Doctoral Program in Economics

UNIVERSITÀ DEGLI STUDI
FIRENZE

UNIVERSITÀ DI SIENA

UNIVERSITÀ DI PISA

Department of Economics and Statistics

Doctoral Thesis in Economics

Cycle XXXIII

Coordinator Prof. Michelangelo Vasta

Three Essays on the measurement of socioeconomic inequalities and well-being

Scientific-disciplinary sector: SECS-P/01

PhD student: Giovanna Scarchilli

Supervisor: Prof. Paolo Brunori

Academic Year: 2020-2021

Three Essays on the Measurement of Socioeconomic Inequalities and Well-being

Dr. Giovanna Scarchilli

Submitted to the Department of Political Economy and Statistics, University of Siena, in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Economics.

Abstract

The complex transmission mechanism of socioeconomic inequalities takes place in several spheres of life. This Doctoral Thesis, composed of three essays, focuses on the characterisation of some components of inequalities and their spread through social groups. In the three contributions, innovative techniques have been exposed and empirically assessed to extend the literature on the measurement of well-being and the study of social inequalities. The first essay represents a study on teenagers' leisure time activities distribution and how it relates with income and subjective well-being realisations. Taken from the German Socioeconomic Panel (SOEP), the information on leisure time activities has been processed with a network-based technique to build a multidimensional index proxying well-being. The second essay presents an evolutionary analysis of cumulative deprivation for the Italian working-age population between 2007 and 2018. A rank-based multidimensional approach is applied for the identification of the cumulatively deprived people. Therefore, an assessment of the statistical multidimensional dependence lying across the identified deprivations is provided following a copula-based technique. The third essay contains a focus on the transmission of health inequality through the socioeconomic background of people. A machine-learning technique is used to derive the population partitioning into social groups and to define the different opportunity backgrounds. Furthermore, the study provides insights regarding the varying effect of individual health-related behaviours on the health status. The 2011 sample of UK Household Longitudinal Study data is used for the empirical application.

Keywords: socioeconomic inequalities, multidimensional indicators, economic complexity, copula function, inequality of opportunity.

Contents

List of Figures

List of Tables

List of Abbreviations

A-F Alkire and Foster, (2011)

AROPE At Risk of Poverty and Social Exclusion

ARP At Risk of Poverty

DDI Diagonal Depenendence Index

DU Direct Unfariness

EU-SILC European Union Survey on Income and Living Conditions

F-H Fréchet-Hoeffding Bounds

F&S Fleurbaey and Schokkaert, (2009)

FG Fairness Gap

FMM Finite Mixture Models

ICI Individual Complexity Index

IOP Inequality of Opportunity

LCA Latent-Class Analysis

LWI Low Work Intensity

MOB Model-Based recursive Partitioning

MoR Method of Reflections

PCA Principal Component Analysis

RCA Revealed Comparative Advantage

SAH Self-Assessed general Health

SDG Sustainable Development Goals

SMD Severe Material Deprivation

SOEP German Socio-Economic Panel

UKHLS UK Household Longitudinal Survey

Chapter 1

Introduction

Perhaps stimulated by the disillusionment that the financial and economic crisis of 2008 has put forward, within the last decade the topics of poverty and well-being have gained increasing attention both in the academic world and in the public debate. On the policy side, we have witnessed at many examples of institutional and governmental initiatives to observe and monitor the social and economic conditions of people. In the EU, among the most famous relevant policy stimuli can be found the European Union's 2020 growth strategy (2010), and the Sarkozy Commission on the Measurement of Economic Performance and Social Progress (2008), led by J-P Fitoussi and the former Nobel prizes J. Stiglitz and A. Sen.

In the following years, the OECD's project on Measuring the Progress of Societies (2013), the European Pillar of Social Rights - with a constant update of 'social scoreboards' officially entering the European Semester of economic policy coordination since 2017 -, and the UN's Agenda 2030 for Sustainable Development (SDGs, 2013) have been some of the outcomes of the aforementioned contributions.

Despite the initial enthusiasm which characterised the period of programming policies for realising and implementing the European targets, the ten-year-after analysis on the fulfilment of the EU2020 strategy came across with some delusion. Furthermore, the recent COVID-19 pandemic has newly exacerbated old vulnerabilities, wiping out years of economic recovery and social progress. A sign of this significant impact has been the recent rise of the inactive and unemployed population as being primarily composed of self-employed, young, females, and temporary-contract owners. On top of that, we are witnessing at the surge of new forms of weaknesses. For example, the lockdown caused an abrupt increase in the school dropout incidence across the countries' most economically poor areas. This phenomenon will strongly contribute to the increase of inequalities due to different opportunities.

Given the situation illustrated above, the academic and non-academic stakeholders have always had a keen interest in studying, on the one side, the phenomenon of socioeconomic inequalities, and, on the other, the related policies and their outcomes. The academic attention towards the studies of socioeconomic inequalities can be traced back to the seminal contribution of A. Sen (1980; 1987). Sen stressed on the necessity of studying societies going beyond the sole income measure and considering societal welfare in a broader sense. In so doing, he gave a notable contribution to the spreading of studies related to well-being. These studies led the researches to employ multidimensional approaches to be able to tackle them. Beside the theoretical component, the increased coverage of life dimensions in survey data and the growing number of computational techniques to aggregate and process the data-collected information gave a valuable contribution to this subject matter.

From a political and philosophical point of view, there is no unique and all-encompassing definition of the concept of ‘well-being’. Nevertheless, there is general agreement that well-being is the outcome of a complex system of interrelationships between the social and economic sphere, including subjective and objective aspects. It is precisely the word *multidimensional* that represents a common thread across the three essays displayed in this thesis.

From an analytical perspective, social scientists have dealt with the measurement of well-being using a variety of statistical tools and models - each and every variety is motivated by given assumptions. The underlying assumptions of the models are both related to prescriptive or normative decisions of the researcher, and to the various necessary decisions related with the empirical application. Despite the clear distinction between the definition of the two types of assumptions, there can be some overlap in the implications of both, when applied. As a result, it is noticeable that among applied researchers in socioeconomic inequalities and well-being two subgroups have emerged: on the one hand, those who favour purely normative assumptions, whilst on the other, those who rely on data-driven techniques to make a decision on certain assumptions. The differences emerge on several stages of the studies on well-being, poverty, and inequalities measurements - among them, the definitions of poverty and well-being, the construction of relevant population groups which are the target of a policy, the setting and validation of relevant parameters for the empirical analysis.

This PhD thesis belongs to that "group" that employs data-driven techniques. However, it does not actually reject to make normative considerations on the phenomena under study. Indeed, it is believed that data-driven methods does not represent a perfect substitute to the normative choice of a policy maker. While the data-driven methods are able to provide clear information on what *is* happening in the society, normative positions state what *should be* happening in the society. Therefore, the former could be seen as a support to address traditional normative issues, which did not find a universal ethical agreement. Moreover, the most relevant innovation brought by data-driven methods is the capacity to recognise new matters and find new solutions, which could be unexpected when approaching those studies that take normative decisions. Furthermore, data-driven methods could be very useful to solve some relevant practical issues related to traditional empirical approaches, which are acknowledged to be determining serious biases in the results.

Combining ideas from economic theory, philosophy, sociology, and data science this thesis provides innovative tools to interpret socioeconomic inequalities and measure well-being. Despite their "distance" regarding the means used in each empirical application, the essays together attempt to provide a set of tools to enrich with practical computational examples specific parts of the literature on well-being and inequalities. The recurring innovative element of the thesis is precisely the exploration and adaptation of new multidimensional measurement techniques. Furthermore, each contribution attempts to contextualise the novelties within the well-being and inequalities' literature domain. However, it should be emphasised that such approaches all refer to the individual sphere, and therefore they do not take into account trends in inequalities related to different observation units, e.g. households and territorial.

This dissertation consists of three essays, each of them is presenting an innovative technique to study well-being measurement, which supported by an illustrative empirical application of the new approaches' advances and limitations. The chapters follow the structure of academic papers. Each essay is self-contained and can be read independently.

In the first essay, in Chapter 2, a network-based approach is used to build a multidimensional index of well-being for teenagers. The well-being status of people is proxied by processing information about several leisure activities. This study aims to assess how the information on the use of time, if observed at youth, can add valuable information on future well-being realisation.

The specific multidimensional index technique employed for this experiment is the "Economic Complexity Index" of Hausmann and Hidalgo (2014). This technique uses the information provided by the network which maps the links between two entities, namely individuals and everyday life activities, in order to provide a ranking of the sample considered. The everyday activities are different in terms of the required effort and cost to be sustained. For this reason, there is high variability in the distribution of people's time employment.

Behind the adoption of the "Economic Complexity Index", which provides a ranking of people according to their use of time, is that, not only specialisation matters, but also diversity is the key to measure human complexity. Firstly, the choices and capabilities of individuals are identified through the observation of the "specialisation" of people in a specific activity, i.e., whether such activity is considerably present in the overall activity set of a person. Second, each activity is defined by its "sophistication". Third, the eclecticism of individuals in terms of the multiplicity of their activities and interests, is considered. The data used for computing the Complexity index comes from the special module dedicated to 17 years old respondents of the German Socio-economic Panel (SOEP), which contains information on their weekly activities. An attempt to use the complexity index as a predictor of subjective and material well-being as recorded in later waves of the survey is proposed, despite the strong sample attrition. From the exercise it emerged that a high complexity for individuals is associated with social activities. Very ubiquitous activities, such as watching TV, are instead associated with low-ranked people in the complexity score. Very specialised activities, e.g., playing an instrument, are instead quite rare and associated with mid-complex people. The complexity ranking are correlated with current subjective well-being perception, and with the economic conditions of the individual in the future.

As stressed within the literature on poverty and social exclusion, there are many forms of deprivation which tend to come together in societies. The evolution of cumulative deprivation is addressed in Chapter 3, regarding the working-age population in Italy between 2007 and 2018. Cumulative deprivation is characterised by disposable income, health status, housing quality, job conditions and educational attainment. All the dimension-related outcomes are observed at the individual level in each single year using the cross-sectional EU-SILC data.

In this paper, a copula-based technique is adopted to estimate the dependence lying among the multiple dimensions of cumulative deprivation. Copulas are used in statistics to evaluate the degree of dependence within a rank-based multidimensional framework; therefore, they have been gaining attention in social studies for inspecting the properties of interrelations taking place among different unit variables. A very recent contribution by Decancq (2020) offers a toolkit to address the analysis of the dependence at the extremes of the distributions' multiple dimensions of well-being: the diagonal dependence index.

In the period considered, the cumulatively deprived population in Italy shows a growing trend, amounting approximately to one million individuals in 2018. A visible peak of the phenomenon, with respect to the total sample, emerges in 2014 and 2015, highlighting a visible correlation with the trend of the estimated dependence index for the empirical multidimensional copula. The presented index of multidimensional dependence could be interpreted as measuring the degree of association between the various forms of deprivations taken into consideration. Given its proximity to the concept of poverty, cumulative deprivation has been contextualised with respect to the current estimates of relative and absolute poverty provided by the Italian National Statistical Institute (ISTAT). A descriptive comparison is provided between the maximum income of cumulatively deprived people by household type and geographic location, with two poverty income thresholds (the AROPE and the ISTAT's Absolute poverty estimated thresholds).

The last essay, in Chapter 4, provides a zoom on a narrower well-being aspect, the individual health. The COVID-19 pandemic's disruption has put under the spotlight the highly unequal distribution of health characterising current societies. Furthermore, health deprivation has shown strong links with deprivation of other facets of life. Individual health is hereby conceived as an objective status of well-being summarising a series of biomedical characteristics of the person. The health inequality is studied from a multidimensional perspective, more precisely, referring to the theory of Inequality of Opportunity in order to assess its relation with other socioeconomic inequalities. The socioeconomic drivers of inequality of opportunity (IOP) in health are investigated and some light is shed on the methodological progress that characterises the IOP models. IOP in health is assessed controlling also social group-specific trends of health-related behaviours in the determination of the health outcome. Despite the health-related behaviour is considered as a proxy of effort, its connection with the social background is kept into consideration within the model framework. The application introduces a new methodology – the Model-Based Recursive Partitioning (MOB) – to derive the population groups while estimating, within each group, the relation between the health status and the effort variable. This study represents an empirical application of the measure of the "direct unfairness" and the "fairness gap", as proposed by Fleurbaey and Shokkaert (2009).

The empirical application is conducted using the UK Household Longitudinal Panel Survey. This dataset, in wave 2, contains data nurse-recorded on a sub-sample of the whole database regarding physical biomarkers. This information has been aggregated into a general index defining the physiological health condition of each individual. The evidence coming out by the adoption of the MOB technique shows a significant role of the socioeconomic background of people in determining health outcomes. Furthermore, it emerges clearly that, the behaviours are significantly affecting the health status with a different magnitude according to the social group of belonging. Despite the lower return to efforts that we observe among the most disadvantaged social groups, the distribution of behaviours show a slightly higher average effort for those with poorer socioeconomic background.

Chapter 2

Measuring the complexity of leisure time: a new methodological proposal to study how the use of leisure time relates with well-being

2.1 Introduction

> *"The quality of life depends on people's objective conditions and capabilities."* Stiglitz et al. (2009)[1]

Although per capita income is still, by far, the most popular measure of well-being, there have been various proposals to extend the horizon of its measurement. Inspired by the works of Fleurbaey et al. (2008); Stiglitz et al. (2009), many scholars have studied well-being from a multidimensional perspective, involving both material and subjective aspects of life. Overall, well-being can be either measured as a multidimensional composite indicator through the aggregation of multiple items (Costanza et al., 2016; Deutsch and Silber, 2005; Peña-López et al., 2008; VanderWeele, 2017), or as a single-domain measure such as material, or subjective well-being (Diener, 2009; Kahneman and Krueger, 2006).

While studying individual well-being - both material and subjective - a growing attention has been paid to the observation of the use of time. In their detailed study on the use of time across countries, Esteban Ortiz-Ospina and Roser (2020), for the project

[1] This statement is one of the twelve recommendations of the Report of the Commission on the Measurement of Economic Performance and Social Progress written by Stiglitz et al. (2009)

"Our World in Data", stated: *"Studying how people spend their time represent an important perspective for understanding living conditions, socioeconomic opportunities, and general well-being"*.

