regression trunk approach within the STIMA algorithm. Both methods can deal with continuous or categorical subject-specific covariates, even if our algorithm does not deal with nominal covariates. Furthermore, as in the Wiedermann model, in the STIMA algorithm, it is possible to distinguish between focal predictors and partitioning covariates, choosing the treatment variable as the first split variable. Dealing with paired comparisons, our approach combines the extended log-linear Bradley-Terry model, including subject-specific covariates with the regression trunk. Thus, the proposed model is named *Bradley-Terry Regression Trunk (BTRT)*. It produces an estimated generalized linear model with a log link and a Poisson distribution presenting the main effects part and an interaction effects part, the latter being composed of a restricted number of higher-order interactions between covariates that are automatically detected by the STIMA algorithm. The interaction effect part can be graphically represented in a decision tree structure, called trunk, because few terminal

nodes usually characterize it. Hence, BTRT allows observing the preference scale in each trunk node and evaluating how the probability of preferring specific objects changes for different groups of individuals. The final result is a small tree that represents a compromise between the interpretability of interaction effects and the ability to summarize the available information about the judges' preferences.

3.2 The Bradley-Terry model

The model proposed by Bradley and Terry, 1952, is the most widely used method for deriving a latent preference scale from paired comparison data when no natural measuring scale is available (Strobl et al., 2011). It has been applied in psychology and several other disciplines. Recent applications include, for example, surveys on health care, education, and political choice (Dittrich et al., 2006) as well as psychophysical studies on the sensory evaluation of pain, sound, and taste (Choisel and Wickelmaier, 2007) or in prioritization of balance scorecards (Rodríguez Montequín et al., 2020). The paired comparison method splits the ordering process into a series of evaluations carried out on two objects at a time. Each pair is compared, and a decision is made based on which of the two objects is preferred. This methodology addresses the problem of determining the scale values of a set of objects on a preference continuum that is not directly observable.

Let $\pi_{(ij)i}$ denote the probability that the object i is preferred in the comparison with j. The probability that j is preferred is $\pi_{(ij)j} = 1 - \pi_{(ij)i}$. The basic Bradley-Terry model can be defined as in Agresti, 2002, p. 436-439

$$\pi_{(ij)i} = \frac{\pi_i}{\pi_i + \pi_j}, \tag{3.1}$$

where π_i and π_j are non-negative parameters (also called worth parameters) describing the location of objects on the preference scale.

The BT model can be expressed as a logistic model for paired preference data. Suppose to have a set of n_o objects to be judged. The BT model can be defined as a quasi-symmetry model for paired comparisons with object parameters λ_i^O such that

$$logit(\pi_{(ij)i}) = \log\left(\frac{\pi_{(ij)i}}{\pi_{(ij)j}}\right) = \lambda_i^O - \lambda_j^O, \tag{3.2}$$

where λ_i^O and λ_j^O are object parameters related to π's in Equation (3.1) by $\lambda_i^O = \frac{1}{2}\ln(\pi_i)$. The superscript O refers to object-specific parameters. Thus, $\hat{\pi}_{(ij)i} = \frac{\exp(\hat{\lambda_i}^O - \hat{\lambda_j}^O)}{1+\exp(\hat{\lambda_i}^O - \hat{\lambda_j}^O)}$, where $\pi_{(ij)i} = \frac{1}{2}$ when $\lambda_i^O = \lambda_j^O$. The model estimates $\binom{n_o}{2}$ probabilities, which is the number of paired comparisons with n_o objects. Note that the logit model in Equation (3.2) is equivalent to the model in Equation (3.1). In addition, identifiability of these two formulation requires a restriction on the parameters related on the last object n_o such as $\lambda_{n_o}^O = 0$ and $\sum_i^{n_o} \pi_i = 1$.

For each pair $i \geq j$, let n_{ij} be the number of comparisons made between object i and j, $y_{(ij)i}$ denotes the number of preferences of i to j and $y_{(ij)j} = n_{ij} - y_{(ij)i}$ denotes the number of preferences of j to i. Assuming that n_{ij} comparisons are independent and have the same probability $\pi_{(ij)i}$, the $y_{(ij)i}$ are binomially distributed with parameters n_{ij} and $\pi_{(ij)i}$.

The Bradley-Terry model can also be fitted as a log-linear model (Dittrich et al., 1998; Fienberg and Larntz, 1976; Sinclair, 1982). Among these authors, Sinclair (1982) introduced a different approach: in comparing object i with object j, the random variables $y_{(ij)i}$ and $y_{(ij)j}$ are assumed to follow a Poisson distribution.

Let n_{ij} be the number of comparisons made between object i and j, and $m(y_{(ij)i})$ be the expected number of comparisons in which i is preferred to j. Then, using the respecification proposed by Sinclair and the notation for log-linear models for contingency tables, $m(y_{(ij)i}) = n_{ij}\pi_{(ij)i}$ has a log-linear representation

$$\begin{aligned} \log(m(y_{(ij)i})) &= \mu_{ij} + \lambda_i^O - \lambda_j^O \\ \log(m(y_{(ij)j})) &= \mu_{ij} - \lambda_i^O + \lambda_j^O, \end{aligned} \tag{3.3}$$

where the nuisance parameters μ are defined by

$$\mu_{ij} = n_{ij} - \ln\left(\sqrt{\frac{\pi_i}{\pi_j}} + \sqrt{\frac{\pi_j}{\pi_i}}\right), \tag{3.4}$$

and they can be interpreted as interaction parameters representing the objects involved in the respective comparison, therefore fixing the corresponding n_{ij} marginal distributions. In total, $2\binom{n_o}{2}$ expected counts are estimated.

This approach allows synthesizing the information about all preferences in a unique design matrix. The design matrix is composed of column vectors representing the responses $y_{(ij)}$, the nuisance parameters μ_{ij}, and the object parameters λ_i^O. For example, given three objects (A B C), an example of a design matrix is given in Table 3.1.

Table 3.1: Design matrix with one judge and three objects: The first column indicates if the object i is preferred ($y_{ij} = 1$) or not ($y_{ij} = 0$) in a certain preference for each pair of objects ij. The second column serves as an index for the $n \times (n-1)/2$ comparisons. Finally, preferences are expressed in the last three columns. For example, the first line shows that object B is preferred to A since $y_{ij} = 1$, $\lambda_B^O = 1$, and $\lambda_A^O = -1$.

Response	μ	λ_A^O	λ_B^O	λ_C^O
$y_{AB} = 1$	1	-1	1	0
$y_{AB} = 0$	1	1	-1	0
$y_{AC} = 1$	2	-1	0	1
$y_{AC} = 0$	2	1	0	-1
$y_{BC} = 1$	3	0	1	-1
$y_{BC} = 0$	3	0	-1	1

When $y_{(ij)}$ assumes values of +1 and -1 instead of 1 and 0, the linear predictor η for the basic log-linear Bradley-Terry model is the following (Hatzinger and Dittrich, 2012)

$$\eta_{y_{(ij)i}} = \log(m(y_{(ij)i})) = \mu_{ij} + y_{(ij)i}(\lambda_i^O - \lambda_j^O). \tag{3.5}$$

The log-linear formulation allows extending the model with multiple subject-specific covariates.

3.3 The extended Bradley-Terry model with subject-specific covariates

In some cases, it could be interesting to analyze the variation of preferences according to subject-specific characteristics. The Bradley-Terry model can be extended to incorporate categorical or continuous covariates. For a categorical covariate S, let $m(y_{(ij)i,l})$ be the expected number of preferences for i compared with j, among individuals classified in covariate category l, with $l = 1 \dots L$, where L represents the total number of levels of the covariate. The Bradley-Terry model is then specified as

$$\begin{aligned} \log(m\left(y_{(ij)i,l}\right)) &= \mu_{ij,l} + \lambda_i^O - \lambda_j^O + \lambda_l^S + \lambda_{i,l}^{OS} - \lambda_{j,l}^{OS} \\ \log(m\left(y_{(ij)j,l}\right)) &= \mu_{ij,l} - \lambda_i^O + \lambda_j^O + \lambda_l^S - \lambda_{i,l}^{OS} + \lambda_{j,l}^{OS}. \end{aligned} \tag{3.6}$$

The parameter λ_l^S represents the main effect of the subject-specific covariate S measured on its l-th level; $\lambda_{i,l}^{OS}$ and $\lambda_{j,l}^{OS}$ are the subject-object interaction parameters describing the effect of S observed on category l and concerning the preference for object i and j, respectively. The model parameters of interest $\lambda_{i,l}^{OS}$ and $\lambda_{j,l}^{OS}$ can again be interpreted in terms of log-odds and as a log-odds ratio

$$\log\left(\frac{\pi_{(ij)i,l}}{\pi_{(ij)j,l}}\right) = 2(\lambda_i^O + \lambda_{il}^{OS}) - 2(\lambda_j^O + \lambda_{jl}^{OS}). \tag{3.7}$$

If the covariate S has no effect on the preferences of the judges, then $\lambda_{i,l}^{OS} = 0$. It means that the model collapses into the previously described basic BT model, and there is just one log-odds for the comparison of two specific objects. However, if there is a covariate effect so that there is at least one interaction parameter between the individuals and the subject-specific covariate that is significantly different from 0, we must distinguish different log-odds for each comparison and each significant subject-object interaction parameter (Hatzinger and Dittrich, 2012).

When continuous subject-specific covariates are included, it is necessary to build up a separate contingency table for each judge, and each different value of the covariate. Table 3.2 shows an example in which two judges, with different ages, express their

preferences regarding three objects.

Table 3.2: Design matrix with two judges, three objects, and one continuous subject-specific covariate: The first column indicates if the object i is preferred ($y_{ij} = 1$) or not ($y_{ij} = 0$) in a certain preference for each pair of objects ij. The second column serves as an index for the $n \times (n-1)/2$ comparisons. Preferences are expressed in the next three columns, and finally the age covariate is showed in the last column.In this example, the two judges express opposite preference, BCA and ACB respectively

Response	μ	λ_A^O	λ_B^O	λ_C^O	*age*
$y_{AB} = 1$	1	-1	1	0	23
$y_{AB} = 0$	1	1	-1	0	23
$y_{AC} = 1$	2	-1	0	1	23
$y_{AC} = 0$	2	1	0	-1	23
$y_{BC} = 1$	3	0	1	-1	23
$y_{BC} = 0$	3	0	-1	1	23
$y_{AB} = 0$	1	-1	1	0	24
$y_{AB} = 1$	1	1	-1	0	24
$y_{AC} = 0$	2	-1	0	1	24
$y_{AC} = 1$	2	1	0	-1	24
$y_{BC} = 0$	3	0	1	-1	24
$y_{BC} = 1$	3	0	-1	1	24

Hence, the LLBT equation for the h-th judge and objects i and j is

$$\log(m\left(y_{(ij)i,h}\right)) = \mu_{ij,h} + y_{(ij)i,h}(\lambda_{i,h}^O - \lambda_{j,h}^O). \tag{3.8}$$

The parameter $\lambda_{i,h}^O$ can be expressed through a linear relation

$$\lambda_{i,h}^O = \lambda_i^O + \sum_{p=1}^{P} \beta_{ip} x_{p,h}, \tag{3.9}$$

where $x_{p,h}$ corresponds to the value of the x_p-th continuous covariate ($p = 1...P$) observed for judge h. The parameters β can be interpreted as the effect of the covariates on object

i, whilst λ_i^O acts as intercept and indicates the location of object i in the overall consensus ranking.

Following this approach, it is possible to compute the deviance of the model as the deviance of a fitted Poisson regression

$$D = 2 \sum_{h=1}^{H} y_{ij,h} \times \log \left(\frac{y_{ij,h}}{m(y_{ij,h})} \right), \tag{3.10}$$

where $y_{ij,h}$ represents the observed values of each comparison ij for each judge h, and $\hat{y}_{ij,h}$ are the predicted values based on the estimated model parameters. This measure indicates how well the model fits the data. If the model fits well, the $y_{ij,h}$ will be close to their predicted values $m(y_{ij,h})$.

3.4 STIMA and trunk modeling

The Bradley-Terry model can be applied to preference data by specifying a regression model for paired comparisons. In this paper, this specification is aimed at estimating in an automatic and data-driven mode the main effects part of the model as well as, if present, its interaction effects part. For this purpose, we resort to the STIMA framework extended with the use of GLM in Conversano and Dusseldorp, 2017, and combine the extended Bradley-Terry model including subject-specific covariates with the regression trunk methodology (Dusseldorp and Meulman, 2004). The main feature of a regression trunk is that it allows the user to evaluate in a unique model and simultaneously the importance of both main and interaction effects obtained by first growing a regression trunk and then by pruning it back to avoid overfitting. The interaction effects are hereby intended as a particular kind of non-additivity, which occurs if the individual effects of two or more variables do not combine additively (Berrington de González and Cox, 2007) or when over and above any additive combination of their separate effects, these variables have a joint effect (Cohen et al., 2013, p. 257).

The implementation of STIMA is based on the integration between generalized

linear models - GLM (McCullagh and Nelder, 1989) and Classification And Regression Trees (CART) (Breiman et al., 1984). A binary splitting algorithm with an ad-hoc defined splitting criterion and a stopping rule is used to model interaction terms in GLM. The estimated model, including main effects and threshold interactions, is equivalent, in its form, to a standard GLM with both random and systematic components and a link function. Usually, this model is used when the analyst has no exact a priori hypotheses about the nature of the interaction effects. For example, regression trunks have been successfully applied in the framework of tourism website evaluation (Conversano et al., 2019). STIMA allows overcoming the problems related to regression models' additive nature and the lack of main effects in tree-based methods. Typically, regression models are hard to interpret when higher-order interactions are arbitrarily included. In contrast, CART-like decision trees quickly identify complex interactive structures but, when data also includes linear main effects, they "would take many fortuitous splits to recreate the structure, and the data analyst would be hard-pressed to recognize them in the estimated tree" (Hastie et al., 2009, p. 313).

Notationally, the generalized linear model estimated by STIMA assumes that a response variable y observed on n subjects has an exponential family density $\rho_y(y;\theta;\phi)$ with a natural parameter θ and a scale parameter ϕ. The response y depends on a set of P categorical and/or continuous covariates x_p $(p = 1,\ldots,P)$ and its mean $\mu = E(y|x_1,\ldots,x_P)$ is linked to the x_ps via a link function $g(\cdot)$:

$$g(\mu) = \eta = \beta_0 + \sum_{p=1}^{P} \beta_p x_{p,h} + \sum_{t=1}^{T-1} \beta_{P+t} I\{(x_{1,h},\ldots,x_{P,h}) \in t\} \tag{3.11}$$

Equation (3.11) refers to a standard GLM presenting a linear predictor η such that $\mu = g^{-1}(\eta)$ (μ is an invertible and smooth function of η). The first P parameters concern the main effects part of the model estimated in the root node of the trunk via standard GLM. In contrast, the other $T-1$ parameters define the interaction effects part of the model obtained by partitioning recursively in a binary way the n cases in order to add additional interaction terms defined by the coefficients β_{P+t} and the indicator variables $I\{(x_{1,h},\ldots,x_{P,h}) \in t\}$. Since a tree structure with T terminal nodes is derived recursively, the so-called trunk, $I\{(x_{1,h},\ldots,x_{P,h}) \in t\}$ with $(t = 1,\ldots,T-1)$ refers to the subset of

cases belonging to the terminal node t of the trunk. The interaction effect of the T-th terminal node is not considered as this node serves as a reference category for the other interaction effects. Being obtained by a sequential binary splitting of the original data, the interaction effects correspond to threshold interactions since the values/labels of the splitting predictors leading to a specific terminal node can be considered as thresholds that partition the predictor space in order to correctly identify a GLM with interaction effects that maximizes goodness of fit by controlling for overfitting.

In a generic iteration of STIMA, adding a new threshold interaction effect in the model means adding a new binary split to the trunk. This happens when the candidate split maximizes the effect size of the model. The search of the additional interaction effect is conducted by considering for each predictor x_p all possible split points for each current terminal node. An additional interaction effect is included if the effect size between the model estimated before the current split and that including the candidate interaction originating from the current split is maximized. Once the split is found, all regression coefficients in the model are re-estimated.
In the case of a continuous response, $g(\cdot)$ corresponds to the identity function, and the effect size is computed as the relative increase in variance-accounted-for. The resulting model is the standard regression trunk model (Dusseldorp et al., 2010). Whereas, if one assumes that observations are independent realizations of Binomial random variables, the link function corresponds to the Logit function, and the effect size is computed as the relative increase in the log-likelihood R^2 observed when passing from the model which does not include the candidate interaction effect to the one that includes it. The resulting model is the logistic classification trunk (Conversano and Dusseldorp, 2017). In all cases, STIMA works by first growing a full trunk, corresponding to the maximum number of splits $T-1$, and then pruning it back using V-fold cross-validation with the c standard error rule ($c \cdot SE$ rule). The constant c varies between 0 and 1, and the higher its value, the more the tree is pruned back.

3.5 The Bradley-Terry Regression Trunk (BTRT)

In the following, we introduce the Bradley-Terry Regression Trunk (BTRT) model to analyze preference data. It combines the extended log-linear Bradley-Terry model including subject-specific covariates introduced in Equations 3.8 and 3.9 with the STIMA-based trunk model specified in Equation 3.11. The resulting model is still a log-linear model aimed at modeling the pairwise comparisons of objects i and j (Equation 3.8) through a different specification of the linear components describing the consensus expressed for the objects (see for example Equation 3.9 for object i). In particular, using the regression trunk approach and considering the possible effect of subject-specific covariates x_p, the estimated consensus expressed for object i by the judge h is

$$\hat{\lambda}_{i,h} = \hat{\lambda}_i + \sum_{p=1}^{P} \hat{\beta}_{i,p} x_{p,h} + \sum_{t=1}^{T-1} \hat{\beta}_{i,P+t} I\{(x_{1,h}, \ldots, x_{P,h}) \in t\} \tag{3.12}$$

Again, the term $\sum_{p=1}^{P} \hat{\beta}_{i,p} x_{p,h}$ is the main effects part assessing the effects of covariates on the consensus for object i. The interaction effects part is estimated by $\sum_{t=1}^{T-1} \hat{\beta}_{i,P+t} I\{(x_{1,h}, \ldots, x_{P,h}) \in t\}$ and is derived from the terminal nodes of a regression trunk that searches for possible threshold interactions between the P covariates assuming they have a joint effect on the consensus expressed for object i besides their individual (main) effect. Thus, the regression trunk has T terminal nodes and for each terminal node t an additional parameter $\beta_{i,P+t}$ is estimated. It expresses the effect of the threshold interaction between the covariates $x_1, \ldots, x_P$ whose split points lead to t. The estimated intercept term $\hat{\lambda}_i$ measures the average consensus about object i in the root node of the trunk whilst the estimated intercept for the terminal node t is $\hat{\lambda}_i + \hat{\beta}_{i,P+t}$. Note that the subscript O is left out from the notation of the $\hat{\lambda}$ parameters for readability reasons.

The estimation procedure of BTRT is framed within the STIMA algorithm, but some steps are different. Once a set of paired comparisons is given, a preliminary data processing step is required to obtain the design matrix of the Bradley-Terry model. In our framework, ties are not included, but the model can be extended by incorporating undecidedness parameters. The final design matrix is composed of $n = n_o \times (n_o - 1) \times H$

rows, where H indicates the number of judges. The total number of rows is equal to the product between the number of comparing objects, that is 2, the number of paired comparisons $(n_o \times (n_o - 1)/2)$, and the number of judges, resulting in $2 \times (n_o \times (n_o - 1)/2) \times H$.

In the above-described framework, estimating a BTRT model needs three essential ingredients: a splitting criterion, a stopping rule, and a pruning procedure.

3.6 Growing the trunk

In each step of STIMA, a generalized linear model with a Poisson link is fitted to the data. To discover the main effects, it is only necessary to fit the model in the root node. The first estimated model consists of P coefficients β that describe the probability distribution of preferring a particular object to another one, given a set $(x_1, ..., x_P)$ of judges' characteristics. STIMA searches for a split among all the values for each continuous covariate. In each step of the regression trunk building procedure, splitting a parent node means finding a dichotomous variable $z^*_{ijp,t}$ that updates the indicator function $I(\cdot)$ introduced in Equation (3.12). For each terminal node t of the trunk, the number of dichotomous variables $z^*_{ijp,t}$ is equal to the number of splits leading to t. The interaction effects part of Equation (3.12) contains $T-1$ terms since one terminal node is treated as the reference group. The search of the best split of the trunk at each iteration is made by taking into account all the available terminal nodes at that step. For a particular terminal node and based on paired comparisons, for each covariate x_p, with $(p = 1, \dots P)$, we consider each unique value of x_p as a candidate split point. Specifically, a Bradley-Terry model is estimated for each of the possible pairs of candidate values $ij \in [1, n_o]; i \neq j$, by discretizing x_p and creating the associated dichotomous variable z_{ijp}.

Next, the split point associated with z^*_{ijp} maximizing the decrease in deviance is computed for the goodness-of-fit test based on the deviance of a Poisson regression model introduced in Equation (3.10). Thus, it is considered the "best" split point, and the node is split according to the specific value of the discretized variable x_p. The splitting criterion of BTRT is based on maximizing the decrease in deviance when moving from a parent

node to the two possible child nodes defined by splitting on z_{ijp}. This is equivalent to comparing the fit of two nested models, one simpler and one more complex, and could lead to a profile log-likelihood ratio test of the hypothesis that the extra parameter β_{P+t} is zero.

