UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXVIII

Partial Order Theory for Synthetic Indicators

Direttore della scuola: Ch.ma Prof.ssa Monica Chiogna

Supervisor: Prof.ssa Giovanna Boccuzzo

Co-Supervisor: Prof. Marco Fattore

Dottorando: Giulio Caperna

January 31th, 2016

Acknowledgements

I miei ringraziamenti istituzionali vanno alla Fondazione CARIPARO, per aver concesso i fondi a questa ricerca e al dipartimento di Scienze Statistiche di Padova per avermi accolto: in particolar modo ringrazio la Professoressa Giovanna Boccuzzo, l'irrinunciabile Patrizia Piacentini, la direttora del dottorato Monica Chiogna e i miei amati colleghi (amati colleghi means beloved colleagues). Inoltre ringrazio il Prof. Marco Fattore per i consigli in questi due anni e I Professori Brüggemann, De Baets, Rademaker e Maggino. Infine ringrazio il Contact center dell'ISTAT per la grande offerta di informazione che mette a disposizione gratuitamente. I miei ringraziamenti personali avranno solo copia cartacea.

Abstract

Given a big set of data with several variables, the aim is the evaluation of each unit with a method that produces a synthetic measure describing a complex and non-observable concept; this goal is achieved respecting the characteristics of the variables, specially the measurement scale. The information gathered with partially ordered sets (poset) reflects this aim, because posets depends only on the order relations among the observations, and allows to handle ordinal and dichotomous variables fairly. In this setting, the vector of variables observed on a unit is handled as a unique entity called profile and not as a group of different variables that need to be aggregated.

Starting from recent developments in poset theory, this thesis is organized in three parts. The first proposes to obtain a unique indicator combining the values given by the severity measures for evaluation, derived from the fuzzy identification method. The second contribution is the HOGS (Height Of Groups by Sampling) procedure, which is aimed to estimate the average rank of groups of units of a big population. HOGS is a step forward the statistical estimation of the average rank of a profile; furthermore it allows the estimation of the effect of external variables on the synthetic measure.
The last results are two new R functions: the first computes the approximated average rank for large data sets overcoming the usual sample sizes considered by the software usually used until now, the second implements the information given by the frequency of profiles in the computation of approximated average rank, making its use more profitable for social sciences.

Abstract

Data una grande popolazione osservata su diverse variabili, ci si pone l'obiettivo di valutare le singole unità con un metodo che sia in grado di produrre una informazione sintetica per la descrizione di un concetto complesso e non osservabile; in questa tesi si vuole raggiungere questo scopo rispettando le caratteristiche dei dati, specialmente la scala di misura di questi.

Gli insiemi parzialmente ordinati (poset) si adattano a questo scopo; questo tipo di insiemi sono costruiti unicamente sulle relazioni d'ordine tra le osservazioni e quindi consentoto di trattare le variabili ordinali e dicotomiche in modo adeguato alle loro caratteristiche. Nella letteratura dei poset, il vettore di variabili osservate su una unità è chiamato profilo e trattato come un oggetto unico senza procedure di aggregazione.

Questa tesi si connette ai più recenti sviluppi nella teoria dei poset ed è organizzata in tre parti principali. La prima propone una sintesi dell'informazione fornita dalle misure di severity, derivate dal metodo di fuzzy identification. Il secondo e principale contributo è la procedura HOGS (Height OF Groups by Sampling), che ha lo scopo di stimare l'average rank di gruppi di unità da grandi popolazioni. HOGS permette di avvicinarsi alla stima statistica dell'average rrank dei singoli profili ed inoltre fornisce un metodo per studiare l'effetto di variabili esterne sulla misura sintetica.
L'ultima parte contiene le funzioni che sono state sviluppate in R: la prima calcola l'average rank approssimato per grandi moli di dati, la seconda implementa l'informazione data dalle frequenze dei signoli profili nella popolazione osservata, rendendo questo metodo più spendibile nelle scienze sociali.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The definition of tools for evaluation and comparison of different alternatives in a multi-criteria framework is a warmly debated topic.
In this context, the computation of a synthetic measure is one of the most interesting solutions, because the output is a very simple and understandable measure, also for an inexperienced audience [Nardo et al., 2005]. It is often easier to interpret a single coherent synthetic measure than to identify the effect of many elementary indicators [Saltelli, 2007].

The simplification of complex and unobservable concepts is the main reason why this kind of methodology has gained a lot of interest in applied sciences, where it led to different approaches of formalization. For instance, often in psychometry research is aimed to model abstract constructs (intelligence, empathy, ...), for this reason the construction of indicators is based on strongly correlated elementary variables that represent the complex concept (*reflective* approach); on the other hand, in sociology and economy composite indicators are commonly used to measure unobservable phenomena (quality of life, sustainability, ...), which are defined aggregating several variables that should be less correlated, with the aim of ¨composing¨ the complex concept (*formative* approach); generally speaking, synthetic measures are useful in the framework of decision making, where the evaluation is the first step of every approach.

Despite the use of synthetic indicators is widespread, there are many methodological issues that should be managed.
This thesis focuses on the issues arising from the application of synthetic indicators to individuals (elementary units). Especially in social science, many concepts are observable through subjective variables (satisfaction, opinion, evaluation, etc), often measured with ordinal or dichotomous scale. In such a framework, the combination of the information given by the data is not straightforward.

The thesis aims to deal with some of these issues it develops a methodology and a related mathematical formalization to define synthetic measures starting from ordinal and dichotomous variables.
In this work we want to: **avoid** the use of scaling procedures on the original variables; **accomplish** the specific intent to develop a methodology based only on the data structure; and **obtain** a final measure related to each individual. Our final aim is to obtain this results with a model that produce easy readable results.

Hence, we focus on the theory of Partially Ordered Sets (poset), because it is a mathematical approach to define space based only on the reciprocal order of observations. This approach is introduced in chapter 2.

Theory of poset [Davey and Priestley, 2002] contains several tools for the formalization and solution of the problem under analysis. In this setting, the vector of variables observed on a unit is handled as a unique entity called **profile** and not as a group of different variables that need to be aggregated.

The information on poset's structure is carried out by the use of *linear extensions*, that can be defined as the atomic level of partial order information, since the set of all linear extensions of a poset corresponds to all the possible orders that the poset contains. One of the solutions derived from poset theory is the computation of the rank of a unit inside a partially ordered set. The theoretical rank is called average rank (**AR** or medium height) and provides a simple and easy-to-read synthetic measure representing all the information on order relations between the units of the set.

The computation of **AR** is almost impossible in real case studies, because of the computational time that grows too fast with the number of units. Usually such a growth is faster than exponential but slower than factorial. Because of

the practical limitations, two main approaches have been proposed in order to handle poset information:

Sampling of Linear Extensions, consists in the analysis of the poset by sampling of its linear extensions, for the estimation of average rank Lerche and Sørensen [2003] or for the Fuzzy Identification method for evaluation [Fattore, 2015]. The latter is conceived for social sciences, particularly for the measurement of deprivation.

Approximation of AR, with the proposals of: Extended Local Partial Order Model "LPOMext" [Brüggemann and Carlsen, 2011], and Mutual Probability approximation [De Loof et al., 2011].

In Chapter 3, starting from the results of the method proposed by Fattore [2015] and Fattore and Arcagni [2014], we developed and applied a proposal for a synthetic measure, completely based on posets. It uses the concepts of *severity* in order to derive a unique indicator to describe the complex concept, is has been used in the thesis to study the concept of life satisfaction.

In the approximation approach, Brüggemann and colleagues started to develop the information derived form posets in a framework of evaluation, taking advantage of the concepts of comparability and incomparability to describe the poset [Brüggemann and Patil, 2011]. These results are used in order to gather information for: the approximation of average rank with the *LPOM* methods [Brüggemann and Carlsen, 2011; Brüggemann et al., 2004], the mutual rank probability approximation [Brüggemann et al., 2003] and the description of the data. In the meanwhile, the concept of stochastic ordering [Lehmann and Romano, 2011] has been used to define a multi-criterion ranking approach [Patil and Taillie, 2004]. From a Belgium research group come some advances in the formalization of posets, with the PhD thesis of Loof [2009], which develops some strategies based on poset of sub-posets ordered by inclusion and called *lattice of ideals*, and the approximation procedure based on mutual rank probabilities [De Loof et al., 2011].

Moving from these results, the thesis proposes a new approach for the approximation of average rank of individuals and the evaluation of the effect of external

variables on the average rank, we call this method HOGS (Height Of Groups by Sampling) and is presented in Chapter 4.

The last chapter of the thesis (Chapter 5) describes the computational improvements that we developed for the computation of the approximated average rank. In this chapter we present two achievements. The first is the new algorithm for LPOMext, developed in R; it allows to compute the approximated average rank for tens of thousand of units. The second achievement is the implementation of frequencies in the LPOMext formula; this improvement changes the original concept of average rank, taking into account the frequency distribution of profiles.

1.2 Main contributions of the thesis

1.2.1 Work hypothesis

Given a set of data with several variables, the aim is the evaluation of the observations with a method that produces a synthetic information about a concept measured through some variables. This aim has to be achieved respecting the characteristics of the data, first of all the measurement scale.
The synthesis of information gathered from linear extensions reflects this scope, because linear extensions depend only on data and their order relations.
We assume that: every profile, has a corresponding value (or a set of values) on the range of a latent variable, that is not observable and consequently is measured by the synthetic indicator (see Figure 1.1 for an intuitive representation). This hypothesis is the starting point of this thesis, and an efficient computation of a synthetic measure is the cornerstone.

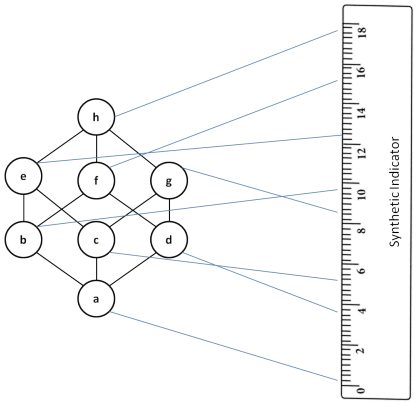


Figure 1.1: Profiles of a poset ordered according to the synthetic indicator

1.2.2 Description of the original contributions

Starting from poset theory, the thesis focuses on both the approximation and the sampling approach, proposing a solution for some of the issues related to the une of posets in social science (a scheme can be found in Figure 1.2):

Approximation Approach

A - Large Datasets. The software for computation of Approximated AR is not meant to handle more than some hundreds of observations, and no results have been proposed to study the effect of explanatory variables on the poset.

1. A new method for the approximation of the AR in large data sets is proposed. We call it *HOGS*: Height Of Groups by Sampling. It is based on a sampling procedure of small groups of observations, and the estimation of AR for sub-groups identified by external/explanatory variables (Chapter 4);
2. The method LPOMext is implemented in order to account for frequencies of profiles, allowing a definition of AR based on observed frequencies (Chapter 5);
3. The HOGS method is applied to the evaluation of life satisfaction, a typical unobservable multidimensional concept in social sciences.

B - Reduction of weak orders. In the approximation approach it is common to observe weak orders with many profiles listed in the same group:

1. The use of frequencies (implemented in the computational part of the thesis) impose a *weighting effect* on the poset, this effect depends on the frequency distribution of the elementary variables.

Sampling Approach

The sampling approach is considered in relation to the method of Fuzzy Identification [Fattore, 2015]. For this approach we present the following proposals:

C - Guidelines. We studied some of the properties connected to the shape and length of the threshold for the method of Fuzzy Identification (Chapter 3). Results come from simulation studies on several posets, different in dimension and structure;

D - Synthetic Measure. A Synthetic Measure, derived by the severity measures proposed by Fattore [2015] and Fattore and Arcagni [2014] is presented. This is an intuitive method to build a synthetic measure (Chapter 3).

1.2.3 Computational developments

E - High performing software. We developed functions pursuing the ability to handle large sets of profiles and managing tens of thousands of observations instead of hundreds. We developed novel functions in R to implement all of the previous proposals:

1. A new function to compute the approximated average rank [Brüggemann and Carlsen, 2011] is presented in Chapter 5;
2. In Chapter 4, the methodological proposal described in A.1 is realized. The function allows to estimate the mean of ranks among subgroups;
3. A new function to implement the frequencies of profiles in the approximated AR (as described in point A.2) is described in Chapter 5.

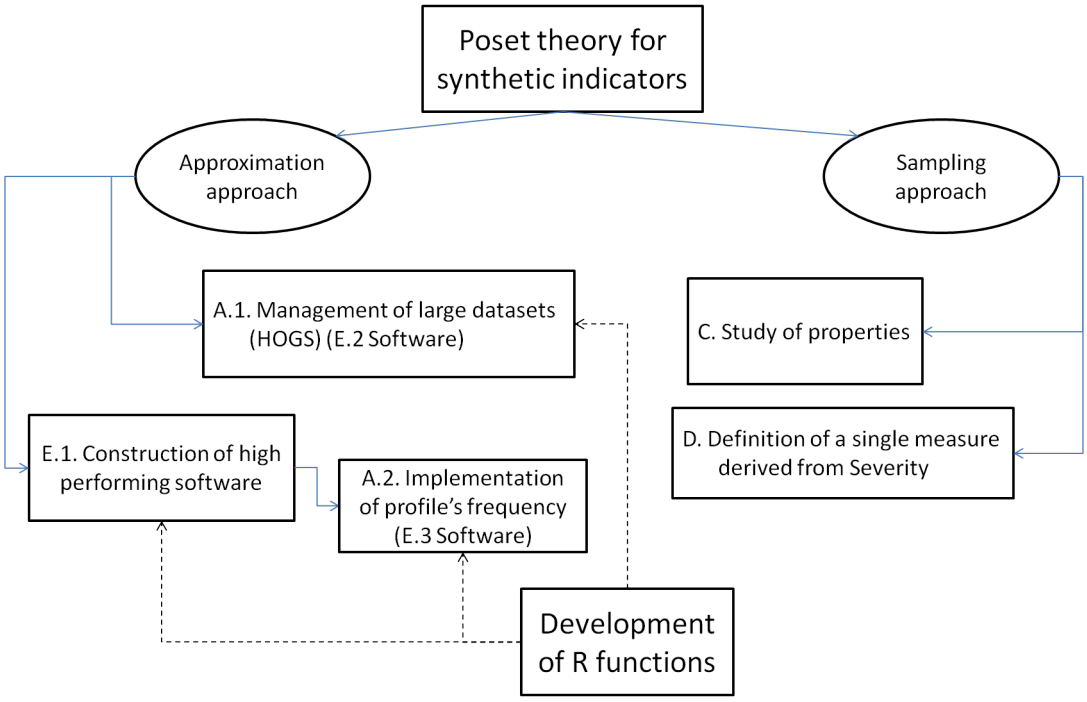


Figure 1.2: Main contributions of the thesis

Chapter 2

Introduction to poset theory

The simplification of complex and unobservable concepts is the main reason why this kind of methodology has gained a vivid interest in applied sciences.

The fundamental definitions and concepts used in this thesis for a brand new approach come from the theory of Partially Ordered Sets (poset), which is a mathematical approach to define space according to the order relations between observations and not to their position in an Euclidean multidimensional space. Theory of poset [Davey and Priestley, 2002] contains several tools for the formalization and solution of the problem under analysis.

This chapter contains four sections: the first is an introduction to the mathematical theory of partial order, together with the limits it has respect to the aim of the thesis. The second part contains the motivating idea of the thesis. The third and fourth sections are devoted to the description of the literature background, to the approximation methods and the sampling approaches respectively [Brüggemann and Carlsen, 2011], [De Loof et al., 2011] and [Fattore, 2015].

2.1 Why Partial Order Theory?

Poset theory is a mathematical formalization of discrete space. Poset theory has been used in the field of data analysis to compare elements of small population;

it is adapt for both quantitative and qualitative data. The best use of poset for quantitative data comes out in the case of information which cannot be interpreted in the classic euclidean sense. As instance, if the effect of a variable on a response does not follow any function, or the description of this function is impossible. In chemometrics, for example, the aim is the extraction of information from chemical systems by mean of the analysis of data; often in this scientific field it is not possible to assess the effect of chemicals respecting quantitative scales, and sometimes is not even possible to compare two chemical mixture. In social statistics, one can find many similar cases, with particular quantitative data like years of schooling, that are completely quantitative in the sense of time spent, but not in the sense of ¨amount of education¨. Moreover, this approach is able to describe those cases where there is no quantity but just order, like ordinal and dichotomous data. This premise implies the usefulness of such a theory in the field of social statistics, where the presence of qualitative data is massive.
Poset theory allows to use the order information contained in data, also if information on distance is absent, or it is reasonably better to avoid using it.
All the definitions about poset theory that are presented in this section, can be find in the text: ¨Introduction to Lattices and Order¨ [Davey and Priestley, 2002]

2.1.1 What is an Order

The elements of a family can be ordered with respect to some criteria, for instance the order: ¨father $\geq$ mother $\geq$ older sister $\geq$ younger brother¨ could be the order in a common family, if we consider the age as order criterion. The order is the relation between the elements of the group (*set*) that respects some properties.
Let P be a set, an **order** on P is a relation ($\leq$) between two elements of P such that, for all $x, y, z \in P$ the following properties hold:

reflexivity $x \leq x$,

antisymmetry $x \leq y$ and $y \leq x$ imply $x = y$

transitivity $x \leq y$ and $y \leq z$ imply $x \leq z$.