Regarding the literature on well-being and time use, on the one side, there are studies focusing on analysing social inequalities and the allocation of leisure time activities (Aguiar and Hurst, 2007; Burchardt, 2008; Lippe et al., 2010; Merz and Rathjen, 2014). On the other side, scientific production is widening following the work of Kahneman and Krueger (2006), which proposed to combine the use of time with affective ratings to analyse a societal well-being score and a rank the activities observed. In both cases, the time use is valuable information for proxying well-being. The latter studies aim at giving more consistency to the notion of subjective well-being and life satisfaction, while assuming the intrinsic value of an activity that can be defined by the emotions reported by the individual. The former studies keep the analysis perspective on a less personal level and they analyse how the use of time relates not only with subjective well-being but also with other notions of well-being.

With this study, the information on the distribution of leisure time use across the population is analysed to produce an index that could rank activities and individuals invoking the various notions of life-satisfaction and material well-being, by gathering them into a broader concept of human flourishing (VanderWeele, 2017).[2] The study presented introduces a data-driven methodology to collect and process information on leisure time use data, and to rank activities performed in the leisure time and individuals observed in a certain period of their life.

The aim of this paper is that to introduce a way to study a specific social group according to the use of leisure time observed, and furthermore, that to contextualise such information in the broader box of well-being notions. The main underlying assumption of this methodology is that the distribution of the use of time across individuals and leisure activities create a complex network of capabilities and opportunities, which can evaluate both the single activity and the individual with respect to the others. In the context of activity-related capabilities, specialisation is a plus. However, specialisation is not enough to observe multifaceted abilities or social integration. In order to observe whether ones' abilities are valuable and enrich the individual well-being, it is possible to refer to the *diversity* of specialisations and the *ubiquity* of the activities. These two "ingredients" characterise the so-called *individual complexity* and the *activity complexity*. The information extrapolated through the complexity rankings is used to proxy the means and capabilities acquired by the person and necessary to study

[2] VanderWeele (2017) defines human flourishing as a multi-domain concept involving the psychological, material, health both mental and physical, social, and meaning sphere.

future realisations of material well-being and life satisfaction.

With this paper, individual complexity is defined and its relation with well-being is explored. The considered leisure time activities provide a wide range of information due to their quantitative and qualitative heterogeneity. The subjects of the empirical application correspond to the sole group of teenagers. The decision to focus on a specific age group is that the variability of activities and leisure time use strongly depend on the age distribution. Moreover, the choice of this specific cohort represents a way to proxy individual capabilities formation and both relate them with current socioeconomic conditions and with future well-being realisations.[3]

As already anticipated, the main innovation of the study is the adoption of a new technique for processing information on the time use distribution. "Borrowed" from the macroeconomic literature on trade and growth, the adopted measure is the Economic Complexity Index. The Economic Complexity represents a measure of the capability of a country to be the major exporter of a diverse series of goods which are rare and highly required, and it has been successfully translated in the added value of knowledge accrued by a country through development and economic growth. This widely notorious measure, in the specific purpose of this study, is given the name of *Individual Complexity Index*, when referred to the people, and *Activity Complexity Index*, when referred to the activities observed.

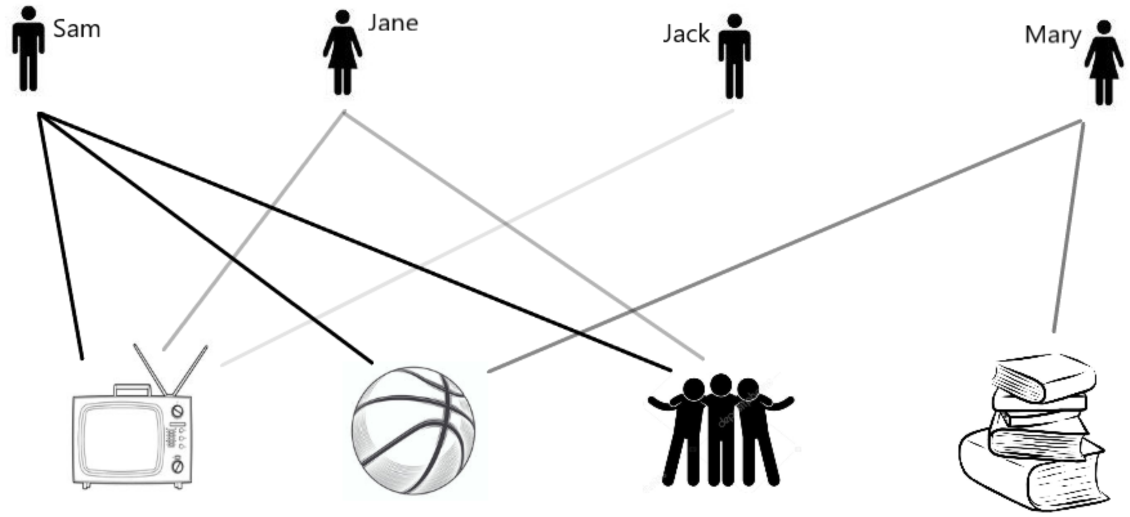
This indicator – which is hereby believed to have a good potential within the use of cross-sectional micro-economic data – aggregates the data making use of the Method of Reflections (MoR), an algorithm presented by Hausmann and Hidalgo (2014) for creating their Economic Complexity Index. The main information necessary for the implementation of the new Individual Complexity Index is provided by the network which maps the links between two entities: the individual and the activities he makes in his free time. The algorithm's inputs are the *diversity* of the people's activity sets and the *ubiquity* of every single activity across the population. The MoR is a recursive algorithm that repeatedly corrects one measure with the other to enlighten the in-depth information on each person and activity, which does not emerge at first sight.

Figure 2.1 is a graphical representation of the concepts of *diversity* and *ubiquity* through a very simple example of individual-activity network.

As it is possible to see from the figure, Sam watches TV, plays basketball and spends time with friends. Therefore, one could say that he has a diversity score of 3, whilst Jane and Mary have a diversity of 2 and Jack only of 1. There are in total three people watching TV, thus this activity has a ubiquity score of 3. Despite Mary and

[3] This conceptualisation takes inspiration to the work of Sen (1993) only in the form of the philosophical definition of capabilities.

Figure 2.1: The Individual-Activity network, an example



Jane have the same diversity, Jane is doing an activity that is very ubiquitous, whereas Mary reads a lot of books, which is quite rare.

The algorithm of Hausmann and Hidalgo (2014) captures all this information in an iterative way and produces a *complexity* ranking of all the people and the activities.

With this application, it is possible to observe a wide set of leisure activities carried out by a population of 16/17-years-old individuals: the data used for the computation come in fact from a special model built on respondents aged 16/17 of the German Socio-Economic Panel (SOEP). The network linking people and individuals is computed the observation of the activity-specific distribution of time use across people. The link between a person and the activity is indicating that the person devotes a considerably high amount of time to that activity with respect to the rest of the people. The aim of this application is the exploitation and the exploration of the informative power of Individual Complexity and the evaluation of its capacity to proxy individual well-being in various forms.

The complexity index is assumed to be useful to describe the well-being of people. More specifically, a positive relation is expected between complexity and well-being in various forms. The impact of material household well-being is assumed to be positive as well on the current complexity.

Therefore, the analysis is run over two binaries. On one side, the complexity outcome is contextualised with respect to the leisure time activities observed and with the current subjective and material well-being of the teenager. On the other side, an analysis on the interrelation between individual complexity and future realisations of well-being is provided. The analysis presented does not investigate causation rules but simply correlation patterns.

The empirical application represents an initial exploration of the possible adaptation of this technique to a different context. From a technical perspective, the use of the MoR's aggregation technique contextualised within multidimensional well-being, enables to contribute to the literature on composite indicators regarding data-driven weighting procedures.[4] Within the family of data-driven methods, the frequency techniques are widely used among multidimensional poverty indices. Given the association of the frequency of (no-)deprivation concept with the ubiquity and diversity, the presented complexity index could be placed within the data-driven weights classification, in particular linking it to the frequency techniques.

The paper is structured in the following sections. Section 2.2, contains an illustration of the methodological background of the Economic Complexity Index. In Section 2.3 follows a description of the data used for the empirical application. In Section 2.4, the empirical results and interpretations are provided. The last Section reports a conclusive discussion on the empirical application.

2.2 Methodology

The MoR computation's first step is the construction of the individual-activity network, which is created by following a relative frequency rule. The (bipartite) individual-activity network can be associated with a binary bi-adjacency matrix M. The links of the bipartite network are determined by an index transformation of the original quantity of unit time that an individual spends averagely on the specified activity.

The applied transformation represents a weighting procedure of the *importance* of the activity in terms of, i) the whole activity set of the specified individual, ii) the total amount of the activities considered, which, afterwards is converted into a binary matrix. This transformation can provide a measure of Revealed Comparative Advantage, which is hereby interpreted as a measure of Revealed Comparative Affection (Revealed Comparative Advantage (RCA)) of an individual regarding a certain activity. The so-called **RCA**, or "Balassa Index", has been adopted by Hausmann and Hidalgo to generate the binary bi-adjacency matrix necessary to implement the algorithm for the complexity computation, the Method of Reflections.

The rationale behind the RCA computation is to derive the value of the match

[4] See Decancq and Lugo (2013) for an extensive discussion on the role of weights in multidimensional composite well-being indicators.

between all individuals and the activities basing on the person's revealed comparative affection towards each activity. The matches define a bipartite network that filters the links between the individual and the activity, keeping only the individual affections. By means of this, the concept of well-being which can be extrapolated is adapted to this exercise of grasping the differences across peoples' opportunities, stimuli and interests, and, consequently, at ranking them.

The input for the RCA computation is the original data matrix containing the time spent per activity, $\underset{n\times m}{\mathrm{P}}$, where $i = 1, ..., n$ and $a = 1, ..., m$ respectively the number of the individuals and of the activities. The value for each match between individual i and activity a in the $RCA_{i,a}$ results from the following operation:

$$\boldsymbol{RCA_{i,a}} = \frac{\boldsymbol{s_{i,a}}}{t_a} \tag{2.1}$$

where the value $s_{i,a}$ represent the time spent by individual i on activity a with respect to the total time spent on leisure activities by individual i. This value is collected in a matrix $\boldsymbol{S}$ obtainable by dividing the total time each person spend in the activities considered $\sum_a p_{i,a}$ to the original $\boldsymbol{P}$ matrix. This matrix represents a scaling factor for the time-unit spent by the individual i in each activity as a ratio of total $(a = 1, ..., m)$ activities done. The value t_a represents the sum of all the time each individual has devoted to a specific activity a, its corresponding collection for all activities is the vector T.

$$\underset{n\times m}{\mathrm{S}} = \underset{n\times m}{\mathrm{P}} / \sum_a x_{i,a} \ , \ \underset{1\times m}{\mathrm{T_a}} = \sum_i \underset{n\times m}{\mathrm{S}} \tag{2.2}$$

Dividing $\boldsymbol{S_{i,a}}$ by the element T_a, a further scaling of the individual amount of time per activity is provided with respect to the other people. The $\boldsymbol{RCA}$ matrix from Eq.2.1 illustrates, for every individual, the revealed comparative affection towards each activity. In other words, the RCA shows which individual is devoting a "greater than usual" portion of time in a certain activity.

When $RCA_{i,a} > 1$, it means that an individual i does an activity a for a portion of time that is greater than the average amount of time dedicated to such activity by other people. Furthermore, such activity is a considerable part of the overall i^{th} individual activity set. Defining the bi-adjacency matrix as $\boldsymbol{M}$, the value of the match in the case of an $RCA_{i,a} > 1$, is $m_{i,a} = 1$. Alternatively, the value of the match is $m_{i,a} = 0$, if the portion of activity done by an individual is very small compared to its activity set and the total amount of consumption of this good, $(RCA_{i,a} \leq 1)$. Therefore, single value of the bi-adjacency matrix, $\boldsymbol{m_{i,a}}$, is used to build the individual-activities bipartite network as follows:

$$M_{i,a} = \begin{cases} 1\; if\; RCA_{i,a} > 1 \\ 0\; if\; RCA_{i,a} \leq 1 \end{cases} \tag{2.3}$$

The M matrix has size $(n \times m)$.From now on, any reference to the activity per person coincide with the match as shown in Eq.2.3 in which the individual shows a greater than 1 revealed comparative affection towards a specific activity.

The RCA represents the first step of the weighting scheme between dimensions of leisure time. At this stage, the weight is converted into a binary variable dividing the population in two groups, namely the individuals who keep the dimension as essential and those who do not. This procedure attaches "same weight" equal to 1 to all the dimensions with a high RCA score. Otherwise, it equals to 0. All the activities are considered as perfect substitutes and they are measured with the same unit, i.e. time. Thus, the activities are observed through a frequency approach.

Afterwards, the information retained is the degree of the nodes within the individual-to-activity network $M_{i,a}$. The node degrees in this bipartite network are respectively the measures of *diversity* and *ubiquity*: $k_{i,0}$ and $k_{g,0}$. The node degree of $individual_1$ is equal to the sum of all the edges that start from a particular vertex, the $individual_1$, connecting another vertex.

The *diversity* measure of a specific individual is the sum of all the activities done by an individual:

$$k_{i,0} = \sum_{a} M_{i,a} \tag{2.4}$$

While the *ubiquity* measure of a specific activity is the sum of all the individuals who consume it:

$$k_{a,0} = \sum_{i} M_{i,a} \tag{2.5}$$

The diversity vector $\underset{n\times 1}{\mathbf{k_{i,0}}}$, and the *ubiquity vector* $\underset{1\times m}{\mathbf{k_{a,0}}}$, contain respectively the measure computed for each individual and good.

The Index of Individual Complexity, can be obtained by the standardisation of the resulting vector $\boldsymbol{k_{i,N}}$ where the subscript represents the N^{th} iteration of the Method of Reflections for the individual i. Symmetrically, the Activity Complexity can be obtained by the standardisation of the resulting vector $\boldsymbol{k_{a,N}}$. [5]

[5] This number is approaching an even number around the 20^{th} iteration for the individual com-

The Method of Reflections is an iterative algorithm combining the information provided by the individual-activity network properties. Each iteration computes a value representing an approximation of the conditional probability to move through the network's links to reach a specific individual starting from a different one. Before explaining its implicit meaning, it is worth to present its algebraic construction.