This split search procedure is repeated by searching for each splitting node t the best split point so that, once found, the new dichotomous variable $z^*_{ijp,t}$ is added to the model, and an additional interaction effect is included. When the split is found, all regression coefficients in the model are re-estimated.

Preliminarily, the user is required to choose between two main approaches that could be followed in BTRT:
a) *One Split Only (OSO)*, where the splitting covariates already used in the previous splits are not considered as candidate splitting variables for the current split;
b) *Multiple Splitting (MS)*, where the whole set of covariates is considered to split the current node despite some of them having been previously selected to split other nodes. The OSO approach returns a tree in which it is possible to analyze the interaction effects between all the covariates. Following the splits along the tree, we can observe the covariates that interact (two sequent split represent an interaction). In addition, the model output presents the beta coefficients[1] associated with the terminal nodes generated by the splits of the tree and, therefore, by the interaction between the selected covariates. In this case, the final tree might not necessarily return the best model to produce the best goodness of fit (i.e., the maximum reduction in deviance). Besides, following the MS approach, it is possible to achieve the maximum reduction in deviance, but there is a risk of obtaining a tree that utilizes the same covariate (with different values) to split several, even subsequent, nodes. In this case, only the main effects part may be retained, and thus it is not possible to analyze interactions. We compare the two criteria in the actual data application.

At each split step, the estimated regression parameters $\hat{\beta}_{i,P+t}$ measure the probability of preferring a specific object i, given the interaction between different characteristics of a particular group of judges. While some similar methods, such as M5 (Quinlan,

[1] see Section 4.1

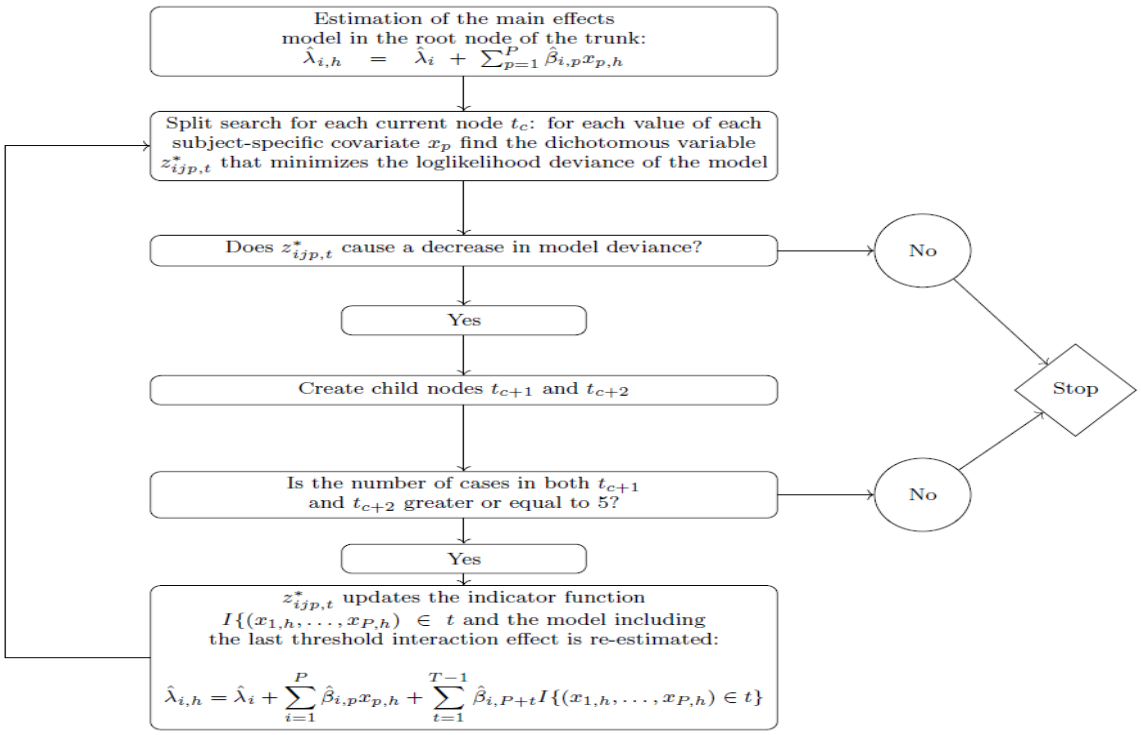


Figure 3.1: STIMA algorithm applied to ordinal data: the Bradley-Terry Regression trunk algorithm before pruning.

1992) and Treed regression (Alexander and Grimshaw, 1996), estimate several linear models, one in each node of the tree, the regression trunk model estimates a single linear model only.

Consistent with standard criteria applied in decision tree modeling, the stopping criterion of BTRT is based on the a-priori definition of the minimum number of observations for a node to be split. The default implementation is based on the requirement that the size of the new nodes should be at least equal to five, even if the minimum bucket size can be modified based on the depth of the tree requested by the user. Figure 3.1 shows a flowchart in which the tree growing procedure is schematically explained.

The final BTRT model estimates the number of parameters equal to the number of intercepts, plus the number of main effects parameters, plus the number of interactions. The summary of the number of parameters can be expressed as follows

$$(n_o - 1) + [P \times (n_o - 1)] + [(T - 1) \times (n_o - 1)]. \tag{3.13}$$

3.7 Pruning the trunk

When the final estimated trunk model presents many higher-order interactions, it may be challenging to interpret the results, and the overfitting problem might occur. However, growing the full expanded trunk is necessary since a small trunk may not capture the natural interactive structure of the data if the splitting process ends too early. For this reason, BTRT considers a pruning procedure operated after the trunk grows. In particular, a V-fold cross-validation of the BTRT model deviance is computed for each step split of the trunk. The user has to provide the number of subsets V in which the entire data set is divided. To obtain the cross-validated deviance, all the preferences expressed by a particular judge h in the design matrix are randomly assigned to a specific subset and, for V times, the BTRT trunk model estimated in a specific node is trained on $V-1$ subsets while the left-out subset is treated as a test set. At the end of the process, a predicted value $\hat{y}_{ij,h}$ is obtained for each observation in the data matrix. Following this approach, the case-wise cross-validation deviance D^{cv} is

$$D^{cv} = \frac{1}{n}\left[2\sum_{i'=1}^{n} y_{i'j;h} \times \log\left(\frac{y_{i'j;h}}{\hat{y}_{i'j;h}}\right)\right], \qquad (i', j) \in n_o, (i' \neq j), h \in H \tag{3.14}$$

where n equals the total number of rows of the design matrix and i' is its generic row. Note that the number of rows n is greater than the total number of judges H. The standard error of D^{cv} is

$$SE^{cv} = \sqrt{\frac{1}{n}\sum_{i'=1}^{n}\left[y_{i'j;h} \times \log\left(\frac{y_{i'j;h}}{\hat{y}_{i'j;h}}\right) - D^{cv}\right]^2} \tag{3.15}$$

Usually, D^{cv} decreases after the first splits of the trunk and starts to increase next. BTRT uses the same $c \cdot SE$ pruning rule used in STIMA (Dusseldorp et al., 2010). Let $t^* \in [1, T]$ be the size of the regression trunk with the lowest D^{cv}, say $D^{cv}_{t^*}$. The best size of the BTRT trunk t^{**} corresponds to the minimum value of t such that $D^{cv}_{t^{**}} \leq D^{cv}_{t^*} + c \cdot SE^{cv}_{t^*}$. We investigate about the optimal choice of the pruning parameter c in Section 3.8.

3.8 Simulation study: the choice of the pruning parameter

Pruning the BTRT model with the $c \cdot \text{SE}$ rule requires choosing the most suitable value for the parameter c. The optimal value may depend on the characteristics of the data, such as sample size. In this section, a simulation study is carried out to assess the value of the optimal c to select the final BTRT model. For the regression trunk approach used to detect threshold interactions in the linear model, Dusseldorp et al., 2010 reported that most of the time, a value of $c = 0$ results in a regression trunk with too many interaction terms while a value of $c = 1$ gives a small-sized regression trunk with too few interaction terms.

As for BTRT, we compare the performance of seven pruning rules obtained by specifying seven different values of c ranging from 0 to 1, namely: 0.00, 0.10. 0.30, 0.50, 0.70, 0.90 and 1.00. Three different scenarios are considered for the data generating process (DGP):

$$\lambda_{i,h} = \lambda_i + \beta_{i,1} x_{1,h}; \tag{3.16}$$

$$\lambda_{i,h} = \lambda_i + \sum_{p=1}^{4} \beta_{i,p} x_{p,h}; \tag{3.17}$$

$$\lambda_{i,h} = \lambda_i + \sum_{p=1}^{4} \beta_{i,p} x_{p,h} + \beta_{i,5} I(x_{1,h} > 0.00 \cap x_{2,h} > 0.50). \tag{3.18}$$

In the first scenario (Equation 3.16), only one subject-specific covariate (x_1) affects the preferences expressed by the generic judge h on each object i. In the second one (Equation 3.17), four subject-specific covariates are assumed to influence the judges' preferences. These two models present linear main effects only so that the performance metric of the pruning rules is the proportion of times a BTRT model with at least one interaction term is selected (Type I Error). In the third scenario (Equation 3.18), a model including both linear main effects and threshold interaction effects is considered as a threshold interaction term between x_1 and x_2 is added to the main effects part

of the model. In this case, the performance metric of the pruning rule is the Type II Error, obtained by computing the proportion of times the selected regression trunk model omits x_1 and x_2 exactly as the first and only two interacting variables. In all cases, all the covariates x_p are standard normally distributed.

3.9 Design factors and procedure

Three design factors are considered in the simulation study:

- The number of judges H: 100, 200, 300;
- The number of objects n_o: 4, 5. The consensus rankings were set as (A B C D) and (A B C D E), respectively, by using decreasing values of λ_i, namely $(0.9, 0.4, 0.3, 0.0)$ in the first case, and $(0.8, 0.4, 0.2, 0.1, 0.0)$ in the second one;
- The effect size of each covariate x_p on the preferences expressed by the judge h on each object i. Values of the parameters β_i are reported in Table 3.3 for each set of objects, the two possible effect sizes and the three different scenarios.

We only considered the case of 4 and 5 objects as design factors because working on paired comparisons means extending the number of judges' evaluations to 6 and 10, respectively. It seems more realistic if only a few objects are presented to judges when paired comparisons. Furthermore, as the number of objects increases, the size of the design matrix increases, as does the computational cost of searching for the split. However, the computational cost does not increase in the same way when the number of judges increases. For this reason, the BTRT model provides results in good times when the number of judges is high, but the times expand when the number of objects increases. The combination of the three design factors ($n_o \times H \times$ *effect size*) results in 12 different BTRT specifications. For each of them, we generate 100 random samples, so that 1,200 data sets were generated for each true scenario, given in Equations (3.16), (3.17), and (3.18). In each run, a BTRT with a maximum of five terminal nodes ($T = 5$) is estimated.

Once the design factors are set, following Equation 3.1 the values of $\hat{\lambda}_{i,h}$ are estimated in order to obtain the probability that a judge h prefers the object i to j. The latter is computed for each possible comparison as follows

$$\pi_{(ij)i,h} = \frac{\exp\left[2(\hat{\lambda}_{i,h} - \hat{\lambda}_{j,h})\right]}{1 + \exp\left[2(\hat{\lambda}_{i,h} - \hat{\lambda}_{j,h})\right]}; \tag{3.19}$$

The design matrix of the log-linear Bradley Terry model requires the values of y in the first column. The response y is coded as a 0-1 variable depending on whether or not an individual preference occurs for each comparison ij. Thus, we consider $y_{ij,h}$ as the realization of a Bernoulli distribution that assumes the value 1 with probability $\pi_{(ij)i,h}$. The main problem for this kind of coding is that it is possible to obtain combinations of 0-1 values for the same judge that do not verify the transitivity property between the preferences. The number of all possible combinations of two values for each judge is equal to $2^{\frac{n_o(n_o-1)}{2}}$, where the exponent is the number of paired comparisons obtainable from n_o objects. However, when ties are not allowed, the number of permutations of n_o objects equals $n_o!$, which is much smaller than the number of all the possible combinations of two values. When n_o is higher than 3, it is very likely to obtain combinations that do not find a counterpart in the universe of allowed rankings. For instance, when the number of objects is equal to four, there could be 64 combinations of 0-1, of which only 24 are allowed. So, there could be 40 combinations not allowed. We replaced the combinations not allowed with the closest permutation in the universe of $n_o!$ rankings to avoid this problem.

3.10 Results

Results of the simulation study are summarized in Tables 3.4, 3.5 and 3.6. For the first two scenarios, the pruning rules are evaluated for the Type I error (Tables 3.4, 3.5) while for the third scenario, the focus is on the Type II error (Table 3.6). To facilitate the interpretation of the results, the tables for Type II errors show the power of the pruning rules (i.e., 1 - error) rather than the Type II errors. Results are reported for the 9 different

values of the c parameter (0, 0.1, 0.3, 0.5, 0.7, 0.9, 1), as well as for the number of objects (4 or 5), the number of judges (100, 200 or 300) and the effect sizes (Low or High). As conventionally done, a threshold value of 0.05 is used for Type I error so that we are accepting that there is a five percent probability of identifying an interaction effect when there is not one. Hence, higher values are shown in boldface because the error is too high. For power, we used the value 0.8 as a threshold so that if the power is less than 0.8, then the power is too small, and the values are shown in boldface.

Table 3.4 reports the results for the first scenario where only the main effects of the single covariate x_1 are considered. When the number of objects is equal to 4 and the effect of x_1 is low, the pruning rules with $c \geq 0.1$ result in acceptable Type I errors despite the sample size. However, when the effect size increases, the case with $H = 100$ requires higher values of c (i.e., $c \geq 0.3$) for the pruning parameter. When the number of objects is equal to 5 the inverse situation is observed: for small effect sizes higher values of c (i.e., $c \geq 0.5$) are required, whilst for a high effect sizes lower values of c (i.e., $c \geq 0.3$) can be used.

Table 3.5 displays the Type I errors when all the covariates $x_1, ..., x_4$ influence judges' preferences individually (second scenario). In this case, for $n_o = 4$ the values of $c \geq 0.3$ provide acceptable error rates despite the effect size. compared to the situation in which the effect size is high; for $n_o = 5$ and high effect size it would be better to choose a pruning parameter $c \geq 0.5$.

The third scenario reflects the case in which all the covariates $x_1, ..., x_4$ influence the expressed preferences, and the first two covariates interact with each other, as shown in Equation 3.18. The power (1 - Type II error) is displayed in Table 3.6 for each possible value of c. It emerges that for $n_o = 4$ a value of $c \geq 0.3$ is considered as satisfactory despite the effect size (except in case there are 100 judges and low effect size), while for the $n_o = 5$ case with high effect size, it is preferable to increase the value of c up to 0.9.

Recall that low parameter c may return a large tree. The true model does not include interaction between variables in the first two scenarios, so low c parameter values return a too high Type I error. In the third scenario, the true model refers to a minimum

size tree with a single interaction. For this reason, as the effect size of the covariates and the population size increase, higher values of parameter c are required to obtain high power. It follows that the ability of the BTRT model to find the proper interactions between covariates increases when the number of judges and objects increases. In addition, if the judges' characteristics have a high impact on the choices, then the quality of performance of the BTRT model improves considerably.

Summarizing, the simulation study results show that a value of the pruning parameter c between 0.5 and 1 is a good choice in almost all situations. These results are consistent with those reported in Dusseldorp et al., 2010, for the linear regression model and in Conversano and Dusseldorp, 2017, for the logistic regression model.

Table 3.3: Simulated values of β_i for the estimation of the pruning parameter c

N. objects = 4									
Effect-size		Low				High			
object	A	B	C	D		A	B	C	D
				1st scenario (Equation 3.16)					
β_1	0.30	0.20	0.10	0.00		0.90	0.80	0.70	0.00
				2nd scenario (Equation 3.17): add β_2, β_3 and β_4					
β_2	0.20	0.30	0.10	0.00		0.80	0.70	0.90	0.00
β_3	0.10	0.20	0.30	0.00		0.70	0.90	0.80	0.00
β_4	0.30	0.10	0.20	0.00		0.90	0.70	0.80	0.00
				3rd scenario (Equation 3.18): add β_5					
β_5	0.25	0.15	0.35	0.00		0.55	0.65	0.45	0.0

N. objects = 5										
Effect-size		Low					High			
object	A	B	C	D	E	A	B	C	D	E
					1st scenario (Equation 3.16)					
β_1	0.40	0.30	0.20	0.10	0.00	0.90	0.80	0.70	0.60	0.00
					2nd scenario (Equation 3.17): add β_2, β_3 and β_4					
β_2	0.30	0.20	0.10	0.40	0.00	0.80	0.90	0.60	0.70	0.00
β_3	0.20	0.10	0.30	0.40	0.00	0.70	0.60	0.80	0.90	0.00
β_4	0.10	0.20	0.40	0.30	0.00	0.90	0.70	0.60	0.80	0.00
					3rd scenario (Equation 3.18): add β_5					
β_5	0.25	0.15	0.35	0.45	0.00	0.55	0.65	0.45	0.60	0.00

Table 3.4: Results first scenario: Type I error. Error higher than 0.05 in boldface.

N. objects	$n_o = 4$						$n_o = 5$					
Effect size		Low			High			Low			High	
N. judges	100	200	300	100	200	300	100	200	300	100	200	300
c = 0.0	**0.76**	**0.82**	**0.82**	**0.95**	**1.00**	**1.00**	**0.80**	**0.90**	**0.98**	**0.75**	**0.84**	**0.82**
c = 0.1	**0.16**	**0.18**	0.04	**0.62**	**0.51**	**0.58**	**0.60**	**0.58**	**0.60**	**0.30**	**0.38**	**0.26**
c = 0.3	0.01	0.00	0.00	**0.26**	**0.12**	**0.08**	**0.32**	**0.18**	**0.28**	**0.08**	**0.08**	0.00
c = 0.5	0.00	0.00	0.00	**0.08**	0.05	0.02	**0.12**	0.04	**0.10**	0.00	0.02	0.00
c = 0.7	0.00	0.00	0.00	0.03	0.00	0.00	0.04	0.02	0.00	0.00	0.00	0.00
c = 0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00
c = 1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00

Table 3.5: Results second scenario: Type I error. Error higher than 0.05 in boldface.

N. objects	$n_o = 4$						$n_o = 5$					
Effect size		Low			High			Low			High	
N. judges	100	200	300	100	200	300	100	200	300	100	200	300
c = 0.0	**0.88**	**0.86**	**0.98**	**0.95**	**0.94**	**0.98**	**0.97**	**1.00**	**0.98**	**0.91**	**0.96**	**1.00**
c = 0.1	**0.58**	**0.56**	**0.66**	**0.67**	**0.66**	**0.74**	**0.74**	**0.86**	**0.86**	**0.62**	**0.70**	**0.80**
c = 0.3	**0.14**	**0.06**	**0.10**	**0.11**	0.04	**0.10**	**0.09**	**0.14**	**0.12**	**0.16**	**0.28**	**0.18**
c = 0.5	0.04	0.02	0.00	0.01	0.00	0.00	0.01	0.02	0.04	**0.06**	**0.06**	0.02
c = 0.7	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
c = 0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
c = 1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3.6: Results third scenario: Test's power (1-Type II error). Power lower than 0.80 in boldface.

N. objects	$n_o = 4$						$n_o = 5$					
Effect size		Low			High			Low			High	
N. judges	100	200	300	100	200	300	100	200	300	100	200	300
c = 0.0	**0.00**	**0.00**	**0.00**	**0.03**	**0.02**	**0.01**	**0.02**	**0.00**	**0.01**	**0.00**	**0.00**	**0.02**
c = 0.1	**0.45**	**0.52**	**0.28**	**0.30**	**0.20**	0.80	**0.22**	**0.06**	**0.01**	**0.28**	**0.12**	**0.02**
c = 0.3	**0.79**	0.94	0.84	0.84	0.84	0.99	0.82	**0.52**	**0.46**	**0.74**	**0.28**	**0.14**
c = 0.5	0.99	0.99	0.99	0.92	0.94	0.98	0.96	0.96	0.88	0.98	**0.44**	**0.24**
c = 0.7	1.00	1.00	1.00	0.96	0.98	1.00	1.00	1.00	1.00	0.98	0.80	**0.56**
c = 0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90
c = 1.0	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	0.96

Chapter 4

The Bradley-Terry Regression trunk on preference data

4.1 Application on a real dataset

This section shows a practical application of the regression trunk for preference rankings on a real data set following two different approaches. The STIMA algorithm based on the BTRT model has been implemented in the *R* environment (R Core Team, 2021) by using the packages *prefmod* (Hatzinger and Dittrich, 2012) and *BradleyTerry2* (Turner and Firth, 2012).

The analyzed data have been collected through a survey carried out at the University of Cagliari (Italy). In particular, 100 students ($H = 100$) enrolled in the first year of Master Degree in Business Economics were asked to order five characteristics of an ideal professor ($n_o = 5$) based on what they considered the most relevant: clarity of exposition (o_1), availability of teaching material before the lectures (o_2), scheduling of midterm tests (o_3), availability of slides and teaching material accompanying the selected books (o_4), helpfulness of the professor (o_5). These characteristics were ranked with values from 1 to 5, where 1 was assigned to the characteristic considered as the most important, and 5 to the least important one. Students were not allowed to indicate ties. Moreover, for

each student, seven subject-specific covariates have been collected: year of study (x_1), total number of ECTS obtained (x_2), grade point average (x_3), course attendance in percentage (x_4), daily study hours (x_5), gender (x_6), and age (x_7). Table 4.1 reports the key statistics for each subject-specific covariate.