These properties are all fundamental to define the order as it is used intuitively in everyday life. Reflexivity, for example, is necessary because it is impossible to

compare a value with itself if this property does not hold. The relation ¨<¨ is not reflexive.

A set equipped with such an order relation is said to be **ordered**. Examples of Ordered sets are everywhere, every ordinal variable determines an order and any comparison too.
A hierarchic organization is usually a *weak* order, because some of the elements are equal. The weak order is different from the *complete* order, where every element is different and ordered. Following the family example, the criterion ¨age¨ imposed a complete order with a relation equal to: ¨older or equal than¨. On the contrary, if the relation was ¨being responsible for¨, the order would then be like this: {father, mother} $\geq$ older sister $\geq$ younger brother. This relation is different because both the parents are responsible for all the children of the family and each other, so they are comparable and equal thanks to the property of *antisymmetry*.

2.1.2 The Partial Order

¨Who you like the most? Grandma or Grandpa?¨, sometimes it is possible that two elements are neither equal nor ordered, it often happens when more than one criteria are considered simultaneously. The relation can be defined as **partial order** if there exist **incomparable** elements in the set:

incomparability $x \| y \Leftrightarrow x \not\leq y$ and $y \not\leq x, \quad x, y \in P$

This case happens in presence of multiple attributes that are conflicting. Consider the evaluation of pollution level in two different agricultural fields, where one has clean water and high level of lead in the soil, and the second never received pesticides but its water is full of industrial waste. Without a priority scale on the pollutants and a correct measure of contamination severity, it is not possible to order the fields and evaluate which is better. Another example can be observed in Table 2.1, where an individual answers two questions (called q_1 and q_2) about the quality of three objects. In this example it is impossible to define which object is the best, because: $q_2(x) \geq q_2(y)$ but $q_1(y) \geq q_1(x)$, this conflict is called

	q_1	q_2
x	low	high
y	medium	medium
z	medium	low

Table 2.1: Incomparability example

incomparability, and implies the absence of an order between the elements x, y, the same situation happens between x and z but not between y and z because $y \geq z$ respect to every attribute.

In the usual notation two incomparable elements a, b are represented by $a \| b$; in the example of Table 2.1 the following relations can be assess: $x \| y$, $y \geq z$, $x \| z$. The mathematical formalization of Partially Ordered Sets, commonly called Poset, allows to describe some type of data that are commonly used in social and applied science. Every system of ordinal variables can be correctly handled with this approach, that takes care of every order information and avoids the use of euclidean space an quantitative concepts.

2.1.2.1 Covering Relation

To understand the construction of poset representations, it is important to use the concept of *coverage*. One element covers another if there are not other elements between them, in mathematical language:
Given x, y, z in the ordered (or partially ordered) set P,

$$x \text{ is } \textit{covered} \text{ by } y \ (x \lhd y) \text{ if } x < y \text{ and } x \leq z < y \Rightarrow x = z$$

If something is between two elements in a covering relation, it has to be one of these elements. Moreover, if P is finite,

$$x < y \Leftrightarrow \text{a sequence like } x = x_0 \lhd x_1 \lhd \ldots \lhd x_n = y \text{ always exists.}$$

This formalization assesses that the order relation determines and is determined by a list of covering relations.

An object can cover and can be covered by many others.

	q_1	q_2
x	low	low
y	high	medium
z	medium	high

$\Rightarrow \qquad x \lhd y,\ y||z,\ x \lhd z$

2.1.2.2 Hasse Diagram

In the last decades of the 19th century, mathematicians started to represent partially ordered data with an hand-writing technique; they were used to represent every profile with a *node* and every covering relation with an *edge*. Some more rules are useful in order to make the result easy to read in every case, but *direction*, *nodes* and *edges* are sufficient to draw this graph. This representation is called *Hasse Diagram*, from the name of the German mathematician *Helmut Hasse*. Hasse did not invent the representation but made an extensive use of it (Birkhoff [1948]), allowing it to spread.

	q_1	q_2
a	2	0
b	0	2
c	3	3
d	4	5
e	7	6
f	5	7

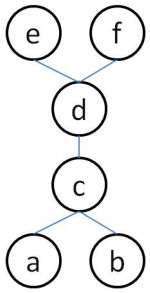


Figure 2.1: Example of Hasse diagram

This representation is an oriented graph, it allows an observable description of the set. Thanks to the properties of covering relations, it describes every relation of order between the nodes drawing only the covering edges.

2.1.2.3 Down-set and Up-set

For every ordered (or partially ordered) set, there are two important families of sets associated. They are central in the development of methodologies based on posets. If P is an ordered (or partially ordered) set and $Q \subseteq P$:

- Q is a **Down-set** if: $x \in Q,\ y \in P$ and $y \leq x$ implies $y \in Q$
- Q is an **Up-set** if: $x \in Q,\ y \in P$ and $y \geq x$ implies $y \in Q$

Consequently, the down-set(up-set) of an element $x \in P$, is the set of all the elements of P that are lower(higher) than the element x itself.

2.1.2.4 Chains and Antichains

There are two extreme cases in the field of partially ordered sets:

Complete Comparability

It occurs when every element of the set is comparable to all the others, and there are not incomparable pairs. Sets like this are usually called **chains** or **linear orders**, because they form a complete order and the shape of their Hasse diagram is clearly linear (Figure 2.2).
Formally a set P is a chain if:

$$\forall x, y \in P, \text{ either } x \leq y \text{ or } y \leq x.$$

Complete Incomparability

It is the extreme case, where every element is incomparable to all the others, no comparison leads to an order. In such a situation, the Hasse diagram looks like a horizontal line of isolated nodes (Figure 2.3) and is called **antichain**.
A set P is defined antichain if:

$$\forall x, y \in P, \quad x \leq y \iff x = y.$$

2.1.2.5 Order-Maps

In the analysis of complex structures such as posets it is fundamental to define a criterion to recognize when two ordered sets are the same in an ordinal sense. This similarity is called *order-isomorphism*. Two sets P, Q are **order-isomorphic**

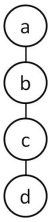


Figure 2.2: Chain

Figure 2.3: Antichain

$(P \cong Q)$, if there exists a map $\varphi : P \to Q$ such that

$$x \leq y \text{ in } P \Leftrightarrow \varphi(x) \leq \varphi(y) \text{ in } Q$$

φ is called **order-isomorphism** and has the role to mirror the order structure of the two sets.
To be bijective is a necessary but not sufficient property of order-isomorphism, and hence is possible to notice that:

$$\begin{aligned} \varphi(x) = \varphi(y) &\iff \varphi(x) \leq \varphi(y) \ \& \ \varphi(x) \geq \varphi(y) \\ &\iff x \leq y \& x \geq y \\ &\iff x = y \end{aligned}$$

If a map is not bijective and respects the order only in a single direction, it can not be defined as order-isomorphism between the sets. Formally:

$$\forall x, y \in P, \text{if } x \leq y \Rightarrow \varphi(x) \leq \varphi(y) \text{ in } Q,$$

such a map is called **order-preserving** map.
In this case the relation is unidirectional, and the order of the set P implies the order of set Q. This concept will be extremely useful in the following.

2.1.3 Linear Extensions

If an order-preserving map relates a partial order to a linear order on the same set of objects, the linear order is defined **linear extension**. This order preserving map is the most important in this work's framework.
Every linear extension $\omega_i(P)$ can be interpreted as the original poset, enriched by much information on comparability of the elements. Naturally, a poset can have more than one linear extension, the number depends on the dimension and structure of incomparable elements in the set. In figure (2.5), an example of poset is drawn next to one of its possible linear extensions. The set of all linear extension of a poset P is called $\Omega(P)$.

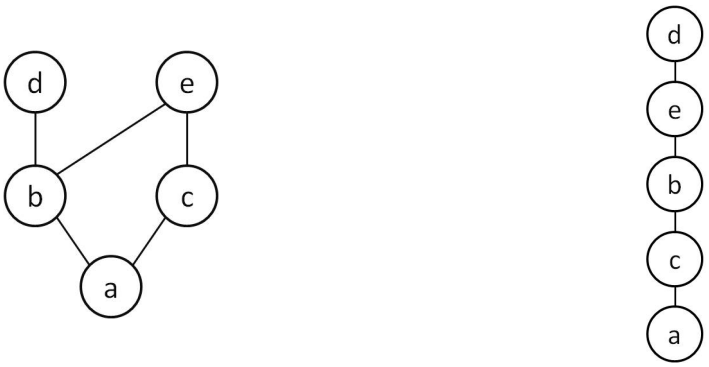


Figure 2.4: A poset and one of its linear extensions

2.1.3.1 Information from Linear Extensions

The relevance of linear extensions is well established by a fundamental result from Schröder (Schröder [2012]), where it is stated that two different finite posets have different sets of linear extensions and that every poset is the intersection of the set of its linear extensions. In other words, a poset coincides with the comparabilities that are common to all its linear extensions.
Given this result, any single linear extension can be considered as the atomic level of order information of a poset, like a single observation in a population. Hence:

- every linear extension describes one of the possible orders of objects,

- the set of all linear extensions of a poset $\Omega(P)$ identifies uniquely the partial order structure,
- the number of linear extensions $|\Omega(P)|$ suggests how complex is the poset and how many pairs of objects are incomparable.

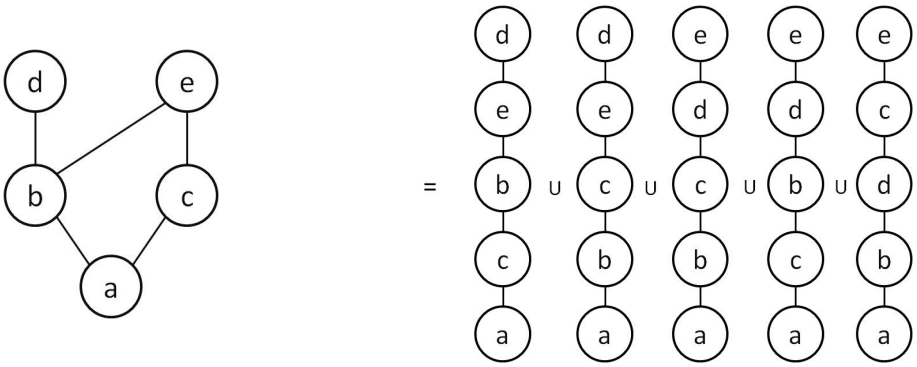


Figure 2.5: A poset and its linear extensions

2.1.4 The average rank of a profile

The rank of an object x on a single linear extension ω_i is called **height** $h_i(x)$ (or simply *rank*); knowing the height of objects among the entire set $\Omega(P)$ (Figure 2.6) allows to define their **Average rank** (or Medium Height): a measure which summarizes all the information about the object's position in the set.

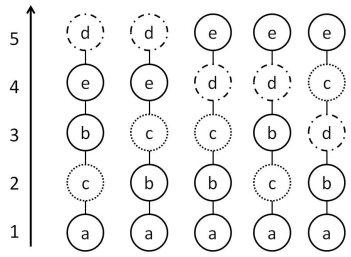


Figure 2.6: Height of elements among all linear extensions

In Table 2.2 the computation of the average rank for the poset of figure (2.5) is reported. Given the structure of the partially ordered set, the element e is

frequently in the top position or, in the worst case, in the second one. The difference between d and e depends on the relation they have with respect to c.

$h(\cdot)$	ω_1	ω_1	ω_1	ω_1	ω_1	$\overline{h}(\cdot)$
e	4	4	5	5	5	4.6
d	5	5	4	4	3	4.2
c	2	3	3	2	4	2.8
b	3	2	2	3	2	2.4
a	1	1	1	1	1	1

Table 2.2: Computation of the average rank

The knowledge of the position of a unit with respect to all the others is the main aim of this work. With the computation of the **Average Rank** an order is obtained, this order could be *Weak* or *Complete* if, respectively, there are some elements with the same value of $\overline{h}(\cdot)$ or every element has a different value with respect to all the others. Nevertheless, the computation of the average rank makes every element comparable.

2.1.5 Computational Issue

The observation of the entire set of linear extensions, is an hard task. In their work, Brightwell and Winkler [1991] prove that the problem of determining the height of an element x of a given poset is #P-complete (pronounced *Sharp p complete*). This assertion refers to the computational complexity of the problem. With a trivial expression we can affirm that the number of linear extensions and, consequently, the procedure of information gathering from them have a computational time that cannot be evaluated in a deterministic way. The best results for the approximation of the number of linear extensions started from the results of Dyer et al. ([Dyer et al., 1991]), but extremely satisfying results can be found also in the results of Karel De Loof's PhD thesis ([Loof, 2009]) .

It is not possible to forecast the time needed to compute the number of linear extensions in a deterministic way, because the number of linear extensions is not directly dependent on the number of elements in the poset; it depends on

the structure of the comparable and incomparable elements. Just to clarify the magnitude of variability; let assume to observe two posets:

I, that is constituted by nineteen, comparable elements and one element incomparable with all the others (*isolated* element);

II, in which the elements are ten, but these are all incomparable;

The number of linear extensions for the first poset is $|\Omega(\mathbf{I})| = 20$, because the isolated element can occupy every position in the chain with length 19. In the second case, every element can take every position, leading to a number of linear extension equal to

$$|\Omega(\mathbf{II})| = 10! = 479001600.$$

In the case of *real* posets, made by more than 10 elements, this number can be extremely high, leading to a clear truth:

¨The set of linear extensions is too big!¨

In order to handle with this lack of information, researchers are following two paths:

Approximation of the average rank: an algorithm is used to find an approximation of the average rank without observing any linear extension [Brüggemann and Carlsen, 2011];

Sampling of linear extensions: only a sub-sample of the linear extensions is observed [Fattore, 2015].

The insights and applications of the two approaches are really different. In the following: first section focuses on the approximation approaches, showing the most used approach and a new proposal that could be an improvement for big posets. The second section is devoted to the sampling approach used in recent works for the evaluation of deprivation.

2.2 Approximation of the average rank

The average rank describes the position of an element respect to the bottom and the top of the poset $(\bot, \top)$. This type of information is what we are looking for respect to our hypothesis (1.2.1).

The computation of the average rank is not influenced by external information such as the importance of the constituting variables or the frequency of profiles; for this reason it is often considered far from the framework of decision making. The relation of this approach with the field of MultiCriteria Decision Aid has been analyzed by [Bruggemann and Carlsen, 2012].

2.2.1 LPOM - Approximation

The explanation of the following method can be found in Brüggemann and Carlsen [2011], while, one of the most extended presentation of posets and their applications can be found in Brüggemann and Patil [2011]. This method is the second generation of Local Partial Order Model, for this reason it is commonly known as *Extended LPOM*.

The idea of average rank comes from the seek of a linear or weak order from poset data (check 2.1.1 for a brief introduction of these concepts), that are attractive for sake of comparison and evaluation. The method that is presented in this part is called *Local Partial Order Model* and is based on some concept we already introduced, anyway it needs a deeper presentation. Let P be a finite poset with order relation $\leq$, denoted by $(P, \leq)$ when P alone could be misinterpreted.
Let $p \in P$ be defined by a vector of attributes or variables $\mathbf{q} \in \mathbb{B}$, and $|P|$ to be representing the cardinality of the set P.
Given that for $x, y \in P$, $x \leq y \Leftrightarrow q_i(x) \leq q_i(y)\ \forall q_i \in \mathbb{B}$ and $\exists i^*$ such that $q_{i^*} < q_{i^*}$, and for $x, y \in P$, $x \parallel y \Leftrightarrow x \not\leq y$ and $y \not\leq x$.
Then, $\forall x \in P$ we can define the following subsets of a poset:

Down Set $O(x) = \{y \in P : y \leq x\}$

Up Set $F(x) = \{y \in P : y \geq x\}$

Incomparables $U(x) = \{y \in P : y \parallel x\}$

It is important to notice that $O(x) \cap F(x) = x$, because of the reflexivity property of the relation $\leq$. The set made out of all the elements equal or smaller than x contains x itself. The same is true for the set made by equal or higher elements.

2.2.1.1 Concepts and Formulas

The concepts moving the Extended LPOM are basically two:

- the position of an element x respect to its upset and downset is fixed, there is no linear extension where an element of the upset $F(x)$ could be lower than x. Hence the average rank of x has to be contained inside an admissible interval;
- the elements $y \in U(x)$ have a range of possible positions to take respect to x (lower or higher), and these positions depends on the relations between y and the sets $O(x)$ and $F(x)$.

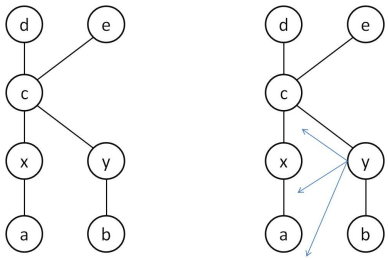


Figure 2.7: Example of LPOMext

Both these insights are essential for the definition of this approximation method. In order to understand this procedure more deeply, we propose the example of Figure 2.7. In this example the height of x has to be at least 2, that is the dimension of its downset, indeed $|O(x)| = |\{x, a\}| = 2$. Moreover, the only source of uncertainty are the elements of the incomparable set $U(x) = \{y, b\}$, which can have only a limited number of positions respect to x; the number of possible

positions for y is 3, two of these positions are below x an one is over it, as one can see from the figure. According to this example y is lower than x in two out of three times 2/3, and so the effect of y on the average rank of x is $\delta_y(x) = \frac{2}{3}$.
In formula:

$$H(x) = |O(x)| + \sum_{y \in U(x)} \frac{|O(x) \cap U(y)|}{|O(x) \cap U(y)| + |F(x) \cap U(y)|}. \tag{2.1}$$

The first part of the formula describes the lowest possible height of x, while the sum computes the contribution of the incomparable elements, adding them together. This formula defines the approximated average rank as if it is composed by two parts, the comparable one and the incomparable one.
The fraction $\delta_x(y) = \frac{|O(x) \cap U(y)|}{|O(x) \cap U(y)| + |F(x) \cap U(y)|}$ defines the effect of every element that is incomparable to x. It represents a proportion of positions; the proportion of position in which y is lower than x (and then increases the rank of x), out of the total number of possible positions.