The MoR performs N iterations simultaneously for the individuals and the activities. The first iteration brings respectively to the average ubiquity of the activities done by individual i (Eq.2.6) and the average diversity associated to activity a (Eq.2.7):

$$k_{i,1} = \frac{1}{k_{i,0}} \sum_{a} M_{i,a} k_{a,1} \tag{2.6}$$

$$k_{a,1} = \frac{1}{k_{a,0}} \sum_{i} M_{i,a} k_{i,1} \tag{2.7}$$

Therefore, by recursively plugging the output in Eq.2.6 in the successive iteration of Eq. 2.7, up to the N^{th} time, it is possible to obtain the following results:

$$k_{i,N} = \frac{1}{k_{i,0}} \sum_{a} M_{i,a} k_{a,N-1} \tag{2.8}$$

$$k_{a,N} = \frac{1}{k_{a,0}} \sum_{h} M_{i,a} k_{i,N-1} \tag{2.9}$$

Plugging the recursion at $(N-1)$ of Eq.2.9 in the recursion at N of Eq.2.8, it is possible to obtain a formula that expresses the diversity recursion as a function of the initial conditions and the diversity of the other individuals (i') who perform the same activities of i.

$$k_{i,N} = \sum_{i'} \tilde{M}_{ii'} k_{i',N-2} \tag{2.10}$$

The new system obtained contains a matrix called $\tilde{M}_{ii'}$.

$$\tilde{M}_{ii'} = \sum_{a} \frac{M_{i,a} M_{i',a}}{k_{i,0} k_{a,0}} \tag{2.11}$$

The Eq.2.10 represents the individual-activity network as a Random Walk process and $\tilde{M}_{ii'}$ is the transition matrix. Eq.2.11 expresses the probability of reaching individual i, starting from individual i' and passing through the activities they carry out.

plexity (Hausmann and Hidalgo, 2014). For what concerns the activity complexity the N^{th} iteration approaches an odd number around 19.

The bigger is the similarity of two individuals' activity set, the higher will be their proximity in the final ranking obtained by the Complexity Index.[6]

We can say that at the N^{th} iteration, $k_{i,N}$ is a linear combination of the elements of the initial step $k_{i,0}$ where the coefficients result by the product of all the degree of nodes lying in the path, which connect the individual i' to individual i (Hidalgo and Hausmann, 2009).

With a higher number of iterations, it is possible to grasp more information from both characteristics. Therefore, the correlation between the initial and the final iterations is decreasing in the number of iterations.

Given that the sequence $k_{i,s}$ converges to a vector with all equal values for $s \to \infty$, the difference between subsequent elements in the sequence of $k_{i,N}$ and its limiting value progressively shrinks ($w_i = k_i - k_{i,N}$). According to the algebraic interpretation of the matrix $\tilde{M}$, the iterative process let the result of Eq.2.10 to converge to the eigenvector of the matrix in Eq.2.11 associated with the second highest eigenvalue of the matrix $\tilde{M}$.[7]

Given that we are working with a bipartite network, the even and odd iterations have different meanings. More precisely, there is a clear interpretative distinction between these two variables (Caldarelli et al., 2012). For what concerns individuals, the even variables ($k_{i,0}, k_{i,2}, ...$) are generalised measures of diversification of their activity sets, while the odd variables ($k_{i,1}, k_{i,3}, ...$), are generalised measures of the ubiquity of the activities. Furthermore, the two sequences $k_{i,N}$ and $k_{a,N}$ are inversely related. This means that people who are highly diversified in terms of activities done, will more likely being doing also activities which are rare (less ubiquitous), while very common activities will tend to be associated with less diversified people.

2.2.1 Concluding remarks on the methodology

The RCA index can extract the information concerning the individual level of affection towards a specific activity defined through a *relativistic* approach. This index collects only information regarding the time-use unit measures, but it attaches a constant substitutability measure to all the activities considered. The value judgements regarding each activity depend only on their distribution across the population, and the frequencies observed imply a certain data-driven outcome on the network.

[6] Section 2.6 illustrates the definition of a Random Walk process in the network theory framework.

[7] (Kemp-Benedict, 2014). An extensive discussion on this statement is presented in the Theoretical Appendix section 2.6

Recalling the inverse relationship between diversity and ubiquity, and contextualising it in the macroeconomic context, it is possible to find the emergence of a contrast between the traditional Ricardian theory of specialisation and the message brought by the Method of Reflections. While Ricardo predicted that specialisation in a narrow sector is the ”virtuous” path for growth of a country, the Economic Complexity conveys a different message. The rationale behind MoR and Economic Complexity is that, not only specialisation matters, but also diversity is a pathway to growth.

Therefore, as far as high complexity is associated with high diversity, low ubiquity will be as well related with high complexity. In order to provide a graphical intuition of this relation, Caldarelli et al. (2012) suggest to represent the individual-activity network - the binary matrix P - having sorted all the people and all the activities by increasing complexity. More precisely, the activities - the columns of the matrix - are sorted in ascending complexity from the left to the right. The people - the rows of the matrix - are sorted in ascending complexity from the bottom to the top.

Fig.2.2 shows the bi-adjacency matrix of the bipartite network of people and activities in 2011, both elements are sorted by increasing Complexity Index scores.[8]

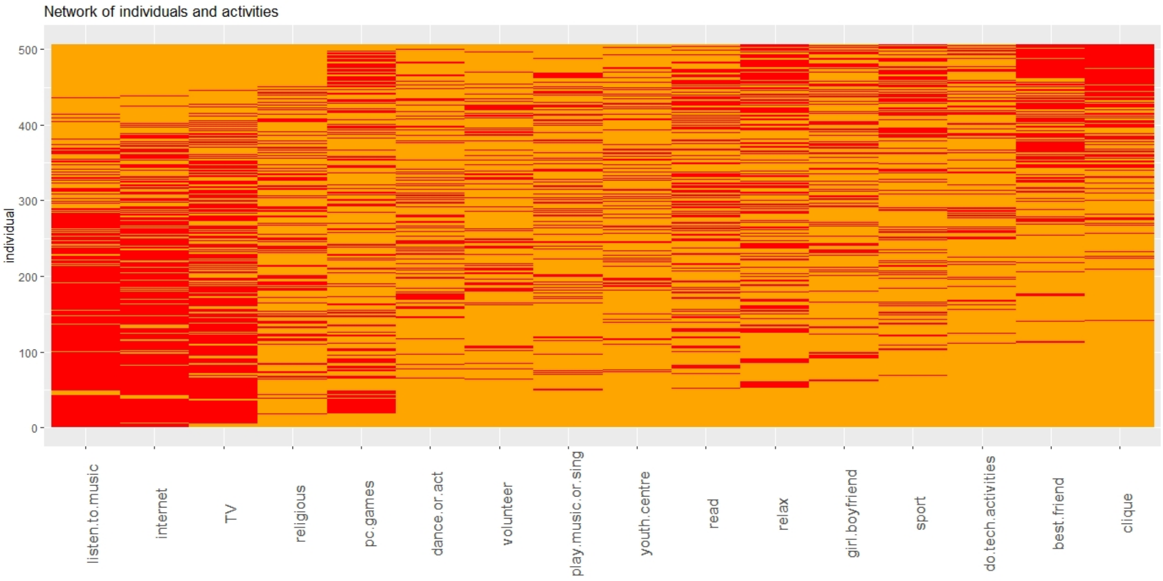


Figure 2.2: Network of individuals and activities sorted by their Complexity

It is worth reminding that the network shows the links between the people and the activities assigned with the RCA index. The dark red spots identify those individuals dedicating a higher quantity of time with respect to the average time spent by other people to a specific activity (RCA index = 1). The more red spots are appearing

[8] It is a binary matrix whose components are only 0, yellow, and 1, red.

alongside a row, the more diversified is the individual activity set. Likewise, the more red spots are appearing alongside a column the more frequently is the single activity appearing across various activity sets. On the contrary, the orange spots show a non-significant consumption amount (RCA index = 0).

People at the bottom of the y-axis of figure 2.2, are associated with the lowest complexity. The higher the complexity of the activities, the more on the right corner a given activity will appear.

If complexity preserves the relation with diversity and ubiquity as explained above, the resulting matrix should look like a triangular matrix. If we would have observed a higher polarisation towards specialisation, this matrix would have looked like a block diagonal matrix, whilst, instead, it is more likely to be an upper triangular matrix. In order to algebraically test whether the diversity and ubiquity contribute to define a complex system, Caldarelli et al. (2012) propose to test the nested structure of the matrix of the bipartite network.[9] In graphical terms, a nested bi-adjacency matrix shows that the number of the links in each row, from the lower to the upper one, is a subset of the previous row.

Being the individual-activity network a nested structure, some characteristics concerning both the activities and the individuals observed are implied. First, a randomly chosen activity appearing in a highly diversified individual's activity set will likely be a non-ubiquitous activity. By contrast, a randomly chosen activity that appears in poorly diversified activity sets will less likely be complex good. Second, a randomly chosen individual who does very frequently a highly ubiquitous activity is not likely to be highly diversified. Conversely, a randomly chosen individual that does most of the time a rare activity is more likely to be highly diversified. The system's nestedness is also a necessary condition of well-behaviour of the algorithm of the MoR adopted to obtain the Complexity Index.

2.3 Data

The empirical application has been realised using data on the use of time from the Youth Questionnaire of the SOEP. The SOEP is a wide-ranging representative longitudinal study of private households. The computation is provided on a constant activity set for population samples referred to different years 2006, 2011 and 2016. The dataset containing the leisure activities questionnaire is referred to young respondents aged 16 or 17. Furthermore, the same individuals who took part to the 2011

[9] The nestedness is a statistical property of bipartite networks when are represented as a triangular matrix. Section 2.6, contains a paragraph on the nestedness test adopted for this paper.

youth questionnaire sample have been sampled from the individual adult dataset of 2016. The following list exposes the SOEP data sets were necessary for the empirical application. The list indicates the number of observed individuals net of the missing values on the leisure activities and on the sample attrition.

- `wjugend`: Individual questionnaire for people aged 16-17, year 2006 (N = 286)
- `bbjugend`: Individual questionnaire for people aged 16-17, year 2011 (N = 506)
- `bgjugend`: Individual questionnaire for people aged 16-17, year 2016 (N = 493)
- `bbh`: Household questionnaire, year 2011 (N = 506)
- `bbp`: Individual questionnaire, year 2016 (N = 188)

The reasons that led to the employment of the youth questionnaire are mainly three. Firstly, there is a trade-off between the specificity of the activities observed and the variety of the individuals' characteristics. For example, the average disposable leisure time for an adult with a family is lower than the one of a teenager. Given that the person-to-activity network is obtained with a frequency approach, the age is a source of heterogeneity for which it is necessary to control. The ideal application would be a computation of the complexity ranking for each specific age class. Furthermore, an age specific computation allows for varying the activity set considered. Secondly, the choice of the data on the teenagers allows to measure the value of leisure time when there is no - or at least lower -, trade-off between it and the income earnings. Lastly, a possible connection with the Equality of Opportunity theory could be done if considering the value of Individual Complexity as an outcome variable whose determinants beyond individual control are represented by the current household economic status. On the contrary, the Individual Complexity is considered to be a determinant of future realisations of material well-being status and both current and future subjective well-being status.

The activities for which the time frequencies are observed are watching television, surfing on the internet, listening to music, playing an instrument or singing, doing sport, dancing or acting, doing tech-activities (coding and programming), reading, spending time with girl/boyfriend, best friends, clique, going to a youth centre, volunteering activity, going to religious initiatives, relaxing (or none of the other alternatives).[10]

The complexity algorithm adopts a relativistic perspective which strictly depends on the distributions observed at time t. Thus, there could be some dependence of

[10] The youth questionnaire of 2016 includes also the time spent on the social networks.

the outcome on the activities distribution characteristics regarding each point in time. The empirical computation is performed for three different years in order to control, at least partly, for the relativistic perspective inherent to the results. The data which have been employed exploit the panel structure of the SOEP - following the same individual throughout time -, and through the observation of different samples in different years.

Each activity is associated with a score, which defines the frequency of use of a person. The frequencies have been normalised on a monthly scale, as shown in the following table:

Survey time definition	Frequency per month
Daily	1
Weekly	10/30
Monthly	4/30
Rarely	1/30
Never	0

The Complexity Index is the result of a recursive bipartite iteration, hence, both the individual and the activities are ranked – in this regard, it is interesting also to observe an activity-specific pattern in the ranking process-.

Therefore, in order to extrapolate some useful insights regarding the activity ranking, the activities have been divided into two groups, whether they are considered as social or non-social activities.

On data attrition

The merge between the 2011 teenager dataset with the individual adult data of 2016 presents the 62% of attrition. Hence from the 506 individuals in the original sample the population observed five years later is composed of 188 individuals (the 37%). Among those people who do not appear in the individual questionnaire five years later, the 53% are people who, having changed their place of residence with respect to the household included in the panel, vanish form the questionnaire. The remaining 47% of missing matches appear only in the households questionnaire. There has been a drastic reduction of the representative power of the outcoming dataset. Therefore, there might be strong biases emerging from the analysis due to the plausible correlation between social and economic conditions of those who could afford leaving parental home before 23 and those who could not. It is plausible that parents support the longer education of their children with a longer cohabitation. This could mean that the people who remain in the panel are mostly people who decided to continue their studies with tertiary education. By contrary, who left the parental home between 18 and 23 might be who economically emancipated through work or who, with the financial support of their parents could leave the parental home for continuing the education elsewhere. Given that most of the general socio-demographic information is not available in the youth questionnaire (the gender is missing as well), it is not easy to identify specific attributes who correlate with the disappearance from the panel, neither to assess which portion of them are missing at random. Therefore, the entity of the attrition could be verified through the mean difference analysis between the attributes observable within the two samples. As it emerges from the table A.4 in the Appendix A, the complexity ranking on the sub-sample of people observed five years later does not show a statistically significant mean with respect to the total sample. Moreover, table A.5, shows that there is no significant change in the means of each activity between the two samples. Hence, although the Complexity index's descriptive parameters are not demonstrated to be significantly varying across the samples, the level of attrition brings with it obvious problems in the ability to generalise what emerges from the time analysis.