Table 4.1: Descriptive statistics of the subject-specific covariates in application.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Year of study	x_1	100	1.18	0.39	1.00	1.10	0.00	1.00	2.00	1.00	1.64	0.70	0.04
ECTS	x_2	100	37.69	40.22	27.00	28.89	5.93	0.00	163.00	163.00	1.90	2.23	4.02
Grade point average	x_3	100	23.02	6.93	24.80	24.49	3.26	0.00	30.00	30.00	-2.36	5.17	0.69
Course attendance	x_4	100	87.37	13.34	90.00	89.53	13.34	40.00	100.00	60.00	-1.22	0.93	1.33
Daily study hours	x_5	100	3.73	1.62	4.00	3.64	1.48	0.25	8.00	7.75	0.48	0.05	0.16
Gender	x_6	100	1.44	0.50	1.00	1.42	0.00	1.00	2.00	1.00	0.24	-1.96	0.05
Age	x_7	100	21.00	3.25	20.00	20.27	1.48	19.00	41.00	22.00	3.16	13.59	0.33

The rankings were converted into ten paired comparisons to apply the Bradley-Terry model. Dealing with a few judges and several covariates, each judge will likely have at least one characteristic that differs from the other judges. In this framework, for each pair of comparing objects, the response variable y is binary and takes 0 and 1. Therefore, 20 observations are obtained for each judge so that the total number of rows n is equal to 2,000.

Once the design matrix is obtained, a Poisson regression model is estimated in the root node. Next, the split search as described in Section 3.6 is performed. In the following, we compare the results obtained for the two splitting options currently implemented for BTRT: the OSO approach and the MS approach.

4.2 One-Split-Only (OSO) approach

The full tree can have a maximum number of splits equal to the number of subject-specific covariates P based on the OSO approach. Thus, the maximum depth regression trunk has seven splits. In this application, the trunk before the pruning is composed of 6 splits and 7 terminal nodes because no more splits respected the minimum bucket condition (i.e., number of judges greater or equal to five).

Table 4.2 reports the node splitting information and the deviance D of the final model estimated in each node (see Equation 3.10). Notice that the deviance of the main effects model is reported in the first row of Table 4.2 while the deviance of the model, including a simple dichotomous variable inducing the first split of the trunk (*bestsplit1*) is reported in the second row. The threshold interactions are specified starting from the third row of the table, i.e., from *bestsplit2* onwards.

Table 4.2: Pruned regression trunk: OSO approach. The table shows the node in which the split is found, the splitting covariate, and its split point together with the deviance associated with each estimated model.

	Node n.	*Splitting covariate*	*Split Point*	*Model Deviance*
	1	main effects (no splits)		1115
bestsplit1	root	x_3 (grade point average)	27.50	1096
bestsplit2	2	x_7 (age)	25.00	1080
bestsplit3	4	x_2 (n. of ECTS)	39.00	1064

The maximum-depth regression trunk is pruned applying the $c \cdot SE$ rule described in Section 3.7 based on both the case-wise 10-fold cross-validation deviance (D^{cv}) introduced in Equation 3.14 and its standard error (SE^{cv}, Equation 3.15). Table 4.3 shows the results of the cross-validation estimates.

Table 4.3: 10-fold cross-validation results with OSO approach: D = model deviance (Eq. 3.10); D^{cv} = casewise cross-validation deviance (Eq. 3.14); SE^{cv} = standard error of D^{cv} (Eq. 3.15).

	D	D^{cv}	SE^{cv}
mod0	1115	0.5957	0.0003
mod1	1096	0.5910	0.0004
mod2	1080	0.5870	0.0005
mod3	1064	0.5858	0.0005
mod4	1058	0.5874	0.0005
mod5	1048	0.5890	0.0005
mod6	1033	0.5894	0.0005

Note that D^{cv} is much smaller than the model deviance D, because we used two

different specifications for these two (see Equation 3.10 and 3.14): D decreases between one model and another, while D^{cv} is decreasing up to the model 3 having four terminal nodes. Applying the pruning rule with the c parameter is unnecessary in this case cause the cross-validation deviance starts to increase from the fourth model ($mod4$). Thus, the pruned trunk corresponds to the model in Table 4.2. The final trunk including three splits and $T = 4$ terminal nodes is shown in Figure 4.1 .

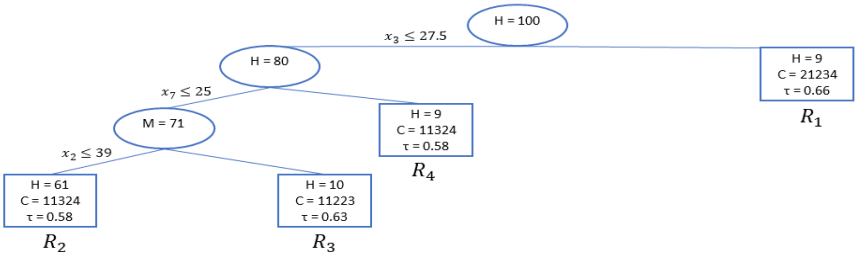


Figure 4.1: Pruned regression trunk: OSO approach on students' survey. The number of judges H is shown for each node. The splitting covariate x_p is presented for each split. The consensus ranking C and the extended correlation coefficient tau_x are shown in each terminal node. The objects are the following: clarity of exposition, availability of teaching material before the lectures, scheduling of midterm tests, availability of slides and teaching material beside the selected book, and the helpfulness of the professor.

Figure 4.1 shows the pruned regression trunk. It reports the number of judges H belonging to each terminal node T. The consensus ranking C is computed by using the differential evolution algorithm for median ranking detection (D'Ambrosio et al., 2017) and the extended correlation coefficient τ_x (Emond and Mason, 2002) within the group. Both measures are computed using the R package *ConsRank* (D'Ambrosio et al., 2019). The consensus ranking reports the positions of the objects ordered from o_1 to o_5. Ties are allowed only for the consensus ranking within the groups so that two tied objects have the same associated value.

4.3 Multiple Splitting (MS) approach

The MS approach allows covariates already used in previous splits for the split search. To compare the MS approach with the OSO one, a regression trunk with the same number of terminal nodes of the OSO trunk is grown for the MS case ($T = 7$). The results associated with the pruned tree are reported in Table 4.4.

Table 4.4: Pruned regression trunk: MS approach. The table shows the node in which the split is found, the splitting covariate, and its split point together with the deviance associated with each estimated model.

	Node	*Covariate*	*Point*	*Deviance*
	1	main effects (no splits)		1115
bestsplit1	root	x_3 (grade point average)	27.50	1096
bestsplit2	2	x_7 (age)	25.00	1080
bestsplit3	4	x_2 (n. of ECTS)	39.00	1064
bestsplit4	8	x_3 (grade point average)	21.00	1050

The pruning procedure is performed using the ten-fold cross-validation estimation of the deviance and its standard error. Table 4.5 shows the results associated with the pruned trunk deriving from the MS approach.

Table 4.5: 10-fold cross-validation results with MS approach: D = model deviance (Eq. 3.10); D^{cv} = casewise cross-validation deviance (Eq. 3.14); SE^{cv} = standard error of D^{cv} (Eq. 3.15).

	D	D^{cv}	SE^{cv}
mod0	1115	0.5957	0.0003
mod1	1096	0.5910	0.0004
mod2	1080	0.5870	0.0005
mod3	1064	0.5858	0.0005
mod4	1050	0.5809	0.0005
mod5	1038	0.5810	0.0005
mod6	1026	0.5809	0.0006
mod7	1018	0.5814	0.0006

The MS approach, for each split, generates a reduction in deviance greater than that obtained with the OSO approach. The cross-validation deviance is decreasing up to model 4. Figure 4.2 compares the two approaches in terms of cross-validation deviance obtained from one split to another. It displays that the MS approach returns a regression trunk capable of better explaining the preferences expressed by the judges.

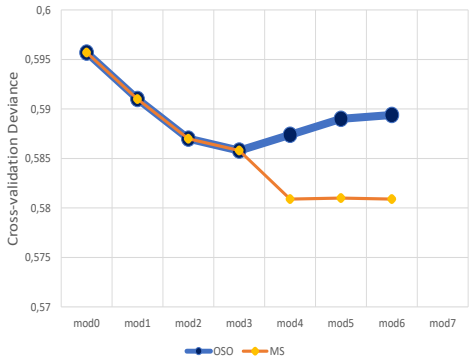


Figure 4.2: Comparison between OSO and MS approaches

Using the information obtained from the simulation study presented in Section 3.8, with $n_o = 5$ and $H = 100$ a possible pruning parameter is $c = 0.5$ so that the the final trunk is that corresponding to model 4 (*mod4*) in Table 4.5 with four splits and five terminal nodes. Figure 4.3 shows the pruned regression trunk.

Note that the professor's quality of exposition (o_1) is always preferred to all the other objects in the pruned tree, except by the judges in Region 1. As expected, the two approaches provide different results: the OSO approach detects the interaction between all the covariates under study (see Figure 4.1) but does not return the best regression trunk in terms of goodness-of-fit. The MS approach returns a trunk that fits the data better, but the final BTRT model may be more challenging to interpret.

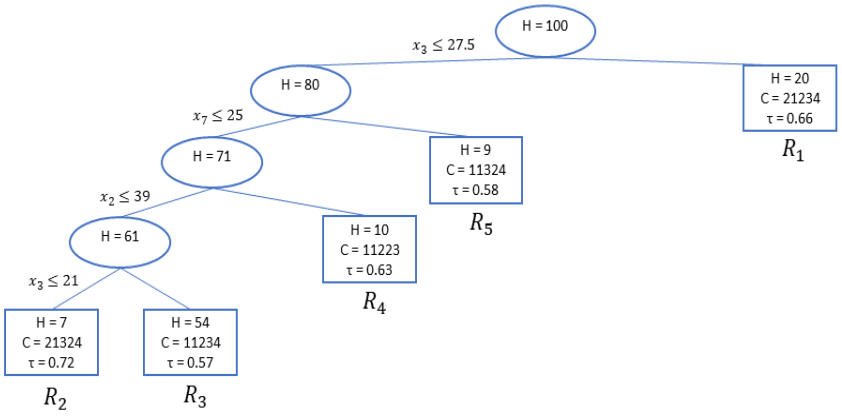


Figure 4.3: Pruned regression trunk: MS approach on students' survey. The number of judges H is shown for each node. The splitting covariate x_p is presented for each split. The consensus ranking C and the extended correlation coefficient tau_x are shown in each terminal node. The objects are the following: clarity of exposition, availability of teaching material before the lectures, scheduling of midterm tests, availability of slides and teaching material beside the selected book, and the helpfulness of the professor.

The model deriving from the MS regression trunk returns the coefficients shown in Table 4.6.

Table 4.6: MS regression trunk final output: the Table shows the estimated coefficients associated to the objects o_1, o_2, o_3, and o_4. The last object o_5 is set as reference level, so that the estimated parameters associated to $\hat{\lambda}_{o_5,h}$ (the professor helpfulness) are automatically set to zero. The standard errors are shown in parenthesis and the stars $'*'$ associated to some estimate coefficients indicate that they are significantly different from zero with a pvalue lower than 0.001 ($'***'$), 0.01 ($'**'$) and 0.05 ($'*'$), respectively.

	$\hat{\lambda}_{o_1,h}$	$\hat{\lambda}_{o_2,h}$	$\hat{\lambda}_{o_3,h}$	$\hat{\lambda}_{o_4,h}$
$\hat{\lambda}_i$	3.36 (1.98)	4.96** (1.68)	3.46* (1.59)	-2.41 (1.72)
$\hat{\beta}_{i,x1}$	-0.90* (0.42)	-0.43 (0.40)	-0.03 (0.40)	-0.56 (0.42)
$\hat{\beta}_{i,x2}$	0.02*** (0.005)	0.009 (0.004)	0.003 (0.004)	0.009 (0.004)
$\hat{\beta}_{i,x3}$	-0.16*** (0.04)	-0.14*** (0.04)	-0.09* (0.03)	-0.01 (0.04)
$\hat{\beta}_{i,x4}$	-0.008* (0.006)	-0.01* (0.006)	-0.01** (0.006)	-0.007 (0.006)
$\hat{\beta}_{i,x5}$	-0.04 (0.06)	-0.07 (0.05)	-0.12* (0.05)	-0.06 (0.05)
$\hat{\beta}_{i,x6}$	0.31 (0.18)	0.29 (0.15)	0.29 (0.15)	0.36* (0.15)
$\hat{\beta}_{i,x7}$	0.17** (0.06)	0.03 (0.04)	0.03 (0.04)	0.15** (0.04)
$\hat{\beta}_{i,R2}$	-2.30*** (0.62)	-1.96*** (0.56)	-1.47** (0.55)	-0.47 (0.59)
$\hat{\beta}_{i,R3}$	-2.86*** (0.58)	-1.37** (0.47)	-0.73 (0.45)	-0.32 (0.46)
$\hat{\beta}_{i,R4}$	-3.56*** (0.67)	-1.47** (0.53)	-1.14* (0.52)	-1.32* (0.54)

The regions $R_2, \ldots, R_5$ obtained from the regression trunk represented in Figure 4.3 are defined as follows:

$$R_2 = I(\text{grade point average} \leq 21, \text{age} \leq 25, \text{n. of ECTS} \leq 39),$$
$$R_3 = I(21 < \text{grade point average} \leq 27.5, \text{age} \leq 25),$$
$$R_4 = I(\text{grade point average} \leq 27.5, \text{age} \leq 25, \text{n. of ECTS} > 39),$$
$$R_5 = I(\text{grade point average} \leq 27.5, \text{age} > 25),$$

The region R_1 plays the role of reference category. It is defined by the indicator

function $I(\text{grade point average} > 27.5)$. From the side of the main effects, looking at the values in Table 4.6 the final model shows that the covariates x_3 (grade point average) and x_4 (course attendance in percentage) have a significant and negative effect on the preferences expressed. In particular, looking at the $\hat{\beta}_{i,x_3}$ coefficients, it can be seen that as the grade point average increases, the tendency to prefer the professor's clarity (o_1) to his helpfulness (o_5) is lower. On the contrary, when the number of ECTS increases, the tendency to prefer the professor's clarity to the professor's helpfulness is higher. These two results might suggest that students looking for a high average grade consider interacting with the professors even outside of class hours. On the other hand, students who have a high number of ECTS may not be interested in a high average grade, but only in obtaining a degree quickly, so they believe it is more important than the teachers are clear during the lessons.

As for the interaction effects, Table 4.6 shows that the last region R_4 has significant and negative coefficients whatever the considered object. In each case, when the students' grade point average is lower than 27.5 and the age is higher than 25, there is a strong tendency to prefer the professor's helpfulness to all other attributes.

4.4 Discussion

This chapter introduced a new Bradley-Terry Regression Trunk (BTRT) model to analyze preference data. BTRT is based on a probabilistic approach in which the judges' heterogeneity is taken into account with the introduction of subject-specific covariates. Combining the log-linear Bradley-Terry model with the regression trunk methodology allows generating, through Poisson regressions, an easy-to-read partition of judges based on their characteristics and the preferences they have expressed. The effects of the judges' characteristics and their interactions on the object choice are estimated simultaneously. BTRT accounts for the drawback of the classic tree-based models when no a priori hypotheses on the interaction effects are available. At the same time, it allows detecting threshold interactions in an automatic and data-driven model. The final result is a small and easily interpretable tree structure, called regression trunk, that only considers the

interactions that significantly improve the main effects model fit.

Simulations showed that the ability of the BTRT model to find the right interactions increases when both the sample size and the number of objects to be judged increase, particularly if the covariates have a high impact on the choices. The results suggest that a value of the pruning parameter c between 0.7 and 0.9 is a good choice in most cases. These values are consistent with those reported in Dusseldorp et al., 2010, for the linear regression model and in Conversano and Dusseldorp, 2017, for the logistic regression model.

The two different approaches introduced for the BTRT model have been used in a real dataset application. It emerges that the One-Split-Only approach aims to verify the interaction effect between all the covariates taken into consideration, and the final result is easier to interpret. On the other hand, the Multiple Splitting approach yields a tree that can capture the most significant interactions between the variables selected by the model. The BTRT model appears well-suited to analyze the probability distribution of preferring a particular object for a specific group of individuals with a specific set of characteristics. For this reason, it can be used for both descriptive and predictive purposes as it allows the user to estimate the impact of each subject-specific covariate on the judges' choices, the overall consensus ranking, and the effect size of the interactions between covariates.

Future research is addressed to consider cases when categorical subject-specific covariates with more than two categories are used as possible split candidates and investigate further model performance and stability concerning (big) datasets presenting a high number of objects, rankings, and covariates. This would allow us to evaluate better the two approaches illustrated in Section 4.1. In addition, an R function is currently under development to allow replications and extensions of the BTRT procedure. At the same time, research efforts will extend the model to cases where missing values (i.e., partial orderings) are allowed. As the number of objects increases, paired comparisons become more challenging to treat. A solution to this issue is furnished in Chapter 6.1, where we present an extension of the BTRT model to analyze ordinal data treated as rankings. This extension is based on the Mallows specification of the BT model.

Chapter 5

The Bradley-Terry Regression Trunk on financial data

5.1 Applitation of BTRT on financial data

Public finance can be described as the study of the government's role in the economic system (Gruber, 2005). Policymakers must respect the government budget balance, which is the overall difference between government revenues and spending. This principle became even more critical following the financial crisis that erupted in 2007 in the US and then turned into a sovereign debt crisis. The public authorities' goal is to find the right balance between government revenues and government expenditures to achieve desirable effects and avoid undesirable ones (Jain, 1989). If we singularly consider these two components, it is pretty clear how they influence the countries' Gross Domestic Product (GDP). By looking at the spending approach, the GDP is directly influenced by the government's public spending, while the taxes generally act in the opposite direction. However, the relationship between tax revenues and public spending and their interaction with economic components is unclear. Manage and Marlow, 1986, focused on analyzing the causal relationship between taxation and central government public expenditure. They showed that taxation causes expenditure at the state level of government but that such causation becomes bidirectional in the short run. However, Anderson et al., 1986, concluded that

government expenditures cause government taxes by conducting the Granger causality test in the same year. On the other hand, there is a discrete interest in the literature about tax revenue determinants. At the basis of this interest, there is a common question: where do the differences in tax revenues between countries come from? According to Kaldor, 1963, underdeveloped countries' tax revenues are much lower than developed countries. This effect can be explained by the fact that taxes can be paid from the income surplus over the population's minimum subsistence needs (Boukbech et al., 2018). Therefore, an emerging country has less room for transforming national income volume into taxes to finance collective needs without creating intolerable social tensions. It is reasonable to assume that the higher the country's level of development, the greater its capacity to raise tax resources (Brun and Diakite, 2016).

Several authors worked on the relationship between socio-economic explanatory variables and tax revenues as response variables through cross-sectional or panel empirical studies. Their goal was to find the tax revenues' determinants. The main findings suggest that the principal factors of tax pressure are represented by the Gross Domestic Product per capita (Gupta, 2007a; Pessino and Fenochietto, 2010), the productive specialization captured by the sectoral composition of the GDP (R. J. Chelliah, 1971; R. Chelliah, 1975; Tait et al., 1979; Piancastelli, 2001; Karagöz, 2013), external factors such as the level of foreign direct investment (FDI) and trade (Cassou, 1997; Gupta, 2007b; R. M. Bird et al., 2008), the level of public debt (Teera and Hudson, 2004) and policy makers' choices, such as exchange rate, inflation rules (Keynes-Oliveira-Tanzi effect) and financial-fiscal policies (Tanzi, 1989. Other works analyzed the role of government efficiency and institutional factors such as political stability and political and civil rights (Martın-Mayoral and Uribe, 2010). On the social side, some researchers study the impact of the educational level (as a share of public expenditure on education), illiteracy rate, and population growth on tax revenues (Bahl and Wallace, 2005). Accountability and civil and political rights are also considered determinants of tax revenue. Also, factors such as corruption, entry regulations, and the rule of law can play a determinant role in defining tax revenues (R. Bird et al., 2004).

These authors used different methodologies to achieve their similar goals. Some

have applied dynamic general equilibrium models (Feltenstein and Cyan, 2013), but others have recurred to econometric techniques, such as the first cross-sectional study on international tax ratios conducted by Lotz and Morss, 1967. They introduced the tax effort concept and determined that per capita income and trade share are determinants of the tax share. On the panel empirical methods side, Pessino & Fenocchietto, 2010, determined a panel version of a stochastic tax frontier model, while others focused on static fixed and random effect models and dynamic panel data techniques that use the generalized method of moments (Gupta, 2007; Martin-Mayoral and Uribe, 2010).

Between the works mentioned above, Teera & Hudson, 2004, and Pessino & Fenocchietto, 2010, applied their methodologies to large samples of countries by considering different countries' geographical locations or income levels. However, there are works based on a restricted sample of countries. Castro and Camarillo, 2014, considered only the 34 countries from the Organisation for Economic Co-operation and Development (OECD) by using lagged values of the tax revenues over 2001-2011. The basis of this choice is that the research for tax determinants may not be significant for a heterogeneous group of countries. The determinants of tax revenue can be different in low, middle, and high-income countries.