This method improves significantly the approximation proposed in the basic model called *LPOM0* Brüggemann et al. [2004], where the entire set $U(x)$ was used as a unique entity, instead of approximating the effect of every $y \in U(x)$ alone.

As confirmed by the authors of *LPOMext* method, the main crucial points in the comparison of *LPOM* to the exact average rank are:

1. the ¨Combinatorial¨ effect, happening when an element $y \in U(x)$ can take more positions than what is expected by *LPOMext*, because of different relation of x and y respect to some ¨covering antichain¨; The authors of the method proposed some criteria to quantify this effect.

2. the ¨One-after-Another¨ effect, which underline the absence of simultaneity in the evaluation of the effects of the elements of $U(x)$;

3. the absence of ¨Restrictions¨, in the sense of limitation of the possible position for an element of $U(x)$ according to the expected position of another element of the same group. For instance, in Figure 2.7, the effect $\delta_x(b)$ and

$\delta_x(y)$ wuold be 2/3 for both of them, but they should be different since they are ordered (and then dependent in the sense of average rank).

Analysis and proposals respect to these crucial points can be found in the original work.
As last point we want to make explicit the relation of this method with weak orders: the *LPOMext* method considers the elements as a mathematical entity. Without taking into account the frequency of them, an element is considered in the computation if it appears at least once; no modification is considered if the same profile appears twice. The application field of this thesis forces us to deal with big populations and then to take into account the frequencies of the profiles.

2.2.2 Mutual probabilities approximation

A recent work [De Loof et al., 2011] proposes a new method to approximate the average rank: the authors use two concepts that have not been introduced yet. First, **rank probability** $P(\text{rank}(x) = i)$ of an element $x \in P$ is defined as the fraction of linear extensions in which element's rank is equal to i. According to this definition the average rank of an element is the expected value of the rank:

$$\overline{h}(x) = \sum_{i=1}^{|P|} i \cdot P(\text{rank}(x) = i).$$

Second, the actual change in the point of view comes from the quantity called **Mutual rank probability** $P(x > y)$ of two elements of the poset, defined as the fraction of linear extensions in which $x > y$.

The proposal depends on a relationship between average ranks and mutual rank probabilities, proven by the authors in the same article with a theorem (we quote it, adapting notation):
Theorem 1 *For a poset $(P, \leq)$ where $P = \{p_1, p_2, \ldots, p_n\}$ and $p_l \in P$, the following relationship holds between the average ranks and the mutual rank probabilities:*

$$\overline{h}(p_l) = \sum_{i=1}^{n} i \cdot P(\mathit{rank}(p_l) = i) = 1 + \sum_{j=1}^{n} P(p_l > p_j).$$

The proof is given in the original article. So the relation between average rank and mutual probabilities is established, and it implies

$$\overline{h}(x) = 1 + \sum_{y \in P}^{n} P(x > y) = |O(x)| + \sum_{y \in U(x)}^{n} P(x > y).$$

This formula is extremely similar to the one used in *LPOMext* but the computation of the mutual rank probability is different. To clarify the next steps we will use the quantities $o(x) = |O(x) \setminus \{x\}|$, and $f(x) = |F(x) \setminus \{x\}|$. Using an approximation of the mutual rank probability proposed by [Brüggemann et al., 2003], for $x \neq y$:

$$P^*(x > y) = \frac{[o(x)+1][f(y)+1]}{[o(x)+1][f(y)+1] + [o(y)+1][f(x)+1]}.$$

This could be enough to improve the previous approximation, but the authors suggest to improve the accuracy with an approximation of the values $o(x)$ and $f(x)$ that takes into account the incomparable elements in this way:

$$\begin{aligned} \tilde{o}(x) &= o(x) + \sum_{z \in U(x)}^{n} P^*(x > z) \\ \tilde{f}(x) &= f(x) + \sum_{z \in U(x)}^{n} P^*(x < z). \end{aligned}$$

With the improved information, they introduced the formula

$$\rho(x) = o(x) + 1 + \sum_{y \in U(x)} \frac{[\tilde{o}(x)+1][\tilde{f}(y)+1]}{[\tilde{o}(x)+1][\tilde{f}(y)+1] + [\tilde{f}(x)+1][\tilde{o}(y)+1]}. \tag{2.2}$$

For a poset with dimension n, the time complexity of this approximation method is $\mathcal{O}(n^2)$. Exactly the same of Local Partial Order Models.
The simulations carried out by the authors of this method show that the estimator $\rho(x)$ has a smaller mean squared error than the *LPOMext* model, it is true in randomly generated poset with $n > 4$ and in most of the real dataset used to test them.

2.2.3 The approximation approach applied to the social science

In the last two sections, we described the most used procedures for the approximation of the average rank, but, beside the level of error coming from approximation procedures, there are some other limits in this approach that emerge if we analyze the data coming from social surveys.
Our point of view is the social statistics, where the observation of big datasets is frequent, in this case we need a methodology `ad hoc`.

Dimension of the dataset. All the approximation methods are developed in the framework of small datasets, this implies some drawback respect to the use of big datasets, typically used in social studies:

Frequency of profiles. The *LPOMext* and the *Mutual Probability* method are supposed to handle the poset as a set of distinguished elements of a set. If two elements are equal respect to the vector of attributes, these are considered as an equivalence class and treated like one element in the computation of approximation. It is straightforward that this is an important drawback for social statistics, where probably every profile is observed at least once, but some are more relevant than others because of their frequency.

Complexity of the poset. The structural complexity of a poset depends on the observed elements, in the framework of big datasets it is probable to observe almost every combination of the elementary variables, generating a challenging complexity.

Software Limitations. All the software developed in the last years are limited to one or two hundreds observations, whereas a typical social survey is based on a sample with thousands of observations.

Weight of Variables One of the main features of the poset approach for synthetic indicators is the complete absence of external information about the relevance of the elementary variables. This is a great strength for the construction of a synthetic indicator based on the data driven approaches (see

[Decancq and Lugo, 2013]), but, on the other hand, it could be a limit when the importance of variables should be taken into account.

2.3 Sampling of linear extensions

Despitew, the estimation of average rank by sampling of linear extensions has been proposed in literature Lerche and Sørensen [2003], in this thesis we propose a method that uses the sample of linear extension, in a different way. We will refer to this approach as *Evaluation Method* for posets. It was actually developed for the measurement of deprivation in the case of ordinal variables, nevertheless, we propose some interpretations in the framework of our hypothesis.

2.3.1 A method to evaluate Posets

This subsection is devoted to remind and quote some of the basic concepts constituting the method proposed by Fattore [2015]. We suggest the reading of the original work in order to understand completely the mathematical formalization. A specific note should be devoted to underline the different aims between this chapter of the book and the original work proposed in Fattore [2015] and Maggino and Fattore [2011]: the cited work contains the conceptual and mathematical formalization for the measurement of social concepts in poset data, specifically the most recent one is explicitly focused on the measurement of deprivation. On the other hand, the proposal of the next chapter is the utilization of the tools given by these works in order to define a procedure to obtain synthetic indicators out of ordinal or dichotomous data.

2.3.1.1 Product order of variables

In the original formalization of this method, the partially ordered set contains all the possible values that are observable considering the starting variables' structure. In partial order theory, it is called *product order*, because it is made by the interaction of the linear order determined by the single variables. As instance, two dichotomous variables define their own linear orders made by two levels $\{0, 1\}$. The product of these two orders produces a poset made by the elements: `00`, `01`,

10 and 11. Then, the elements of the poset need to be described by a sequence of values (e.g. 10), and this sequence is called **profile**, called **p** in the following.

2.3.1.2 Setting a threshold

The poset derived from a product order is a mathematical structure without information about the meaning of the variables constituting it. In order to give a ¨meaning¨ to this structure, is possible to define a **threshold** τ which contains external information defined by the researcher. Because of the multidimensional framework, it is possible to define a multidimensional threshold, that is a list of profiles, that respects two requirements:

1. Every element of the threshold must be considered completely deprived;
2. It has to be made by incomparable profiles (i.e. to be an *antichain*, introduced in 2.1.2.4).

The scope of the threshold is to cover the elements that are ¨deprived¨, acting like a frontier between the low and the high level of the poset. It can be only defined externally. Deeper information about the threshold and its meaning are presented in the next chapter.

2.3.1.3 Identification step

The role of the *identification function* $idn(\cdot)$ is to assign a *deprivation membership* score to every element of the poset, this score is contained in the interval $[0, 1]$:

$$\begin{aligned} idn &: P \mapsto [0,1] \\ &: \mathbf{p} \mapsto idn(\mathbf{p}). \end{aligned}$$

The construction of the identification function is inspired to the principle of *reduction to linear extensions*, previously introduced in (2.1.3.1). Each linear extension l is interpreted as a binary classifier where a profile **p** is classified deprived or not. A small notation improvement is necessary in order to understand the procedure. In every linear extension l there will be a top element of the threshold τ_l, that is better than all the other elements of τ. For every linear extension,

a profile $\mathbf{p}$ is defined as deprived in that linear extension and the value of the identification function is 1 if this profile is lower or equal than τ_l; otherwise the value is 0:

$$idn_l(\mathbf{p}) = \begin{cases} 1 & \text{if } \mathbf{p} \trianglelefteq \tau_l \\ 0 & \text{otherwise} \end{cases}$$

The position of the profile $\mathbf{p}$ and the value of τ_l can change in different linear extensions; the aggregation of the results among the observed linear extensions gives the value of the identification function:

$$idn(\mathbf{p}) = \frac{1}{|\Omega(P)|} \sum_{l \in \Omega(P)} idn_l(\mathbf{p}).$$

At the end of the procedure, every statistical unit will be associate with its identification value, that will be:

- $idn(\mathbf{p}) = 0$ if the profile is always not deprived in every linear extension,
- $idn(\mathbf{p}) = 1$ when the profile is under the threshold in every extension,
- $idn(\mathbf{p}) \in (0, 1)$ if the profile is ambiguously defined in the middle.

Following the derived information it is possible to define three subsets:

- Non deprived profiles $W = \{s \in P : m(s) = 1\}$

- Deprived profiles $D = \{s \in P : m(s) = 0\}$

- Ambiguous profiles $A = \{s \in P : 0 < m(s) < 1\}$

For deeper explanation and properties we suggest again to refer to the original work, of [Fattore, 2015].

2.3.1.4 Severity

The measure of *severity* defines the intensity of deprivation of the deprived or ambiguous elements (subsets D and A), by assigning a numerical value to each profile of $D \cup A$:

$$\begin{aligned} svr &: D \cup A \mapsto \mathbb{R}^+ \\ &: \mathbf{p} \mapsto svr(\mathbf{p}). \end{aligned}$$

In every linear extension, the measure of interest is the *distance* between a profile and the first element higher than τ_l called $\mathbf{q}_l$. The distance is computed respect to the rank of the two objects:

$$svr_l(\mathbf{p}) = \begin{cases} r_l(\mathbf{q}_l) - r_l(\mathbf{p}) & \text{if } \mathbf{p} \trianglelefteq \tau_l \\ 0 & \text{Otherwise} \end{cases}.$$

The severity value is computed only on those linear extension where the profile is considered as deprived.
The *deprivation severity* of a profile is then obtained aggregating all the results observed on the linear extensions:

$$svr(\mathbf{p}) = \frac{1}{|\Omega(P)|} \sum_{l \in \Omega(P)} svr_l(\mathbf{p}).$$

The Wealth function *wea* is a function complementary to *svr*; it measures the concept of intensity in the opposite direction, the direction of *non deprivation*. The computation is the same of *svr*, but it is oriented to the positive side and is evaluated on the subset $(W \cup A)$ [Fattore et al., 2011].
The *R* package devoted to the computation of these functions is called `PARSEC`, and was developed by Fattore and Arcagni [Fattore and Arcagni, 2014].

2.3.2 Limits of Evaluation Method

In the last section, we described a method to evaluate the elements of a poset, avoiding the concept of average rank. Despite the augmentation of information handled by this approach, there are still some limitations we would like to underline.
The method for the evaluation of a poset is not conceived to define a synthetic measure. Then, some of the drawback we are going to underline are referred to our aim and not to the method itself.

Computational Limitations. The R package devoted to the computation of this method is optimized. On the other hand, the number of linear extensions to observe in order to get the results grows too much with the

dimension of the poset. The procedure is already time consuming in the case of five dichotomous variables (this case requires to observe at least 123 million linear extension).

Effect of the threshold. The threshold has an evaluation effect on the original variables (as we will explain deeply in the next chapter), but this effect is not strong enough to impose a sort of priority system on the poset.

Weight of Variables. As well as the approximation approach, excluding the threshold, there is complete absence of external information on the relevance of elementary variables.

Chapter 3

A synthetic measure from the sampling approach

3.1 A step towards a synthetic indicator

In this chapter a simple procedure to construct a synthetic indicator is proposed. Taking the lead of the proposals contained in Fattore [2015], Maggino and Fattore [2011] and Fattore and Arcagni [2014], which have been presented in the last part of the previous chapter (see Section 2.3.1).
The definition of the synthetic indicator is the topic of the next paragraph (3.1.1). In the following section the central concepts of the Evaluation method are described in their implications: the threshold is the only source of exogenous information, its implications will be presented with examples. The last part (3.3), contains the study of *Life Satisfaction* and its relations with many socio-economical factors.

3.1.1 The Height of a profile

The evaluation procedure proposed in [Fattore, 2015], is constituted by the identification function and the severity functions. These functions has been previously introduced in 2.3.1.3 and 2.3.1.4. In the following we propose to combine the information given by the method to obtain a unique measure.
The combination of the results transforms the meaning of the measure: the func-

tions of identification and severity are conceived to describe the membership of a profile to a group, the group of deprived individuals for instance, and the intensity of this membership. The aim and meaning of the combined values is different, because it tries to represent the position of the profiles respect to a complex concept mimed by the poset.

The measure we suggest consists in the combination of the indexes of severity produced in Fattore and Arcagni [2014]:

$$I_H(p) = wea(\boldsymbol{p}) - svr(\boldsymbol{p})$$

Starting from the values of the absolute *severity* and *wealth* measures, it is possible to obtain a unique value by their difference. Given that:

$$\begin{aligned} wea(\boldsymbol{p}) - svr(\boldsymbol{p}) &= \frac{1}{|\Omega(\Pi)|}\sum_{l\in\Omega} wea_l(\boldsymbol{p}) - \frac{1}{|\Omega(\Pi)|}\sum_{l\in\Omega} svr_l(\boldsymbol{p}) \\ &= \frac{1}{|\Omega(\Pi)|}\sum_{l\in\Omega}(wea_l(\boldsymbol{p}) - svr_l(\boldsymbol{p})), \end{aligned}$$

the index I_H is the mean of the difference $\Delta_l = wea_l - svr_l$ on the set of all linear extensions.

Meaning: The value Δ_l represents a sort of height of the profile, *evaluated* with respect to the defined threshold; in every linear extension the profile is compared to the highest element of the threshold, it gets positive value if it is higher, negative otherwise.
The value of Δ_l is called *evaluated height* because, respect to the concept of Average Rank, it contains the information given by the threshold τ. Two elements of the poset could be equal with respect to the average rank and different with respect to I_H or viceversa; the difference depends on the threshold and specifically on its *length* and *shape*.

In the following section, the method of Fattore [2015] is the center of the discussion. Some thoughts about the meaning of the threshold, and the criteria

to define it are presented. These consideration are valid also for the indicator we proposed here.

3.2 About the meaning and the use of the threshold

The threshold τ proposed in the *Evaluation* method of Fattore is the cornerstone of the procedure; it inserts information that are exogenous respect to data, impressing an evaluative meaning. This kind of information is essential in a multivariate evaluation framework, where an absolute best does not exists, and the relative most acceptable option is the aim (Munda [2008] and Arrow and Raynaud [1986]).

But, what is the best way to choose the threshold? Is it correct to look at elementary variables or is it better to choose an expert defined set of profiles? The aim of this section is to help the new users, giving some tips to manage this decision with more consciousness.
This section is divided in four parts: in the first one, the evaluation procedure and the indicator I_H are compared to a more classic method based on the poset theory. The argument of interest is the value added by the user defined threshold. In the second part the effect of the *shape* of the threshold is investigated, trying to understand how it can impress values and priorities on the elementary variables. The third part is devoted to show the effect of the threshold's dimension on the discrimination skill of the identification function an the indicator I_H. Finally, in the last part, we propose some short guidelines for the definition of a threshold, pulling together the deductions presented in this entire section.