2.4 Empirical application

Economic growth models traditionally conceive the specialisation as a strength. The primary reference may be the Ricardian model of specialisation. The experiment of Hausmann and Hidalgo (2014) shows however, that this prediction is not correct: the "diversity" factor also plays a central role in the development of a country. Although this study's subject switches from the country to the individual, the same principle stands for the latter unit of observation. The higher is the diversity and the complexity of people, the better we would expect to be their well-being status.

By contrast, the more a specific activity is spread out across the population, the less complex will be the individuals who mostly spend their time on it. As a result, there is an inverse relationship between the complexity of activities and their ubiquity scores.

The illustration of the complexity outcome is presented initially from the point of view of the activities, afterwards, from the point of view of the individuals. This is due to the fact that understanding how complexity relates with certain activities helps to understand as well the individual complexity measure as a function of the leisure time activities.

A complete picture regarding the activities' complexity could be provided by the observation of the relation between complexity and ubiquity. Therefore, figures 2.3, 2.4 and 2.5 depict the relation identifying the activities' characteristics for the three years considered. It is possible to consider these figures as able to provide a sort of "identikit" of the complex activities and to evaluate whether it emerges a stable relationship between the observed activities and their complexity ranking across time.

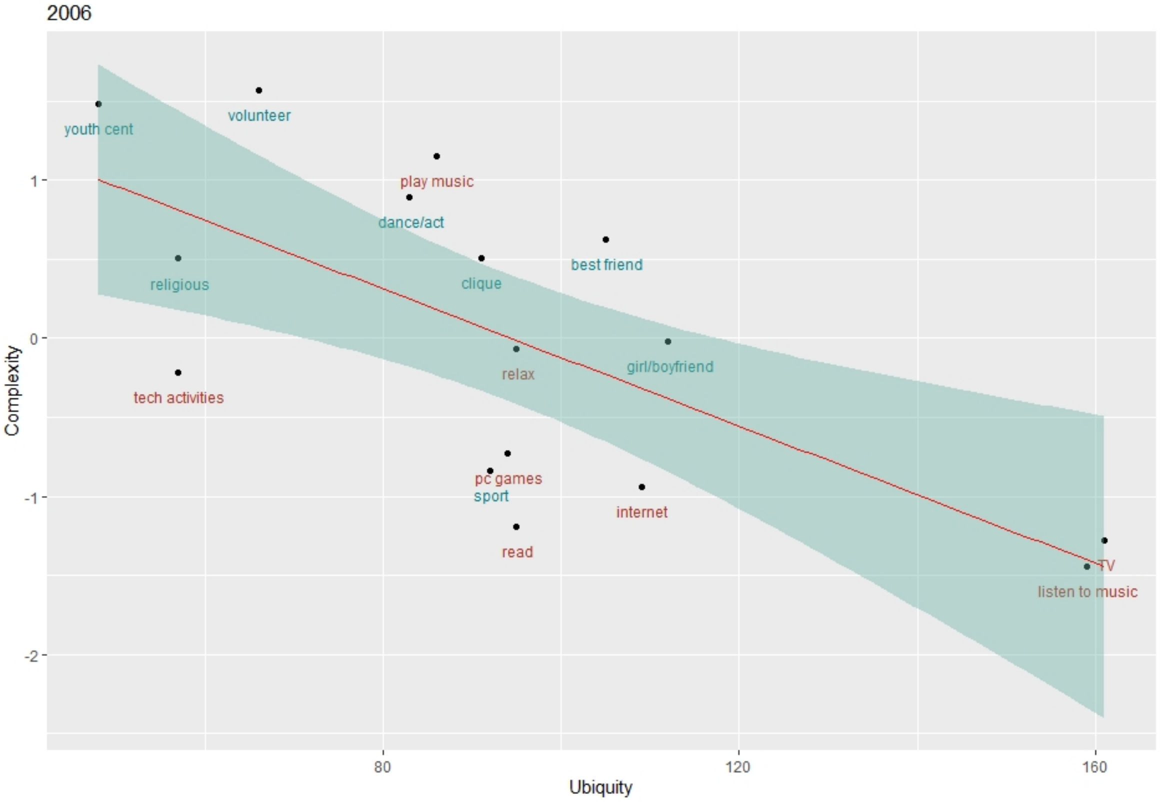


Figure 2.3: Complexity VS Ubiquity - 2006

Figures 2.3, 2.4 and 2.5 show the relation between the complex activities and their ubiquity score. As expected, there is an inverse relationship between the complexity of an activity and its ubiquity. From a year-by-year perspective, it emerges that, despite not all the single observations fall in the confidence interval, the rankings assigned to the different activities in terms of complexity and ubiquity of activities is somehow stable.

Dividing the activities into two groups according to whether these are social or individual activities, it emerges that social activities are mostly associated with higher complexity. This relation results visibly stable across years. The "social" activities are those including people of similar age spending time with each other. In the plots, it is possible to distinguish the social from the non-social activities by the colour of the labels. In all the three years, a higher concentration of social activities is evident

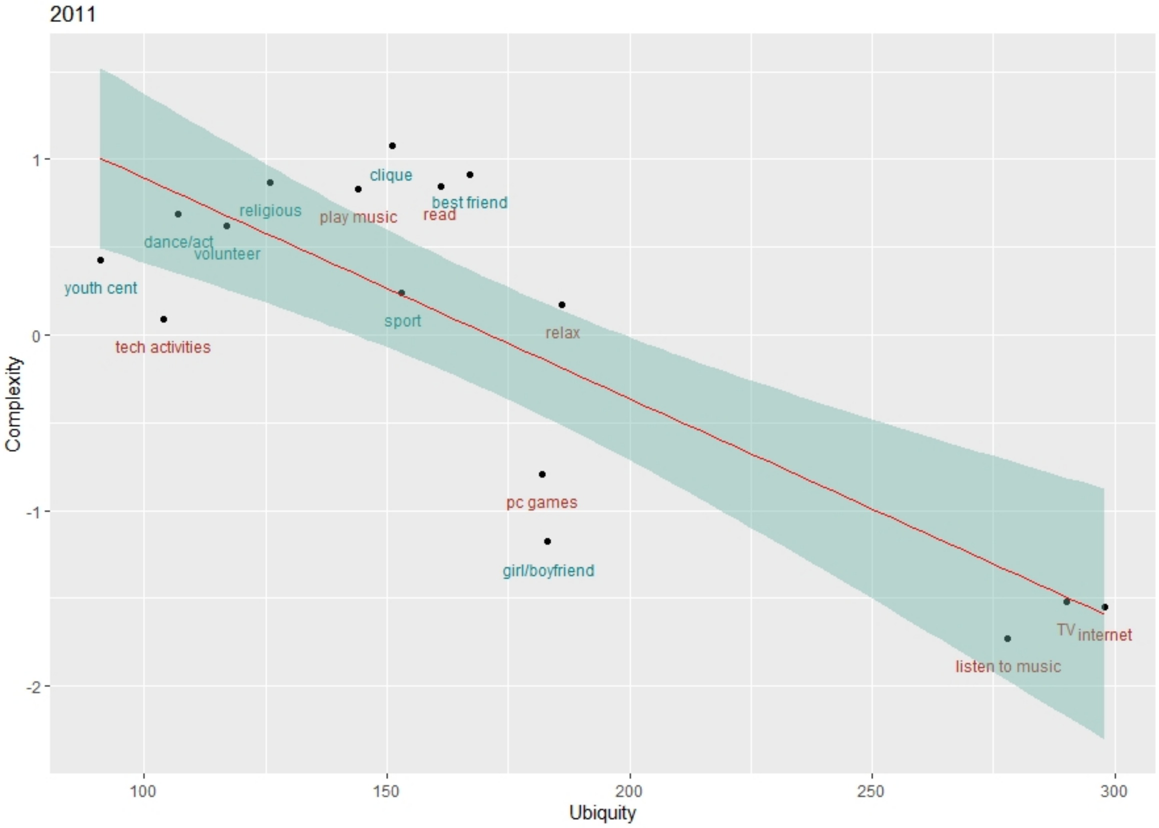


Figure 2.4: Complexity VS Ubiquity - 2011

in the graph's top-left corner, where the complexity score is high. The worst ranked activities are stable across the years: spending time watching TV, listening to music and, for 2016, being on the social networks. These activities are widespread across a population of 17-years-old individuals and, for this reason, they clearly end up being the less complex.

Which is a plausible interpretation of the Individual Complexity ranking? In which way is such ranking related to a human flourishing notion? How can be the top-ranked complex people described? We discuss the answer providing some outstanding results.

As seen from the previous figures, there is a pattern which emerges spontaneously across each yearly complexity VS Ubiquity plot, the distribution of social activities. Indeed, it emerges that social activities are non ubiquitous and always associated with high complexity.

The relationship between the scores of the individual complexity and the intensity of social activities is a notable support of the intuition concerning the role of social activities.

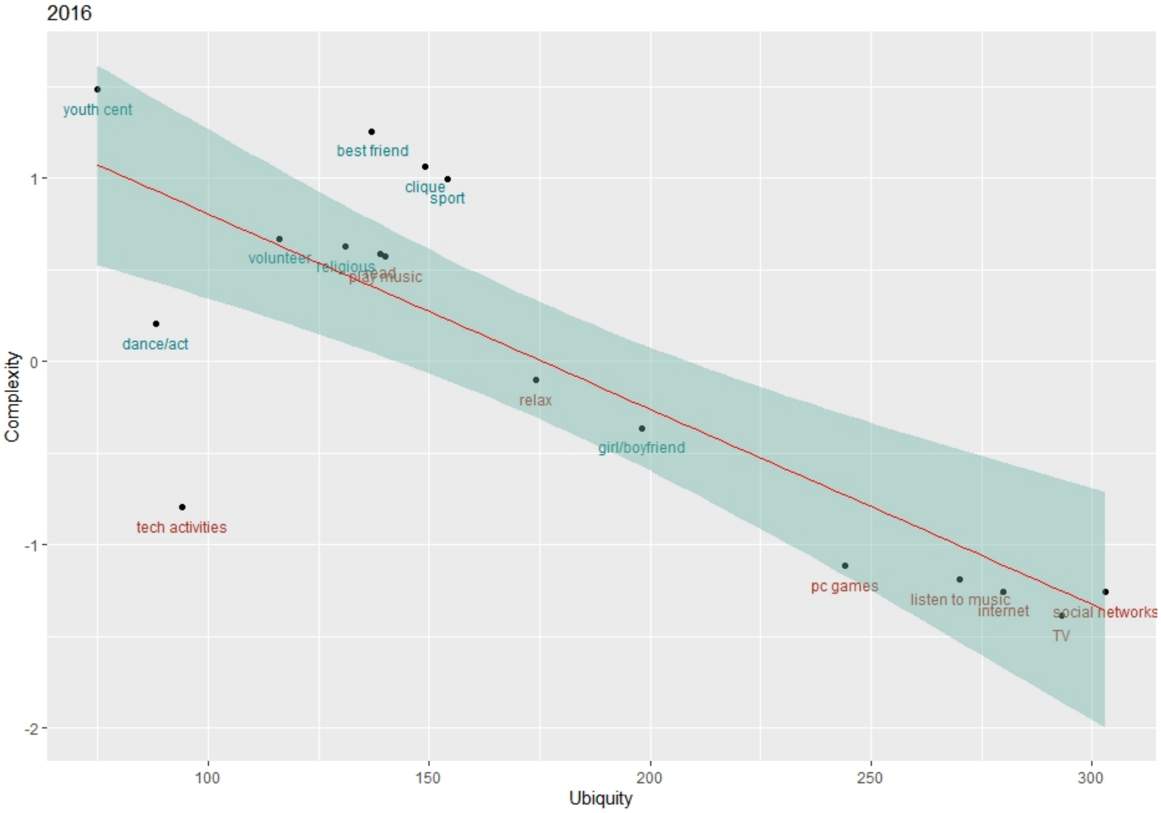


Figure 2.5: Complexity VS Ubiquity - 2016

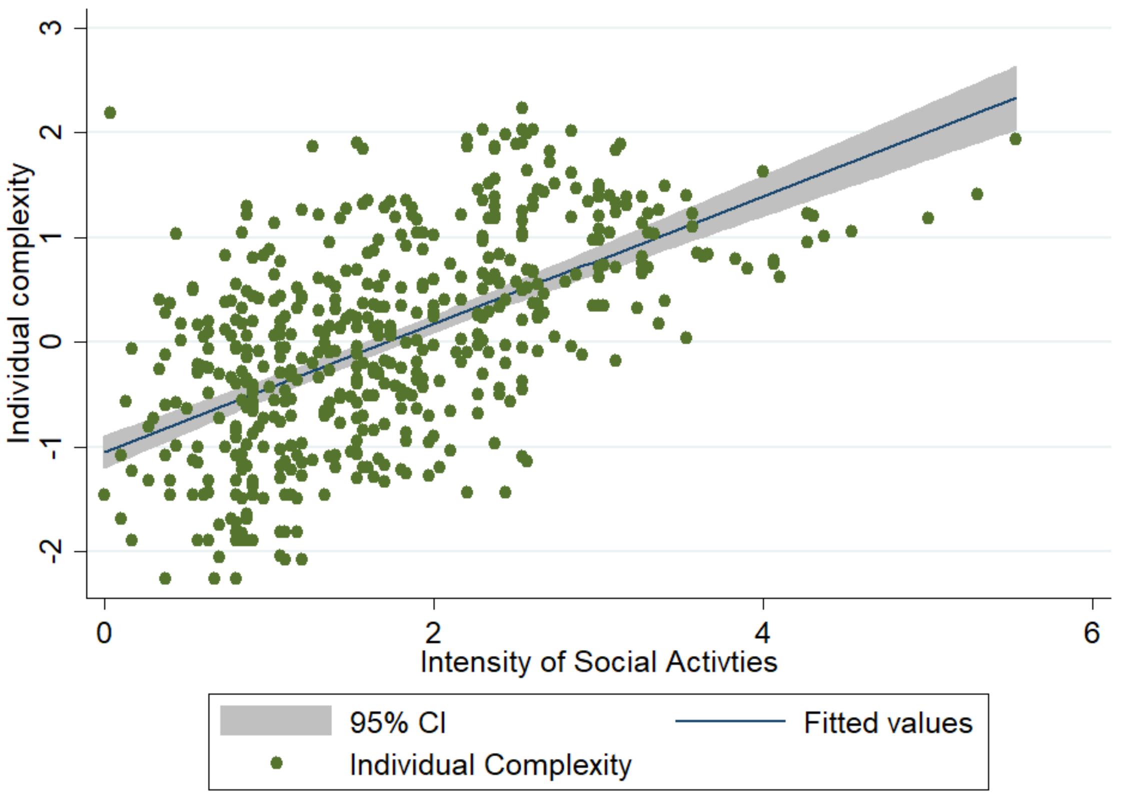
Figure 2.6, shows the fitted relation between the individual complexity and the intensity of social activities appearing in the time use set of people.