Our work matches the needs mentioned above: Finding the tax revenues components by accessing the heterogeneity of countries and focusing on the relationship between tax revenues and government expenditure. We present an application that differs from those made previously. Unlike the works cited, we decided to study the determinants of tax revenues by decomposing them into taxes on income, social security contributions, taxes on property, and taxes on goods (Organisation for Economic Co-operation and Development classification). In this way, it is possible to simultaneously consider the effect of socio-economic variables on different tax categories. In addition, the tax categories are paired compared on their size, which means that the tax revenue categories have been ordered according to their size and then compared to each other. Passing from continuous data to categorical ones (from numbers to rankings and then to paired comparisons), we can apply the Bradley-Terry model for matched pairs, using the log-linear formulation with subject-specific covariates in order to capture and to quantify the effects of each

country socio-economic feature on their tax revenues category. In addition, the transformation from continuous data to rankings seems reasonable when comparing countries with different fiscal systems. The Organisation for Economic Co-operation and Development (OECD) tax revenues classification results from a big effort to obtain a common ground to compare data from different countries. Our transformation allows us to work on sizes and not on precise continuous values of tax revenues so that the problem of comparability becomes easier to overlook Here, we used data from different databases by combining data from OECD, International Monetary Fund (IMF), and World Bank for the year 2018. The heterogeneity in the model is also taken into account by expressing variables in terms of Gross Domestic Product (GDP) and applying a particular partitioning model for the Bradley-Terry model. Specifically, the Bradley-Terry Regression Trunk (BTRT) is chosen to investigate the interactions between covariates that most affect the comparisons between taxation items. The result is a small regression tree that creates a partition of countries and provides valuable information about each terminal node's main effects, interaction effects, and estimated tax ordering.

The proposed model provides a solution to discover interaction effects when no a-priori hypotheses are available. It produces a small tree, called trunk, representing a fair compromise between a straightforward interpretation of the interaction effects and an easy-to-read partition of countries based on their socio-economic characteristics and the order of their tax revenues. This model is also justified because there are relatively few works in the classification community for paired comparisons data, especially in public finance studies. Then, we decomposed the government expenditure by following the Classification of the Functions of Government (COFOG) classification to capture the effect of each type of government expenditure on each tax revenues category.

5.2 Data

We use a cross-section dataset that covers 100 countries and their associated tax revenues by category for 2018. These initial data are taken from the Global Revenue Statistics Database (OECD, 2018), where the OECD classification of taxes is used. This ensures

consistency across countries and provides a high granularity of tax revenue categories (Constructing the global revenues statistics, 2018). According to OECD classification, taxes are classified by the base of the tax: income and profits (heading 1000), compulsory SSCs (heading 2000), payroll and workforce (heading 3000), property (heading 4000), goods and services (heading 5000), other taxes (heading 6000). All these categories are expressed in terms of the level of taxation through the tax-to-GDP ratio, calculated by the ratio of nominal tax revenue of a country h and its nominal GDP for the year 2018. The tax-to-GDP ratio is well suited for cross-country research studies aiming to compare tax levels across countries with different development degrees.

At the initial stage of our analysis, we transformed tax revenue categories from continuous data to rankings by assigning values from 1 to 6 to each category, where 1 represents the higher tax category, and 6 is the lower one. This transformation aims to obtain paired comparisons between categories that are the basis of the BTRT model. Once rankings are obtained, we calculated the consensus ranking by maximizing the extended correlation coefficient τ_x. Following this specification, the consensus ranking is the best compromise between a set of rankings. It is the solution to a maximization problem: the consensus ranking is that ranking in the permutation space maximizes the sum of correlations between itself and all the other rankings. We used the R package *ConsRank* for this calculation, and the result is the following: goods and services $>$ income and profits $>$ compulsory SSCs $>$ property $>$ workforce $=$ other taxes. The categories taxes on workforce and other taxes are ranked in the last position. In most cases, in the original continuous data, these two categories present values equal to 0, which conducted us to exclude them so that our analysis covers 100 countries and four tax revenue categories.

Once the data cleaning on tax revenue categories is operated, we focused on collecting predictors to determine the tax revenues as independent variables. We followed the suggestions from the previous literature mentioned in the introduction Section 5.1, and we considered a high number of socio-economic covariates by combining information from IMF, OECD, and World Bank databases. All covariates refer to the year 2018, which is the last year for sufficient financial data we needed, and they are almost all

expressed in terms of GDP.

In the specific, before the feature selection process, the country-specific covariates we considered in our analysis are the sequent: current account balance; employment and unemployment rate; general government gross debt; general government net lending/borrowing; gross fixed capital formation; percentage of change in the gross domestic product; gross national savings; the volume of exports and imports of goods and services; account ownership at a financial institution; subsidies and other transfers; interest payments; compensation of employees; value-added of agriculture, services, industry, and manufacturing; population density; final consumption expenditure; banking non-performing loans; claims on central government; households consumption; government consumption; trade volume; environmental performance index; government expenditure for military, education, health, and others. The latter covariate was constructed by difference, subtracting military expenses, education expenses, and healthcare expenses from the total government expenditure. This step was necessary due to the lack of data on the other public expenditure items included in the COFOG classification.

In addition, for replacing missing values, we also considered the location variable from the OECD database. It classifies the $H = 100$ countries into four categories: 36 OECD countries, 29 countries from Africa, 15 countries from Asia, and 21 countries from South America. Missing values for each covariate were replaced with the median value conditioned to the location of each specific country.

5.3 Bradley-Terry-Luce Lasso for covariates selection

When dealing with BT models, the inclusion of covariates yields models with a high number of parameters. Therefore, in applications like ours, it would be better to select the most relevant terms to reduce the complexity of the model. Generally, not all the explanatory predictors likely impact the objects' ordering in a high-dimensional dataset. For this reason, we applied a subject-specific covariates selection through the Bradley-

Terry-Luce Lasso (BTLL) proposed by Schauberger and Tutz (2019). The authors refer to this methodology as BTLL regression since the BT model is strongly connected to the axiom formulated by Luce, 1959, which states that other objects do not influence the decision between two objects. Luce's axiom is usually defined as independence from irrelevant alternatives. This methodology is based on the maximization of the penalized log-likelihood as follows

$$l_p(\xi) = l(\xi) - \Lambda J(\xi), \tag{5.1}$$

where $l(\xi)$ is the classic log-likelihood with (ξ) indicating a vector with all the parameters. The penalty term is represented by $J(\xi)$, with Λ as a tuning parameter that quantifies how seriously the penalty term must be taken. When $\Lambda = 0$, the classic ML estimate is obtained.

The penalty for subject-specific covariates yields for all the covariates x_p that share the same effect and can be formulated as follows

$$P(\beta_{i,1}, ..., \beta_{i,P}) = \sum_{p=1}^{P} \sum_{i<j} |\beta_{i,p} - \beta_{j,p}| \tag{5.2}$$

If $\Lambda \to \infty$ all the effects of x_p are merged to one single cluster, it is eliminated from the model as all effects tend to zero. On the contrary, when $\Lambda = 0$, the model gives coefficients equal to a classic BT model with subject-specific covariates. It derives that for a finite value of Λ some of the covariates are eliminated, while the remaining are still identified in the model.

The penalty terms can be internally weighted according to the principle of adaptive lasso (Zou, 2006), and finally, they can be combined as

$$J(\xi) = \sum_{l=1}^{L} \psi_l P_l, \tag{5.3}$$

where ψ_l represents penalty-specific weights. This formulation allows combining all

penalty terms into one joint penalty controlled by the tuning parameter Λ to make the optimization procedure easier. The comparability of different penalty terms requires two conditions. First, all subject-specific covariates have to be scaled to compare their effect sizes. Second, the weights ψ_l have to be assigned according to the number of penalties and free parameters they include (Bondell and Reich, 2009; Oelker et al., 2015).

We used the R package $BTLLasso$ (Schauberger, 2015) that applies L_1 penalties to the Fisher scoring. This package allows implementing a 10-fold cross-validation procedure to find the optimal level of tuning parameters Λ. By applying the $cv.BTLLasso$ R function before starting the regression trunk building procedure (i.e., in the root node), the cross-validation procedure detected $\Lambda = 1.23$ as optimal value for the size of penalties applied to our $\beta_{i,p}$ coefficients. As a result, we removed the covariates that presented an effect size equal to zero on two or more different tax revenue categories. The selected variables with a significative main effect on tax revenues are the sequent: employment rate, general government gross debt, gross national savings, the volume of exports and imports of goods and services, interest payments, value-added of agriculture and services, final consumption expenditure, bank non-performing loans, claims on central government, government consumption, EPI, and expenses on the military, education, health, and other expenses. In total, 13 out of 30 subject-specific covariates are removed from our dataset. At the end of the feature selection, our dataset is composed by $H = 100$ countries, $n_o = 4$ tax revenue categories, and 17 x_ps subject-specific covariates. It is quite interesting that all the covariates related to the public expenditure remain in our analysis, meaning that their size main effect on the composition of tax revenues can not be overlooked. Table 5.1 shows a summary of the descriptive statistics for each covariate x_p chosen after conducting the BTLL future selection.

Missing values reported in Table 5.1 were replaced with the median value of each covariate conditioned to the location membership of the country (i.e., OECD, Africa, South America, and Asia).

Table 5.1: Key descriptive statistics for subject-specific covariates after selection through BTLlasso

covariates	x_p	*missing*	*mean*	*sd*	*median*	*min*	*max*	*range*	*skew*	*kurtosis*	*se*
Employment rate	x_1	2	0.59	0.09	0.59	0.39	0.85	0.45	0.05	0.27	0.01
Gov gross debt	x_2	1	0.59	0.35	0.50	0.09	2.38	2.29	2.12	6.96	0.03
Savings	x_3	5	0.21	0.08	0.21	0.02	0.46	0.44	0.53	0.48	0.01
Exp	x_4	13	0.04	0.06	0.05	-0.36	0.17	0.54	-3.59	22.69	0.01
Imp	x_5	11	0.04	0.06	0.05	-0.23	0.18	0.41	-1.66	6.12	0.01
Interests	x_6	21	0.08	0.05	0.07	0.00	0.29	0.29	1.34	2.98	0.01
Agric added-value	x_7	2	0.09	0.09	0.05	0.00	0.45	0.45	1.69	2.79	0.01
Services added-value	x_8	21	0.57	0.10	0.57	0.29	0.79	0.50	-0.40	-0.20	0.01
Final consumption	x_9	8	0.78	0.11	0.79	0.43	1.12	0.69	-0.49	1.88	0.01
Bank NPloans	x_{10}	16	0.04	0.06	0.03	0.00	0.42	0.42	4.77	25.75	0.01
Claims centr gov	x_{11}	4	0.12	0.18	0.08	-0.14	1.42	1.56	3.85	23.62	0.02
Gov consumption	x_{12}	8	0.16	0.05	0.16	0.04	0.39	0.35	0.63	2.42	0.01
EPI	x_{13}	0	0.59	0.16	0.59	0.00	0.87	0.87	-0.93	2.21	0.02
Military spending	x_{14}	2	0.01	0.01	0.01	0.00	0.05	0.05	1.44	3.72	0.00
Education spending	x_{15}	33	0.04	0.01	0.04	0.01	0.08	0.07	0.72	0.25	0.00
Health spending	x_{16}	0	0.07	0.03	0.07	0.02	0.17	0.15	0.74	0.96	0.00
Other spending	x_{17}	9	0.20	0.10	0.20	0.03	0.76	0.73	1.77	7.89	0.01

5.4 Results

The final result of the BTRT model is a tree that represents a compromise between an easy-to-read partition of countries and an effective capture of the main effects and interaction effects that have the most significant impact on the order of magnitude of the four tax revenue categories.

The main results in terms of the node in which the split is found, best split covariate, best split point, and model deviance are shown in Table 5.2. We used the same nodes coding procedure as used in CART. These values are referred to the pruned regression trunk because the pruning procedure has already been applied to the full tree.

The algorithm tried to split the root node by considering all the values of each subject-specific covariate in Table 1. As a result, the first best split covariate is represented by the Environmental Performance Index (EPI) with a value of 0.7. This covariate

Table 5.2: Pruned regression trunk. The table shows the node in which the split is found, the splitting covariate, and its split point together with the deviance associated with each estimated model. The codification of the nodes follows the same scheme as in the CART algorithm.

	Node n.	*Splitting covariate*	*Split Point*	*Model Deviance*
	1	main effects (no splits)		313
bestsplit1	1	x_{15} (Environmental Performance Index)	0.70	271
bestsplit2	3	x_2 (Gross debt)	0.40	238
bestsplit3	2	x_4 (Health spending)	0.06	213
bestsplit4	5	x_8 (Employment rate)	0.51	190
bestsplit5	4	x_8 (Environmental Performance Index)	0.56	171

has been scaled in data cleaning steps to obtain an index between 0 and 1. It is well known that EPI is highly related to income, wellness, and human development (Lai & Chen 2020). The BTRT results confirm the expectations of EPI as a key covariate by selecting it as the first choice and best split covariate. The 100 countries in node 1 are splitted according to the threshold EPI $= 0.7$ so that 76 countries go left to node 2 (EPI ≤ 0.7) and 24 go right to node 3 (EPI > 0.7). The second-best split covariate is the GDP change in percentage (x_2), which splits node 3 according to the value 0.4. The government health spending (x_4) is the third split covariate, with a value equal to 0.06. It splits node two and generates nodes 4 and 5 with 46 and 30 countries. The employment rate (x_8) splits node 5 with a value equal to 0.51. The employment rate is unaffected by voluntary changes in labor force participation so that it can be considered a useful indicator of current labor market conditions. The unemployment rate is affected by the size of the labor force (e.g., it may fall if workers give up looking for work, and as the labor market is recovering, unemployment can rise because more people are entering the labor force). The last best split of the pruned tree is still the EPI, with a value of 0.56.

Table 5.3 shows the deviance, cross-validation deviance, with associated standard error, for each step of the BTRT building process. Note that the results referred to *model*6 (*mod*6) are reported even if the pruned tree stops at *model*5 (*mod*4). We also reported the model results for one more split to show how the pruning procedure works. In our case, by adding another split to the tree, the model deviance D and the cross-

validation deviance D^cv decrease. The D^{cv} values are much smaller than the D values because they are mean values of deviances calculated on each row of the design matrix (Equation 3.15), while D is calculated as a sum (Equation 3.14). The pruning back procedure consists of applying the $c \times SE$ rule as for the STIMA algorithm and BTRT model. Here, we use a value $c = 0.5$ as suggested by simulations results of the BTRT model in Chapter 3.8, and $model6$ is pruned ($0.3565 + 0.5 \times 0.0014 > 0.3567$). Then, the final tree corresponds to model5 ($mod5$) with five splits and $T = 6$ terminal nodes.

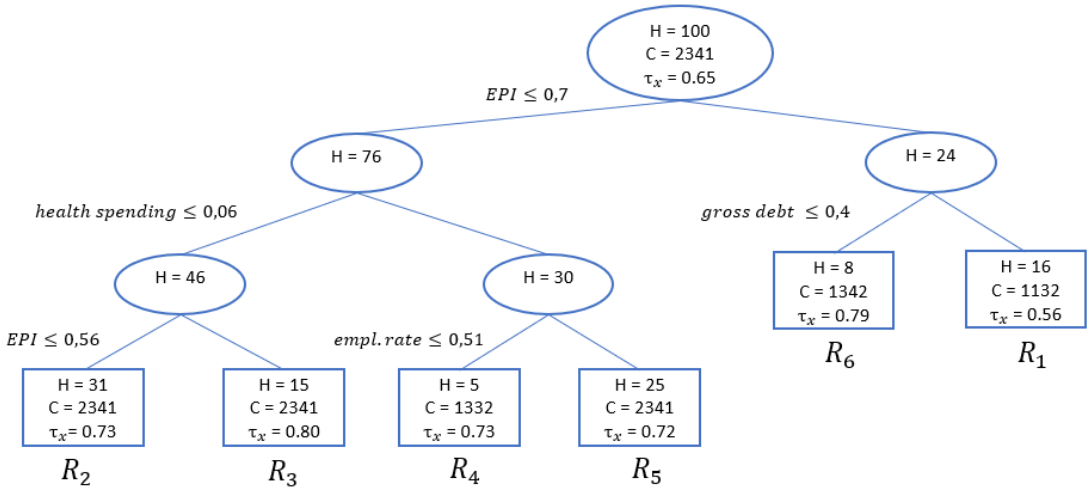
Table 5.3: 10-fold cross-validation results: D = model deviance for a Poisson distribution; D^{cv} = casewise cross-validation deviance (Eq. 3.14); SE^{cv} = standard error of D^{cv} (Eq. 3.15).

	D	D^{cv}	SE^{cv}
mod0	313.0671	0.3775	0.0010
mod1	271.9198	0.3682	0.0010
mod2	238.8955	0.3659	0.0011
mod3	213.5469	0.3617	0.0011
mod4	190.8353	0.3594	0.0013
mod5	171.2580	0.3567	0.0013
mod6	136.5742	0.3565	0.0014

Figure 5.1 shows the pruned regression trunk. In each terminal node, we report the number of countries H belonging to T, the consensus ranking C, and the associated correlation coefficient τ_x. The consensus rankings are calculated by maximizing the correlation τ_x between rankings inside the nodes, where τ_x refers to the extended correlation coefficient. These values are reported just as descriptive statistics since they do not derive from the estimated parameters of the BTRT model. The consensus ranking C and the associated correlation coefficient τ_x are calculated using the DECOR function within the *ConsRank* *R* package.

As a result, the BTRT model found the first-order interaction between the EPI and health spending and the second-order interaction between EPI, health spending, and employment rate. It seems that in countries with higher health expenditure, the structure and development of the labor market, associated with the degree of environmental performance, have a substantial effect on the composition of tax revenues categories for

Figure 5.1: Pruned regression trunk: In each terminal node $T1...T6$ and in the root note, the number of countries H, the consensus ranking C, and the correlation coefficient τ_x are shown. The order of taxes is the sequent: taxes on income, social security contributions, taxes on property, and taxes on goods. The consensus rankings C assign values to this object, where 1 corresponds to the highest and 4 to the lowest.



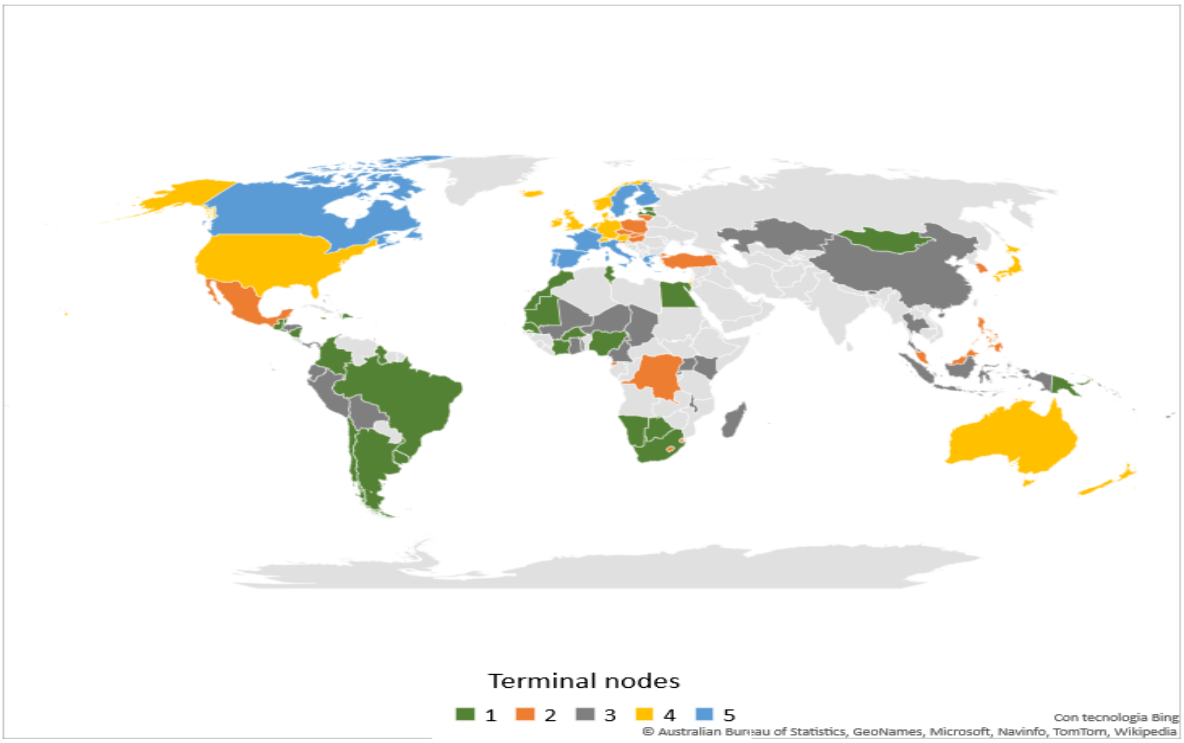
the sample of countries in our dataset.

For a visualization of the composition of the terminal nodes deriving from the application of the BTRT model on financial data, Figure 5.2 shows the world map with the final breakdown of the countries in the respective terminal nodes $T1, \ldots, T5$.

The results in terms of coefficients are summarized in Table 5.4 for taxes on income o_1, compulsory social security contributions o_2, and taxes on property o_3. For taxes on goods o_4, the estimated parameters are automatically set to 0 as this revenues source acts as the reference category.