3.2.1 The shape of *threshold*

The possible interpretation of the threshold are multiple, but it is clear that the purpose of the proposer is to define it as a set covering every deprived element of the poset; all the elements that are lower than the threshold are deprived. Without the meaning imposed by the threshold, two variables constituted by the

same number of level cannot be distinguished, neither conceptually nor mathematically. In order to perceive this limit, the case of a poset generated by two ordinal variables made by three levels is sufficient. It can be observed in Figure 3.1.

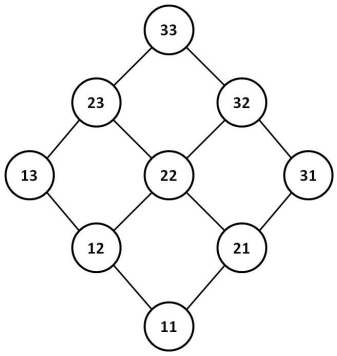


Figure 3.1: Poset derived by two variables with three levels

Without an externally defined preference system, the constituting variables are completely interchangeable. For instance, these variables could represent the degree of appreciation for an ice-cream's taste or the personal self-perceived safety in a residential zone; considering only the structure of the poset, the two variables are identical. Adding a threshold to this structure, we impose a system of values and discriminate among the variables.

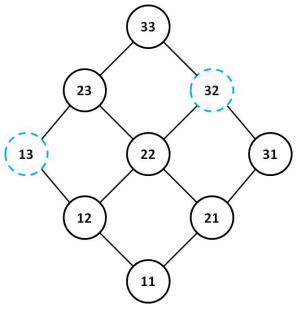


Figure 3.2: Asymmetric threshold

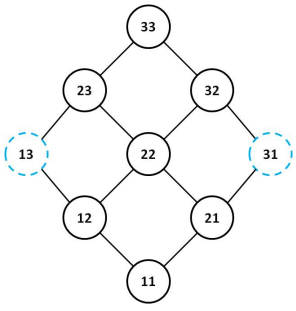


Figure 3.3: Symmetric threshold

Figures 3.2 and 3.3 look very similar, and yet they impress a very different meaning to the structure. In order to make it simple: in Figure 3.2 the

constituting variables are non exchangeable! In this setting profiles 32, 22 and 31 are always deprived. The same cannot be confirmed for 23, since it could be higher or lower than 32 depending on the linear extension. Therefore, there is an implicit assertion about the importance of the variables on the classification.

On the contrary, in the poset represented in Figure 3.3, we can perceive a sort of symmetry among the variables, the first variable is equated to the second. Thus, the poset is evaluated in a manner that is more similar to the computation of the average rank. Thanks to this property of the threshold, it is possible to impress a system of preferences and weights, these ¨weights¨ are implicit and non linear, because they change intensity in different positions of the poset. So, it is not fair to interpret this preference effects in the same way of Composite Indicators weights.

In order to underline the relevance of such a property, it is sufficient to imagine a poset made by several dichotomous variables, like the one represented in Figure 3.4. This type of variable has only two levels, by definition, therefore the levels are not much descriptive; nevertheless two dichotomous variables could represent a deeply different meaning. For example, this is important in deprivation studies, where the ownership on different goods is considered: owning a fridge and be owner of a house are conceptually different information; these variables are different in their meaning and distribution but not in their level structure. In an European country like Italy, to be deprived of the fridge is much more meaningful than not having the property of an entire house.

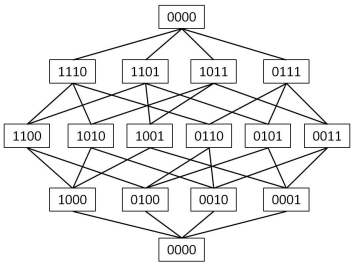


Figure 3.4: Poset derived by four dichotomous variables

3.2.2 The length of the threshold

If the shape can influence the importance of the variables, the length of the threshold can change the *fuzzyness* of the profiles that are non comparable to the threshold itself. With the word *length*, we mean the dimension of the threshold, that is, the cardinality of the set τ.
Indeed, if the threshold is defined by a single element p, only the elements in the downset of p will be certainly deprived and the inverse result apply to the elements of the upset; all the other profiles will be included in the group of ambiguous elements. On the contrary, if the length of the threshold is big, the amount of uncertainty will decrease in intensity.

This concept is particularly intuitive in the case of *identification* function: in order to illustrate this effect, two evaluation functions have been computed on the same poset, made by four variables measured on four levels (256 nodes), using the following thresholds:

Single threshold composed by the single profile: `2232`, that is quite central in the poset;

Extended threshold composed by the profiles: `2232`, `1233`, `2133`, `2223`, `1242`, `2142`, `2241`, `3132`, `3222`, `3231`, `1332`, `2322`, `2331`, actually all the profiles that are obtained from the *single threshold* and changing every elementary variable by one unit at time.

The results of the identification function show the effect of the threshold's length on the classification. In the *single* case (Figure 3.5), more or less forty profiles out of 256 (15.6%) are in an intermediate position around value 0.5, and only 67 have a value of deprivation close or equal to one. On the other hand, when the threshold is extended (Figure 3.6), the number of certainly deprived elements increases to more than one hundred units, because it now contains nodes that were ambiguous in the single-threshold case. Moreover, one out of three elements previously defined as non deprived (Identification = 0), is now less certain identified, and are distributed in the interval $(0.2; 0.4)$. Judging the whole picture, there is more deprivation, and the class of *absolute ambiguity*, identifiable with

the values around 0.5, disappears.

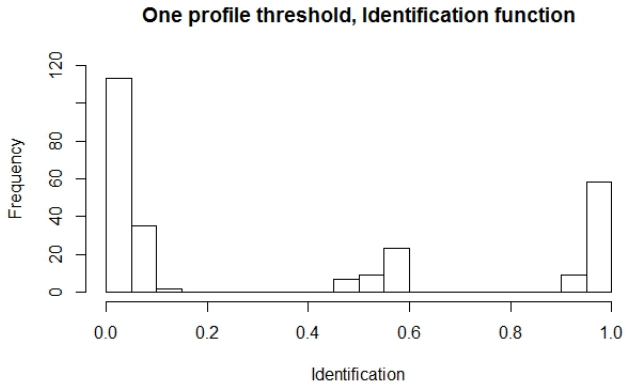


Figure 3.5: Identification function with Single threshold

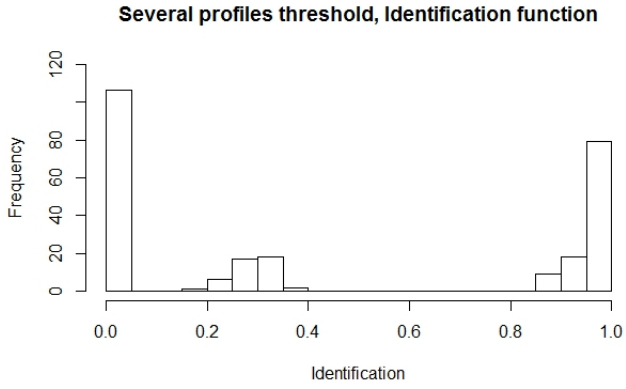


Figure 3.6: Identification function with Extended threshold

The identification function is oriented to recognize deprived profiles because of it definition, which takes into account the highest element of the threshold in every linear extension (τ_l).

According to this property, the enlargement of the threshold's dimension can only imply the equality or the increment of the deprivation level of the poset's elements.

$$if\ \tau^s \subset \tau^e \Rightarrow idn^{\tau^s}(p) \leq idn^{\tau^e}(p),\ \forall p \in P.$$

The same results apply to the severity function.

3.2.3 Comparison between the evaluation approach and the average rank

A common tool developed in poset theory, is the computation of the average rank of an element of the poset. As described in paragraph (2.1.4), this method allows to compute the medium height of observed units, making possible to order them reciprocally, and in our assumption, defining their position respect to an unobservable dimension. The main interests of these section are the similarities and differences between the average rank and the evaluation method previously presented.

The evaluation method is constituted by two different measures: the *identification function*, which defines the degree of membership of the profiles to one of the opposite sides of the poset (the bottom or the top); and the *severity function*, that measures the depth of a profile, describing how much it moves towards on of the two poles. The use of two different measures is motivated by the suggestion of Sen [1976], which proposed that a methodology for measuring multidimensional poverty is made up of an identification method and an aggregate measure.

Moreover, the user defined threshold impress a meaning of centrality to all the element composing it. These elements may be neither higher nor lower than the threshold τ. In this way, the poset gains some reference points, which do not exist in the definition of the average rank.

So, the definition of the threshold gives a specific meaning to the results of the evaluation function: if a profile is always higher than this reference group (resulting in an identification function equal to 0), it can be classified as non deprived (using the terminology shared by Fattore [2015] and [Alkire and Foster, 2011]). Similarly, if the profile is always lower the deduction is the opposite, but

it actually is an evaluated classification, not simply a score. As instance, if the poset is obtained from variables about the satisfaction of individuals, defining a good threshold will enable to discriminate between satisfied and unsatisfied subjects.

This result is not possible in the case of average rank, in that case, the definition of a limit value between two different poles will result in something with a different meaning, since:

- If the *identification* function is equalt to 1, there is no linear extension where the profile is greater than every element of τ. In other words, there is at least one element of the threshold that is better than the profile, in every linear extension. This information is not accessible in the average rank approach;
- Using the average rank, the classification of a profile as deprived implies that its *average value* is lower than a predetermined value, and hence it means that the profile is sufficiently low in *average*. Because of the properties of the arithmetic mean, this result tells nothing about the distribution of ranks among the linear extensions, because there could be some linear extensions influencing the average rank more than others. In this sense, the average rank can be defined as less robust than the evaluation method.

The difference between these two approaches is conceptually important, especially in the definition of a classification procedure; indeed these approaches imply two different definitions of classification. The simple fact of imposing a subset of the poset as threshold impresses a meaning on the constituting variables. Instead in the average rank method the variables are handled as mathematical entities, according to the observed levels, and nothing else.

From the *severity* point of view, the added value is determined by the shape and position of the threshold, because it can influence the severity measure observed on profiles. The point we want to stress is the centrality of the threshold; without that, average rank and evaluation method are not so different. Indeed,

if the complexity of the threshold increases, the difference between the two approaches increases too. The length and shape of the threshold is deeper examined in the following part.

The last relevant difference between these approaches relies in the profiles that are considered: the average rank method is based only on the observed profiles and its complexity depends on the relation structure of the set, for this reason it is used for small samples or it is performed with the help of approximation procedures ([Brüggemann and Carlsen, 2011], [De Loof et al., 2011]). On the other hand, the evaluation procedure is commonly based on all the possible profiles that can be observed starting from the elementary variables (2.3.1.1), hence, it does not depend on the numerical dimension of the population. Recent developments (Fattore [2015]) allowed the use of a smaller set of profiles in the evaluation method, avoiding the evaluation of those profiles that are not observed in the population.

3.2.4 Conclusions about the threshold

The thoughts proposed in this section are not meant as rules, but more as highlights. According to our aim, we propose three criteria for the choice of the threshold:

- **Meaning** The threshold can be dependent on the univariate distribution of the elementary variables or be defined externally by experts who evaluate the entire profiles;

- **Shape** The structure impressed on the poset by the threshold have to be taken into account, an asymmetric threshold could be recommended in the case of variables with different importance;

- **Length** The number of elements of the threshold influences significantly the results, especially the identification function. As a rule of thumb: if the information useful to define a large threshold is available, then is better to use it entirely.

3.3 Study of life satisfaction in Italy

The aim of synthetic indicators is the measurement of complex and non-observable concepts; in this work the complex concept under study is *Life Satisfaction* as a proxy of well-being. It is a multidimensional concept that cannot be defined objectively, because it depends on both life and socio-psychological conditions. It is fair to say that: the same conditions, (assuming the possibility of identical conditions among humans), are evaluated in different ways by different subjects. Such a variability depends on culture, society and psychology interactively, so we can not measure effective satisfaction but its perception.

In order to analyze satisfaction, we use the indicator construction method proposed in 3.1.1. $I_H(x)$ represents the position (height) of profile p with respect to the threshold τ. Once one got the value of such an indicator for every observed unit, it is possible to use this information as a variable. The values of $I_H(p)$ come out from an average of ranks among linear extensions, this construction procedure needs to be taken into account, in order to use and interpret the results.

3.3.1 Data

Data comes from a survey carried out by the Italian National Institute of Statistics (ISTAT). This survey is part of the Multi-Purpose surveys ¨Aspetti della vita quotidiana¨, literally *Aspects of Daily Life*.
This survey is extremely useful because of its longevity and complexity; it collects information on many life aspects such as: work, health, safety perception, social inclusion, society and much more. In the year 2012 more than 40 thousands individuals have been interviewed.
In this work the focus is oriented on the concept of Life Satisfaction and its determinants. In order to produce a measure of Life Satisfaction, seven variables have been considered, each one describes the satisfaction on a single aspect of life: Economical Situation, Health Status, Family Relations, Friendship Relations, Free Time, Work Conditions and Environment. Three of these variables are not taken into account in the following, because:

- **Work Condition** is observed only among the individuals who were occupied at the time, for sake of precision, occupied and homemakers. The study is intended for all people.

- **Environment** is observed only since recent time in the surveys, and this application is meant to be part of a wider, time-crossing, application.

- **Friendship Relations** is too much associated to family relations and free time, showing a small amount of original information.

Hence, at the end of the selection procedure the elementary variables are: Economical Situation (*Economy*), Health Status (*Health*), Family Relations (*Family*) and Free Time (*Time*). All the satisfaction variables are measured on a four-levels ordinal scale: *A Lot, Enough, A Little, Not at All*. The poset produced by these variables is composed by $4^4 = 256$ nodes. Apart from specific reasons, the most limiting criterion in the selection of starting variables is computational. The number of linear extensions required from the *Identification* method in a poset made by 256 elements according to Karzanov and Khachiyan [1991] is around 6.1×10^{12}. Adding one single variable with four levels, this number increases to 7.8×10^{15}, more then one thousand times bigger, making the computation too long for any research purpose.

3.3.2 Construction of the satisfaction Indicator

The procedure to determine the values of the satisfaction indicator has been described in Section 3.1.1. In this application the established threshold is made by a unique profile: `2232`, so it is the first deprived profile and the less severe one (among the deprived). The profile `2232` means: level 2, *A Little satisfied*, on *Economy*, *Health* and *Time*, and level 3, *Enough* satisfied on *Family*.
This profile has been selected taking into account the univariate distribution of the elementary variables (Table 3.1). In the threshold only the *Family* variable has a value equal to *Enough Satisfied*. This threshold assesses that: the border between satisfaction and dissatisfaction is located in this exact combination of levels. If a profile will show a *Family* values lower than 3, it will implies that

Variable	4.%A lot	3.%Enough	2.%A Little	1.%Not at all	Total
Economically	2.6	41.9	39.3	16.3	100
Health status	18.5	64.2	1.8	4.4	100
Family relations	37.3	55.8	5.6	1.4	100
Free time	16.2	51.5	25.7	6.6	100

Table 3.1: Univariate distribution of the satisfaction variables

profile to be lower or incomparable to the threshold; the incomparability case is obtained only, at least, one of the other variables is higher than 2.
This threshold takes into account the higher frequency of the highest levels of satisfaction observed on *Family*; the same care has not been repeated for *Health*, because of the low frequency of the top level of satisfaction for this variable.

After the computation of the value of the indicator $I_H(\cdot)$ for every individual, the data have been scaled with the use of *MIN-MAX* method. The scaling is important because the aim is the definition of an intuitive indicator, and the use of absolute values (that in this case have a range equal to $(-90; 166)$) is not a straightforward solution for the first description of the phenomenon. In the following the scaled indicator is called $S(\cdot)$, to represent satisfaction.
Observing the indicator I_H, there is no need to focus on the level `zero` as it happen for the interpretation of the $svr(\cdot)$ index. The satisfaction is full if the indicator is 1, in the opposite case it is 0.

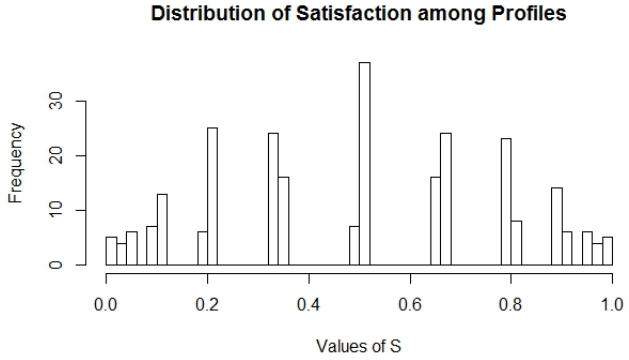


Figure 3.7: Distribution of satisfaction on the 256 nodes, without frequencies

The construction of the indicator defines a value for each one of the 256 nodes of the poset, these values are represented in Figure (3.7). The highest concentration is on the middle values, because the poset is larger in the central part, in the sense of number of profiles, the central nodes are evaluated around those values.

The graph is symmetric; this feature will be useful in the following, because it implies that every asymmetry in the observed distribution is attributable to the distribution of satisfaction in the population.
In the following, the distribution of profiles is called *theoretical* because it does not depend on the amount of population in every profile.

The observed distribution of S in the population of 2012 is represented in Figure 3.8. The picture shows a strongly skewed distribution, with a high amount of individuals assessing good levels of satisfaction. The number of elements with a value lower than 0.4 is extremely low, this graph is actually the evaluated distribution of *Satisfaction* among the Italian population in 2012. The interviewed are quite satisfied, their answers show a population which is highly pleased with their level of satisfaction.