Based on the amount of social activities every person declared to be doing on a weekly or daily basis, figure 2.6 shows a dispersion plot throughout complexity scores. From the plot emerges a strictly positive relation between the two series, a further element in the support of the interpretation which links individual complexity with human flourishing.

In which way is Human Flourishing - intended as a fertile network of social connection and active social participation - positively linked to both subjective and objective dimensions of well-being?

Although it has been already depicted how any evolutionary analysis may not be fully reliable due to the panel data attrition as illustrated in Section 2.3, it will follow a discussion on how the complexity index correlates with contemporary and future realisations of income and life-satisfaction. Table 2.1 illustrates the Spearman rank correlations between the Individual Complexity and both current and future material and subjective well-being.

Figure 2.6: Interpreting of Individual Complexity as Human Flourishing



Unique rank of:	Unique rank of Complexity
Life satisfaction 2011	0.126**
Life satisfaction 2016	-0.099
Household net income 2011	0.143**
Individual net income 2016	0.213**
Observations	529

* ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 2.1: Spearman Correlations table

As it emerges from table 2.1, there are significant rank correlations between the complexity and the material conditions observed in 2011 and 2016, and between the complexity and the contemporary individual life-satisfaction. Future subjective life-satisfaction appears to be uncorrelated with the ranking of past complexity, yet it turns out to be negatively related with complexity. This is somehow not surprising given the strong role of temporary feelings involved in the subjective well-being measures, which make it hardly comparable across time.

Despite the analysis presented does not provide a basis for a causation relation, the association of complexity with contemporary and future well-being metrics is intended to justify two distinct empirical studies. On one side, a more accurate definition of teenager complexity may result from the role of contemporary household characteristics and well-being. On the other side, individual complexity could be studied form the perspective of a future well-being determinant.

The contemporary relation between complexity, life-satisfaction and household income is further explored in the next tables. Tables 2.2 and 2.3 respectively provide a purely descriptive relation between the average complexity and by levels of life satisfaction and by income deciles.

More accurately, tables 2.2 and 2.3 show the average individual complexity levels in 2011 for life-satisfaction levels and current household income deciles.

	Individual Complexity		
Life satisfaction	(mean)	(sd)	(N)
1	-1.288	.	1
3	0.222	1.051	10
4	-.5385	1.027	14
5	0.0446	0.890	30
6	-0.147	0.957	36
7	-0.064	1.001	80
8	-0.0589	1.013	166
9	0.104	0.978	125
10	0.316	1.021	43
Total	0.002	0.998	505

Table 2.2: Average Complexity by life satisfaction level - 2011

A visible increase in the average complexity in the population observed emerges in the group of higher-income deciles and the higher life-satisfaction levels. However, especially by focusing on life-satisfaction, it is noticeable that some groups are dramatically small, therefore they lack of valuable generalising power over the whole

	Individual Complexity		
Income deciles	(mean)	(sd)	(N)
1	-0.309	0.984	49
2	-0.185	0.889	56
3	-0.127	1.014	54
4	-0.134	0.907	41
5	0.152	1.013	57
6	0.151	0.929	50
7	-0.016	0.968	58
8	0.217	1.113	44
9	0.029	1.077	48
10	0.239	1.055	47
Total	-0.001	1.002	504

Table 2.3: Average Complexity by household income decile - 2011

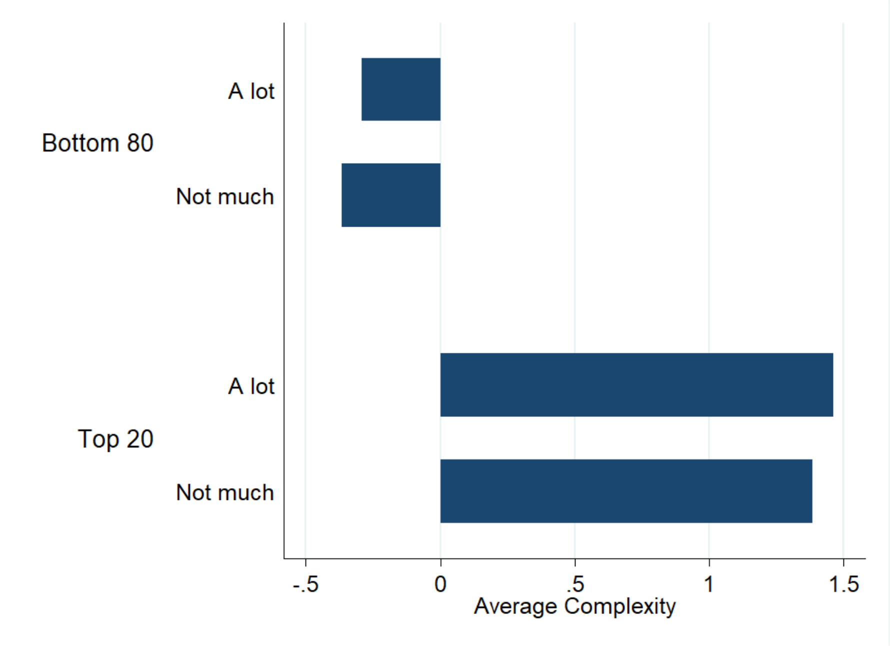
population.[11] Due to the small sample size, which gets even tighter in merging the data set in the future years, it is not possible to provide any parametric insight into the complexity-to-well-being relation.

In order to keep an opportunity-oriented perspective, the distribution of teenagers' complexity has been compared through the observation the parental support provided on studies and life choices. Figure 2.7 provides a descriptive illustration of the parental support variation through two groups of individual complexity. The average level of peoples' complexity is observed by the subjective perception of parental support in education and by distinguishing the individuals from belonging to the top fifth-ranked complex people and the rest.

Even though the role of parental support in education is positively affecting individual complexity, it does not emerge any heterogeneity in the impact between the top 20 complex people and the bottom 80.

[11] Figure A.1, in Appendix A, summarises the relation between complexity and life-satisfaction levels across years, illustrating the heterogeneous sizes of each life-satisfaction level group.

Figure 2.7: Average complexity by parental support - Top 20 VS bottom 80 complex people.



2.5 Conclusions

The Individual Complexity Index is a multidimensional composite indicator summarising information on the use of time which has been used to enrich the description of individual well-being. The employment of this particular index within micro-data could represent innovation for processing a hidden part of the information available in the data. The MoR, the weighting function adopted, aggregates information following a data-driven approach that evaluates the activity set in terms of two particular characteristics, namely the diversity of the activity sets and the ubiquity of all the activities across the sample. The Complexity Index is computed by applying the *Method of Reflections*, an algorithm generally used in macroeconomic studies, which provided interesting interpretations on country-growth rankings.

The motivation behind the employment of data on the use of time is that, it is considered to be adding valuable information to the picture of well-being status, both from the material and subjective point of view. Indeed, good material conditions of the household can amplify the possibilities to invest in non-remunerative activities, as well as the subjective perception of life status is qualitatively determined by the leisure activities.

By observing the use of time distribution throughout leisure activities, it has been possible to both rank the activities and the people. Therefore, the complexity scoring has been used as a new perspective to observe people's well-being status and material outcomes.

From this empirical application, the individual complexity could be building up a definition of well-being as *human flourishing.* The concept of human flourishing could be associated to the ancient Greek adjective *poly-tropon*, used in the Odyssey to describe Ulysses. This word, which literally means ”many wayed”, is a metaphorical adjective which stands for a multi-faced deep complexity. Ulysses is a person with thousands of resources; his diversity represents his ability to adapt and survive in various situations.

From the point of view of the existing literature on the Economic Complexity, new possible scenarios of application of this methodology has been presented translating its macroeconomic original interpretation to an individual-based context.Besides the empirical outcome, this measure could represent an attractive data-driven approach within the context of standard well-being index construction.

Nevertheless, the empirical analysis presents numerous shortcomings concerning both the methodology and the data used. Regarding the former, this computational method strongly depends on the sample specific network structure. For this reason, the ability to use the results to make more general considerations is somewhat limited. In order to partly overcome this shortfall it has been provided a computation for different years and samples.

Despite the reduced representative power of the data, it emerged the presence of a significant relation between Complexity and selected metrics of both material and subjective conditions. Besides that, it has been already argued that the main limitation of this empirical application is the data set attrition, which did not allow to provide a complete picture of the Complexity score’s predictive power with respect to future well-being dimensions.

Future developments of the study could consider the extension of the benchmark well-being dimensions to provide a more comprehensive definition of Individual Complexity. Additionally, in the light of the serious attrition in the sample when constructing a panel, it might be more appealing to develop a parametric analysis on the explanatory power of Individual Complexity over contemporaneous material and subjective well-being dimensions. Last, in the light of the problematic sample attrition for realising the analysis with respect to future well-being realisations, the exploration of other panel data sources could be taken into account for observing the impact of Individual Complexity on future realisations of well-being dimensions.

2.6 Theoretical Appendix

2.6.1 Random Walk definition

In a network, the probability of moving to vertex j after having done t steps to reach vertex i, is given by the following Markov chain[12]:

$$\pi_{j,t+1} = \sum_{i|j \in N(i)} \frac{1}{d_i} \pi_{i,t} \tag{2.12}$$

The i_{th} node of the graph has a certain degree, d_i that represents the number of links that depart from it. The Eq. 2.12 shows that this probability is obtained by the summation of all the N links starting form i and going backward. This probability will be different from zero if the edge j belongs to one of those N links. Going from the single step representation to a full network, all the links can be represented by the transition matrix P, whose elements are $p_{i,j} = \frac{1}{d_i}$. If there is no connection between e.g. i and the n^{th} vertex, then $p_{i,n} = 0$. It is possible to represent the probability that the process at vertex i transits to vertex j at next step in matrix notation using the transition matrix:

$$\boldsymbol{\pi^T_{t+1}} = \boldsymbol{\pi^T_t P} \tag{2.13}$$

The Eq. 2.13 for $N \to \infty$ represents a probability distribution of the process until N steps. Going to infinity, the distribution for $(N+k)$ steps it will not vary.

2.6.2 Algebraic interpretation of the Method of Reflections

The connection between the Method of Reflections and the eigenvector of the matrix $\tilde{M}$ can be done through Eq. 2.13 substituting the elements of such system with the elements of the $\tilde{M}$ matrix of Eq. 2.10. Therefore, 2.10 is represented as a problem of linear algebra (Kemp-Benedict, 2014).

$$\boldsymbol{k_{i,N}} = \boldsymbol{\tilde{M} k_{i,N-2}} \tag{2.14}$$

Given that the diversity corresponding to individual i at the N^{th} iteration converges to a constant number, we can write this concept as:

$$k_i = \lim_{N \to \infty} k_{i,N} \tag{2.15}$$

[12] Such a process is called Markov chain because is a process which depends strictly on the starting position, the position at time t, and it is independent on which vertex will be reached at step $t+1$.

The same statement holds for the row vector of diversity $\boldsymbol{k_i}$. Now, given that for $N \to \infty$ $\boldsymbol{k_{i,N}}$ and $\boldsymbol{k_{i,N-2}}$ are indistinguishable, we can rewrite the Eq. 2.14 as:

$$\boldsymbol{k} = \boldsymbol{\tilde{M}k} \tag{2.16}$$

Matrix $\tilde{M}$ is the *transition matrix* of the Random walk, i.e., all its columns the coefficients are $0 \leq m_{i,j} \leq 1$ and their sum column-wise is equal to 1; hence we can assess that is also *row-stochastic.*

The linear system in (2.14), for a high number of iterations is equal to(2.16) and coincides with the concept of eigenvector centrality of the $\tilde{M}$ matrix. More precisely, last equation is equivalent to the eigenvector centrality of a row stochastic matrix (Mealy et al., 2018). The eigenvector centrality of a matrix is the row vector corresponding to the highest eigenvalue λ. If $\boldsymbol{\tilde{M}}$ is row stochastic, its highest eigenvalue is $\lambda = 1$ and the associated eigenvector will have the same value on each component[13]. Let $\boldsymbol{\tilde{M}}$ be a squared invertible matrix, its eigenvector associated to the eigenvalue λ, is a vector $\boldsymbol{k}$ such that the following equality holds:

$$\boldsymbol{\tilde{M}}\ \boldsymbol{k} = \boldsymbol{\lambda}\ \boldsymbol{k} \tag{2.17}$$

The last two equations are equivalent. The Perron-Frobenius theorem implies that, given the presented properties of the $\tilde{M}$ matrix, the power iteration shown at Eq. 2.14 converges to the eigenvector associated with the highest eigenvalue of $\tilde{M}$ that have been defined in Eq. 2.16.

What follows from the Perron-Frobenius theorem is that, Eq. 2.14 coincides with the power iteration that, starting from an initial value $k_{c,0}$ that is not orthogonal to $\boldsymbol{k_{c,N}}$, will converge to $\boldsymbol{k_{c,N}}$ as the iteration grows.

Therefore, the linear system described at Eq. 2.14 converges to this constant eigenvector, that is exactly the eigenvector centrality of $\tilde{M}$. Furthermore, given that for a high number of iterations the heterogeneity shrinks, there is no need to look at the eigenvector centrality to obtain the Individual Complexity Index (ICI), but we need to look at the sequence $k_{i,N}$ when there is still some variability in order to be able to draw a cardinal ordering of the values.