The regions $R_2, ..., R_6$ are defined as follows

Figure 5.2: Countries' distribution map in each terminal node



$$R_2 = I(\text{EPI} \leq 0.56, \text{health spending} \leq 0.06),$$

$$R_3 = I(0.56 < \text{EPI} \leq 0.7, \text{health spending} \leq 0.06),$$

$$R_4 = I(\text{EPI} \leq 0.7, \text{health spenging} > 0.06, \text{empl. rate} \leq 0.51),$$

$$R_5 = I(\text{EPI} \leq 0.7, \text{health spenging} > 0.06, \text{empl. rate} > 0.51),$$

$$R_6 = I(\text{EPI} > 0.7, \text{gross debt} \leq 0.4)$$

Note that the region $I(\text{EPI} > 0.7, \text{gross debt} > 0.4)$ does not appear neither in model final output nor in the regions mentioned above. The reason is that it acts as a reference region for model specification needs.

By looking at the coefficients of the BTRT model output, some interesting results emerge:

- For the first object o_1, taxes on income, the level of military spending x_{14}, and

Table 5.4: MS regression trunk final output: the Table shows the estimated coefficients associated to the objects taxes on income o_1, social security contributions o_2, taxes on property o_3, and taxes on goods and services o_4. The last object o_4 is set as reference level, so that the estimated parameters associated to $\hat{\lambda}_{o_4,h}$ are automatically set to zero. The standard errors are shown in parenthesis and the stars $'*'$ associated to some estimate coefficients indicate that they are significantly different from zero with a pvalue lower than 0.001 ($'***'$), 0.01 ($'**'$) and 0.05 ($'*'$), respectively.

	$\hat{\lambda}_{o_1,h}$	$\hat{\lambda}_{o_2,h}$	$\hat{\lambda}_{o_3,h}$
$\hat{\lambda}_i$	13.15* (5.33)	-33.63** (11.83)	-50.09 (115.79)
$\hat{\beta}_{i,x1}$	-4.16 (2.77)	1.24 (4.62)	8.73 (4.54)
$\hat{\beta}_{i,x2}$	0.71 (0.93)	-5.28** (1.94)	-2.99 (1.98)
$\hat{\beta}_{i,x3}$	5.45 (3.68)	20.95** (7.56)	15.78 (8.15)
$\hat{\beta}_{i,x4}$	-16.68** (5.58)	-5.02 (9.13)	-32.94* (12.96)
$\hat{\beta}_{i,x5}$	14.89* (5.43)	-11.30 (7.38)	5.96 (8.51)
$\hat{\beta}_{i,x6}$	-15.22* (7.38)	-2.61 (10.94)	4.06 (12.10)
$\hat{\beta}_{i,x7}$	-10.12* (4.48)	2.40 (10.70)	-15.69 (12.72)
$\hat{\beta}_{i,x8}$	-8.05* (3.78)	16.25* (6.72)	11.85 (7.51)
$\hat{\beta}_{i,x9}$	2.82 (3.39)	5.23 (5.50)	7.94 (6.69)
$\hat{\beta}_{i,x10}$	-4.54 (4.09)	16.92* (7.60)	-0.19 (10.47)
$\hat{\beta}_{i,x11}$	0.33 (1.60)	7.91** (3.05)	9.32* (4.04)
$\hat{\beta}_{i,x12}$	-11.06 (7.49)	-28.28* (12.56)	-51.68** (16.17)
$\hat{\beta}_{i,x13}$	-6.54* (2.65)	22.30*** (6.52)	24.17** (7.61)
$\hat{\beta}_{i,x14}$	-25.63 (25.38)	6.32 (33.37)	137.84** (53.16)
$\hat{\beta}_{i,x15}$	-3.98 (21.42)	-65.68 (38.44)	-45.98 (43.92)
$\hat{\beta}_{i,x16}$	10.26 (13.88)	20.80 (17.44)	5.51 (25.97)
$\hat{\beta}_{i,x17}$	-4.15 (2.63)	22.81** (7.01)	33.23*** (8.64)
$\hat{\beta}_{i,R2}$	-3.21* (1.32)	0.83 (1.88)	14.30 (115.09)
$\hat{\beta}_{i,R3}$	-2.25* (1.06)	0.50 (1.42)	11.58 (115.09)
$\hat{\beta}_{i,R4}$	-0.77 (1.32)	4.87* (2.22)	19.19 (115.10)
$\hat{\beta}_{i,R5}$	-3.24*** (0.96)	3.31* (1.29)	6.00 (153.15)
$\hat{\beta}_{i,R6}$	7.93 (201.22)	-4.26* (1.73)	8.35 (115.10)

exports x_4 have a substantial impact on the size of taxes on income. In particular, the higher the level of military spending or exports, the lower the probability that taxes on income are higher than the taxes on goods o_4 (i.e., the reference category). On the contrary, the import levels and health expenditures positively affect the size of taxes on income. Then, all the regions harm taxes on income except for R_6. Note that all the countries in this region are OECD members;

- About the second object o_2, compulsory social security contributions, the covariates savings x_3, EPI x_{13}, and other spendings x_{17} have a positive and strong impact on the size of o_2. For instance, the higher the EPI, the lower the log-odds that social security contributions are higher than taxes on goods. The EPI has a positive and high impact on o_2, contrary to what happens for income taxes, for which the EPI has a negative effect. It seems that in the most developed countries, it is very likely that social security contributions are higher than taxes on goods. In addition, it is interesting that the government consumption x_{12} has a strong and negative effect on social security contributions. Finally, the region R_4 and R_6 have a significative effect on o_2. The first has a positive impact, the latter a negative one;

- In the end, the third object o_3, taxes on property, has a strong tendency to be the last object ranked as the intercept is the lowest one. Then, about the main effects, the covariates export levels x_4 and the government consumption x_{12} have a strong and negative effect on the size of taxes on property. On the contrary, military spending x_{14} and other spending x_{17} positively impact this tax category. Regarding the interaction effects, all the regions found by the BTRT algorithm positively impact the comparison between taxes on property and taxes on goods and services.

5.5 Comparing BTRT with the basic LLBT model and BTtree

The following section shows an application to our data of the LLBT model without subject-specific covariates and the "BTtree" model proposed by Strobl (Strobl, Wickelmaier, & Zeileis, 2011).

The LLBT model without subject-specific covariates applies a generalized linear model with log-link and Poisson distribution intending to calculate the λ_i values that act as intercepts in the BTRT model. Once this value has been obtained, the worth parameters π_i are easily calculated through the inversion shown after Equation 3.2. This model has been applied using the *prefmod R* package , which does not allow the integration of numerical subject-specific covariates. For this reason, the analysis was carried out assuming that the covariates do not affect the order of magnitude of the tax revenue categories.

An application of this type can be considered an analysis aimed at finding consensus ranking, with the difference that a probabilistic approach is followed here. In fact, in addition to the estimated values of the lambda coefficients, the standard errors and the significance of the coefficients themselves are provided. The output of the LLBT model without subject-specific covariates is shown in Table 5.5.

Table 5.5: Log-linear Bradley-Terry model without subject-specific covariates: results

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.61	0.05	67.24	0.00
o1	-0.20	0.07	-2.72	0.00
o2	-0.73	0.08	-9.24	0.00
o3	-1.26	0.08	-15.37	0.00
o4	0.00	0.00	0.00	0.00

From these values we get the worth parameters π_i for each object so that $\pi_1 = 0.33$, $\pi_2 = 0.11$, $\pi_3 = 0.04$, and $\pi_4 = 0.50$. These values are in line with those shown in Figure 2. If we observe the consensus ranking within the root node of the tree, we obtain the

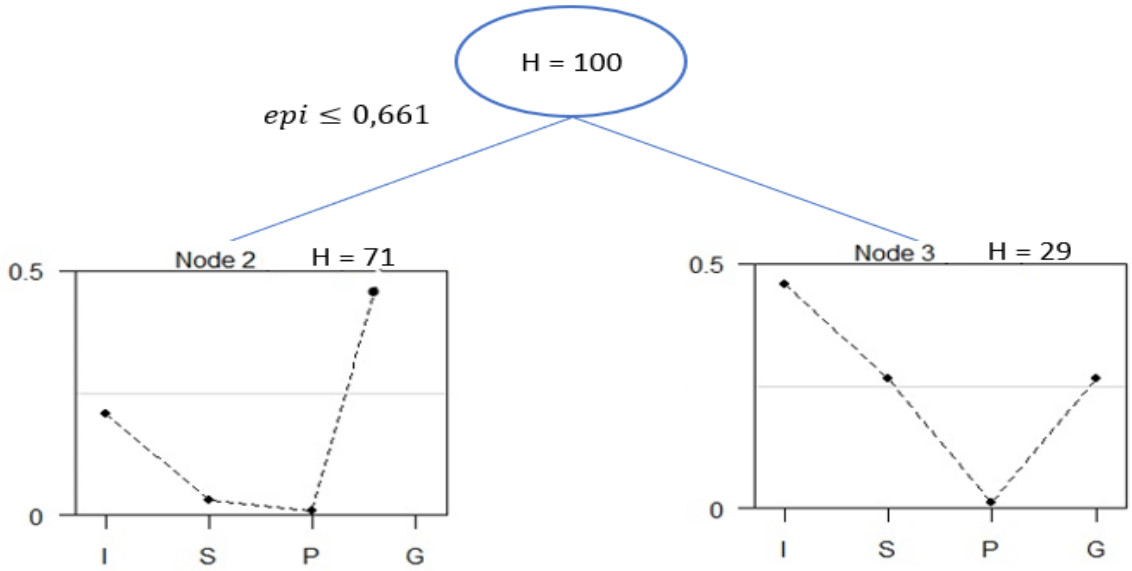
same ordering: the fourth object (taxes on goods) is in the first position, the first object (taxes on income) is in second place, the third object (taxes on property) is in the third position and the second object (social security contributions) is in the last position.

This approach is the same one followed by Dittrich. However, it is based on the use of the *prefmod* *R* package, in which the possibility of regressing the response variable on numerical subject-specific covariates is not implemented. Furthermore, since it is not a tree-based division procedure, different results are not provided based on the characteristics of the individuals, but the average value for each of them of the lambda coefficients is calculated. The BTRT model, on the other hand, provides a different composition of the tax revenues ordering based on the main effects and the interaction of the covariates that have the most significant effect on the ordering itself.

Strobl et al., 2011, have developed a model capable of applying a tree model to the BT model. They determined a model-based recursive partitioning where splits are selected with a semi-parametric approach by looking for instability of the basic Bradley-Terry model object parameters. The model is implemented within the *psychotree* *R* package by using the *bttree* function. The drawback of this approach is that the final output contains less information than those contained in BTRT. The final result of BTtree provides the preference scales in each group of the partition that derives from the order of object-related parameters, but it does not offer information about how the subject-specific covariates affect the judges' preferences. Therefore, this semi-parametric model returns beta coefficients neither for the main effects nor for the interaction effects between the covariates. In addition, it is essential to underline that the split selection procedure is carried out differently than in the BTRT model. As pointed out by the authors, the testing procedure for the split search can be very challenging (Wiedermann et al., 2021). They use the M-fluctuation test that is a score-based procedure (Zeileis and Hornik, 2007; Zeileis et al., 2008) to research the best split covariates, while our method is based on the easy-to-compute decrease in deviance by following the regression trunk approach within the STIMA algorithm.

Applying the *bttree* function to our dataset, we obtain the result shown in Figure 5.3. The visualization of the composition of the two terminal nodes was obtained by

Figure 5.3: Application of BTtree proposed by Strobl et al, where H is the number of judges, I = taxes on income, S = social security contributions, P = taxes on property, and G = taxes on goods and services. The two graphs indicate the estimated worth parameters for both terminal nodes.



plotting the results of the *bttree* function. In the Strobl model, EPI is the first and only covariate used to split countries. Although this first result is in line with that obtained by the BTRT model, this application's main difference is even more evident. The Strobl model does not find interactions between covariates as it works differently than the BTRT: after fitting a BT model, one wonders if the order presented in the root node is the same for all observations in the dataset. Then the covariates that cause greater instability in the ordering and, therefore, in the estimated lambda parameters are selected. Using EPI as the first split variable, it is possible to obtain two different orders, as shown in Figure 5.3. For countries with an EPI lower than or equal to 0.661, the revenues related to income taxes are lower than those on goods. Conversely, the exact opposite occurs when EPI is greater than the indicated value.

Income taxes are in the second position in the terminal nodes to the left of the first split, while taxes on goods are in the first position. The situation is reversed in the terminal nodes to the right of the first split. However, through the application of the BTRT model, it is further possible to differentiate the subgroups based on the ordering of their income by taxation, thanks to the effect of the interactions.

In conclusion, the model of Strobl and the BTRT respond to different needs. While the first searches for the variables that cause greater instability in ordering objects, the second searches for the variables that improve the model's goodness-of-fit and have the most significant impact on ordering objects by considering their main and interaction effects simultaneously.

5.6 Discussion

The analysis of financial data could be challenging when considering different countries worldwide. We deviated from the usual way these data are treated in literature by analyzing the problem from a different point of view: how the size of tax revenues by category of different countries is affected by the characteristics of the countries themselves? For this purpose, we create a new dataset by combining data mainly from OECD, IMF, and World Bank databases for the year 2018. First, the tax revenues by four different categories (OECD classification) have been ordered by size for every 100 countries in our sample. These orderings represent the starting point to approach this problem by following an innovative probabilistic approach for preference rankings, the Bradley-Terry Regression Trunk model. Most studies in literature focus on the determinants of tax revenues, but few investigate the composition of tax revenues. Moreover, there are few studies in which countries that are not part of the same economic organization/area (e.g., OECD) are analyzed simultaneously. This study, therefore, makes it possible to analyze tax categories through the paired comparison system for a high number of countries worldwide.

The BTRT model fits well when the need is to partition individuals, splitting the observations based on the orderings associated with them and the causal relationship between the ordering and the subject-specific characteristics. It is reasonable to assume that the ordering of tax revenue for a country depends on its socio-economic characteristics. Hence, for each country, we collected a group of covariates generally used in the literature to study the determinants of tax revenues. In addition, we also tried to contribute to that branch of literature that deals with the study of the relationship between tax revenues and government expenditures. Four representative expense items have been

considered for this purpose: spending on military, education, health, and others (residual category for all the other government expenses).

Given the large number of subject-specific covariates that make up the dataset, we applied the Bradley-Terry-Luce Lasso selection model. It is suitable when the goal is to choose the covariates that significantly impact each object of an order. It allowed us to considerably reduce the number of parameters in the BTRT model. At this point, the orderings constitute the basis for creating a design matrix in which all orderings are expressed through the paired comparison system. This matrix contains all the information the model needs, and the tree-based algorithm starts its work: countries are partitioned by choosing the best split covariates in terms of model deviance reduction until the search for a different interaction effect causes an improvement in the information value of the tree itself. The final result is a small trunk tree, representing a fair compromise between ease of interpretation of results and effectiveness in capturing the best interaction effects between the covariates.

In our case, the final result is a partition of 100 countries into six terminal nodes. The splitting covariates are represented by the Environmental Performance Index (EPI, used as a proxy of development degree), the gross government debt, health spendings, and employment rate. The algorithm selected two first-order interactions and one higher-order interaction effect. The firsts are EPI-health spending and EPI-government gross debt interactions, while the higher-order interaction is EPI-health spending-employment rate.

The results section shows the most significant coefficients for each object composing the tax revenues orderings. They contain information about the effect size of each tax revenue category's most critical main effects and interaction effects, which constitutes the main strength of this model: to provide a probabilistic measure of the causal relationship between subject-specific covariates and objects presented in paired comparisons. In Section 5.5, we compared our model with those used in the field of preference data for the partitioning or analysis through the Bradley-Terry model. Compared to Dittrich model, ours allows a tree-based algorithm to break down the analyzed sample. This approach allows for an easy-to-read representation of the final result. Compared to

Strobl semi-parametric model, however, ours returns an estimate of the effect size of the subject-specific covariates on each object of the paired comparison.

In conclusion, this chapter presented an application in public finance that can represent a point of reflection for policymakers. It makes it possible to estimate the effects of an economic shock on the structure of tax revenues. This type of application can also represent a point of reflection for countries that want to change the composition of their tax revenues. The final results show that it is possible to obtain a variation in the size of the tax revenues by operating on one's socio-economic characteristics. In this way, it may not be necessary to initiate legislative processes for regulatory changes in taxation, which in many cases take a very long time.

Future research aims to interpret the final output even more intuitively through a simplification of the structure of the dependent variable. Our case derives from the outcome of each pairwise comparison between objects. Instead, a solution might be to deal directly with rankings rather than converting them into pairwise comparisons. The Mallows extension for the Bradley-Terry model can be a solution to reduce the high number of parameters of the final BTRT model, which is the main BTRT model drawback.

Chapter 6

Advances on the Bradley-Terry Regression Trunk model: the Mallows extension

6.1 Mallows-Bradley-Terry model

The government budget, also called public fiscal balance, is a flow variable. It is calculated as the overall difference between government revenues and spending. If the outcome is positive, the budget is surplus; otherwise, there is a deficit. Positive net lending allows governments to provide financial resources to the economic sector, while negative net lending indicates governments need to apply financial resources to other sectors. The budget can be referred to the central government or local municipalities, and one of the critical roles for policy-makers is to maintain these balances positively, avoiding undesirable effects (Jain, 1989). This need was increased after the 2007 financial crisis erupted in the U.S. real estate market, which turned into a sovereign debt crisis.

More generally, a country's GDP is negatively influenced by tax revenues and positively influenced by the government's public spending. Nevertheless, the relationship between tax revenues and government expenditures is unclear. In literature, several

alternative hypotheses try to explain it. For instance, Peacock and Wiseman, 1979, followed by several other authors (Anderson et al., 1986; Von Furstenberg et al., 1986), support the spend-and-tax model by demonstrating that spending leads to revenues. On the contrary, Friedman, 1978 suggests that taxes lead to government spending because the second has to adjust to the level supported by taxation. The Friedman theory was supported by other authors, such as Manage and Marlow, 1986, Ram et al., 1988, and Blackley, 1986. Meltzer and Richard, 1981, finding the middle ground between these two theories, demonstrate that spending and taxes are determined simultaneously, supported by authors such as Miller and Russek, 1990, Bohn, 1991, and Jones and Joulfaian, 1991. All three theories are the result of empirical tests on time series data. This fact led to critiques due to the concern that the time series data are stationary, with the risk of leading to spurious results (Hondroyiannis and Papapetrou, 1996). Most of the studies conducted in the literature focus on the tax revenues determinants by considering a single country or countries in the same organization (e.g., OECD countries) in a specific period. This choice is usually considered as a solution to the big issue affecting comparability across heterogeneous groups of countries: for instance, there are countries where hospitals are classified as public corporations instead of the government sector. In addition, there are cases in which the concept of public ownership is not clear. However, this issue is partially overcome if a unique classification for tax revenues and government expenditures is followed.

This chapter focuses on the determinants of tax revenues by using government expenditures as country-specific covariates. We built a new dataset composed of 100 countries worldwide, and for each of them, the tax revenues by four categories are collected for 2018. The heterogeneity among countries in our dataset is taken under control by collecting data from the OECD database, which follows the same classification for tax revenue categories in different countries. The main innovation is that tax revenues are differentiated by four categories, based on the tax source (i.e., taxes on income, social security contributions, taxes on property, and taxes on goods and services). The critical point of our analysis is that those continuous and numerical data were ordered by their size for each country to obtain 100 rankings. This transformation ensures consistency in comparing data regarding different countries with different fiscal systems. Even by

following the OECD classification for tax revenues, there could be inconsistency issues. The transformation into rankings can be a solution. The size of tax revenue categories is respected, but now we work on ordinal data instead of continuous ones. In our case, it can be reasonable to assume that the tax revenues size depends on the characteristics of a country (IMF).

Differently from the application in Chapter 5.1, once we obtained rankings, we applied the Mallows-Bradley-Terry (MBT) to our data. The Mallows specification extends the BT model for paired comparisons to the case of rankings. This model is typically used for treating rankings and discovering the causal effect of specific covariates, also called subject-specific covariates. When the log-linear Bradley-Terry model (LLBT) is applied, the causal effect is estimated through a GLM. In our case, we introduce country-specific covariates, which are four government expenses categories: health expenditure, military expenditure, education expenditure, and other expenses. We collected these data by following the COFOG classification, allowing comparability among countries adopting different accounting systems.

In summary, the primary purpose of this application is to obtain a partition of countries based on the size of their tax revenue categories and the relation between taxes and government expenses. Hence, the MBT model is combined with the regression trunk within the STIMA algorithm. This combination determines the new model presented in this paper, the Mallows-Bradley-Terry Regression trunk (M-BTRT). Chapter 5 demonstrated that government expenses and other socio-economic characteristics impact the size of tax revenues differentiated by categories. In that application, we partitioned countries through the BTRT model and used a high number of country-specific covariates. However, one of the significant drawbacks of the BTRT model is the interpretation of the results. This model has a high number of parameters to be estimated. The Mallows extension reduces this number and makes interpretation of the final results easier. Hence, the M-BTRT can be considered the BTRT model's natural advance. The reasons why the Mallows extension is easier to interpret are shown in Chapter 6.3, dedicated to our methodology. We show an application on financial data that picks up the BTRT application discussed above. The main differences are:

- we directly focus on the relationship between tax revenues and government expenses without considering other socio-economic characteristics;
- we use rankings as input data instead of paired comparisons;
- the model is based on the Mallows extension to the basic BT model.