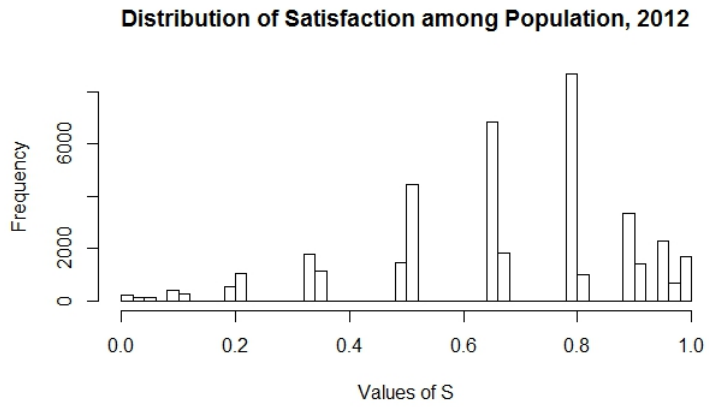


Figure 3.8: Distribution of satisfaction according to the observed frequencies

This trend is more observable in the representation of deciles (Figure 3.9): the first 10% of the population has a satisfaction level in the interval $(0; 0.33)$.

Mean	0.671	Std Deviation	0.230
Median	0.668	Interquartile Range	0.306
Mode	0.782	Coeff Variation	34.334

Descriptive Statistics of the Index in the population

The difference respect to the distribution of theoretical profiles is large, it means that a really small amount of people define itself as severely dissatisfied on every aspect. The difference between theoretical and observed frequencies is reabsorbed in the interval $(0.5; 0.90)$, actually between the median and the nineth decile. The highest levels of the index $(0.9; 1)$ are distributed in theoretical and observed data similarly.

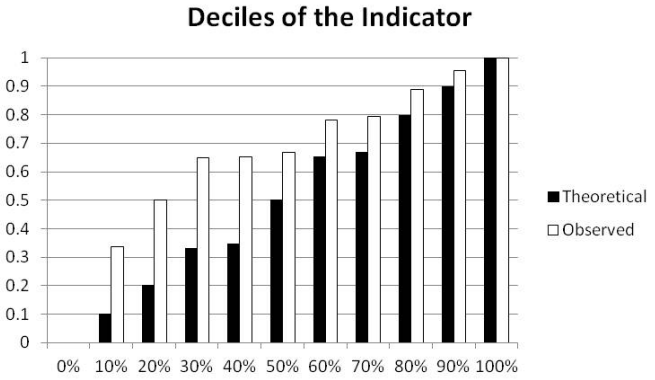


Figure 3.9: Deciles of the indicator, theoretical and with the frequencies

The amount of individuals with a *Satisfaction* lower than 0.5 is less than 20% of the population.
Just to have a criteria for comparison, between these results and the basic *Identification* approach, one can consider that the profile 2232 has a value of $I_H(2232) = -1$, while its *Satisfaction* value is $S(2232) = 0.35$. It means that every value higher than 0.35 should be considered as satisfied in the identification point of view. The population of Italy can be defined as mainly satisfied, since four individuals out of five are assessing a S bigger than 0.35.

Effect of the elementary variables on the indicator The relation between the *satisfaction* and the single variables is an important information to understand the behavior of the indicator; the regression tree method has been used to explore these internal relations. Given the construction procedure of the indicator, the observed effects are dependent on the distribution of the elementary variables.

Group	Mean $S(\cdot)$	Time	Health	Economy	Family
1	0.121	1+2	1+2	1	—
2	0.324	1+2	1+2	2+3+4	—
3	0.486	1+2	3+4	1+2	—
4	0.582	3+4	—	1+2	1+2+3
5	0.714	1+2	3+4	3+4	—
6	0.785	3+4	—	3+4	1+2+3
7	0.812	3+4	—	1+2	4
8	0.937	3+4	—	3+4	4

Table 3.2: Groups identified by the regression tree

$N = 39571$ $\overline{S}(\cdot) = 0.671$	$Time \leq 2$ $N = 12781$ $\overline{S}(\cdot) = 0.481$	$Health \leq 2$ $N = 3530$ $\overline{S}(\cdot) = 0.257$	$Economy = 1$ $N = 1161$ $\overline{S}(\cdot) = 0.122$	Group 1
			$Economy \geq 2$ $N = 2369$ $\overline{S}(\cdot) = 0.324$	Group 2
		$Health > 2$ $N = 9251$ $\overline{S}(\cdot) = 0.567$	$Economy \leq 2$ $N = 5983$ $\overline{S}(\cdot) = 0.486$	Group 3
			$Economy > 2$ $N = 3268$ $\overline{S}(\cdot) = 0.714$	Group **5**
	$Time > 2$ $N = 26790$ $\overline{S}(\cdot) = 0.761$	$Family \leq 3$ $N = 16001$ $\overline{S}(\cdot) = 0.681$	$Economy \leq 2$ $N = 8225$ $\overline{S}(\cdot) = 0.582$	Group **4**
			$Economy > 2$ $N = 7776$ $\overline{S}(\cdot) = 0.785$	Group 6
		$Family = 4$ $N = 10789$ $\overline{S}(\cdot) = 0.880$	$Economy \leq 2$ $N = 4874$ $\overline{S}(\cdot) = 0.812$	Group 7
			$Economy > 2$ $N = 5915$ $\overline{S}(\cdot) = 0.927$	Group 8

Table 3.3: Regression tree with the original variables

The results of the regression tree are represented in tables (3.3) and (3.2). The procedure defines 8 groups, with increasing medium value of the indicator $S(\cdot)$; they show different levels of the elementary variables.
The *Family* variable is not considered in the first three groups, it means it is not useful to discriminate among the low levels of the indicator; this elementary variable seems to be more influential at the highest values. Indeed, in the two top levels, the value of the family is four (*A lot*), the most satisfied group with *Family* lower than four has a mean value of the indicator equal to 0.785.
The role of *Health* is complementary respect to the *Family*'s one: it shows its importance in the lower levels, where the the medium value is 0.324 or less; in those groups the value of *Health* is equal or lower than *A Little* (1 or 2). The real key variable of the severe dissatisfaction is *Economy*, indeed, in the lowest group every observation has value *Not at All* on it.
The seventh group is peculiar, the elements of this group have a high satisfaction, but the level of *Economy* is 1 or 2. Therefore, if the family relations are completely satisfying and the free time is at a good value, the level of satisfaction could be very high, despite a low level on the economical point of view.

3.3.3 Life satisfaction in society

In the following the original indicator of satisfaction I_H is used as a response variable in a quantile regression procedure without scaling it. In every regression model, the aim is the estimation of some characteristics of the response variable, conditioned to the values of the explanatory variables. In the quantile regression the aim is the estimation of a quantile of the response variable conditionally to the effect of explanatory variables (Koenker [2005], Koenker and Bassett Jr [1978]). This construction makes quantile regression highly recommended when the response variable is not normally distributed, and this is the reason why we are applying it to this data.
The estimated quantile could be the median or the quartiles, but also every other quantile. Therefore, this method is particularly suggested when the effect of the explanatory variables is supposed to change along quantiles. Focusing on order statistics such as quantiles, there is no need of assumption on the distribution of

the response variable in quantile regression. This property makes quantile regression significantly different from many other approaches.

The method of quantile regression has been applied by defining the indicator H_τ as the response variable and 14 socio-economical variables as explanatory ones. In Table 3.4, the explanatory variables that have been selected are shown as well as the estimation of the parameters at the 50th quantile. For example, the estimated parameter for men (with respect to woman) is 2.84.

In this case, the estimated value of the median of our indicator of satisfaction can be easily computed; keeping fixed the subject's qualitative variables to the reference levels (the values between brackets), the estimated quantile for a median age individual in a family with dimension 2 is

$$83.23 - 0.56 \cdot 49 - 1.57 \cdot 2 = 52.95,$$

where 49 and 2 are the values of the variables *age* and *N. of family members*. On the other hand, if the individual is a man, the estimated value has to be increased by the *gender* parameter 2.84, reaching the value of 55.79.

The formalization of the model allows a simple approach of interpretation: the men population has a higher median satisfaction than the women population, and hence they are more satisfied, given all the other explanatory variables in the model (at least at the median).

The table of estimates at the median can be used as reference to study the differences among the explanatory variables. The direction of results is not surprising, but their dimension is quite interesting. If we exclude the age, the most important variables are: Economical change, and Geographical Partition. The economical change agree with the well known concept of relative satisfaction, which assess how the satisfaction is a measure relative to the peers; in this case the effect comes from the comparison with the past. If the situation is worse then the previous year, the level of satisfaction decreases strongly. The comparison between equal and better level gives non-significant results before the 70-th quantile.

Variable *(Reference level)*	Level of variable	Estimate	$Pr > \|t\|$
Intercept		83.23	<.0001
Sex (*Female*)	Male	2.84	<.0001
Marital Status	Unmarried	7.96	<.0001
(*Divorced/Widower/Separ.*)	Married	8.22	<.0001
Education	University or higher	7.78	<.0001
(*Middle or lower*)	High school	4.76	<.0001
Work Conditions	Occupied	-11.27	<.0001
(*Retired/Student*)	Unemployed/Housekeep	-14.90	<.0001
Smoking (*Non Smoker*)	Smoker	-5.45	<.0001
Religious Practice	Often	13.38	<.0001
(Rarely)	Some Times	10.42	<.0001
Politic Discussion	Often	4.42	<.0001
(Rarely)	Some Times	3.67	<.0001
House contract	Rent	-6.79	<.0001
(Free use without property)	Property	2.75	0.059
Economical changes	Better	19.88	<.0001
(Worse)	Equal	19.72	<.0001
House Type	Distinguished	16.12	<.0001
(Rural/Public)	Average city house	11.53	<.0001
Geo. Partition	North	22.57	<.0001
(South)	Center	13.57	<.0001
Age		-0.56	<.0001
N. Family members		-1.57	<.0001

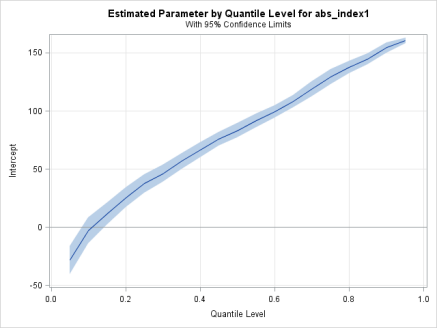
Table 3.4: Quantile regression's parameters at the median of I_H

The geographical partitions probably acts as a wide summary of relevant factors for the determination of life satisfaction. Indeed, The southern partition has the most negative effect of the quantiles of satisfaction. The comparison between the center and the north underlies a statistical difference, assessing the northern partition as the best environment for satisfaction improvement.

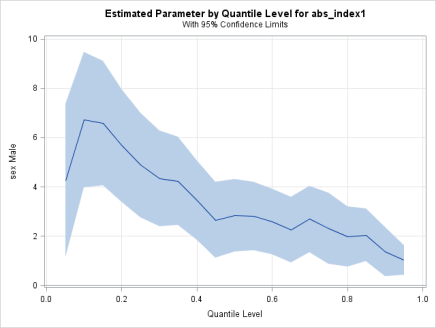
Education level is surprising in the opposite point of view. Its effect is smaller that 10 from the third to the tenth decile. The university and high school levels are statistically different, by a really small amount (2.5) for the middle-high quantiles.

Moreover, it is useful to evaluate the quantile regression's estimates among many percentiles. In the following, those trends are represented: every graph represents the estimation at the percentiles from the 5-th to the 95-th with steps of 5 quantiles.

Figure 3.10: Quantile regression of *Intercept* and *Gender*



a. Parameter of Intercept

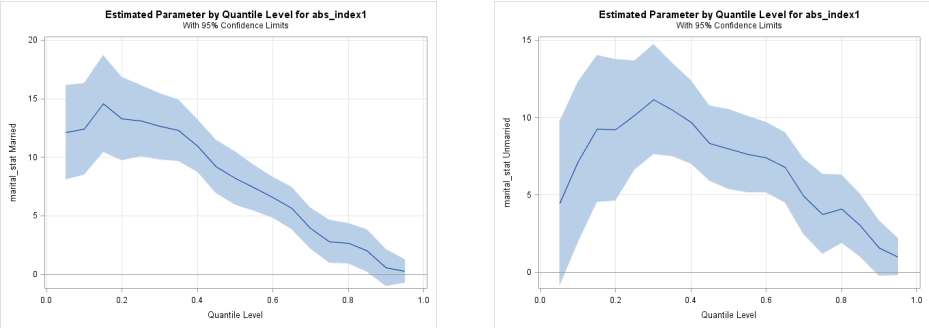


b. Param. Males vs Females

The graph shown in Figure 3.10*a* represents the estimated value of the intercept. It follows the trend of the indicator, and hence it constantly increases among the quantiles. The original range of the indicator is close to 250, precisely $(-90; 166)$. The values of the intercept moves in the interval $(-28; 160)$, almost

linearly with respect to the increase of quantiles. In Figure 3.10*b* the estimated parameters for the men group is plotted. There is strong evidence of the higher level of satisfaction for the male population with respect to the female one. For instance, in the lower quantiles, the men' satisfaction is between 4 and 7 points higher. Furthermore, the level 95% confidence interval computed among the estimates proves the male parameter to be significantly higher than zero in every quantile. At this point, we begin highlighting the shrinking effect on the estimated values, which is a trend that can be found in many cases under study. The positive effect of being male is reduced in proximity of highest satisfaction to less than half of the observed maximum.

Figure 3.11: Quantile regression of variable *Marital Status*

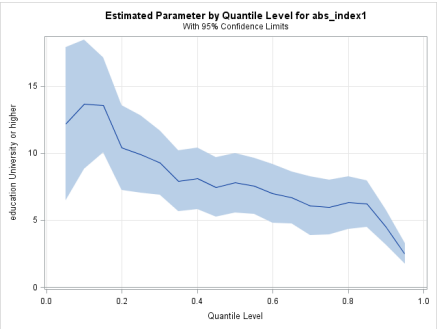


*a.*Parameter of Married *b.* Param. Unmarried

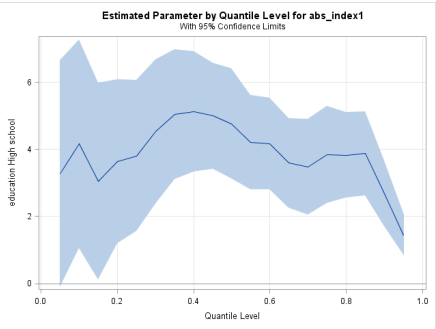
Most of the parameters are significantly different from zero, especially because of the large number of observations used. The effects of the *Marital Status* variable are not surprising: married and unmarried individuals show a higher level of satisfaction with respect to the class of individuals whose relations were interrupted. The difference between married and unmarried individuals is small (Figures 3.11*a* and 3.11*b*). From this data, notice that the group of unmarried people is composed by both singles and individuals that are too young for marriage. From the graphs, it is possible to see how the status of unmarried has a different effect on the satisfaction among the quantiles: it is maximum around the 30-th percentile, where the value is 12.6, then, it decreases constantly to

0.25 where the estimation is not significantly different from zero. This type of fluctuation of the effect is visible only with the use of quantile regression. In this application, we get particular information by this method: at the lower levels of satisfaction, the *Unmarried* effect behaves differently from the *Married* one.

Figure 3.12: Quantile regression of variable *Education*



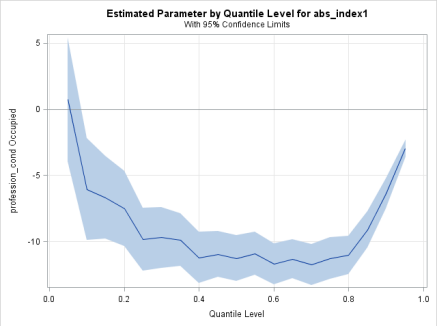
a. Parameter of University



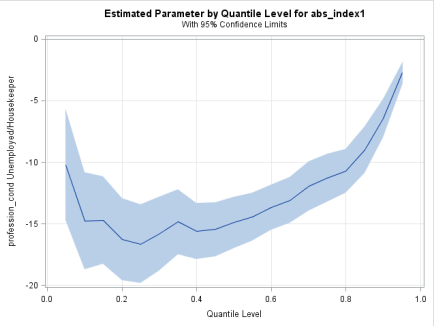
b. Param. High School

The effect of education is sharper: with the increase of education, the median satisfaction grows too. The parameters are evaluated with respect to the level of *middle school* or lower. In figures 3.12*a* and 3.12*b*, the effects are represented showing that *university* is stronger than *high school*, which confirms the ordinal nature of the variable *Education*. The university effect is stronger in the low levels, which means that the less satisfied graduated individuals are significantly (around 14 points) more satisfied than the less satisfied people with a middle or low education. In addition, the differences among the quantiles are visible, but these are certainly not as big as we experienced with *Gender* and *Marital Status*; it probably means that education is useful as a tool for satisfaction at every level and not only for the low level of general satisfaction. The effect of a high school education has the same direction as the graduation, but it has a smaller intensity. Furthermore, the parameter of *High school* reaches its maximum around the 40-th quantile showing a different shape from the *University*'s effect. Hence, the education is certainly a good investment to gain good results and increase satisfaction for everyone.