In the limit of large N, these deviations are proportional to the eigenvector of $\tilde{M}$ with the largest eigenvalue less than one. That is, they are proportional to the eigenvector associated with the second largest eigenvalue v_2. The eigenvector associated with the **second** highest eigenvalue of the linear system in Eq. 2.16 defines the direction of the

[13] *Networks: an Introduction*, Newman (2010)

system convergence when there are still some differences among the individuals. Such eigenvector coincides with the algebraic definition of the ICI.

The method to find the Complexity Index requires on one side to compute the vector of the $N \simeq 20$ iteration and to standardise it. On the other hand, it is necessary to derive the $\tilde{M}$ matrix and compute the eigenvector associated with its second-highest eigenvalue. The standardisation of such eigenvector will be exactly equal to the standardised vector obtained at the N^{th} iteration.

$$ICI = \frac{\boldsymbol{k} - < \boldsymbol{k}}{>} stdev(\boldsymbol{k}) \tag{2.18}$$

Since the ICI vector has as many rows as the number of individuals involved it is possible to rank them according to the coefficient's scores associated with each of them. The computation of the Activities Complexity Index, ACI, is equivalent to the procedure exposed above.

2.6.3 Nestedness test

We perform a *nestedness* test for the matrix shown in Figure 2.2. The test is provided from the software package FALCON[14] (Beckett et al., 2014).

Our output statistics has a p-value equal to 0.02. The test's output statistics is the τ-temperature, the ratio between the nestedness measure of the input matrix and the average nestedness of the null-models.

$$\boldsymbol{T} = \frac{Nest_M}{< Nest_{null} >} \tag{2.19}$$

The null models are computed by the software and represent similar matrices to the case under study. The estimator of nestedness is the z-score of the τ-temperature. If the τ-temperature tends to 0, it means that the model tested is highly nested compared to the null-models. The threshold chosen to accept the null hypothesis of nestedness is $p < 0.05$.

[14] `https://github.com/sjbeckett/FALCON`

Bibliography

M. Aguiar and E. Hurst. Measuring trends in leisure: The allocation of time over five decades. *The Quarterly Journal of Economics*, 122(3):969–1006, 2007.

S. J. Beckett, C. A. Boulton, and H. T. Williams. Falcon: a software package for analysis of nestedness in bipartite networks. *F1000Research*, 3, 2014.

T. Burchardt. Time and income poverty. 2008.

G. Caldarelli, M. Cristelli, A. Gabrielli, L. Pietronero, A. Scala, and A. Tacchella. A Network Analysis of Countries' Export Flows: Firm Grounds for the Building Blocks of the Economy. *PLOS ONE*, 7(10):1–11, 10 2012. doi: 10.1371/journal.pone.0047278. URL `https://doi.org/10.1371/journal.pone.0047278`.

R. Costanza, L. Daly, L. Fioramonti, E. Giovannini, I. Kubiszewski, L. F. Mortensen, K. E. Pickett, K. V. Ragnarsdottir, R. De Vogli, and R. Wilkinson. Modelling and measuring sustainable wellbeing in connection with the un sustainable development goals. *Ecological Economics*, 130:350–355, 2016.

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

J. Deutsch and J. Silber. Measuring multidimensional poverty: An empirical comparison of various approaches. *Review of Income and Wealth*, 51(1):145–174, 2005.

E. Diener. Assessing subjective well-being: Progress and opportunities. *Assessing Well-being*, pages 25–65, 2009.

C. G. Esteban Ortiz-Ospina and M. Roser. Time use. *Our World in Data*, 2020. https://ourworldindata.org/time-use.

M. Fleurbaey, E. Schokkaert, and K. Decancq. What good is happiness? 2008.

R. Hausmann and C. Hidalgo. *The Atlas of Economic Complexity: Mapping Paths to Prosperity*, volume 1 of *MIT Press Books*. The MIT Press, January 2014. ISBN ARRAY(0x40b326a0). URL `https://ideas.repec.org/b/mtp/titles/0262525429.html`.

C. A. Hidalgo and R. Hausmann. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575, 2009.

D. Kahneman and A. B. Krueger. Developments in the measurement of subjective well-being. *Journal of Economic perspectives*, 20(1):3–24, 2006.

E. Kemp-Benedict. An interpretation and critique of the method of reflections. 2014.

T. Lippe, J. Ruijter, E. Ruijter, and W. Raub. Persistent inequalities in time use between men and women: A detailed look at the influence of economic circumstances, policies, and culture. *European Sociological Review*, 26, 03 2010. doi: 10.1093/esr/jcp066.

P. Mealy, J. D. Farmer, and A. Teytelboym. A new interpretation of the economic complexity index. *Alexander, A New Interpretation of the Economic Complexity Index (February 4, 2018)*, 2018.

J. Merz and T. Rathjen. Time and income poverty: An interdependent multidimensional poverty approach with german time use diary data. *Review of Income and Wealth*, 60(3):450–479, 2014.

M. Newman. *Networks: an Introduction*. Oxford University Press, 2010.

I. Peña-López et al. Handbook on constructing composite indicators: methodology and user guide. 2008.

A. Sen. Capability and well-being. *The Quality of Life*, 30, 1993.

J. Stiglitz, A. Sen, and J. Fitoussi. Report of the commission on the measurement of economic performance and social progress (cmepsp). 01 2009.

T. J. VanderWeele. On the promotion of human flourishing. *Proceedings of the National Academy of Sciences*, 114(31):8148–8156, 2017.

Chapter 3

The evolution of cumulative deprivation in Italy between 2007 and 2018: a multidimensional copula-based approach

3.1 Introduction

> *"The interest in multidimensional poverty arose initially out of a concern that monetary poverty measures were not sufficiently capturing the multiple and overlapping deprivations experienced by the poor".* Alkire (2018)

There is a broad agreement that there exist strong interrelations among different dimensions of life and that their interaction defines the overall well-being status of a person.[1] In the last decade, the measurement of individual well-being, together with the qualification of poverty conditions, have been focusing on their multidimensional nature (Atkinson and Bourguignon, 1982; Bourguignon and Chakravarty, 2019; Deutsch and Silber, 2005).

The standard empirical approaches to measuring and interpreting individual well-being provide two types of outcomes. One is a dashboard of indicators associated with each specific dimension. The other one is a synthetic indicator aggregating all the dimensions together. The former is mostly adopted by governments and international institutions to observe population's socioeconomic conditions. With the *dashboard approach*, it is possible to define and measure separately the single contribution of each

[1] An extensive discussion is presented by Stiglitz et al. (2009)

dimension involved. Both methods generally aggregate data across the population. Among the most notorious examples of the dashboard of indicators on well-being are the Sustainable Development Goals (SDG) proposed by the United Nations in the Agenda 2030. In the specific case of Italy, the national statistical institute provides an integrated description of the main economic, social and environmental phenomena through the report on the *Benessere Equo e Sostenibile*[2] (equal and sustainable well-being).

The construction of synthetic multidimensional indicators is also prominent in the field of socioeconomic studies. Multidimensionality in well-being and poverty implies to working on a wide range of data types, from continuous and cardinal data to discrete and qualitative data. This issue is addressed by the aggregation function chosen by the researcher. Techniques for multidimensional indices are generally appearing in the measurement of well-being with data that are firstly aggregated across the population and then in the dimensions. Some examples are the work on the Human Development Index proposed by the United Nations [3] or the Better Life Index proposed by OECD[4]. Micro-founded studies are flourishing for the poverty measurement instead. Following the work of Sen on defining multidimensional poverty, Alkire and Foster (2011), Alkire and Foster, (2011) (A-F), and Bourguignon and Chakravarty (2019) propose methods to identify the poor without aggregating the single-dimensional information. They have observed various types of deprivation and constructed a rule which determines the characteristics held by a poor or deprived person. A notorious poverty indicator inspired by the A-F technique is the At Risk of Poverty and Social Exclusion (AROPE) rate computed by the Eurostat.

As suggested by Decancq and Schokkaert (2016), when analysing well-being, there are some fundamental aspects to follow: *i)* the within-dimension distributional characteristics, *ii)* the identification of situations of cumulative deprivation in across various dimensions, *iii)* the between-dimension interrelations. The former aspect requires focusing the attention on the heterogeneity that lies across the population concerning the outcomes in a specific dimension. The latter two aspects, more frequently neglected, highlight the correlation between the dimensions characterising a specific well-being or poverty condition. A simple example can help in understanding why it is so.

Let assume a researcher wants to study the welfare conditions of two different societies through the observation of two socioeconomic dimensions: income and health. Furthermore, let assume that the observation is done on two representative individuals

[2] Link to the BES report: https://www.istat.it/it/archivio/rapporto+bes

[3] Link to the UNHD reports website: http://hdr.undp.org/en

[4] Link to the OECD BLI reports website: http://www.oecdbetterlifeindex.org/

of each society.

The following tables describe the two societies in terms of the percentage scores, on a 0-100 scale, of each person in each dimension.

Table 3.1: Comparing two hypothetical societal multidimensional distributions

(a) Society A

Individual	**Income**	**Health**	$I(x_i^J)$
i_1	10	90	**50**
i_2	90	10	**50**
$\bar{\mathbf{x}}_\mathbf{j}$	**50**	**50**	

(b) Society B

Individual	**Income**	**Health**	$I(x_i^J)$
i_1	10	10	**10**
i_2	90	90	**90**
$\bar{\mathbf{x}}_\mathbf{j}$	**50**	**50**	

If the researcher would address a distributional study on the well-being of the people in the two societies through a dashboard approach, she would end up with the result presented in the last rows of both tables 3.2a and 3.2b which represent a simple descriptive parameter for each dimension distribution, i.e. the mean. The conclusion drawn from this comparison would be that the two societies are equal in terms of average income and health.

Alternatively, following a multidimensional composite indicator approach she could end up with the last columns presented in the two tables. The outcomes of the computation coincide with a simple column-wise average aggregation of the individual-specific scores in each dimension adopting equal weights. What would emerge from the analysis is a completely different picture with respect to the previous result: the two societies are very different and the society represented by table 3.2b might be more unequal in terms of income and health distribution.

Although the two approaches presented through this example depict fundamental pieces of information to describe the socioeconomic status of the population of a country, they are lack an element to enrich the comparison between two societies. Indeed, the weights assigned to the two dimensions in the aggregation may vary, changing the outcome index score used for the two societies' comparison. Furthermore, both approaches miss the important fact that there might be non-randomness in the repeated low outcomes appearing in all the welfare dimensions of society *b*. In other words, the low outcomes recurring for the same individual are a symptom of a condition of *cumulative deprivation.* This possible outcome could explain the high concentration in the tails of multidimensional well-being joint distributions already discussed in the literature (Stiglitz et al., 2009; Tkach and Gigliarano, 2020).

In order to study the phenomenon of cumulative deprivation and assess the relation the observed society's have with the distributional settlement, it is necessary to investigate the within-dimensional dependence.

The degree of dependence between dimensions of people's well-being (or people's poverty, when observing the multidimensional deprivations) could be a supportive instrument to understand how a welfare state works to overcome the linkages between features of social exclusion and socioeconomic inequalities. For this reason, the dependence structure can be enriching the current framework of studying socioeconomic inequalities.

The nature of dependence is not a universally defined concept. In statistics, the term *dependence* might be associated with the presence of association within two random variables. The most known scale-invariant Kendall's $\boldsymbol{\tau}$ and Spearman's $\boldsymbol{\rho}$ provide a measure of the concordance between two random variables. The copula-based statistical techniques can help to derive multidimensional correlation indices and allow the investigation on the dependence structure within the dimensions of a multivariate distribution function. An interesting aspect of such a technique is that it is rank-based and, consequently, neutral to the unit of measurement of the dimensions involved. For this reason, such statistical technique can be useful for studying the interdependence lying across different dimensions of well-being. The shortfall of being rank-based is that it can be used only as a relative measure for defining socioeconomic inequalities. As other socioeconomic relative indicators computed with country-specific data, such as the At Risk of Poverty (ARP) from the Eurostat[5], it does not provide a fully informative between-country description of the absolute conditions of poor and cumulatively deprived people. The copula function is a joint cumulative probability distribution function with standard uniform marginal distributions (Nelsen, 2007). Assuming to observe a set of dimension-specific outcomes for a sample of individuals in the society, the copula function of this multidimensional set will return the proportion of individuals in the society who are ranked less than a specific combination of outcome positions. With the copula function, by mapping the society in terms of the joint positions in each well-being outcome dimension, it is also possible to investigate the level of inter-dependencies across these dimensions.

There is a growing literature which adopts copula-based methods to measure the dependence among socioeconomic variables. Quinn (2007) adopted the copula framework for measuring the association between health and income. A contribution of Aaberge et al. (2018) focused on the changing dynamics of the composition of top incomes across time. Decancq (2014), Pérez (2015) and Pérez and Prieto-Alaiz (2016) used copulas for the measurement of global dependence among the dimensions of the Human Development Index, respectively using the Russian Panel data and data from the Human Development Report. More closely related to multidimensional poverty

[5] Which is computed as the 60% of the median equivalent household income of a country.

studies is the contribution of Tkach and Gigliarano (2020), that adopted the information regarding the dependence structure emerged from three poverty dimensions as a data-driven technique to derive weights in multidimensional poverty measurement. Furthermore, the recent paper of García-Gómez et al. (2020) propose to study multidimensional poverty through the dependence assessment of the three AROPE index dimensions.

All the presented studies are using the copula techniques to investigate the overall dependence of the chosen features which are describing a multidimensional phenomenon. This type of association measurement is also named global dependence. Global dependence is a profoundly interesting tool, but it does not provide information on the different patterns of dependence throughout the distributions of the dimensions considered. The correlation across the dimensions of well-being is non-trivial and can vary as well, along with the distribution of each dimension considered. With this paper, the joint-distributional properties of multidimensional poverty are investigated through the assessment of inter-dimensional dependence across cumulatively deprived people.

The copula-based technique is not used here to assess global dependence but to quantify the dependence occurring occurring with respect to a particular set of outcomes. Decancq (2020) presented a technique which perfectly fits with the necessity of investigating the dependence lying at the tails of the multidimensional copula function, introducing the Diagonal Dependence Index.