The final result is still a small regression tree, called trunk, built by considering the main and interaction effects of government expenses on the rankings of tax revenues. The proposed model shares the same advantages of the regression trunk framed into the STIMA algorithm and the BTRT model, with the further advantage of easier-to-read model output.

The rest of the Chapter introduces the main characteristics of our dataset, the Mallows extension to the basic Mallows-Bradley-Terry model, the benefits related to this extension, and how the M-BTRT model works. The final partition of the countries in our data is shown in Chapter 6.4 with a world map and an easy-to-read regression tree visualization. Finally, results and future research steps are discussed in Chapter 6.5.

6.2 Data

Our dataset is cross-sectional and is composed of 100 countries worldwide. For each country, we report tax revenues by category for 2018—data derived from the Global Statistics Database (OECD, 2018). The OECD classification ensures consistency across countries and provides a high granularity of tax revenue categories (Constructing the global revenues statistics, 2018). Taxes are then classified as income and profits (heading 1000), compulsory social security contributions (heading 2000), payroll and workforce (heading 3000), property (heading 4000), goods and services (heading 5000), and other taxes (heading 6000). We consider only four out of six tax revenue categories. We exclude the categories "payroll and workforce" and "other taxes" from our analysis. This choice is justified because they are both ranked in the last position for almost all the countries in our sample. In addition, these two categories present values equal to zero for almost

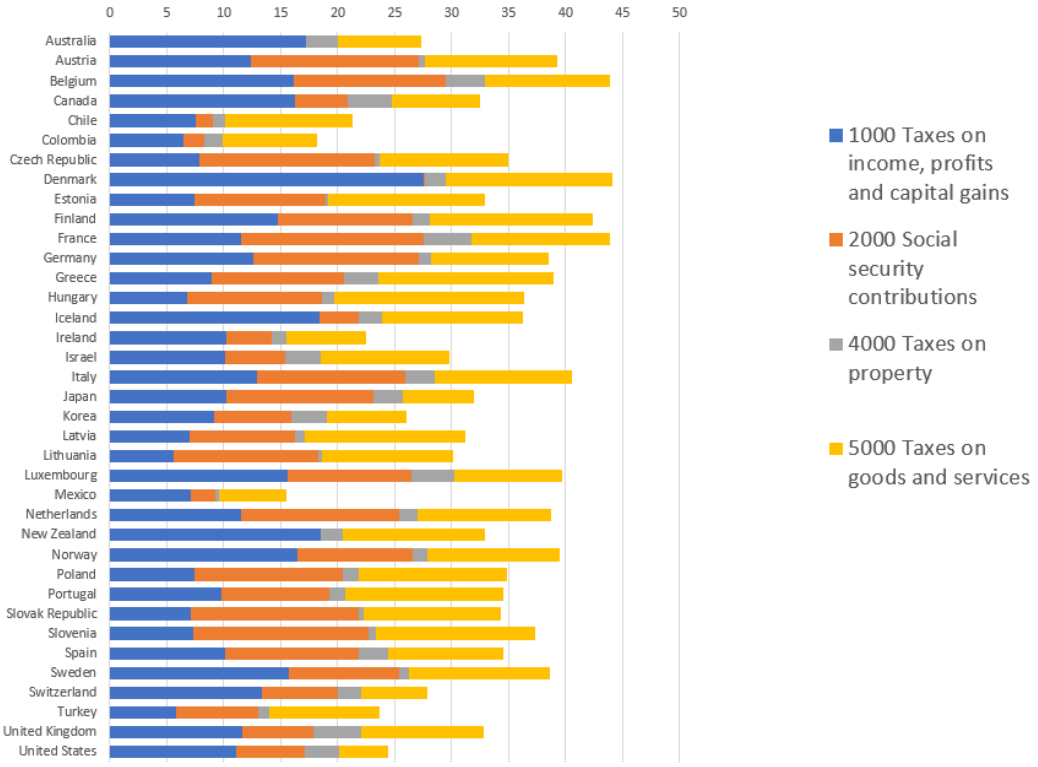


Figure 6.1: OECD: Tax revenues composition

all the cases. The economic size and the development of each country are taken into account by expressing tax revenues as a tax-to-GDP ratio, which is the ratio of nominal tax revenue and nominal GDP for the year 2018. Expressing tax revenues in terms of GDP aims at comparing tax levels across countries with different development degrees.

Figures 6.1 . . . 6.4 show the composition of tax revenue categories for each country in our dataset. Countries are shown based on their economic and geographical position: OECD countries, Africa, Asia, and South America. Data for the last three geographical locations derive from the OECD's respective tables.

The work conducted by the OECD to classify the tax revenues is affected by the comparability problems mentioned above. For this reason, we transformed tax revenue categories from continuous data to rankings. Rankings are numerical vectors that assign to each tax category values from 1 to 4, where 1 represents the higher value and 4 is the lower one. This transformation allows considering the tax data by their size instead of their original value. Given that we started using numerical and continuous data, there

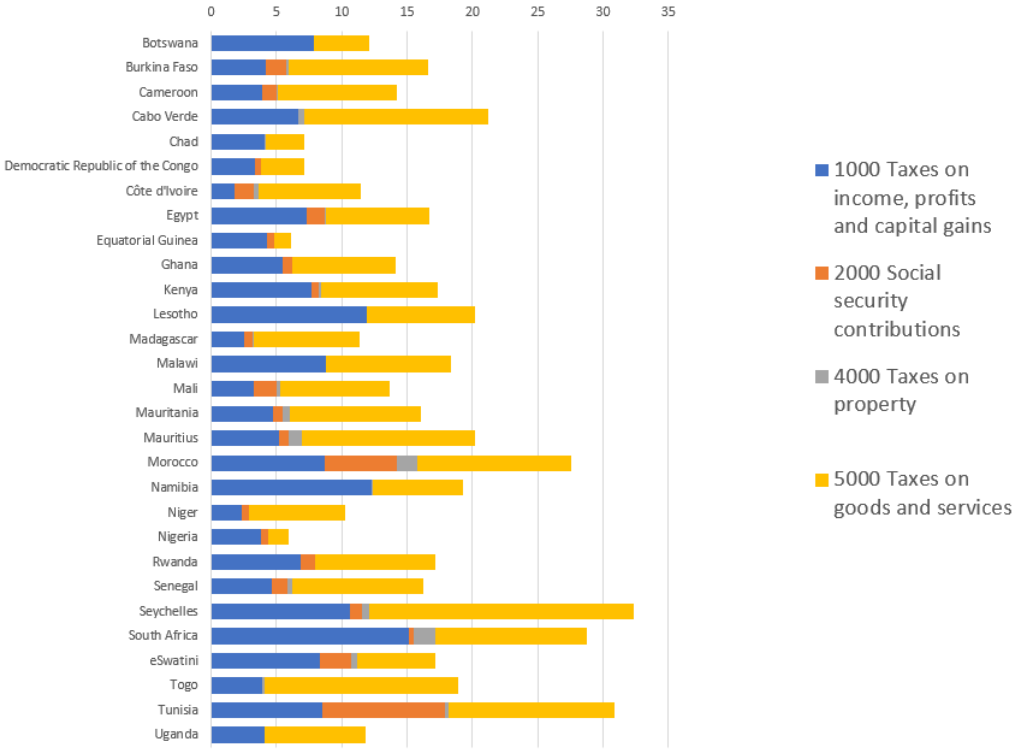


Figure 6.2: Africa: Tax revenues composition

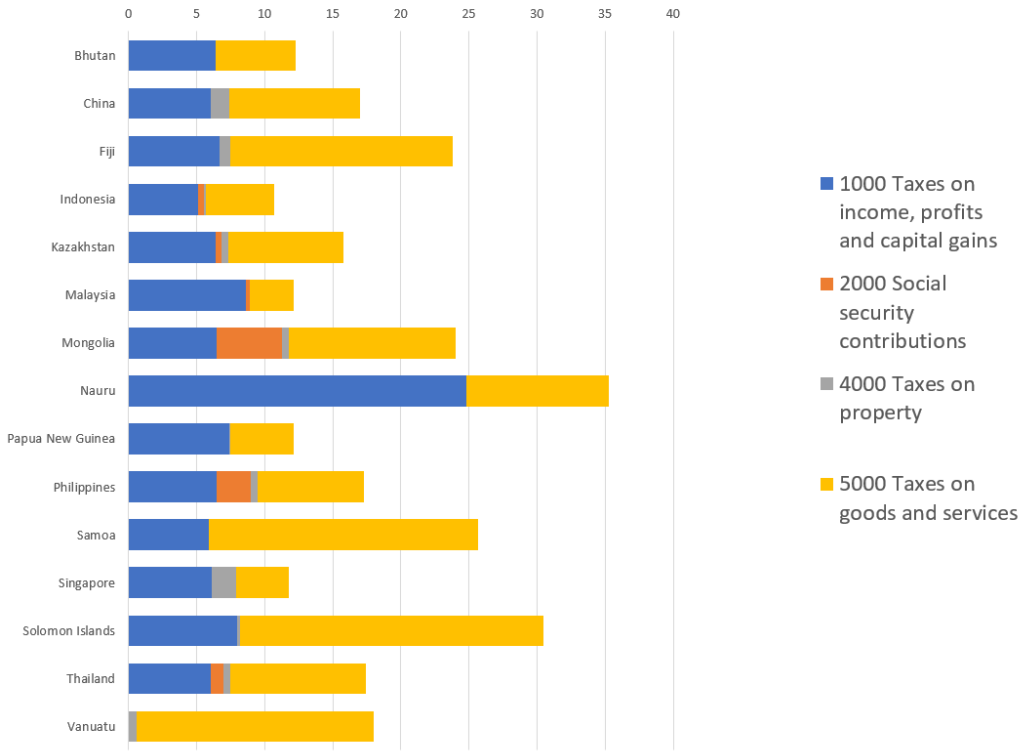


Figure 6.3: Asia: Tax revenues composition

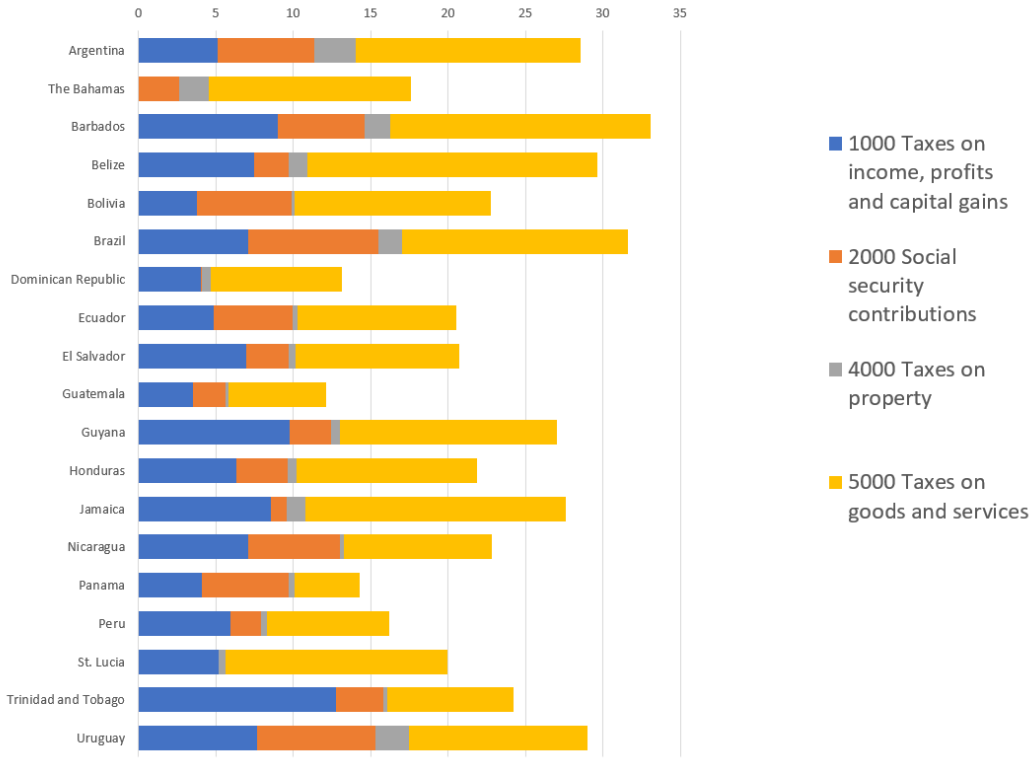


Figure 6.4: South America: Tax revenues composition

are no ties between two or more tax revenue categories in our dataset.

We aim to analyze the causal effect of four government expenditure categories on each tax revenue category by finding subgroups of countries that share a similar pattern in the tax revenue composition. Then, the government expenditure categories represent our country-specific covariates. They refer to the year 2018, the same year tax revenues are referred, and the data source is the IMF database. Specifically, we consider government expenses on military, education, health, and others. The latter is a residual category calculated as the difference between the total government expenditure of the country m (OECD source) and the sum of the other expenses associated with that country. The expenses data are comparable across countries, given that they are collected by following the COFOG classification. Again, the government expenditure categories are expressed in terms of GDP. Missing values for each object and covariate were replaced with the median value conditioned to the location of each specific country. With the term of location, we mean the classification adopted by OECD, which is an economic and geographical location. Specifically, the countries in our dataset are assigned

to four locations: 36 OECD countries, 29 countries from Africa, 15 from Asia, and 21 from South America.

Table 6.1 reports the first six rows of our dataset, where the four objects (i.e., the tax revenue categories) are now expressed as ordinal values. For instance, in Australia, tax revenues on income are the highest, and those on social security contributions are the lowest. Then, the government military expenses are two percent of the GDP. Australia's total government expenditure in 2018 was approximately 37% of the total GDP so that the category "others" was obtained by the difference between the total government expenditure and the sum of the expenses on military, education, and health.

Table 6.1: Dataset head: First six rows. We report the rankings of tax revenue categories associated to each country and the four government expenditures as country-specific covariates.

Countries	income	social	property	goods	military	education	health	others
Australia	1	4	3	2	0.02	0.05	0.09	0.21
Austria	2	1	4	3	0.01	0.05	0.10	0.32
Belgium	1	2	4	3	0.01	0.06	0.10	0.35
Canada	1	3	4	2	0.01	0.04	0.11	0.25
Chile	2	3	4	1	0.02	0.05	0.09	0.09
Colombia	2	3	4	1	0.03	0.05	0.07	0.14

As mentioned in Chapter 2, one of the most common analyses when dealing with rankings is the research of consensus ranking. This solution can be found through distance-based approaches or probabilistic models. In this application stage, we calculated the consensus ranking in our dataset by maximizing the extended correlation coefficient τ_x. We found the solution of the typical aggregation problem by using the R package *ConsRank* and the function $DECoR$. The consensus ranking shows the following order of objects, from the biggest to the lowest in size: Goods and services, income and profits, compulsory social security contributions, and property. Table 6.2 reports a summary of the key statistics for the covariates that will be used as predictors in the M-BTRT model.

Table 6.2: Summary of the kew statistics for our country-specific covariates (i.e., government expenses)

Expenses	vars	mean	sd	median	min	max	range	skew	kurtosis	se
military	x_1	0.01	0.01	0.01	0.00	0.05	0.05	1.44	3.72	0.00
education	x_2	0.04	0.01	0.04	0.01	0.08	0.07	0.72	0.25	0.00
health	x_3	0.07	0.03	0.07	0.02	0.17	0.15	0.74	0.96	0.00
others	x_4	0.20	0.10	0.20	0.03	0.76	0.73	1.77	7.89	0.01

6.3 Mallows-Bradley-Terry Regression Trunk

The Mallows-Bradley-Terry model derives from the specification of the Babington Smith probability model for rankings. This specification is based on the assumption that all the ranking structures can be obtained from the pairwise comparison (Dossou-Gbété et al., 2009). Starting from $n \times (n-1)/2$ paired comparisons parameters, the probability that the ranking r occurs is given by

$$p(r, \theta) = c(\theta) \prod_{i,j} \theta_{ij}^{I(r(i)<r(j))}, \tag{6.1}$$

where $\theta_{ij} \in]0, 1[, 1 \leq i < j \leq n]$ and $I[A]$ is the indicator function of the event A so that $I[A] = 1$ when A occurs and $I[A] = 0$ otherwise. The $c(\theta)$ is the normalizing constant and θ_{ij} is the probability that the object i is ranked lower than the object j.

The Babington Smith model (Kendall and Smith, 1939) involved a number of parameters equal to the number of paired comparisons, which may be too large to deal with. Mallows proposed a sub-model of the Babington Smith model for ranks without tie by assuming a Bradley-Terry model and constraints on the parameters space. This extension reduces the number of Bradley-Terry and Babington Smith model parameters to obtain a more interpretable model.

The Mallows-Bradley-Terry model assumes that the probability $p(r, \pi)$ is proportional to the product $\prod_{i=1}^{n} \pi_j^{n-r(j)}$ so that the model can be written as (Dossou-Gbété et al., 2009)

$$p(r, \pi) = C(\pi) \prod_{j=1}^{n} \pi_j^{n-r(j)}. \tag{6.2}$$

When the constraint $\sum_{j=1}^{n} \pi_i = 1$ is assumed for the model identificability, the model can be reparametrized and the relationship between π_js and the θ_js is $\pi_j = \frac{exp(\theta_j)}{\sum_{l=1}^{n} exp(\theta_l)}$ with $j = 1 : n$ and $\theta_n = 0$. The parameters θ_l refers to the number of paired comparisons parameters $l_1, \ldots, n-1$.

Vitelli et al., 2018, proposed new methods for Bayesian inference in Mallows models that work with any right-invariant distance. This method performs inference on the consensus ranking, also when dealing with partial rankings or pairwise comparisons. In addition, for cases with subject-specific covariates, the authors proposed a mixture model for clustering. Critchlow and Fligner, 1991, proposed treating the Mallows-Bradley-Terry model as a GLM with a log link function and a multinomial family. The MBT model is implemented by the *eba* (Wickelmaier and Schmid, 2004) and $prefmod$ *R* packages. Here we used the $prefmod$ package to estimate the MBT model as a special case of pattern models, where a pattern is considered a set of paired comparisons considered simultaneously (Turner et al., 2020). In this case, the response variable is $y = (y_{12}, \ldots, y_{J-1,J})$ and the probability of observing a pattern is defined as follows

$$p(y) = p(y_{12}, \ldots, y_{J-1,J}) = c \prod_{j<k} \left(\frac{\sqrt{\pi_j}}{\sqrt{\pi_k}}^{y_{jk}} \right) \tag{6.3}$$

The pattern model can be expressed as a log-linear model introducing subject-specific covariates. In addition, several extensions (ties, object-specific covariates, position effects, and missing values) are proposed by Hatzinger and Dittrich in their *R* packages $prefmod$. The model is estimated through a GLM with log link and Poisson distribution. The response variable is the number of times a certain ranking is observed. The design matrix is the input for the application of the glm *R* function. When no subject-specific covariates are added to the model, the design matrix is only composed by a number of rows equal to the number of rankings in the permutation universe when no ties are allowed. This number equals $n!$, and the response variable y counts the number of times each ranking occurs in our input rank data. When subject-specific covariates

are introduced, a different design matrix is built for each country m. An example of a typical design matrix for LLBT pattern models is shown in Table 3.2 in Chapter 3.1. In this case, the number of rows dramatically increases to $n! \times m$. This is the main difference with the typical design matrix for the LLBT model when the orderings are expressed as paired comparisons. In this case, when no subject-specific covariates are introduced, the design matrix is composed by a number of rows equal to the number of paired comparisons $n \times (n-1)/2$, and the response variable y indicates the number of times a specific object i is preferred to another one j.

We apply the log-linear MBT model as a particular class of pattern model to our dataset by combining this model with the regression trunk approach. The result is the M-BTRT model that is an advance of the BTRT model. The algorithm for the split research is the same procedure followed by the BTRT model and synthesized in 3.1. The main difference is that we are now working on rankings instead of paired comparisons to formulate the model the same way as the BTRT. The model can be represented by a single formulation as follows

$$\hat{\lambda}_{i,h} = \hat{\lambda}_i + \sum_{p=1}^{P} \hat{\beta}_{i,p} x_{p,h} + \sum_{t=1}^{T-1} \hat{\beta}_{i,P+t} I\{(x_{1,h}, \ldots, x_{P,h}) \in t\}, \tag{6.4}$$

where $\hat{\lambda}_{i,h}$ are the object-parameter for each object i and each country h in the node t. The first part of the formula refers to the main effect part and it is specified by $\sum_{p=1}^{P} \hat{\beta}_{i,p} x_{p,h}$. It can be interpreted as the main effects of our country-specific predictors on the size of tax revenue category i for the country m. The second part of the equation represents the interaction effects part. The interactions are estimated by $\sum_{t=1}^{T-1} \hat{\beta}_{i,P+t} I\{(x_{1,h}, \ldots, x_{P,h}) \in t\}$ for each group of individuals in each terminal node T. One terminal node has to be treated as reference group, so that we estimate $T-1$ interaction terms, which is equal to the number of splits of the final tree. The estimated intercept $\hat{\lambda}_i$ quantifies the overall location of object i for all the individuals of the trunk.