Figure 3.13: Quantile regression of variable *Professional Condition*



a. Parameter of Employed

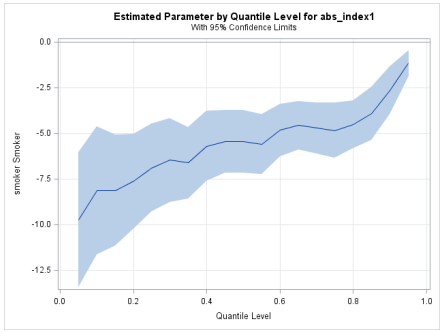


b. Unemployed/Housekeeper

The curves of *professional condition* behave differently from many of the other results: these are not monotone, especially in the case of the *Employed* group (Figure 3.13*a*). This variable has the reference value in *inactive* individuals, mainly retired and a minority part of students. The employed individuals defines themselves as less satisfied than the reference group, and this difference is stronger between the 20-th and the 80-th quantile. Similarly, the group of *Unemployed/Housekeeper* has a lower value of satisfaction, even stronger than the employed group. Maybe it can be interpreted as a positive effect of being retired or a student, simply because it implies being out of working age, with all the sources of stress that can be avoided only with a low or advanced age. Moreover, *Employed* individuals are not significantly more satisfied then the reference group in the first quantile (5-th).

Studying the habits of the individuals, one result shows up significantly: the individuals who belong to the group of smokers show a generalized reduction of the level of satisfaction with respect to non-smokers and those who stopped smoking (Figure 3.14). The smoking habit is not a direct cause of dissatisfaction, but the number of cigarettes is tightly related to satisfaction and happiness. Moreover, this behavior has been found to be a proxy for socio-economical level in many works. In addition, the group of *Non-Smokers* receives a satisfaction boost from the subset of people who actually stopped smoking recently and who experienced the physical and psychological effect. These results are completely in line with the

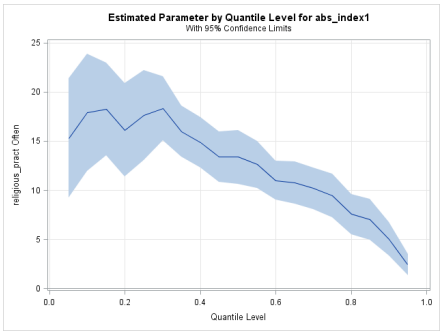
Figure 3.14: Quantile regression of variable *Smoking*



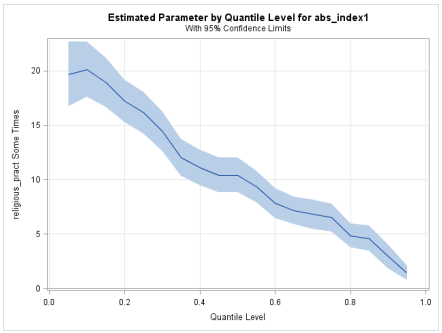
Quant. Param. of Smoker

literature. In their work, Grant et al. [2009] confirmed the association between life satisfaction and health-promoting behavior that is likely to be bidirectional. Regarding ¨quitters¨, in contrast to continuing smokers, an improved subjective well-being has been reported that could be the motivation for the quit attempts by individuals [Piper et al., 2012].

Figure 3.15: Quantile regression of variable *Religious practice*



a. Frequent practice

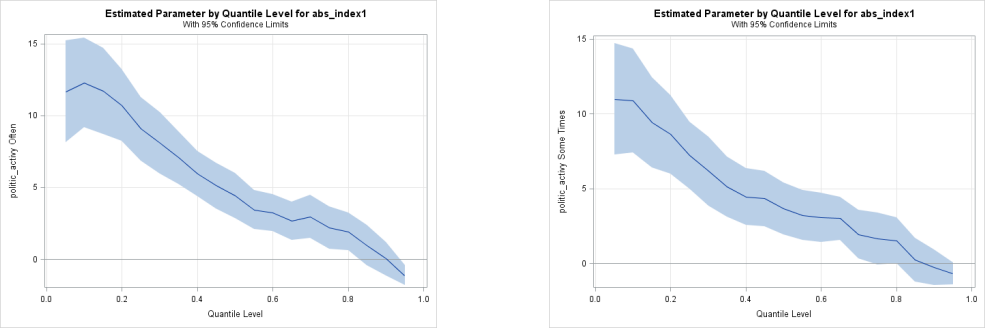


b. Saltuary practice

The effect of religious participation is plotted in figures 3.15*a* and 3.15*b*. The main goal of this type of information is to catch the connection between religiosity and satisfaction, but the limits of a single variable in the description of such a complex concept are straightforward. Hence, using this variable, together with

the *Political Participation* variable, we want to present some perspectives of social inclusion and participation. People who attend the religious services *Often* and *Sometimes* are both more satisfied about life than those who rarely practice. But the rigorous practicing individuals are not more satisfied than the interviewed who visit the religious place once a week or less but more than once a month, as one would expect. To be exact, the confidence intervals of the groups are overlapping at many quantiles.

Figure 3.16: Quantile regression of variable *Political participation*



a. Frequent polit. activity

b. Saltuary polit. activity

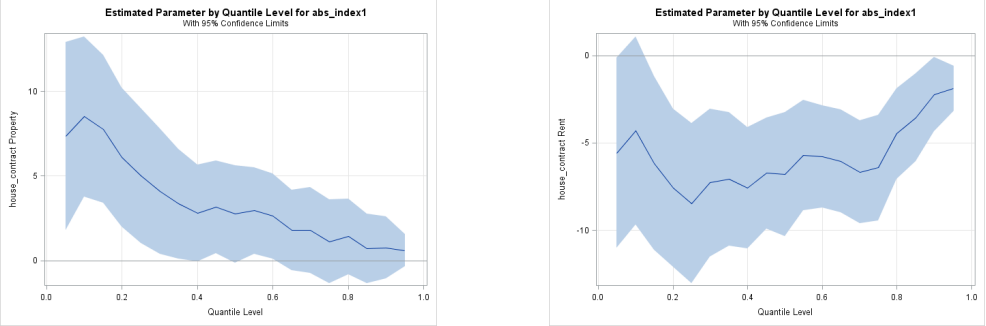
The effects of political participation have the same shape as religious practice but are less intense. The *Often* level (Figure 3.16*a*) and the *Sometimes* level (Figure 3.16*b*) are almost overlapping. This variable is obtained by asking how often the individual talks about politics. Hence, it makes it possible to understand partially the involvement of the individual in society's changes. Differently from religious habits, political activity seems to become lightly negative for satisfaction in the top levels, but the final value is not significant.

This variable together with *Religious practice* are perfect examples to appreciate the utility of quantile regression method: both show the satisfactory effect of participation which is far more important at the lower levels and decreases with the improvement of satisfaction.

There are three explanatory variables that help to model the economic situation of the subjects: the *House type*, the *House contract*, and the *Economic*

changes. The effects of the first variable are represented using the *Public houses* and *Rural Houses* as the reference level. The plotted effects of *Distinguished* and *City houses* are so similar in shape that probably we should use the information to say something about the reference level. In the low level of housing, the subjects declare a lower satisfaction; this difference is stronger between the second and the forth deciles. We can compare average housing with the distinguished housing only by recognizing the larger dimension of the effect observed for the houses of the higher level. The value of the estimates for the group of average city houses is not significantly different from zero in the highest evaluated quantile.

Figure 3.17: Quantile regression of variable *House contract*



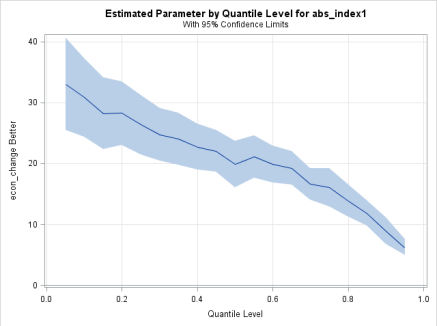
a. Property of the house

b. House for Rent

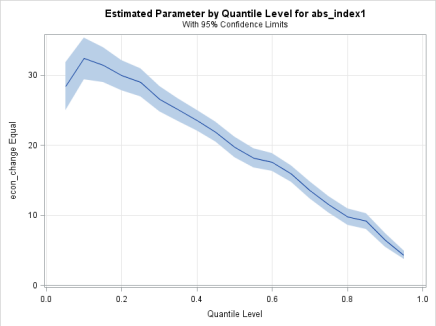
The results given by the *House contract* are more easy to interpret. In Table 3.4, the value *Free use without property* takes the reference role. This group is made by the individuals who own the right to use the house without paying (because it is free for them or because they sold the property with a delayed transaction). The satisfaction of house owners is higher, but this result is significant only in the first three deciles (Figure 3.17*a*). The effect of being a tenant is negative, with a quite constant trend (Figure 3.17*b*); in this case, the significance level is reached only after quantile 0.10.

In the *Economic Change* question, subjects were asked to say if they considered the economic situation of the family improved, stable, or deteriorated with

Figure 3.18: Quantile regression of variable *Economical Change*



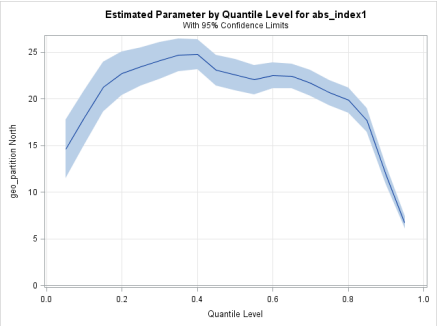
a. Economically improved



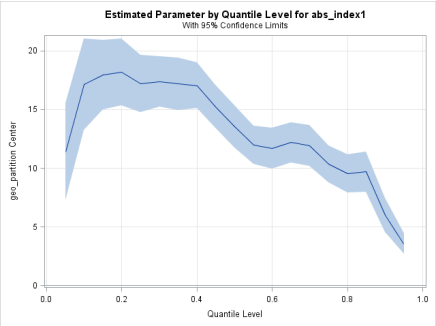
b. Economically equal

respect to the previous year. The reference level is *Worse*, so it is possible to observe the estimated effect of an improved situation, which is presented in Figure 3.18*a*, and the effect of a stable situation (Figure 3.18*b*). The estimated trends for *Better* and *Equal* are almost overlapping in intensity and shape, which means that the real difference is observable between them and the group who perceives its situation as *Worse*. Indeed, the members of this negative class are generally less satisfied, by a large gap, and represent the strongest observed in this work.

Figure 3.19: Quantile regression of variable *Geographical Partitions*



a. Northern Partition

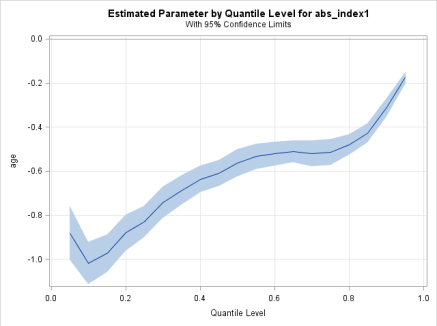


b. Central Partition

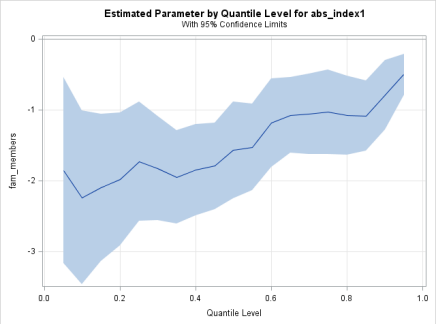
The *Geographical partitions* are a cornerstone in the interpretation of Italian data; the characterization is always strong, and this case follows the rule. The

southern partition is the reference point and the comparison with the *Northern* one is clear (Figure 3.19*a*): the inhabitants of the North are more satisfied among the entire spectrum of observation. Especially in the middle-low levels, the value of the *Northern* parameter is 10% of the entire range of the indicator H_τ, exactly 25 out of 256. People living in central Italy are still more satisfied than the southerners, but the effect is mitigated with respect to the northerners (Figure 3.19*b*). This result respects the knowledge about the socio-economical situation of these partitions.

Figure 3.20: Quantile regression of variables *Age* and *Family Members*



a. Quantile param. of Age



b. Param. of Family Members

As last issue, we want to take into account and show two demographic pieces of information. The *Age* and the number of *Family members* are quantitative variables, hence they are not evaluated with respect to a reference value. The estimated value is the gain/loss of satisfaction experienced by an individual when the considered variable is increased by one unit.
The effect of *Age* is negative on *Life Satisfaction* (Figure 3.20*a*); the age has a range of 90 units, then, in the 1-st decile, where the estimate of the parameter is -1.02, being the oldest of the population would mean to lose more than 90 points with respect to the youngest (since 90 years is the difference between the minimum and the maximum age).

The number of members in the families has a negative effect too, according to

Figure 3.20*b*; in the lowest level, every added member decreases the satisfaction by a value around 2.

3.4 Conclusion

Considering the results of Fattore [2015], we studied some of the features of this method; particularly the threshold, which stands in the center of this model innovation. Starting from the available information a procedure for the construction of a synthetic indicator has been proposed and then used to obtain a measure of a complex and unobservable concept such as *Life Satisfaction*. The result is a unique indicator, that we used to represent the concept under study. The properties of the indicator, allow to study the complex concept with the use of a method such as quantile regression. In this practice, the estimation of the effect of many socio-economical variables on life satisfaction have been enhanced, confirming many result of literature about Satisfaction. It has been possible to estimate the effects of the explanatory variables on the different quantiles of the satisfaction, proving that the external variables act differently at different levels of the response variable.

Chapter 4

HOGS - Height Of Groups by Sampling

This chapter has two important contents: first of all we propose a new approach for the computation of a synthetic measure of an unobservable concept out of ordinal (or mixed) variables, then we proceed with a model to study the effect of explanatory (external) variables on the different levels of such a concept.
These developments start from several issues, that are specific needs due to our field of application: The implementation of profiles' frequency in order to take the distribution into account, the complexity of the poset structure in the case of big datasets, and the dimensional limitation of recent software respect to this topic (see also Sections 2.2.3 and 2.3.2).

In particular, we look for a method able to study a complex concept and to define the relations of the concept with external information. We are interested in the evaluation of the effect of socio-economical variables on the level of the complex concept described by the poset built on the population. To do so, a measure of the concept is the fundamental step, and a criterion to define a relation between this measure and the explanatory variables is the second pillar. Finally, the ability to deal with large sets of data is an important property of this proposal.

4.1 What is HOGS

HOGS is an approach conceived to handle a big set of observations using poset theory with the help of a sampling criterion. This procedure computes the mean of the average ranks of a profile or group of profiles among different samples, and allows the investigation of the relations between the average rank and the explanatory variables. This method is limited to ordinal or nominal explanatory variables at the moment, but only because of computational parsimony.

4.1.1 The HOGS procedure

Let assume to have a population $\mathcal{P}$ represented by the matrix $\mathbf{P}(n, m)$, where n units are observed on m variables $(q_1, \ldots, q_m)$. We organize the first k variables as *internal* variables ($\mathcal{I}$), that are the variables containing the information about the poset. The first k columns of the row i define the $i-th$ **profile** p_i. Moreover, it is common practice to assume all the *internal* variables to be oriented in the same direction; to low values of the variables correspond low values of the latent variable (and consequently of the synthetic indicator). This assumption allows to expect the cograduation between the variables to be always positive or at least null. In the following, we will refer to the remaining $m - k$ variables of $\mathbf{P}$ as *external* ($\mathcal{E}$), because they are not relevant in the definition of the complex concept of interest, but could be, for instance, explanatory variables ($\mathcal{E}$-variables). According to these definitions, the main aim is to investigate the effect of external variables on the poset defined upon the internal ones.

If n is very big, before our research it was not possible to compute the average rank because of computational limits, because many of the available software were developed for small posets cases (see Chapter 5). The approximation *LPOMext* was possible only with the use of a program written in Python called PyHasse [Bruggemann and Voigt, 2009]. The procedure of approximation used from the local partial order models finds its weakness in the complexity of the structure of the poset. The poset observed on a huge amount of units is probably very complex, because it contains with high probability almost every possible profile that

can be found observing the *internal* variables (the concept of ¨possible profile¨ is intended as the product order of internal variables measured on ordinal scale). Computing the average rank (even the exact one) on a poset that contains every possible node, leads to a weak differentiation among the profiles of the same level, therefore the faster function that we developed for the approximation (see Chapter 5) is still not enough. In the following, the concept of complexity of a poset will be widely used. In order to share the meaning of the structure's complexity we propose a general example. Let the number of observations to be constant, and consider different posets with their number of incomparabilities $u = |U(P)|$. Then, complexity is very low in complete orders (like chains) where $u = 0$, and very high in the case of $u = n(n-1)/2$, that corresponds to the anti-chain case. Clearly, the observed complexity is never so neat, all the intermediate shades of complexity can be observed.

In the following, we explain every step of the HOGS procedure. Finally, the algorithm is described with a pseudo code.

To recap, when n is too big, we will refer to sub-samples. We need to define n^*, the optimal number of units in every sample. At the moment n^* is a value defined in the interval $(100; 200)$ following the usual limitation followed by other existing software, but, thanks to the development achieved during this research, these limits are going to be largely relaxed. One of the next steps in this research will be the definition of an optimal value for n^*.
The HOGS procedure follows the following steps:

i Define whether n is too big for *LPOMext* (Is $n >> n^*$?);

ii Sample n^* units from the population with a simple random sampling procedure without replacement to form the sample s_i. The same units can be sampled in different samples;

iii Observe the poset based on s_i and compute the average rank of every profile in the sub-sample;

iv Divide the sample s_i in groups, according to the levels of a grouping criterion

(the criterion is an external variable, a profile or a group of profiles like a cluster). Estimate the *medium average rank* ($\mathcal{H}_\}$), that is the mean of the average ranks of the elements belonging to each group.
The definition of the external grouping criteria has a huge relevance.

v Repeat steps $ii - iv$ until a stopping criterion is reached and at least all observations have been observed once;

vi Compare the results of the groups obtained by all the samples, by means of statistical tests.