The empirical application focuses on cumulative deprivation in Italy between 2007 and 2018. The evolution of cumulative deprivation is used to observe multidimensional poverty and to provide insights on the dependence structure emerging among the variables involved in the analysis. The dimensions of well-being considered are selected based on the recommendations of Stiglitz et al. (2009) in the *Report of the Commission on the Measurement of Economic Performance and Social Progress*. These dimensions are believed to be fundamentally, but non-exhaustively, describing the presence of multidimensional poverty. The EU-SILC database is used in order to construct the indices for each dimension. The dimensions are income, job conditions, educational attainment, health status and housing quality.

The paper is structured as follows. In Section 3.2, a broad presentation of the copula methodology and the diagonal dependence index. In Section 3.3, the data and techniques used for constructing indices for each of the chosen dimensions. In section 3.4, the results are presented and discussed. In Section 3.5, conclusive thoughts and ideas for further extensions of the study are presented.

3.2 Methodological framework

3.2.1 Copula Function

The copula function is a particular multivariate distribution function with uniform univariate margins. Therefore, it is also described as a multivariate function which aggregates all its marginal univariate components.

In this section, the intuition of what is a copula function and of the utility in using it in the field of multidimensional well-being measurement is supported by a detailed illustration of its statistical properties. The phenomenon of individual well-being can be statistically ascribed by a random vector of multiple dimensions (e.g. income, housing conditions, education, job conditions and health).

Let the d-dimensional random vector $X = (X_1, ..., X_d)$ describe the distribution of all the $(1, ..., d)$ dimensions of well-being across individuals in a society. Given a set of realisations observed from the sample fro each dimension $x = (x_1, ..., x_d)$, the joint cumulative distribution of the set of dimensions is defined as

$$F(x_1, ..., x_d) = P(X_1 \leq x_1, ..., X_d \leq x_d) \tag{3.1}$$

Let $F_j(x_j)$ be the j^{th} marginal distribution of $F(x_1, ..., x_d)$, for $j = (1, ..., d)$. Given the joint cumulative distribution function, it is possible to count the proportion of people in the society who have less than or exactly x_j in every j^{th} dimension of well-being. Equivalently, the proportion of individuals in the society who have strictly more than x_j in all d-dimensions is given by the survival function $\bar{F}(x_1, ..., x_d)$, where $\bar{F}(x_1, ..., x_d)$ is the multidimensional complement to one of the *cdf* $F(x_1, ..., x_d)$. In other words, the survival function contains all the outcomes of the random variables which are jointly larger than the set of values x.

The individual who exactly has the amount x_j in dimension j, has a position p_j referred to the j^{th} marginal distribution of the *cdf* $F_j(x_j)$. The positional outcomes of a person on all the d dimensions is the set $p = (p_1, ..., p_d)$ and corresponds to all the ranks that a certain individual has along each single dimension. The positional outcomes statistically coincide with the marginal distributions of the *cdf* F.[6] As follows from the probability integral transform theorem, the normalised rankings of each dimension, the univariate margins of $F(x_1, ..., x_d)$, are uniformly distributed as $U(0, 1)$.

Recalling the second definition of the copula function provided by Nelsen (2007), *"[..] copulas are functions that join or "couple" multivariate distribution functions*

[6] See section 3.6 for a more extensive explanation of the link between the positional outcomes and the marginal distributions of F

to their one-dimensional marginal distribution functions.", the ranking distributions represent the "ingredients" of the copula function of X.

The formal definition of copulas can be derived from the Sklar (1996, 1959).

The Sklar's Theorem For any d-dimensional distribution function F_X with univariate margins $F_1, ..., F_d$, there exist a copula $C : [0,1]^d \longrightarrow [0,1]$ such that, for all $x = (x_1, ..., x_d) \in \mathbb{R}$

$$F_X(x_1, ..., x_d) = C_X(F_1(x_1), ..., F_d(x_d)). \tag{3.2}$$

And, if all F_j for $j = 1, ..., d$ are continuous and strictly increasing, then C_X is uniquely defined in the unit hypercube $[0,1]^d$, and it is uniquely determined as $\Pi_{j=1}^{d} RangeF_j$.

Given that the inverse of a continuous and strictly increasing *cdf* F is $F^{-1} = F^{\leftarrow}$, the copula function of $F(x_1, ..., x_d)$ can be uniquely defined as follows:

$$C(p) = F(F_1^{\leftarrow}(p_1), ..., F_d^{\leftarrow}(p_d)) \tag{3.3}$$

and it is determined on the positional set $p = (Range(F_1) \times ... \times Range(F_d))$.

The Sklar's Theorem combines precisely the univariate marginal densities to form a d-dimensional joint distribution. This theorem depicts the reason of the use of copulas in statistical applications to study dependence between components of a random vector.[7]

Given the random vector X and the position vector P representing the set of distributions of the ranked outcomes observed, being $F_j(X_j)$ the marginal distribution of the j^{th} dimension of X, its copula function is a multivariate distribution C_X defined as:

$$C_X(p_1, ..., p_d) = Pr[F_1(X1) \leq p_1 \text{ and ... and } F_d(X_d) \leq p_d] \tag{3.4}$$

C_X expresses the proportion of individuals in the society who are outranked by the specific position set $p = p_1, ..., p_d$. Equivalently, the survival copula $\bar{C}_X(p_1, ..., p_d)$ represents the proportion of individuals who are outranking the same position set.

Therefore, the observations described by the components of the random vector x can be converted into pseudo-observations p (Charpentier et al., 2007) applying a simple ranking to all the series. The pseudo-observations keep the information about the relative position of individuals in the distribution and ignore the information concerning the absolute values describing the phenomenon.

[7] A simplified illustration of the Sklar's Theorem is presented by Hofert et al. (2019).

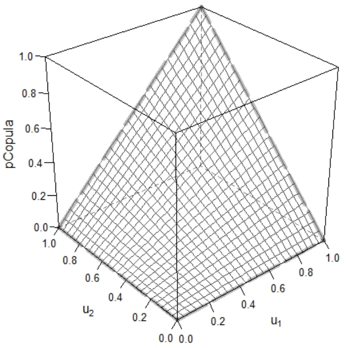


Figure 3.1: Independence Copula (two dimensions)

Intuitively, in case of absence of any type of dependence among the dimensions, the copula function would simply be the product of the d-margins. This example is presenting the simplest copula (shown in figure 3.1). For a random vector $\boldsymbol{P} = (P_1, ..., P_d)$ with $P_1, ..., P_d \sim^{ind} U(0, 1)$, the *independence copula* is

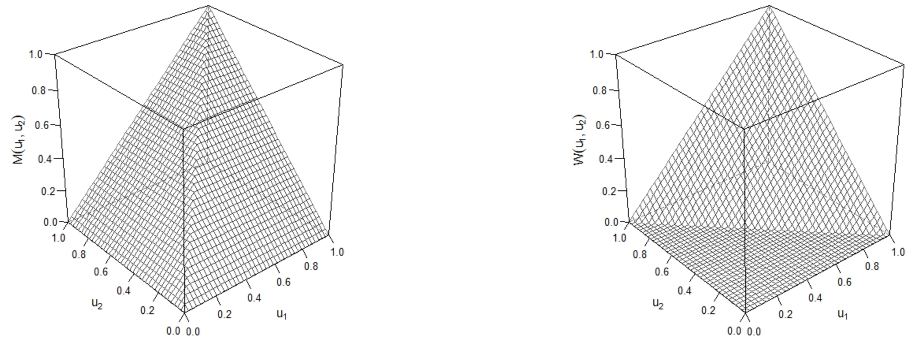
$$\Pi(\boldsymbol{p}) = \prod_{j=1}^{d} p_j, \;\; \boldsymbol{p} \in [0,1]^d \tag{3.5}$$

In absence of interrelation across the dimensions, the aggregation function is a simple product. There are other two types of copulas that need to be mentioned because they represent two extreme cases. They are known as the Fréchet-Hoeffding Bounds (F-H), and they represent the lower and upper bound of every copula. They are respectively $W(\boldsymbol{p}) = max\big\{ \sum_{j=1}^{d} p_j - d + 1, 0\big\}$ and $M(\boldsymbol{p}) = min_{1 \leq j \leq d} \{p_j\}$, for $\boldsymbol{p} \in [0,1]^d$. For any given d-dimensional copula C, the theorem of Hoeffding (1940) and Fréchet(1951) states that any copula C is point-wise bounded from below by a lower bound W, and from above by an upper bound M. The relation among them is the following:

$$W(\boldsymbol{p}) \leq C(\boldsymbol{p}) \leq M(\boldsymbol{p}), \;\; \boldsymbol{p} \in [0,1]^d \tag{3.6}$$

These two extreme cases are referred to the type of dependence taking place among the dimensions. We could interpret the lower bound as representing complete counter-monotonicity among the dimensions, and the upper bound as the complete co-monotonicity among the dimensions.

Figures 3.2 and 3.1 present a bi-variate illustration of, respectively, the copula's extreme bounds and the independence case.[8]. In these figures, the lower bound density is $W(u_1, u_2)$, the upper bound density is $M(u_1, u_2)$.



(a) Upper bound

(b) Lower bound

Figure 3.2: Fréchet-Hoeffding Bounds (F-H)

3.2.2 The measures of dependence and the copula sections

The copula function is useful to investigate the dependence or association between random variables. The dependence, or association, between dimensions of a multivariate copula function is studied taking into consideration two of the extreme cases already introduced. Namely, the co-monotonic case, which implies maximal dependence, and the independence case in which there is a randomly determined association between the elements of the copula. As already anticipated, the dependence can be investigated as a global aspect characterising the whole distribution of the random multivariate vector, or as a phenomenon that varies across the distribution. Since the intention is to investigate the dependence across the well-being dimensions at low levels of their distribution, the concept of tail monotonicity and, more specifically of tail dependence, depicted by Nelsen (2007) can be very useful.

The tail monotonicity is a property of the copula when it emerges a stronger association between its dimensions along the left and the right quadrant of their joint

[8] The plots are computed with randomly generated data using the plot commands of the R-Studio package named *Copula* See Hofert et al. (2019) for further copula plot examples.

distribution. Let X and Y be random variables, if

$$P[Y \leq y|X \leq x] \geq P[Y \leq y], \tag{3.7}$$

which could be written also with the following inequality:

$$P[Y \leq y|X \leq x] \geq P[Y \leq y|X \leq \infty], \tag{3.8}$$

and if the conditional distribution function $P[Y \leq y|X \leq x]$ is a non-increasing function of x, then Y is *left tail decreasing in* X. In other words, for small values of x the conditional distribution function is associated to a high probability, while its value is not increasing in higher values of x.

Within our specific case, the left tail dependence is expressed by the conditional probability that an individual having a position $p_j \leq p$ in the distribution of dimension j, has a position $p_i \leq p$ in the distribution of dimension i. Hence, the left tail monotonicity implies positive quadrant dependence (PQD):

$$P[X \leq x, Y \leq y] = P[X \leq x]\, P[Y \leq y|X \leq x] \geq P[X \leq x]P[Y \leq y] \tag{3.9}$$

With the left tail dependence in the copula framework, important insights for quantifying the phenomenon of cumulative deprivation in well-being are shown. To provide an intuitive geometric interpretation of tail dependence, it is necessary to refer to the *sections* of the estimated copula function.

The sections of a bi-dimensional copula are three: the horizontal, the vertical and the diagonal section. In mathematical terms, the sections of a bi-dimensional copula, $C(u_1, u_2)$, are two-dimensional planes that slice the surface of the copula density function and fall perpendicularly on the (u_1, u_2) plane. While the vertical and the horizontal sections are planes that slice the copula density at a fixed point of u_1 or of u_2, the diagonal section is the function in the $[0, 1]$ interval defined as $\delta_C(p) = C(u_1 = p, u_2 = p)$ where p can be any value in the $[0, 1]$ interval. The *diagonal* section is particularly interesting for the issue of this study.

The following figure shows the contour plot of an independent bi-dimensional copula as shown in figure 3.1. The diagonal section of that copula is a plane falling perpendicularly on the plane of the contour diagrams, and cutting it in the points in which $u_1 = u_2$ (where is the 45° line).

The diagonal section show us the density function of the copula at a specific combination of points of all its dimensions. Precisely, that combination of positions u_1 and u_2 both being equal each other.

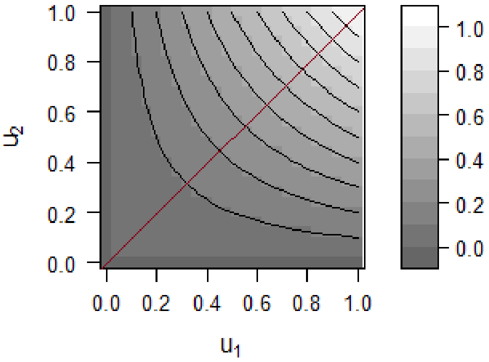


Figure 3.3: Contour plot of the Independence copula

Decancq (2020) proposes a simple and intuitive way to observe the diagonal section of the copula function with the construction of the *diagonal dependence diagram*. In the diagonal dependence diagram there are two curves, the *Downward Diagonal Dependence Curve* representing the diagonal section of the copula function and the *Upward Diagonal Dependence Curve* representing the diagonal section of the survival function. They are respectively showing the proportion of population that is outranked by a specific positional combination set and the proportion of population which is outranking a specific positional combination.