As for every tree-based model, the pruning procedure avoids the overfitting case. The M-BTRT model follows the same pruning procedure as the BTRT model. We apply the pruning back procedure once the full trunk is grown using the V-fold cross-validation

with the c standard error rule ($c \cdot SE$ rule). The constant c varies between 0 and 1, and the higher its value, the more the tree is pruned back. The standard error is applied to the cross-validation deviance calculated for each tree split. The cross-validation deviance is obtained by training on $V-1$ subsets the estimated trunk model in a specific node. The left-out subset is trained as a test set, and the predicted value $\hat{y}_{ij,h}$ is obtained for each observation in the typical design matrix built by using the function *patt.design* from the *prefmod* R package. The case-wise cross-validation deviance D^{cv} is then expressed as in Equation 3.14, and its standard errors as in Equation 3.15. Usually, the cross-validation deviance follows a typical pattern: it decreases after the first splits of the trunk and starts to increase starting from a specific sequent split. The $c \cdot SE$ pruning rule is then applied as in the BTRT model. Let $t^* \in [1, T]$ be the size of the regression trunk with the lowest D^{cv}, say $D^{cv}_{t^*}$. The best size of the trunk t^{**} corresponds to the minimum value of t such that $D^{cv}_{t^{**}} \leq D^{cv}_{t^*} + c \cdot SE^{cv}_{t^*}$. The optimal choice of the pruning parameter c is investigated in Chapter 3.8.

6.4 Application

We start our application by building the design matrix with the *patt.design* R function. It creates a table with information about the rankings contained in our input rank data. When the government expenditures covariates are not included yet, the design matrix is composed of 24 rows, which is the number of rankings in the space of permutation when the number of objects is equal to four and no ties are allowed. The response variable y indicates the number of times a certain ranking occurs in our input rank data. Table 6.3 shows the design matrix without country-specific covariates yet. In the first stage, we estimated a GLM model considering the design matrix as input data and the tax revenue categories as unique parameters of the model. By applying a GLM with log link and Poisson errors, we obtain the $\hat{\lambda}_i$ values of the basic BT model when the Mallows extension is applied to the case of rankings. From the $\hat{\lambda}_i$ we obtain the worth parameters $\hat{\pi}_i = \frac{exp(2\hat{\lambda}_i)}{\sum_{i=1}^{n}(exp(2\hat{\lambda}_i)}$. After that, we plot the worth parameters through the function "patt.worth" in Figure 6.5.

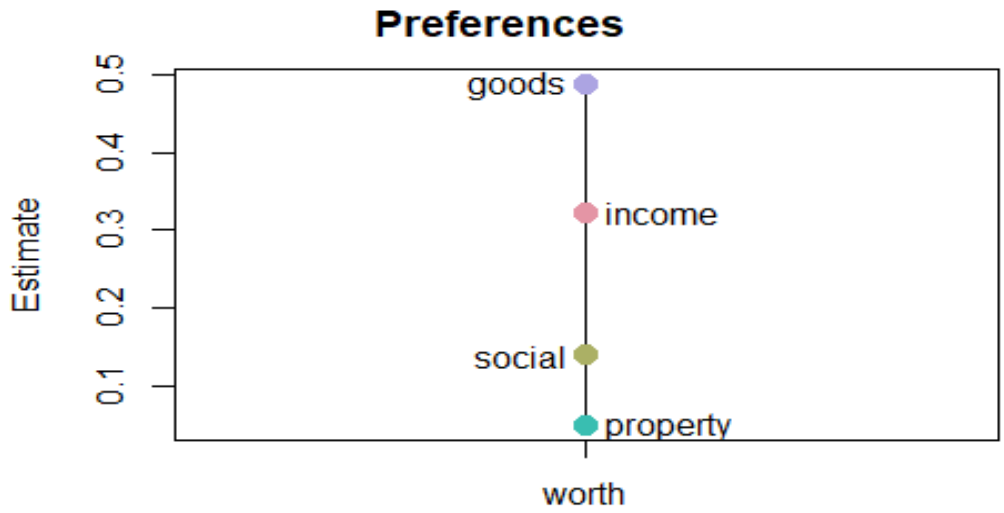


Figure 6.5: Plot of the worth parameters estimated with a GLM when no country-specific covariates are included in the model, and the objects are used as unique parameters. The order shown in the figure respects the consensus ranking. However, here we can quantify the distance among the objects by looking at the value of the estimated worth parameters.

The results of the basic MBT model with the pattern model specification bring the same results given by the consensus ranking research. The estimated ranking for the tax revenues is taxes on goods and services, taxes on income, social security contributions, and taxes on the property. However, we can quantify the distance between the objects in the estimated rank. This is the main advantage of the probabilistic models compared to the simple calculation of the consensus ranking, which only offers information about the order of objects. When the government expenditures are included as country-specific covariates, the design matrix expands for each country if there are no countries with the same characteristics and rankings of tax revenues size. Here, we are using continuous country-specific covariates so that no countries have the same government expenditures values. The design matrix is composed of 2,400 rows, which is the product of the number of permutations 24 and the number of countries 100.

Once obtained the design matrix with our country-specific covariates, the M-BTRT model can be applied for the split research. In Chapter 4.1 two different approaches are proposed for the research of splits: The One-Split-Only (OSO) approach, which does not allow the use of the same covariate for more than one split, and the

Multiple Splitting (MS) approach, which has no restrictions on the covariate to consider a candidate to split. It is demonstrated that the first approach can be helpful when the priority is to find the interaction between all the covariates under investigation, while the second approach restitutes better results in terms of model and cross-validation deviance. The MS approach considers all the values of all the covariates in the dataset for each step of the tree building procedure. For this reason, it considers a higher number of candidates in each split research so that better results are obtained. For this reason, we choose to follow the MS approach for our application.

The final tree after pruning has five splits and six terminal nodes. The best split covariates and the respective best split points are reported in Table 6.4. The model with only main effects has deviance equal to 383. The best candidate covariate for the first split is "health expenses" with a value equal to 0.23, and the model deviance is now decreasing at 362. The first interaction is found with the second split "bestsplit2", covariate "military expenses", with a value equal to 0.02. As we can notice, the covariate x_4 "other expenses" is never chosen by the algorithm for splitting the tree. Table 6.4 shows the results of the final tree after pruning, but the STIMA algorithm applies the pruning only after building the entire tree until a stopping rule verifies. Here, we obtain a tree with five splits because the cross-validation deviance increases starting from the sixth split. Table 6.5 shows the deviance, cross-validation deviance, and standard errors for each model. In Table 6.5 $mod0$ is the model with only main effects. The first interaction is added starting from $mod2$, where we have two splits, three nodes, and the interaction between health expenses and military expenses.

The pruned regression tree is shown in Figure 6.6. For each node, we report the number of countries H, and for each terminal node, we also show the consensus ranking within the group C and its associated extended correlation coefficient τ_x. Note that we report the consensus ranking as a summary measure of the node, but it results from a simple maximization problem without considering the characteristics of the countries (i.e., the government expenditures) in each node. On the contrary, we follow a probabilistic approach, and the results are shown as plots for each terminal node. The plots report the worth parameters π_i and are obtained by applying the formula in Equation 6.4 for each

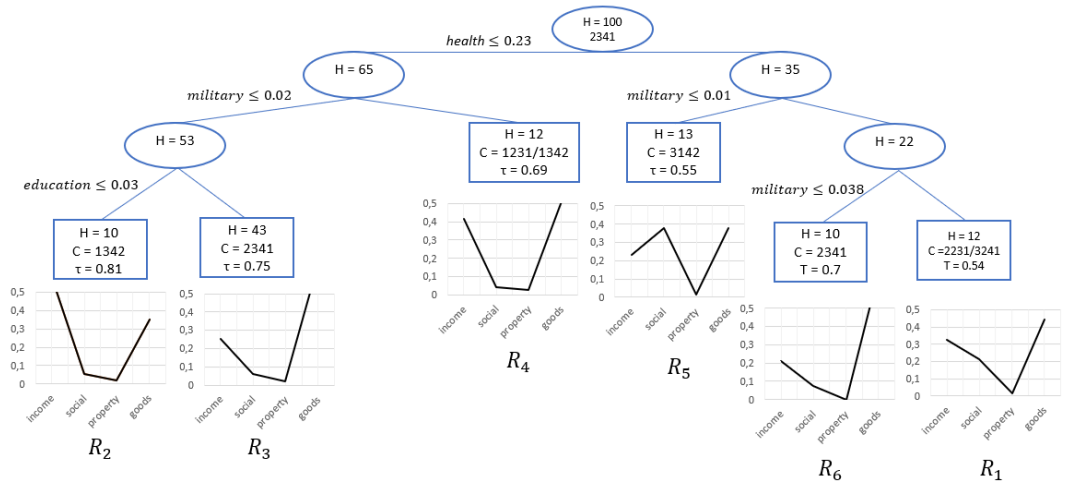


Figure 6.6: Pruned regression trunk: MS approach. The final trunk reports the number of countries H in each node of the tree. For the terminal nodes, the consensus ranking C and its associated correlation coefficient tau_x are reported for the sake of completeness. Next, the worth parameters are obtained through the estimation of the mean value of $\hat{\lambda}_{i,h}$ for each object across all the countries in each terminal node. Finally, the plots of the estimated worth parameters are shown for each region $R_1, \ldots, R_6$.

country terminal node. After that, we calculate the mean value for each object across all the countries inside the node. Finally, we calculate the worth parameters using the relationship between π_i and λ_i.

The regions $R_1, \ldots, R_T$ indicate the regions created by the final tree. They are defined as follows

$$R_2 = I(\text{health expenses} \leq 0.23, \text{military expenses} \leq 0.02, \text{education expenses} \leq 0.03),$$
$$R_3 = I(\text{health expenses} \leq 0.23, \text{military expenses} \leq 0.02, \text{education expenses} > 0.03),$$
$$R_4 = I(\text{health expenses} \leq 0.23, \text{military expenses} > 0.02),$$
$$R_5 = I((\text{health expenses} > 0.23, \text{military expenses} \leq 0.01),$$
$$R_6 = I((\text{health expenses} > 0.23, 0.01 < \text{military expenses} \leq 0.038),$$
$$R_1 = I((\text{health expenses} > 0.23, \text{military expenses} > 0.038),$$

where the last region R_1 is our reference terminal node. The final regression trunk

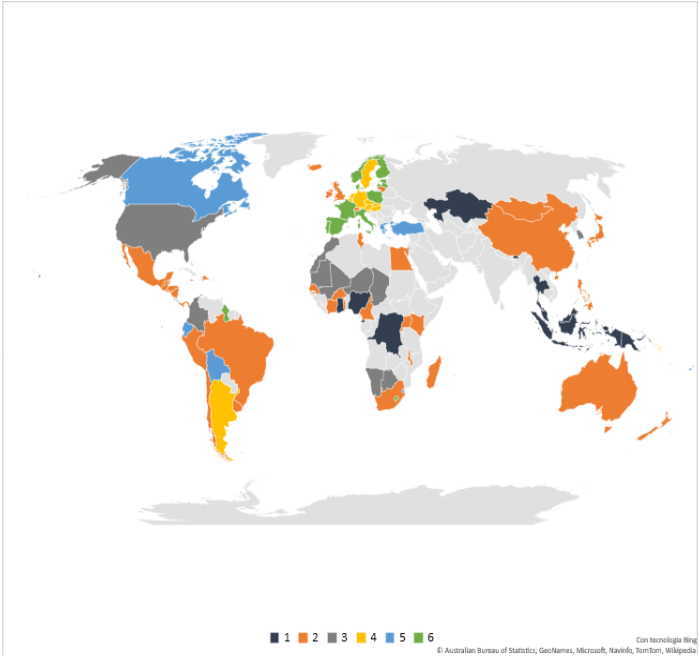


Figure 6.7: Partition of countries based on memberships to the terminal nodes.

individiduate two types of interactions: the first one is a higher order interaction between health, military, and education government expenses (i.e., regions R_2 and R_3); the second is a first order interaction between health and military expenses (i.e., R_1, R_4, R_5, and R_6).

We report the countries composing each terminal node in a map shown in Figure 6.7. The first group comprises the ten countries that constitute the region R_2. These countries are mainly the USA and countries of central Africa. They are characterized by higher revenues for income taxes than taxes on goods, social security contributions, and property. Note that the European countries are all partitioned in the last three groups, regions R_5, R_6, and R_1 in Figure 6.6. They are all characterized by higher government health expenses.

After pruning the tree, the final model reports the coefficients shown in Table 6.6. The coefficients refer to the first three objects (i.e., taxes on income, social security contributions, and taxes on property) because the last one, taxes on goods and services, is our reference object. For each of these three objects, the intercepts $\hat{\lambda}_i$ are reported in

the first row. The main effects are represented by $\hat{\beta}_{i,x_p}$, while the interaction effects are expressed by $\hat{\beta}_{i,R_t}$.

By looking at the coefficients of the BTRT model output, some interesting results emerge:

- For the first object o_1, taxes on income, the level of military spending x_1 has a substantial impact on the size of taxes on income. In particular, the higher the level of military spending, the lower the probability that taxes on income is higher than the other taxes. On the contrary, health expenditures positively affect the size of taxes on income. Then, all the regions increase taxes on income except for R_5 composed by those countries with health spending greater than 0.23;
- About the second object o_2, compulsory social security contributions, the covariate x_3, health spending, have a positive and strong impact on the size of o_2. For instance, the higher the government spending on health, the higher the log-odds that social security contributions are higher than the other tax revenue categories. In addition, it is interesting that the government spending on education, x_2, has a positive effect on social security contributions. Finally, all the regions except for R_5 have a negative effect on o_2;
- In the end, the third object o_3, taxes on property, has a strong tendency to be the last object ranked as the intercept is the lowest one. Then, about the main effects, the covariate military spending, x_1, has a strong and negative effect on the size of taxes on property. On the contrary, all the other covariates present a positive coefficient. The coefficients' signs are the opposite of those observed for social security contributions regarding the interaction effects. All the regions except for R_5 have a positive effect on the size of taxes on property revenues.

6.5 Discussion

The analysis of tax revenues across different countries could be challenging for comparability reasons. We focused on how government spending influences tax revenues using

a magnifying glass. Tax revenues constitute the analysis's object, but we differentiated from the most common literature because tax revenues are decomposed into four tax revenue categories by following the OECD classification. Comparing different countries with different fiscal and accounting systems can be challenging even when adopting a classification like the OECD's. For this reason, we transformed into rankings the numeric and continuous data about the four different tax revenue categories, obtaining ordinal values that represent the size of tax revenues for each country in our dataset. In this way, the size of tax revenues is preserved, and the comparability issue becomes easier to address. In addition, working on rankings allows the adoption of rankings models with the introduction of subject-specific covariates. It is reasonable to assume that the ordering of tax revenue for a country depends on its socio-economic characteristics. In our analysis, the subjects are represented by countries, and the covariates by four government spending categories: military spending, health spending, education spending, and other spending. The causal effect of each covariate is considered as the main effect and interaction effect, even when no a priori information is known about the interactions to include, through a new version of the STIMA algorithm. This new version is based on the BTRT model for preference data but adapted to rankings expressing financial information. The BT model can be extended by following the Mallows specification that works on rankings instead of paired comparisons. The Mallows extension is based on the concept that the probability of observing a specific ranking is proportional to the product of worth parameters associated with each object ranked.

We created a new dataset merging information from different databases (i.e., OECD, IMF, and World Bank). The dataset is composed of 100 records (countries), four objects (i.e., revenues on taxes on income, social security contributions, taxes on property, and taxes on goods and services), and four subject-specific covariates (i.e., spending on military, health, education, and others). The heterogeneity in our dataset is observed through the Mallows-BTRT model so that we obtain a partition of countries based on the size of their tax revenues and the causal relation with the central government expenditure. In our case, the final result is a partition of 100 countries through five splits and six terminal nodes. The splitting covariates are represented by the government spending on health, military, and education. The residual spending category (others)

does not appear to be the best split covariate in the final regression trunk. The algorithm selected one first-order interaction (military spending-health spending) and one higher-order interaction (education-military-health).

The application section shows the most significant coefficients (Table 6.6 for each tax revenues category (except for taxes on goods and services, which represent the reference level). They contain information about the effect size of each tax revenue category's most critical main effects and interaction effects, which constitutes the main strength of this model: to provide a probabilistic measure of the causal relationship between country-specific covariates and objects presented in rankings. Converting the estimated values by the models into worth parameters furnishes elements for predictive analysis. It is possible to investigate the most probable size of tax revenues for countries, given their government expenditure. In addition, this analysis represents a useful tool for policy-makers who want to change their tax revenues' structure by operating on government spending instead of proposing time-consuming tax reforms. In this way, it may not be necessary to initiate legislative processes for regulatory changes in taxation, which in many cases take a very long time.

Future research aims to create an R package with the BTRT and MBTRT functions to partition individuals based on paired comparisons and rankings of objects. The function implemented in R is not generalizable yet and requires modifications for a better user experience. Next, the MBTRT model can be extended for dealing with missing values (e.g., partial rankings), object-specific covariates, and order effects. In all the cases, this model has to be considered a tool to find interaction effects when all the main effects are still considered and when the analysis is composed of individuals, objects, and predictors.

Table 6.3: Design matrix for pattern models generated with "patt.design" *R* function. The input data are $n_o = 4$ tax revenue categories and $H = 100$ countries. The response variable y indicates the number of times a specific ranking is observed in the input data.

y	income	social	property	goods
0	3	1	-1	-3
0	1	3	-1	-3
0	1	-1	3	-3
0	3	-1	1	-3
0	-1	3	1	-3
0	-1	1	3	-3
30	1	-1	-3	3
16	3	-1	-3	1
9	-1	3	-3	1
10	-1	1	-3	3
4	3	1	-3	-1
4	1	3	-3	-1
0	-1	-3	3	1
0	-1	-3	1	3
0	3	-3	1	-1
0	1	-3	3	-1
13	1	-3	-1	3
12	3	-3	-1	1
0	-3	3	1	-1
0	-3	1	3	-1
1	-3	1	-1	3
0	-3	3	-1	1
0	-3	-1	3	1
1	-3	-1	1	3

Table 6.4: Pruned regression trunk: MS approach. The table shows the node in which the split is found, the splitting covariate, and its split point together with the deviance associated with each estimated model.

	Node n.	*Splitting covariate*	*Split Point*	*Model Deviance*
	1	main effects (no splits)		383
bestsplit1	1	x_3 (health expenses)	0.23	362
bestsplit2	2	x_1 (military expenses)	0.02	349
bestsplit3	3	x_1 (military expenses)	0.01	336
bestsplit4	4	x_2 (education expenses)	0.03	325
bestsplit5	7	x_1 (military expenses)	0.03	317

Table 6.5: 10-fold cross-validation results with MS approach: D = model deviance (Eq. 3.10); D^{cv} = casewise cross-validation deviance (Eq. 3.14); SE^{cv} = standard error of D^{cv} (Eq. 3.15).

	D	D^{cv}	SE^{cv}
mod0	383	0.1683	0.0001
mod1	362	0.1613	0.0002
mod2	349	0.1567	0.0002
mod3	336	0.1524	0.0002
mod4	325	0.1473	0.0002
mod5	317	0.1453	0.0002
mod6	312	0.1461	0.0002

Table 6.6: MS regression trunk final output: the Table shows the estimated coefficients associated to the objects o_1, o_2, o_3, and o_4. The last object o_5 is set as reference level, so that the estimated parameters associated to $\hat{\lambda}_{o_5,h}$ (the professor helpfulness) are automatically set to zero. The standard errors are shown in parenthesis and the stars $'*'$ associated to some estimate coefficients indicate that they are significantly different from zero with a pvalue lower than 0.001 ($'***'$), 0.01 ($'**'$) and 0.05 ($'*'$), respectively.