Every step needs a specific care.

i) **Dimension of the population**

The dimensional problem, given by the number of observations in a poset, has been introduced in 2.2.3 and 2.3.2: the number of observed elements determines the width of a poset and, consequently, increases the probability of incomparabilities. This augmented complexity decreases the correctness of approximation procedures because it increases the amount of complex structures, multiplying the *combinatorial effect* described in the introducing Section 2.2.3 and in Brüggemann and Patil [2011].

ii) **Sampling of sub-sets**

Sampling at random from the original population, the most frequent profiles will be observed more frequently, having more representation, and the less frequent ones will be often absent.

Observing a sample, only a subset of the poset P is observed and the structure is easier. For instance, if we assume to observe the poset represented in Figure 4.1, were the darker profiles are more frequent, sampling $n^* = n/2$ elements; we can expect to observe most of the *black* elements and a few of the *white* ones. Hence, the resulting poset for the sample s_i will have an easier structure (Figure 4.2). Along with the reduced complexity of the poset also the approximation is improved. The difference in the probability of extraction among the profiles will determine a regularity in the structure of the observed sub-posets. An example of poset built on the sample is represented in Figure 4.2, where the degrees of

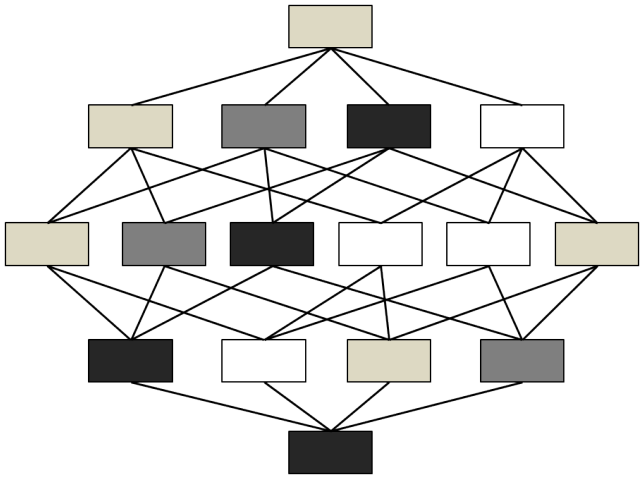


Figure 4.1: Distribution of observations among the poset. Darker implies More frequent

grey are used just to recall the profiles from the previous figure.

This effect of ¨auto-selection¨ of the most frequent profiles is almost sure in every real application where the elementary variables are not homogeneously distributed. Nevertheless, there is a need to formalize exactly the probability of simplification of the sub-poset built on the sample; it is one of the most important future development for this approach.

iii) **Average rank of the sample** Once a sample is observed, the relative poset is mapped onto a linear order with the *LPOMext* method. Therefore, every element $x_j \in s_i$ has an associated value of the approximated average rank (H_i), like the elements in the example (Table 4.1).

s_i	x_1	x_2	$\cdots$	x_j	$\cdots$	x_{n^*-1}	x_{n^*}
$H_i(\cdot)$	3.3	2.7	$\cdots$	5.2	$\cdots$	1.4	4.6

Table 4.1: Example of *LPOMext* on a sample

The results of the approximation procedure can assume values in the range $(1; n^*)$. Unfortunately, according to the properties of the *LPOMext* approach,

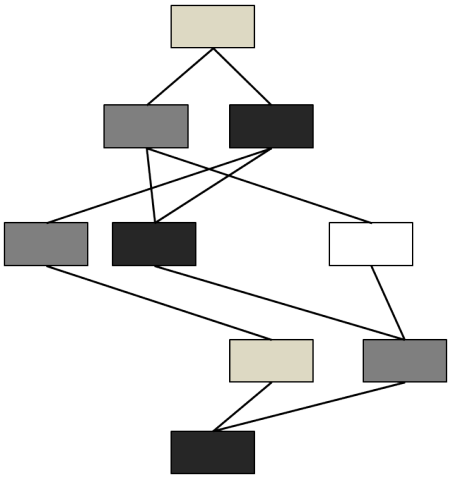


Figure 4.2: Distribution of observations among the sub-poset.

if two units show the same profile, only one representative element is selected: decreasing the maximum observable value for the average rank (see 2.2.3).
This limit can be handled in two ways:

1. Rescaling the values of $H(\cdot)$, to impose the range $(0, 1)$:

$$\tilde{H}(x_j) = \frac{H(x_j) - 1}{(n^*_{s_i})}$$

   where $\tilde{H}(\cdot)$ is the re-scaled value of the approximated average rank and $n^*_{s_i}$ is the number of different profiles observed on the sample s_i.

2. Implementing the frequency of equivalent profiles in the formula of *LPOMext*; in our opinion this is the best solution, and a proposal for this implementation is described in Chapter 5. For sake of simplicity, in the following we will simply assume that every sample contains n^* different profiles.

iv) **Selection of entities for aggregation** In order to select the external information that defines the levels of aggregation, two conceptual paths are pos-

sible: the $\mathcal{E}-$variables and the observed profiles. The choice defines the rules and length of the sampling procedure:

$\mathcal{E}-$Variables. Aggregation by level .

Example: Estimate and compare the average rank of females and males in every sample, respectively $\mathcal{H}_{iF}$ and $\mathcal{H}_{iM}$ ($\mathcal{E}-$variable: gender). More generally, assuming an external variable q_i with e levels: the elements of the sample s_i are divided in e sub groups and the values of $H(\cdot)$ are observed on the elements of every group.
Good features:

- The required time is usually small, if the number of levels is not too high. Since the number of levels is sufficiently smaller that n^*, every single sample will constitute a sufficient population to test the difference of effect among the levels.
- It allows to estimate effect of variables interactions if the level of estimation is defined on their combination (male worker, female worker, male non-worker, female non-worker).

Bad feature:

- The procedure needs to be repeated from the beginning for every external variable of interest, this could be annoying with the current software. This problem is mainly overcame by the computation speed of the HOGS function, thanks to the routines presented in Chapter 5.

s_i	x_1	x_2	$\cdots$	x_j	$\cdots$	x_{n^*-1}	x_{n^*}	$\overline{x}_M = \mathcal{H}_{iM}$	$\overline{x}_F = \mathcal{H}_{iF}$
$H(\cdot)$	3.3	2.7	$\cdots$	5.2	$\cdots$	1.4	4.6	3.6	2.3
gender	M	F	$\cdots$	M	$\cdots$	F	M		

Table 4.2: Example of *LPOMext* by level of gender on a sample

Aggregation by profile .

Example: If the set $\mathcal{I}$ is constituted by two dichotomous variables, the procedure can aggregate the average ranks by the possible profiles: $00, 01, 10, 11$.

If the number of profiles that can be observed according to internal variables is finite, the sampling procedure can be oriented to find an efficient estimation of every profile, with a stopping criterion focused on the convergence of the estimation.
In this case we can assess that the sampling is based on *internal* variables ($\mathcal{I}$).
Good feature:

- The computation of the profiles' value is performed only once, the result of this approach is the average rank of profiles, a value of that profile on the latent variable. It fits well our aim in the research on a synthetic indicator.

Bad features:

- The amount of profiles to estimate could be very large, causing problems to find convergence.
- This approach is limited to $\mathcal{I}$ sets made by ordinal or discrete variables with low number of levels; the use of variables with many levels will determine a problem of estimation, in case of continuous variable it will converge to the unit aggregation.

s_i	x_1	x_2	x_3	x_4	$\cdots$	x_{n^*-1}	x_{n^*}
$H(\cdot)$	3.3	2.7	2.7	5.2	$\cdots$	3.3	4.6
profile	120	111	111	221	$\cdots$	120	112

Table 4.3: Example of *LPOMext* by profile on a sample

The type of comparisons and statistical methods used to analyze the results depends on the chosen level of estimation.
In the following the focus is imposed to the **aggregation by level**, because it fits better for the investigation of the effect of external variables on the ranks of units.

v) **Stop criterion** The stop criterion is an important part of the HOGS algorithm, because the quality of the estimation depends on it.
In the following, the *Level* aggregation is presented; in such a case, the values $\mathcal{H}_{ig}$ are the aim of the estimation, where $g \in \{\text{observable levels of the } \mathcal{E} \text{ variable}\}$. Before any further steps, we want to underline an important result coming from this aggregation procedure. Let x_{ijg} represents the j-th element observed in the i-th sample, and which has been found to show the g-th level of the external variable used for the aggregation; we assume x_{ijg} to be a realization of the random variable X_g.
X_g is the random variable that represents the mean of the average ranks of the elements of the $g-th$ group and n_{ig} is defined as the number of elements of the group g in the sample i.

We assume that $\{X_g^s\} = (X_{1g}, \ldots, X_{ig}, \ldots, X_{sg})$ is a sequence of independent and identically distributed, real valued random variables. Then, if s is sufficiently big, according to the Central Limit Theorem [Polya, 1920]: the distribution of the sample mean is normal, with mean and variance respectively equal to μ_g and σ_g^2/s;

$$\overline{X}_g^s \backsim N(\mu_g, \sigma_g^2/s). \tag{4.1}$$

In this case, we can assess that: if the number of samples (s) large enough, the aggregated value $\mathcal{H}_{ig}$ is a realization of the normally distributed random variable $\overline{X}_g^s$.

The stop criterion could be defined according to this result. Following the widespread heuristic relative to the central limit theorem, the normal distribution of the sample mean is reached if s, the number of samples for every group, is sufficiently big. Usually the number 30 is considered big enough.

Hence, we define 30 as the minimum number of observations for every group g, if a group is not observed in a given sample then we need to observe more samples as long as the number of observations for every group (s_g) is bigger than the heuristic value, $s_g \geq 30, \forall g \in \mathcal{E}$.

Furthermore, the procedure is required to observe every element of the population of the survey under study, because it often represents the entire population. The procedure stops when both these criteria are reached.

The structure of the external variable $\mathcal{E}$ has a central role, its frequency distribution influences directly the computational time of the procedure; this is the reason why we do not suggest to use the HOGS procedure with $\mathcal{E}-$variables measured on the quantitative scale.

vi) **Comparison of results** At the end of the procedure, the results are collected in the HOGS matrix $\mathcal{H}(e, s)$. The dimensions of the HOGS matrix are defined by the number of levels of the external variable (e) and the number of observed samples (s).
So, according to the results on the distribution of the aggregated values $\mathcal{H}_{ig}$, the levels are compared with tests on the difference of means. For instance, in the case represented in Table (4.4), the effects of being male or female can be compared using the $t-$test. Future developments will implement the comparison of levels through the tests on ranks.

s	s_1	s_2	$\cdots$	s_i	$\cdots$	s_{s-1}	s_s	$\mathcal{H}_g$
$\mathcal{H}_{iM}$	3.6	3.7	$\cdots$	5.6	$\cdots$	1.8	4.3	3.8
$\mathcal{H}_{iF}$	3.0	2.9	$\cdots$	4.6	$\cdots$	3.1	3.9	3.5

Table 4.4: Example of the HOGS table $\mathcal{H}$

4.1.2 The algorithm

Here, the HOGS procedure is described in the simplified form of pseudo-code:

Algorithm 1 HOGS algorithm

procedure HOGS($\mathcal{I}, n^*, \mathcal{E} - variable$)
  levs $\leftarrow e$ ▷ Number of levels of the $\mathcal{E} - variable$
  nobs $\leftarrow n$ ▷ Number of units in the population
  UnitCheck $\leftarrow$ Vector(0, nobs) ▷ 1 if a unit has been observed
  GroupCheck $\leftarrow$ Vector(0, levs)
    ▷ Counter, number of samples in which each group is observed
  $\mathcal{H} \leftarrow Matrix(e, s)$
  for $i \leftarrow 1$ to realization of criteria (Units & Groups) **do** ▷ see 4.1.1
    $s_i \leftarrow i$-th sample $\subset \mathcal{I}$ ▷ The samples are subsets of $\mathcal{I}$
    function LPOMext(s_i)
      $H_i \leftarrow$ Approximation on s_i ▷ LPOMext is computed on s_i
    end function
    for $g \leftarrow 1$ to levs **do**
      $\mathcal{H}_{gi} \leftarrow$Mean($h_{gi}$) ▷ $h_{gi} = \{h_i \in H_i | h_i \in$ Group $g\}$
    end for
  end for
  Return $\mathcal{H}$ ▷ which contains s means computed on e levels
end procedure

4.2 Evolution of life satisfaction in Italy

Data are the same presented in Section 3.3.1. As well as before, the set $\mathcal{I}$ is constituted by the four ordinal variables of life satisfaction: *Economy*, *Health*, *Family* and *Time*. Here we present the results of HOGS procedure on the level of life satisfaction with respect to the external variables gender, age and geographical region. According to their definition, all the elements of the HOGS matrix have value in the interval $(0, 1)$.

The results obtained using *gender* as external variable are represented in Figure 4.3: in the picture the time series of satisfaction by gender are plotted. The time series refer to the period covered by the national survey from 1993 to 2012

(2004 is missing).

The line representing females is clearly shifted to the bottom with respect to the male's one. It confirms the result observed in the previous chapter, where the quantile regression shows higher satisfaction among men than women. The confidence bands of the two groups are narrow, showing a statistically significant difference between the groups.

Figure 4.3: Index of life satisfaction by gender. Italy, 1993-2012.

Most of the socio-economical effects are influenced by the age of the respondents. In Figure 4.4, we can see the large difference among four age classes: $(13; 34]$, $(34; 50]$, $(50; 64]$, and $(64; 105]$. These time series proof statistically the dependence of satisfaction by age. The interesting result is in the dimension of the difference between the young individuals and the rest of the population. Furthermore, a strong decrease can be observed in the population of over 64 corresponding to the passage from 2002 and 2003.

With the use of the HOGS procedure, we also estimated the level of satisfaction using the region as external variable.

Figure 4.4: Index of life satisfaction by age class. Italy, 1993-2012.

The Figure 4.5 shows the time series for four Italian regions: Trentino- Alto Adige, Veneto, Lazio, and Campania. *Trentino - A.A.* is the best performing region, this region has an outstanding result respect to all the others. The rest of Italy could be divided in three sections, that mime the usual partition of *North-Center- South*, within the partitions the confidence bands are often overlapping. *Veneto* represents the northern partition, that is the most satisfied. The central partition (represented by *Lazio*) is usually lower than the northern, and only in some years the confidence bands touch each other. It implied an interesting effect of time, that is different among the regions. The time effect is evident in the time series of *Campania*, that shows a generalized lower level of satisfaction among the years, characterized by a strong increment after year 2002, and an even stronger reduction after 2009.

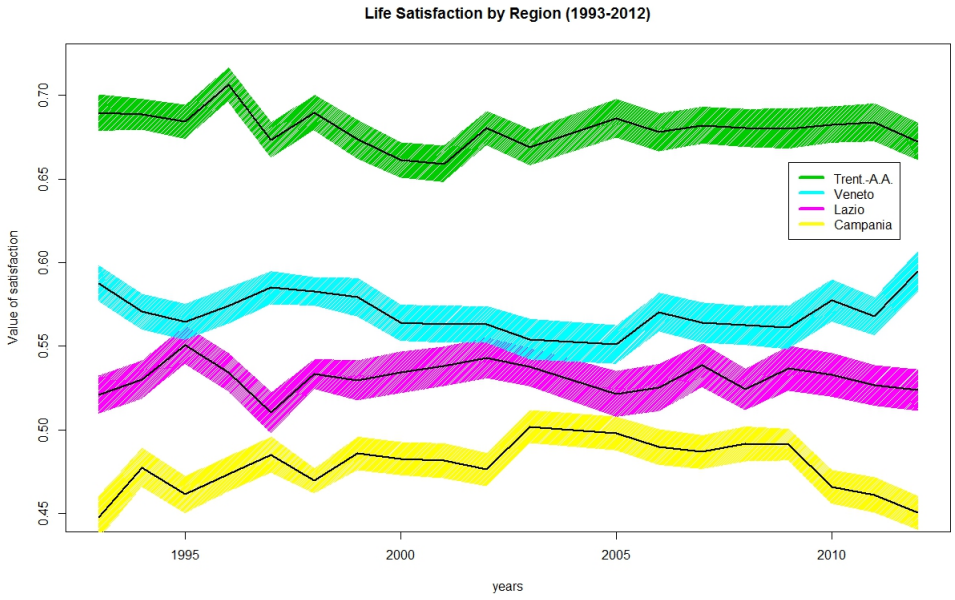


Figure 4.5: Index of satisfaction by region. Italy, 1993-2012.

4.2.1 Conclusion

The use of this new procedure finds its motivation in the large dimension of the datasets. Instead of engorging the basic approximation procedures, that are expected to handle small populations, it decomposes the problem in s samples achieving the following steps:

- the approximation error decreases because of the reduced complexity of the poset referred to the sub-samples;
- the frequency of the profiles is considered and exploited, the most frequent profiles are observed in many different samples, increasing the robustness of the estimates;
- the approximation error depends on the structure of the poset and it is not systematic. Aggregating the results of different samples in a unique test, we obtain a method to measure the distribution of the error and test statistically the observable differences.