	$\hat{\lambda}_{o_1,h}$	$\hat{\lambda}_{o_2,h}$	$\hat{\lambda}_{o_3,h}$
$\hat{\lambda}_i$	-1.24* (0.57)	-0.84 (0.52)	-2.41*** (0.49)
$\hat{\beta}_{i,x1}$	-67.31*** (18.41)	-4.66 (15.25)	-30.35** (10.47)
$\hat{\beta}_{i,x2}$	11.81 (6.85)	-10.14 (5.97)	8.38 (4.47)
$\hat{\beta}_{i,x3}$	12.99*** (3.34)	16.07*** (3.11)	3.12 (2.55)
$\hat{\beta}_{i,x4}$	1.53 (1.24)	-0.70 (1.15)	1.58 (0.91)
$\hat{\beta}_{i,R2}$	0.99* (0.44)	-0.16 (0.41)	0.65* (0.31)
$\hat{\beta}_{i,R3}$	-0.07 (0.31)	-0.84** (0.32)	0.20 (0.19)
$\hat{\beta}_{i,R4}$	1.65*** (0.48)	-0.90* (0.43)	1.14*** (0.32)
$\hat{\beta}_{i,R5}$	-0.41 (0.31)	0.30 (0.32)	-0.03 (0.22)
$\hat{\beta}_{i,R6}$	0.04 (0.36)	-0.72* (0.36)	0.18 (0.20)

Chapter 7

Conclusions

At the end of the reading of this thesis, the reader has learned the main concepts related to preference data. The primary analysis methods are presented and the most used measures of distance and correlation in the literature of preference data. When working on ordinal data, information about the distances between one class and another is lost. If the data is analyzed through distance or correlation measures, positional weights differentiate the positions in a ranking based on their relevance. The simulation study demonstrates that introducing positional weights can cause variations for the consensus ranking based on how far the choice of weights deviates from the typical set of weights suggested in the literature. The first and second chapters of this thesis touch on these points categorized as distance-based approaches. Starting from Chapter 3.1, the thesis is focused on a different way to approach ordinal data through a new probabilistic model. Probabilistic models estimate the probability of observing an order of objects. It is reasonable to assume that the order of objects can depend on some characteristics of individuals who furnish the ranking (e.g., judges express preference rankings) or with whom the ranking is associated (e.g., the ranking of tax revenue categories for countries). The results usually derive from the best model research in terms of goodness-of-fit. We presented a new probabilistic model, the Bradley-Terry Regression Trunk (BTRT), based on the combination of the regression trunk approach with the Bradley-Terry model for paired comparisons. The model is well suited when the aim is to partition individuals based on their preferences and characteristics as covariates. In addition, the algorithm finds the best interactions

between covariates. The model shares the same advantage as decision trees: the easy-to-read visualization of results through a small tree called trunk. The pruning procedure is tested through a simulation study: the model performance in finding the best interaction is investigated in three different scenarios, and the results are furnished in terms of Type I and Type II errors. The BTRT model is then applied to a dataset composed of self-reported data by students at the Università di Cagliari. This application presents two different approaches for interaction research: the One-Split-Only approach and the Multiple Splitting approach. The results of the two approaches are compared, and it results that the Multiple Splitting approach performs better when the aim is to find the best model in terms of goodness-of-fit. Starting from Chapter 5.1, we abandon the concept of preference data in the strict sense and apply the BTRT model to financial data transformed into rankings. This analysis aims to estimate the main effects and interactions of socio-economics covariates on the size of tax revenues for a specific country in the world. We build a new country-sectional dataset from well-known databases (OECD, IMF, and World Bank) where the objects are represented by four tax revenue categories associated with 100 countries worldwide. The tax revenues are converted from continuous data to rankings and finally to paired comparisons for each country. Then, to reduce the number of parameters, we apply a covariates selection through the Bradley-Terry-Luce Lasso to select the covariates that have the most significant linear effect on the size of tax revenues. Out of 30 socio-economic covariates, the algorithm selected 17 covariates. The application of the BTRT model to the dataset after covariates selection brings to a regression trunk composed of five splits and six terminal nodes. The Environmental performance index appears as the first splitting covariate. The algorithm selected two first-order interactions and one higher-order interaction effect. The firsts are EPI-health spending and EPI-government gross debt interactions, while the higher-order interaction is EPI-health spending-employment rate. The results section shows the most significant coefficients for each object composing the tax revenues orderings. They contain information about the effect size of each tax revenue category's most critical main effects and interaction effects, which constitutes the main strength of this model: to provide a probabilistic measure of the causal relationship between subject-specific covariates and objects presented in paired comparisons. It makes it possible to estimate the effects of an economic shock on the structure of tax revenues. This type of application can also

represent a point of reflection for countries that want to change the composition of their tax revenues. The final results show that it is possible to obtain a variation in the size of the tax revenues by operating on one's socio-economic characteristics. In this way, it may not be necessary to initiate legislative processes for regulatory changes in taxation, which in many cases take a very long time.

As the number of objects increases, paired comparisons become more challenging to treat. A solution to this issue is furnished in Chapter 6.1, where we present an extension of the BTRT model to analyze ordinal data treated as rankings. This extension is based on the Mallows specification of the BT model. The Mallows extension is based on the concept that the probability of observing a specific ranking is proportional to the product of worth parameters associated with each object ranked. We used the same dataset with financial data, but we directly focused on the relationship between tax revenues and government spending (military, health, education, and others) without considering other socio-economic characteristics. Then, we work on rankings instead of paired comparisons. The final tree is composed of five splits and six terminal nodes. The splitting covariates are represented by the government spending on health, military, and education. The algorithm selected one first-order interaction (military-health) and one higher-order interaction (education-military-health). Converting the estimated values by the models into worth parameters furnishes elements for predictive analysis. Given their government expenditure, it is possible to investigate the most probable size of tax revenues for countries. In addition, this analysis represents a valuable tool for policy-makers who want to change their tax revenues' structure by operating on government spending instead of proposing time-consuming tax reforms.

Future research is addressed to consider cases when nominal subject-specific covariates with more than one category are used as possible split candidates and investigate further model performance and stability concerning (big) datasets presenting a high number of objects, rankings, and covariates. Simultaneously, research efforts will extend the model to cases where missing values (i.e., partial orderings) or order effects are allowed. Finally, an *R* function is currently under development to allow replications and extensions of the BTRT procedure.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). John Wiley & Sons.

Aledo, J. A., Gámez, J. A., & Rosete, A. (2017). Partial evaluation in rank aggregation problems. *Computers & Operations Research*, *78*, 299–304.

Alexander, W. P., & Grimshaw, S. D. (1996). Treed regression. *Journal of Computational and Graphical Statistics*, *5*(2), 156–175. http://www.jstor.org/stable/1390778

Amodio, S., D'Ambrosio, A., & Siciliano, R. (2016). Accurate algorithms for identifying the median ranking when dealing with weak and partial rankings under the Kemeny axiomatic approach. *European Journal of Operational Research*, *249*(2), 667–676.

Anderson, W., Wallace, M. S., & Warner, J. T. (1986). Government spending and taxation: What causes what? *Southern Economic Journal*, 630–639.

Bahl, R., & Wallace, S. (2005). Public financing in developing and transition countries. *Public Budgeting & Finance*, *25*(4s), 83–98.

Berrington de González, A., & Cox, D. R. (2007). Interpretation of interaction: A review. *Annals of Applied Statistics*, *1*(2), 371–385. https://doi.org/10.1214/07-AOAS124

Bird, R. M., Martinez-Vazquez, J., & Torgler, B. (2008). Tax effort in developing countries and high income countries: The impact of corruption, voice and accountability. *Economic analysis and policy*, *38*(1), 55–71.

Bird, R., Martinez-Vasquez, J., & Torgler, B. (2004). Societal institutions and tax effort in developing countries, ssrn elibrary.

Blackley, P. R. (1986). Causality between revenues and expenditures and the size of the federal budget. *Public Finance Quarterly*, *14*(2), 139–156.

Böckenholt, U. (2001). Mixed-effects analyses of rank-ordered data. *Psychometrika*, *66*(1), 45–62.

Bohn, H. (1991). Budget balance through revenue or spending adjustments?: Some historical evidence for the United States. *Journal of monetary economics*, *27*(3), 333–359.

Bondell, H. D., & Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, *65*(1), 169–177.

Bosch, R. (2006). *Characterizations on voting rules and consensus measures* (Doctoral dissertation) [Pagination: 150]. [s.n.]

Boukbech, R., Bousselhamia, A., & Ezzahid, E. (2018). Determinants of tax revenues: Evidence from a sample of lower middle income countries.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. CRC press.

Brun, J.-F., & Diakite, M. (2016). Tax potential and tax effort: An empirical estimation for non-resource tax revenue and VAT's revenue.

Busing, F. M. T. A., Groenen, P. J. K., & Heiser, W. J. (2005). Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika*, *70*(1), 71–98.

Busing, F. M., Heiser, W. J., & Cleaver, G. (2010). Restricted unfolding: Preference analysis with optimal transformations of preferences and attributes. *Food quality and preference*, *21*(1), 82–92.

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. Shepard, A. Romney, & S. Nerlove (Eds.), *Geometrfic representations of individual preferences* (pp. 105–155). New York: Academic Press.

Cassou, S. P. (1997). The link between tax rates and foreign direct investment. *Applied Economics*, *29*(10), 1295–1301.

Castro, G. Á., & Camarillo, D. B. R. (2014). Determinants of tax revenue in OECD countries over the period 2001–2011. *Contaduría y administración*, *59*(3), 35–59.

Chapman, R. G., & Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of marketing research*, *19*(3), 288–301.

Chelliah, R. J. (1971). Trends in taxation in developing countries. *Staff Papers*, *18*(2), 254–331.

Chelliah, R. (1975). Bass. *HJ,. & Kelly, MR*, 1967–1971.

Choisel, S., & Wickelmaier, F. (2007). Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. *The Journal of the Acoustical Society of America*, *121*(1), 388–400.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Lawrence Erbaum Associates Inc., Mahwah NJ.

Conversano, C., Contu, G., & Mola, F. (2019). Online promotion of UNESCO heritage sites in Southern Europe: Website information content and managerial implications. *Electronic Journal of Applied Statistical Analysis*, *12*(1), 108–139. https://doi.org/10.1285/i20705948v12n1p108

Conversano, C., & Dusseldorp, E. (2017). Modeling threshold interaction effects through the logistic classification trunk. *Journal of Classification*, *34*(3), 399–426.

Cook, W. D., & Seiford, L. M. (1982). On the borda-kendall consensus method for priority ranking problems. *Management Science*, *28*(6), 621–637.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, *57*(3), 145–158.

Coombs, C. H. (1964). *A theory of data.* Wiley.

Critchlow, D. E., & Fligner, M. A. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika*, *56*(3), 517. https://www.learntechlib.org/p/144907

Croon, M. A. (1989). Latent class models for the analysis of rankings. In G. De Soete, H. Feger, & K. C. Klauer (Eds.), *New developments in psychological choice modeling* (pp. 99–121). Elsevier.

D'Ambrosio, A., Amodio, S., & Iorio, C. (2015). Two algorithms for finding optimal solutions of the Kemeny rank aggregation problem for full rankings. *Electronic Journal of Applied Statistical Analysis*, *8*(2), 198–213.

D'Ambrosio, A., Amodio, S., & Mazzeo, G. (2019). *Consrank: Compute the median ranking(s) according to the Kemeny's axiomatic approach* [R package version 2.1.0.]. https://CRAN.R-project.org/package=ConsRank

D'Ambrosio, A., & Heiser, W. J. (2016). A recursive partitioning method for the prediction of preference rankings based upon kemeny distances. *Psychometrika*, *81*(3), 774–794.

D'Ambrosio, A., Iorio, C., Staiano, M., & Siciliano, R. (2019). Median constrained bucket order rank aggregation. *Computational Statistics*, *34*(2), 787–802.

D'Ambrosio, A., Mazzeo, G., Iorio, C., & Siciliano, R. (2017). A differential evolution algorithm for finding the median ranking under the kemeny axiomatic approach. *Computers & Operations Research*, *82*, 126–138.

David, H. A. (1969). *The method of paired comparisons* (A. Stuart, Ed.; 2nd ed., Vol. 12). Charles Griffin & Company Limited.

Deza, M. M., & Deza, E. (2009). Encyclopedia of distances. *Encyclopedia of distances* (pp. 1–583). Springer.

Dittrich, R., Francis, B., Hatzinger, R., & Katzenbeisser, W. (2006). Modelling dependency in multivariate paired comparisons: A log-linear approach. *Mathematical Social Sciences*, *52*(2), 197–209.

Dittrich, R., Hatzinger, R., & Katzenbeisser, W. (1998). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *47*(4), 511–525.

Dittrich, R., Katzenbeisser, W., & Reisinger, H. (2000). The analysis of rank ordered preference data based on Bradley-Terry type models. *OR-Spektrum*, *22*(1), 117–134.

Dossou-Gbété, S., Lafon, D., & Sawadogo, A. (2009). Estimation des paramètres du modèle de Mallows-Bradley-Terry par le maximum de vraisemblance. *41èmes Journées de Statistique, SFdS, Bordeaux*.

Dusseldorp, E., Conversano, C., & Van Os, B. J. (2010). Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics*, *19*(3), 514–530.

Dusseldorp, E., & Meulman, J. J. (2004). The regression trunk approach to discover treatment covariate interaction. *Psychometrika*, *69*(3), 355–374.

Emond, E. J., & Mason, D. W. (2002). A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis*, *11*(1), 17–28.

Feltenstein, A., & Cyan, M. R. (2013). A computational general equilibrium approach to sectoral analysis for tax potential: An application to Pakistan. *Journal of Asian Economics*, *27*, 57–70.

Fienberg, S. E., & Larntz, K. (1976). Log linear representation for paired and multiple comparisons models. *Biometrika*, *63*(2), 245–254.

Francis, B., Dittrich, R., Hatzinger, R., & Penn, R. (2002). Analysing partial ranks by using smoothed paired comparison methods: An investigation of value orientation in Europe. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *51*(3), 319–336.

Friedman, M. (1978). The limitations of tax limitation. *Quadrant*, *22*(8), 22–24.

Garcıa-Lapresta, J. L., & Pérez-Román, D. (2010). Consensus measures generated by weighted Kemeny distances on weak orders. *2010 10th International Conference on Intelligent Systems Design and Applications*, 463–468.

Gormley, I. C., & Murphy, T. B. (2008a). Exploring voting blocs within the irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*, *103*(483), 1014–1027.

Gormley, I. C., & Murphy, T. B. (2008b). A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics*, *2*(4), 1452–1477.

Gross, O. A. (1962). Preferential arrangements. *The American Mathematical Monthly*, *69*(1), 4–8. http://www.jstor.org/stable/2312725

Gruber, J. (2005). *Public finance and public policy*. Macmillan.

Gupta, A. S. (2007a). *Determinants of Tax Revenue Efforts in Developing Countries* (IMF Working Papers No. 2007/184). International Monetary Fund. https://ideas.repec.org/p/imf/imfwpa/2007-184.html

Gupta, A. S. (2007b). Determinants of tax revenue efforts in developing countries. *IMF Working Papers*, *2007*(184).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Verlag.

Hatzinger, R., & Dittrich, R. (2012). Prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software*, *48*(10), 1–31.

Heiser, W. (2004). Geometric representation of association between categories. *Psychometrika*, *69*, 513–545. https://doi.org/10.1007/BF02289854

Heiser, W. J., & D'Ambrosio, A. (2013). Clustering and prediction of rankings within a Kemeny distance framework. In B. Lausen, D. Van den Poel, & A. Ultsch (Eds.), *Algorithms from and for nature and life* (pp. 19–31). Springer International Publishing.

Heiser, W. J., & De Leeuw, J. (1981). Multidimensional mapping of preference data. *Mathématiques et Sciences humaines*, *73*, 39–96.

Hondroyiannis, G., & Papapetrou, E. (1996). An examination of the causal relationship between government spending and revenue: A cointegration analysis. *Public Choice*, *89*(3/4), 363–374. http://www.jstor.org/stable/30024171

Hooley, G. (1993). Multidimensional scaling of consumer perceptions and preferences. *European journal of marketing*, *14*(7), 436–448.

Irurozki, E., Calvo, B., & Lozano, J. A. (2016). PerMallows: An R package for mallows and generalized mallows models. *Journal of Statistical Software*, *71*(12), 1–30. https://doi.org/10.18637/jss.v071.i12

Jain, P. C. (1989). *Economics of public finance*. Atlantic Publishers & Distributors.

Jones, J. D., & Joulfaian, D. (1991). Federal govemment expenditures and revenues in the early years of the American republic: Evidence from 1792 to 1860. *Journal of Macroeconomics*, *13*(1), 133–155.

Kaldor, N. (1963). Taxation for economic development. *The Journal of Modern African Studies*, *1*(1), 7–23.

Karagöz, K. (2013). Determinants of tax revenue: Does sectorial composition matter? *Journal of Finance, Accounting & Management*, *4*(2).

Kemeny, J. G. (1959). Mathematics without numbers. *Daedalus*, *88*(4), 577–591.

Kemeny, J. G., & Snell, L. (1962). Preference ranking: An axiomatic approach. *Mathematical models in the social sciences*, 9–23.

Kendall, M. G., & Smith, B. B. (1940). On the method of paired comparisons. *Biometrika*, *31*(3/4), 324–345.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93.

Kendall, M. G., & Smith, B. B. (1939). The problem of m rankings. *The annals of mathematical statistics*, *10*(3), 275–287.

Lee, P. H., & Yu, P. L. (2010). Distance-based tree models for ranking data. *Computational Statistics & Data Analysis*, *54*(6), 1672–1682.

Lotz, J. R., & Morss, E. R. (1967). Measuring "tax effort" in developing countries. *Staff Papers*, *14*(3), 478–499.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Dover Publications Inc.

Mallows, C. L. (1957). Non-null ranking models. I. *Biometrika*, *44*(1/2), 114–130.

Manage, N., & Marlow, M. L. (1986). The causal relation between federal expenditures and receipts. *Southern Economic Journal*, 617–629.

Marden, J. I. (1996). *Analyzing and modeling rank data*. Chapman & Hall.

Martın-Mayoral, F., & Uribe, C. A. (2010). Economic and institutional determinants of tax effort in Latin America. *Investigación económica*, *69*(273), 85–113.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models, second edition*. Chapman & Hall. http://books.google.com/books?id=h9kFH2%5C_FfBkC

Meila, M., Phadnis, K., Patterson, A., & Bilmes, J. (2007). Consensus ranking under the exponential model. *Proceedings of the Twenty-Third Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*, 285–294.

Meltzer, A. H., & Richard, S. F. (1981). A rational theory of the size of government. *Journal of political Economy*, *89*(5), 914–927.

Meulman, J. J., Van der Kooij, A. J., & Heiser, W. J. (2004). Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 49–72). Sage: London.

Miller, S. M., & Russek, F. S. (1990). Co-integration and error-correction models: The temporal causality between government taxes and spending. *Southern Economic Journal*, 221–229.

Murphy, T. B., & Martin, D. (2003). Mixtures of distance-based models for ranking data. *Computational statistics & data analysis*, *41*(3), 645–655.

Oelker, M.-R., Pößnecker, W., & Tutz, G. (2015). Selection and fusion of categorical predictors with L 0-type penalties. *Statistical Modelling*, *15*(5), 389–410.

Peacock, A. T., & Wiseman, J. (1979). Approaches to the analysis of government expenditure growth. *Public Finance Quarterly*, *7*(1), 3–23.

Pessino, C., & Fenochietto, R. (2010). Determining countries' tax effort. *Hacienda Pública Española/Revista de Economıa Pública*, 65–87.

Piancastelli, M. (2001). Measuring the tax effort of developed and developing countries: Cross country panel data analysis-1985/95.

Plaia, A., Buscemi, S., & Sciandra, M. (2019). A new position weight correlation coefficient for consensus ranking process without ties [e236 sta4.236]. *Stat*, *8*(1), e236. https://doi.org/https://doi.org/10.1002/sta4.236

Plaia, A., & Sciandra, M. (2019). Weighted distance-based trees for ranking data. *Advances in data analysis and classification*, *13*, 427–444.

Quinlan, J. R. (1992). Learning with continuous classes, 343–348.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ram, R. et al. (1988). A multicountry perspective on causality between government revenue and government expenditure. *Public Finance= Finances publiques*, *43*(2), 261–270.

Rodríguez Montequín, V., Villanueva Balsera, J. M., Díaz Piloñeta, M., & Álvarez Pérez, C. (2020). A Bradley-Terry model-based approach to prioritize the balance scorecard driving factors: The case study of a financial software factory. *Mathematics*, *8*(2). https://doi.org/10.3390/math8020276

Schauberger, G. (2015). *Regularization methods for item response and paired comparison models*. Cuvillier Verlag.

Sinclair, C. (1982). GLIM for preference. In R. Gilchrist (Ed.), *Glim 82: Proceedings of the international conference on generalised linear models* (pp. 164–178).

Skrondal, A., & Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings. *Psychometrika*, *68*(2), 267–287.

Strobl, C., Wickelmaier, F., & Zeileis, A. (2011). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics*, *36*(2), 135–153.

Tait, A. A., Grätz, W. L., & Eichengreen, B. J. (1979). International comparisons of taxation for selected developing countries, 1972-76. *Staff Papers*, *26*(1), 123–156.

Tanzi, V. (1989). The impact of macroeconomic policies on the level of taxation and the fiscal balance in developing countries. *Staff Papers*, *36*(3), 633–656.

Teera, J. M., & Hudson, J. (2004). Tax performance: A comparative study. *Journal of international development*, *16*(6), 785–802.

Thompson, G. L. (1993). Generalized Permutation Polytopes and Exploratory Graphical Methods for Ranked Data. *The Annals of Statistics*, *21*(3), 1401–1430. https://doi.org/10.1214/aos/1176349265

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, *34*(4), 273.

Turner, H., & Firth, D. (2012). Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, *48*(9).

Turner, H., van Etten, J., Firth, D., & Kosmidis, I. (2020). Modelling rankings in R: The PlackettLuce package. *Computational Statistics*, *35*. https://doi.org/10.1007/s00180-020-00959-3

Van Deun, K., Heiser, W. J., & Delbeke, L. (2007). Multidimensional unfolding by nonmetric multidimensional scaling of Spearman distances in the extended permutation polytope. *Multivariate Behavioral Research*, *42*(1), 103–132.

Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi Di Rattalma, A., & Arjas, E. (2018). Probabilistic preference learning with the mallows rank model. *Journal of Machine Learning Research*, *18*(158), 1–49.

Von Furstenberg, G. M., Green, R. J., & Jeong, J.-H. (1986). Tax and spend, or spend and tax? *The review of Economics and Statistics*, 179–188.

Wickelmaier, F., & Schmid, C. (2004). A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior research methods, instruments, & computers*, *36*(1), 29–40.

Wiedermann, W., Frick, U., & Edgar, M. (2021). Detecting heterogeneity of intervention effects in comparative judgments. *Prevention Science*. https://doi.org/https://doi.org/10.1007/s11121-021-01212-z

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, 488–508. https://doi.org/https://doi.org/10.1111/j.1467-9574.2007.00371

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*. https://doi.org/https://doi.org/10.1198/106186008X31933

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.