Thanks to its construction, HOGS procedure allows to relate a profile made by multiple variables, to a single value that represents a complex concept without the use of assumptions on the distribution of the variables. With the estimation by level, in particular, it is possible to manage the relation between ordinal explanatory variables and multiple response variables. The further development of this model looks promising and challenging, especially for: the definition of samples size, the estimation of interactions among the explanatory variables and the development of the profiles estimation.

Chapter 5

The computation of approximated average rank in large datasets

5.1 The Average Rank with large datasets

In social science it is common to deal with big amount of data. Therefore we recognize the need of obtaining a fast tool in order to handle tens of thousands of observations without requiring too much memory from the supporting machine. Such a computational tool is the corner stone for the novel methodologies proposed in this thesis and for the research on approximation of average rank for social statistics.

In the present chapter, we deal with the computation of the approximated average rank, which has been introduced in Section 2.2. Among the software for poset-related computations, PyHasse [Bruggemann and Voigt, 2009] is considered one of the most complete. Nonetheless, PyHasse (from *Python* and *Hasse diagram*) has been developed in the framework of chemometrics with a focus on graphical representation of the Hasse diagrams. Hence, it is not meant to deal with many observations.

We developed a function (*R-LPOMext*) with the software `R`, that is able to

perform LPOMext approximation for bigger sets of data.

5.1.1 Characteristics of the procedure

The function R-LPOMext is a simple procedure. Nevertheless, many tools may be obscure for the usual `R`-user, because are obtained from the recently developed package called `PARSEC` (*PARtial order in Socio-EConomics*), that has been presented by Fattore and Arcagni [2014]. This package contains several instruments for the analysis of posets, in the following we list the ones that are part of our procedures.

`pop2prof`: translates the matrix made by internal variables ($\mathcal{I}$) in the correspondent list of observed profiles, the information is enriched with the frequencies of profiles in the population described by $\mathcal{I}$;

`getzeta`: computes the zeta matrix, a squared boolean matrix were the list of profiles describes both the rows and columns. The boolean value is true if the row profile is lower or equal than the column profile;

`downset(upset)`: returns a boolean vector indicating which poset's elements are equal or below(above) at least one element of the function's argument (see Section 2.1.2.3);

`incomparability`: returns a boolean matrix (as the zeta metrix) whose elements are TRUE when row and column profiles are incomparable.

5.1.2 The algorithm of R-LPOMext

R-LPOMext has no dimensional limitations but those imposed by the memory used for function `getzeta`. Every step of the algorithm uses as less memory as possible, because we accept a slower procedure in order to obtain the ability to manage larger datasets; this is the reason why the `apply` functions are missed while simpler and slower `for` and `while` cycles are used.

The R-LPOMext algorithm is described in Algorithm 2. We add here some tips in order to make the reading of the pseudo-code faster.

The main source of confusion in the code is the difference between theoretical profiles, observed profiles, and profiles of the observations.

Once the population is observed, there are n rows in the matrix of internal variable $\mathcal{I}$, every statistical unit has an observed profile that can be equal to the profile of other units. The number of theoretical profiles is equal to the multiplication of the number of levels for every $\mathcal{I}$-variable as introduced in Section 2.3.1.1. So, the observed profiles could be a subset of the theoretical ones, and their frequency depends on the profiles shown by all the observations.

For instance: let observe a population of 7 individuals on two internal variables measured on three levels $\{1, 2, 3\}$. The theoretical profiles are 9, given by every possible coming from the composition of the two internal variables (`11,12,21,22,...`). In Table 5.1 we can see the observed profiles. `11,23,31,12` are the observed profiles, and their frequency is always 2, excluding `12` that has been observed once.

Observation	Observed Profile
x_1	11
x_2	23
x_3	31
x_4	23
x_5	12
x_6	11
x_7	23

Table 5.1: Example of observed profiles

In the R-LPOMext algorithm the frequencies of profiles are not taken into account, because in LPOM approaches the frequencies of profiles are not used for the approximation of the average rank.

In row 2, every observation determines a profile according to its internal variables; the aim of R-LPOMext is to obtain a vector of approximated average ranks (lpom) which elements correspond to the elements of the vector ¨strings¨. The

Algorithm 2 R-LPOMext algorithm in R

```
1: procedure R-LPOMext(J)
2:     strings ← toString(J)                    ▷ Vector of profiles of Observations
3:     lpom ← Vector(n)
4:     poset ← pop2prof(J)                      ▷ It is the poset of the observed population
5:     n_p ← Number of profiles
6:     Z.pop ← getzeta(poset)                   ▷ Builds the zeta matrix of order relations
7:     incom ← incomp(Z.pop)                    ▷ Matrix of incomparability between profiles
8:     for p ← 1 to n_p do
9:         down_p ← downset(p)
10:        up_p ← upset(p)
11:        height_p ← |down_p|
12:        incom_p ← p-th row of incom          ▷ List of profiles incomparable to p
13:        for every element i ∈ incom_p do
14:            effect_i ← π_d/(π_d + π_u)|_i    ▷ see Formula 2.1
15:            height_p=height_p+ effect_i
16:        end for
17:        lpom[p] ← height_p
18:                ▷ height_p is assigned to all the observations with profile p
19:    end for
20:    return lpom
21: end procedure
```

statements down_p, up_p, and height_p are the realization of the different parts of the LPOMext formula. The effect of every incomparable element is evaluated step by step in a recursive sum.

Future steps This procedure is extremely fast and allows the use of datasets made by tens of thousands of observations, but still, it needs improvements. First of all, we want to implement the possibility to use other approximation formulas (see 2.2 as example).

5.2 The use of profiles' frequency

In the recent development of procedures based on poset theory, the repetition of the same profile is usually treated as a single *equivalence class*, without any

information about its frequency. This illusory lack of information is, oppositely, a precise choice, due to the development context of these procedures. Many of the advances in the application of poset theory has been proposed in the science of chemometrics, where small samples of units are measured on precise quantitative scales.

In the context of this research, the ordinal scale of measure and the large amount of information impose the opposite choice; the information given by the frequency of profiles must be used in the evaluation of average rank. This is the aim of this section.

The procedures of approximation of the average rank are based on the observed profiles: if two or more observations show the same one, these are grouped in a unique equivalence class, becoming a unique element in the computation of the average rank.
Example: Lets take a chain made by 3 elements ($a < b < c$), with frequency described in Figure 5.2; in the same table is possible to see the value of the approximated average rank $H(x)$ (in this example it is equal to the exact average rank).

Profile	Frequency	$H(x)$
a	2	1
b	5	2
c	3	3

Table 5.2: Frequency and rank for the example set

The proposal is to take into account the frequencies of the profiles $\{a, b, c\}$, and compute the approximated average rank considering every profile for its dimension. The rank of the elements of a profile is the range of ranks occupied by the profile's members (see Figure 5.1).

According to this approach, we define the average rank of the elements of a profile as the middle value of the range of ranks, as described in Table 5.3.

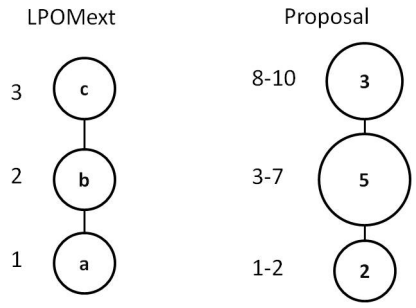


Figure 5.1: Ranks of the elements of the example chain

Profile	Frequency	$H(x)$	new measure
a	2	1	1.5
b	5	2	5
c	3	3	9

Table 5.3: Frequency, LPOMext and new approximation for the example set

The rank of all the elements of a profile in a chain is determined by:

1. The number of observations in the downset (and not the number of profiles in the downset),

2. The number of observations in the profile itself: the observations within the same profile are assumed to be uniformly distributed inside the equivalence class.

The classic LPOMext procedure computes the height of profile x as the average number of lower *profiles*. This proposal defines a slightly different height, given by the average number of *observed units* that are lower than x.
This example may appear too simple, because it refers to a linear order, but the concept of average rank is completely based on linear orders (linear extension); we propose to assume this interpretation of rank to every linear extension, impressing a larger meaning to the concept of average rank.

5.2.1 LPOMext with frequencies

The basic formula of LPOMext is composed by the *starting height* of the profile x given by the cardinality of the set ($|O(x)|$), and the sum of the *effects of incomparable elements* ($\sum_{y \in U(x)} \frac{\eta_d(y)}{\eta_d(y)+\eta_u(y)}$).
In our proposal we want to define a generalization of the LPOMext method, not a brand new approximation procedure.

The starting height is given by the number of observations in $|O(x)|$ and the sum of ¨effects of incomparables¨ is not changed, not even in the indices (that are representing profiles as usual, not the observations). The only difference is in the amount of every addend of the sum, in this proposal the number is not one for every addend (as in LPOMext), it is $f(y)$ that is the number of units observed in the equivalence class of y ($y \in U(x)$).
Then, given the LPOMext formula 2.1;

$$H_{av}(x) = |O(x)| + \sum_{y \in U(x)} \frac{\eta_d(y)}{\eta_d(y) + \eta_u(y)},$$

we propose

$$H^*_{av}(x) = F\{O(x)\}^* + \sum_{y \in U(x)} \frac{\eta_d(y)}{\eta_d(y) + \eta_u(y)} f\{y\}, \tag{5.1}$$

where $F\{O(x)\}^*$ represents the cumulative frequency of all the elements of the downset of x. We call this model *Local Partial Order Model for Observations* (*LPOM-O*). The first part of the formula determines how to measure the starting height of a profile x which frequency is considered only by a half $F\{O(x)\}^* = F\{O(x) \setminus x\} + 1/2 f\{x\}$, for the determination of the starting height. Thanks to the assumption of uniform distribution for the observations inside the profiles, we propose to consider the position of every observation to be equal to the middle point of its profile range of rank. This is the reason why the frequency of the observed profile is reduced by an half in the computation of the cumulative $F\{O(x)\}^*$, because it coincides with the assumption of uniform distribution (see Table 5.3). Clearly, other functions of the downset can be defined for $F\{\cdot\}^*$,

making the LPOM-O formula even more generalizable.

5.2.2 The algorithm for LPOM-O

The LPOM-O algorithm that we developed in `R` is described in Algorithm 3. The general structure of the algorithm is not changed respect to R-LPOMext but, it contains three fundamental differences.
In row 5 the frequency of the profiles is considered. This passage allows the computation of the frequency for every profile, this is the fundamental change. The information of `freq.prof` is used for the definition of starting height (row 12) and the weight of the effect (row 15).
The values of `height` and `effect` are computed according to 5.1, they do not determine some changes respect to the R-LPOMext algorithm's organization.

The result of this algorithm is the vector `lpom-o`, which has length n equal to the number of observations of the dataset. This output is ready to be attached to the data where it comes from because the order of observations is respected. This naive features implements the usability of the function for further analysis.

As a final remark: the information supplied by the frequency depends completely on the characteristics of the internal variables, it impresses to the average rank something unique that is dependent on the population distribution. Actually it is one of our main aims, and a fundamental feature for a method used in social statistics.

Algorithm 3 LPOM-O algorithm in R

procedure LPOM-O($\mathcal{I}$)
  strings $\leftarrow$ `toString`($\mathcal{I}$) ▷ Vector of profiles of Observations
  lpom-o $\leftarrow Vector(n)$
  poset $\leftarrow$ `pop2prof`($\mathcal{I}$) ▷ It is the poset of the observed population
  freq.prof $\leftarrow f(p),\ \forall p \in P$ ▷ Frequencies of Profiles
  $n_p \leftarrow$ Number of profiles
  Z.pop $\leftarrow$ `getzeta`(poset) ▷ Builds the zeta matrix of order relations
  incom $\leftarrow$ `incomp(Z.pop)` ▷ Matrix of incomparability between profiles
  for $p \leftarrow 1$ to n_p **do**
    down$_p \leftarrow$ `downset`(p)
    up$_p \leftarrow$ `upset`(p)
    height$_p \leftarrow F\{O(x)\}^*$ ▷ Frequency of downset
    $U_p \leftarrow p$-th row of incom ▷ List of profiles incomparable to p
    for every element $i \in U_p$ **do**
      effect$_i \leftarrow \frac{\eta_d(y)}{\eta_d(y)+\eta_u(y)} f\{y\}|_i$ ▷ see Formula 5.1
      height$_p$=height$_p$+ effect$_i$
    end for
    lpom-o[p] $\leftarrow$ height$_p$
      ▷ height$_p$ is assigned to all the observations with profile p
  end for
  return lpom-o ▷ A vector containing a value for every observation
end procedure

Bibliography

Alkire, S. and Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of public economics*, 95(7):476–487.

Arrow, K. J. and Raynaud, H. (1986). Social choice and multicriterion decision-making. *MIT Press Books*, 1.

Birkhoff, G. (1948). *Lattice theory*, volume 25. American Mathematical Society, New York.

Brightwell, G. and Winkler, P. (1991). Counting linear extensions. *Order*, 8(3):225–242.

Brüggemann, R. and Carlsen, L. (2011). An improved estimation of averaged ranks of partial orders. *MATCH Commun.Math.Comput.Chem*, 65:383–414.

Bruggemann, R. and Carlsen, L. (2012). Multi-criteria decision analyses. viewing mcda in terms of both process and aggregation methods: Some thoughts, motivated by the paper of huang, keisler and linkov. *Science of the Total Environment*, 425:293–295.

Brüggemann, R., Lerche, D., and Sørensen, P. B. (2003). First attempts to relate structures of hasse diagrams with mutual probabilities. *Order Theory in Environmental Sciences*, page 7.

Brüggemann, R. and Patil, G. P. (2011). *Ranking and Prioritization for Multi-indicator Systems*, volume 5. Springer.

Brüggemann, R., Sørensen, P. B., Lerche, D., and Carlsen, L. (2004). Estimation of averaged ranks by a local partial order model#. *Journal of chemical information and computer sciences*, 44(2):618–625.

Bruggemann, R. and Voigt, K. (2009). Analysis of partial orders in environmental systems applying the new software pyhasse. In *Wittmann Jochen, Flechsig Michael, Simulation in Umwelt-und Geowissenschaften, Workshop Potsdam*, pages 43–55.

Davey, B. A. and Priestley, H. A. (2002). *Introduction to lattices and order*. Cambridge University Press, New York.

De Loof, K., De Baets, B., and De Meyer, H. (2011). Approximation of average ranks in posets. *MATCH- Commun.Math.Comput.Chem.*, 66:219–229.

Decancq, K. and Lugo, M. A. (2013). Weights in multidimensional indices of wellbeing: An overview. *Econometric Reviews*, 32(1):7–34.

Dyer, M., Frieze, A., and Kannan, R. (1991). A random polynomial-time algorithm for approximayion the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17.

Fattore, M. (2015). Partially ordered sets and the measurement of multidimensional ordinal deprivation. *Social Indicators Research*, pages 1–24.

Fattore, M. and Arcagni, A. (2014). Parsec: an r package for poset-based evaluation of multidimensional poverty. In *Multi-indicator Systems and Modelling in Partial Order*, pages 317–330. Springer.

Fattore, M., Brüggemann, R., and Owsiński, J. (2011). Using poset theory to compare fuzzy multidimensional material deprivation across regions. In *New Perspectives in Statistical Modeling and Data Analysis*, pages 49–56. Springer.

Grant, N., Wardle, J., and Steptoe, A. (2009). The relationship between life satisfaction and health behavior: a cross-cultural analysis of young adults. *International Journal of Behavioral Medicine*, 16(3):259–268.

Karzanov, A. and Khachiyan, L. (1991). On the conductance of order markov chains. *Order*, 8(1):7–15.

Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50.

Lehmann, E. L. and Romano, J. P. (2011). *Testing statistical Hypotheses*, volume 5. Springer.

Lerche, D. and Sørensen, P. B. (2003). Evaluation of the ranking probabilities for partial orders based on random linear extensions. *Chemosphere*, 53(8):981–992.

Loof, K. D. (2009). *Efficient computation of rank probabilities in posets*. PhD thesis, University of Ghent.

Maggino, F. and Fattore, M. (2011). New tools for the construction of ranking and evaluation indicators in multidimensional systems of ordinal variables. *Proceedings of the'New Techniques and Technologies for Statistics', Brussels*.

Munda, G. (2008). *Social multi-criteria evaluation for a sustainable economy*. Springer.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., and Giovannini, E. (2005). Handbook on constructing composite indicators. *OECD statistic working paper*.

Patil, G. P. and Taillie, C. (2004). Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking and prioritization. *Environmental and Ecological Statistics*, 11(2):199–228.

Piper, M. E., Kenford, S., Fiore, M. C., and Baker, T. B. (2012). Smoking cessation and quality of life: changes in life satisfaction over 3 years following a quit attempt. *Annals of Behavioral Medicine*, 43(2):262–270.

Polya, G. (1920). On the central limit theorem of calculus of probability and the problem of moments. *Math. J., German*, 8(3):e4.

Saltelli, A. (2007). Composite indicators between analysis and advocacy. *Social Indicators Research*, 81(1):65–77.

Schröder, B. (2012). *Ordered sets: an introduction*. Springer Science & Business Media.

Sen, A. (1976). Poverty: an ordinal approach to measurement. *Econometrica: Journal of the Econometric Society*, pages 219–231.