Given a d-dimensional random vector X with copula function C_X and survival function $\bar{C}_X$, its Downward Diagonal Dependence Curve D_X is defined as:

$$D_X(p) = C_X(p, ..., p); \quad \forall\, p \in [0, 1] \tag{3.10}$$

and its Upward Diagonal Dependence Curve $\bar{D}_X$ is defined as:

$$\bar{D}_X(p) = \bar{C}_X(1 - p, ..., 1 - p); \quad \forall\, p \in [0, 1] \tag{3.11}$$

The graphical representation of the diagonal section of a d-dimensional copula (figure 3.4) is a two-dimensional plot having on its x-axis the points previously shown in figure 3.3 on the main diagonal. This line represents the set containing all the combinations of positions between the d variables $\boldsymbol{p} = (p_1, ..., p_d)$ such that all the positions in one dimension are equal to the positions in the other one. The y-axis represents the proportion of population that is outranked by each position combination in the set. In other words, the y-axis of the Diagonal Dependence Diagram coincides with the copula density. Figure 3.4 is an example of how a Downward Diagonal Dependence Curve of

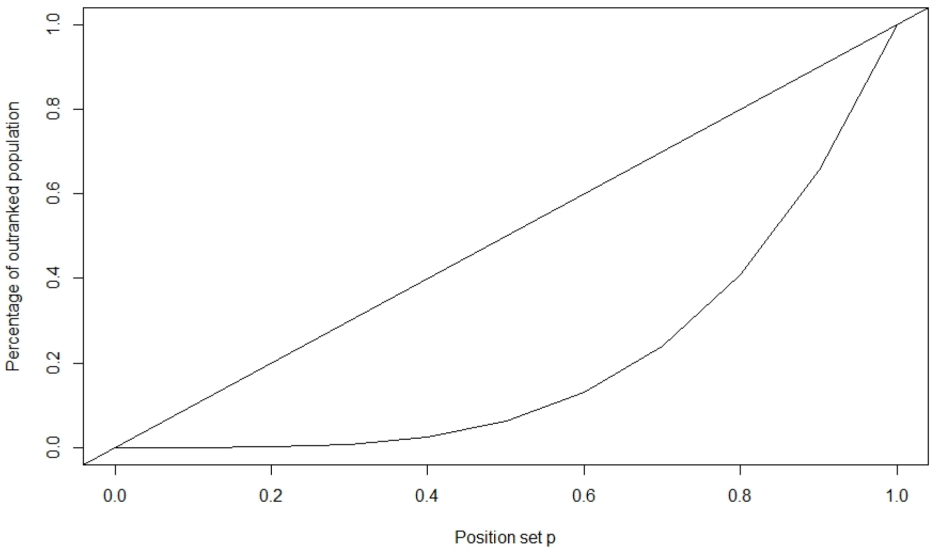


Figure 3.4: Downward Diagonal Dependence Curve

a d-dimensional copula function looks like.

The diagonal dependence diagram is meant to compare the diagonal section of the empirical copula estimated on the multiple dimensions of well-being, with the case of a co-monotonic copula. The co-monotonic copula describes a "feudal" or "cast" society: being poor in one dimension automatically implies being poor in all the other dimensions. Given that the highest density of a two-dimensional co-monotonic copula function lies exactly on the points in which the individual positions on each dimension are the same, the diagonal section of a co-monotonic copula coincides with the 45° line. The comparisons can be done as well between copulas and between survival functions.

Decancq (2020) proposes a way to assess the dominance of a d-dimensional random vector X on another Y according to the downward diagonal dependence orderings if

$$D_X(p) \geq D_Y(p); \;\; \forall \, p \in [0,1] \tag{3.12}$$

and to the upward diagonal dependence orderings if

$$\bar{D}_X(1-p) \geq \bar{D}_Y(1-p); \;\; \forall \, p \in [0,1] \tag{3.13}$$

With both diagonal dependence orderings it is possible to operate a pair-wise comparison of two different copulas with respect to their proximity to the co-monotonic case. In order to observe global dominance within the diagonal dependence orderings, both cases shown in equations 3.12-3.13 have to take place.

Going back to the tail dependence concept, the co-monotonic case will present the highest tail dependence. The tail dependence can be measured with the tail dependence parameters. Let p be the point indicating the joint positions of individuals along the dimensions and $C(p,p) = \delta_C(p)$ the diagonal section of the copula at p. If there is left tail dependence, there exists a limit for p approaching 0 from above, of the conditional probability that Y is less than the percentile p, given that X is less than the percentile p. Analogously, there is right tail dependence if there exist a limit for p approaching 1 from below, of the conditional probability that Y is greater than the percentile p, given that X is greater than percentile p. In other words, it would be necessary to look at the density of the copula along its diagonal section as well as the density of the survival along its diagonal section in order to measure the tail dependence.

Looking at figure 3.4, a parallel with the Lorenz curve arises immediately. Following the principles of tail dependence in order to measure the level of diagonal dependence among the components of a copula C_X, Decancq (2020) proposes to calculate the area underlying both the Downward and the Upward Diagonal Dependence Curves and derive an index measure of downward and upward dependence. The aggregation of these two measures is the Diagonal Depenendence Index (DDI). The Diagonal Depenendence Index (DDI) represents a measure of proximity of the society described by the copula with the society described by the co-monotonic copula. The greater this area will be, the closer the curve to the diagonal line and the more unequal the society will be.
The diagonal dependence index is obtained by averaging the downward and upward diagonal dependence indices that are respectively:

$$\boldsymbol{\delta}_d^-(X) = \frac{2(d+1)\int_I D_X(p)dp - 2}{d-1} \tag{3.14}$$

and

$$\boldsymbol{\delta}_d^+(X) = \frac{2(d+1)\int_I \bar{D}_X(p)dp - 2}{d-1} \tag{3.15}$$

The computation of these two indices can give us an idea of the level of the interrelations lying across the dimensions of the phenomenon described by the sample data. The resulting Diagonal Dependence Index is computed as follows:

$$\boldsymbol{\delta_d}(X) = \frac{\boldsymbol{\delta}_d^-(X) + \boldsymbol{\delta}_d^+(X)}{2} \tag{3.16}$$

As Decancq (2020) had demonstrated in his paper, the diagonal dependence index turns out to be equal to the multidimensional generalisation of the Spearman's Footrule.

3.3 Data

The empirical application uses European Union Survey on Income and Living Conditions (EU-SILC) cross-sectional data for Italy from the year 2007 to 2018. The sample size each year amounts approximately to 40 thousands records but a sub-sample is selected representing the workforce (people aged 25-60). The exclusion of younger individuals is motivated by the necessity of avoiding the count of those who are still in education as people deprived in education.

The sample size is representative of a population which amounts approximately to 30 million people every year. This population is composed of 50,2% females and 49.8% males.

The following five dimensions of well-being have been built: income, working conditions, educational attainment, health status, housing quality conditions. Although these dimensions provide a non-exhaustive description of the total well-being status, they are retained to be relevant for describing individual well-being. The time period and the variable construction was, of course, conditional to the availability of data in the EU-SILC yearly database. Each dimension has been computed separately and for every single year. For every dimension, the data are scaled on a continuous sequence of uniformly distributed values. To do so, the individual dimensional outcomes have been ranked. When the dimension-specific measure presents an elevated number of ties, a second level of ranking has been performed. For all the dimensions the remaining ties have been sorted randomly.

More specifically, the *income* dimension is the disposable household income divided by the EU-SILC specific household equivalence scale. All the within-household ties are assigned a random ranking. The *housing condition* dimension is an index constructed on two steps. First, several dwelling-related features have been summarised in ordered scales (counting the no-deprivations in dwelling furniture, neighbourhood features, household size). Second, for each individual, all the dwelling-related features have been aggregated through a geometric mean to provide a summary measure on housing condition. The use of geometric mean allows for aggregating without imposing full substitutability across the different scores. The *educational attainment* is constructed applying two ranking levels. At first people have been ranked according to the ISCED level; second, according to the years in education. The remaining ties have been ranked randomly. Similarly, for the *job condition* dimension the first ranking level is derived from the work intensity index provided by the SILC data. The second level of ranking is performed according to the contract type of the person. The contract type takes three levels: permanent contract, self-employed, temporary contract.

The *health* dimension is derived from the estimated latent general health status extrapolated from the Self-Assessed general Health (SAH) distribution conditional on some personal characteristics and health-related behaviours available in the data set. Table 3.3 provides a synthetic illustration of the dimension-specific index construction.

Below it is discussed in more detail the health dimension construction. Due to the absence of objective health measures in survey data, the use Self-Assessed general Health (SAH) in social studies is very common.[9]

A pitfall of such a measure is that it provides a very low variability among the respondents and difficulty to observe it from some distributional perspective. Many empirical studies adopted non-linear regression techniques to translate the subjective health categories into a cardinal measure representing the estimated underlying latent health status taking a continuous form.[10] The health dimension in this study represents a latent general health status estimated using data on the single-year declaration of self-assessed general health (a factor which takes five levels indicating increasing health conditions) from the EU-SILC individual data set. The estimation takes into account other health habits/features, and some personal characteristics (presence of any chronic disease, limitations in everyday activity, age and gender).

Calling $\boldsymbol{X'}$ the vector of the regressors presented, the estimated ordered logit model of latent health status is: $h^* = \boldsymbol{X'}\beta + \epsilon$. Then, the predicted latent general health status for each individual is $h^*_i = \boldsymbol{X'_i}\beta$, also named *z-score*.

[9] Idler and Benyamini (1997) have demonstrated that the SAH could be a good predictor of other health measures such as life expectancy and use of healthcare.

[10] These approaches have been validated by the contribution of Van Doorslaer and Jones (2003).

Table 3.3: Description and methods of construction of the selected well-being dimensions

Dimension	Data used	Method to derive the outcome variable
Household income	Equivalised total disposable household income	*Unit:* Household
Health status	General health status, presence of chronic illness, limitations in everyday activities, age, gender	*Unit:* Individual *Method:* Ordered logit estimation *Outcome variable:* *z-score*/latent general health
Housing quality	Rooms per person, material conditions of the household, house location social and economic conditions	*Unit:* Household *Method:* Geometric mean of all the dwelling-related no-deprivations *Outcome variable:* Index of quality of housing
Educational attainment	ISCED level, years of schooling	*Unit:* Individual *Method:* Two levels of ranking: 1) ranking of people with respect to ISCED level 2) ranking with respect to the years of schooling *Outcome variable:* Educational attainment by time
Working condition	work intensity, type of contract (permanent or temporary)	*Unit:* Individual *Method:* Two levels of ranking: 1) ranking of people with respect to work intensity 2) ranking with respect to the type of contract *Outcome variable:* Working condition

3.4 Results and discussion

In this section, it is presented the outcome of the empirical five-dimensional copula density function. Studying what type of multidimensional interaction is associated with the highest density, the phenomenon of cumulative deprivation is depicted and contextualised. The pair-wise yearly dependence comparisons and the empirical diagonal dependence index are presented.

In order to derive the copula density, the population for each of the five dimensions in every yearly sample is ranked. The ranked series of each dimension represents the margins of the copula density. The empirical copula density is derived by grouping the ranked population according to the combination of percentiles in each dimension. Technically, the population is divided into N groups. Being q the number of quantiles of the uni-dimensional ranked series and d the number of dimensions, $N = q^d$. The number of positional groups quantifies the grid size belonging to the $\boldsymbol{I}^d = [0, 1]^d$ set. In other words, N represents all the possible combinations of the given quantiles among d dimensions.

The copula density can be evaluated simply by counting the proportion of society that is falling in each of these groups of positional combinations. Given that it is not necessary to compute an overall dependence measure, there is no need of ordering the positional combinations. The only necessary information is the proportion of population outranked by combination associated with the same quantile in each dimension. In other words, it is necessary to create groups according to the highest observed quantile position among all the dimensions. Therefore, by focusing on those whose maximal position corresponding to the lowest quantile, we can individuate people experiencing cumulative deprivation.

The set of cumulatively deprived people includes who falls in the low-income class, having a bad general health status, experiencing low-quality conditions of housing and job, and not being highly educated.

100-quantiles, or simply percentiles, have been adopted for computing the yearly copula densities and the diagonal sections. For describing the cumulative deprivation condition and incidence across the population a higher quantile size has been employed. More precisely, the quantiles selected for observing the cumulatively deprived population are terciles, quintiles and deciles.[11]

Of course, the lower the size of the quantile, the narrower will be the sample size referred to a single positional-combination. In the case of three quantiles, there are

[11] The terciles imply to group population in three groups, each one representing the 33.33% of the population, the quintiles represent the 20% of total, and the deciles represent the 10% of total.

$3^5 = 243$ groups of people according to the possible combinations of positions per percentile. In the case of hundred quantiles, there can be 100^5 possible combinations. Of course, not all of these combinations do appear within our sample. Both because some combinations are not likely to be realistic; and due to limited size of the sample used. If there was no dependence among these five dimensions, the probabilities for each of the N combinations of the quantiles would always be the same. More precisely, if $q = 3$, the probability of the independence case is $0.004 (= 1/243)$ for every positional group. If $q = 5$, the probability for the independence case of every single positional group is $0, 00032(= 1/3125)$. For q being equal to 10 or 100, the independence case will assign respectively a probability of $1^{(-5)}$ and $e^{(-10)}$ to each positional group.

The cumulative deprivation

In order to have an initial descriptive idea of the "importance" of cumulative deprivation phenomenon, is provided a summary of its frequency of appearance in each yearly grid of combinations. By observing the head-count of people who happen to "fall" in each quantile combination of the five dimensions, we can rank the combinations according to the population group sizes. The intent is to check whether the head-count of who falls in the lowest quantile in all the dimensions is frequently observed over the total possible positional combinations.

By simply ranking all the possible combinations according to the population group sizes, it is provided an identification, for each year, of the quantile size for which the cumulative deprivation falls among the most frequent combinations. Table 3.4 shows, for the selected quantile sizes, the correspondent year in which cumulative deprivation is appearing as one of the three most frequent cases.

Table 3.4: Presence of cumulative deprivation among three most frequent cases

Year	**3 q**	**5 q**	**10 q**	**100 q**
2007	Yes	Yes	Yes	No
2008	Yes	Yes	Yes	No
2009	Yes	Yes	Yes	No
2010	Yes	Yes	Yes	No
2011	Yes	Yes	Yes	No
2012	Yes	Yes	Yes	No
2013	Yes	Yes	Yes	No
2014	Yes	Yes	Yes	No
2015	Yes	Yes	Yes	No
2016	Yes	Yes	No	No

2017	Yes	Yes	Yes	No
2018	Yes	Yes	Yes	No

As it is possible to see from table 3.4, the phenomenon of cumulative deprivation is taking place considerably in the sample if we observe it up to a division of the ranking distributions into deciles (or terciles and quintiles). This phenomenon can already be a signal of the intrinsic interconnection lying among the selected dimensions. Rising the cut points for partitioning the ranked population reduces the incidence of cumulative deprivation both due to the narrower reference population over the total possible groups and also because of the limited representative power of survey data regarding the extremes of the distribution.[12]

Figure 3.5 is the time series of the proportion of population associated with the cumulative deprivation case. The cross-sectional weights have been applied in order to present the actual population quantities. The proportion of Italian population, represented by the EU-SILC data, lying in the bottom tercile (33%) for all the dimensions, is almost 1 million people in 2018.

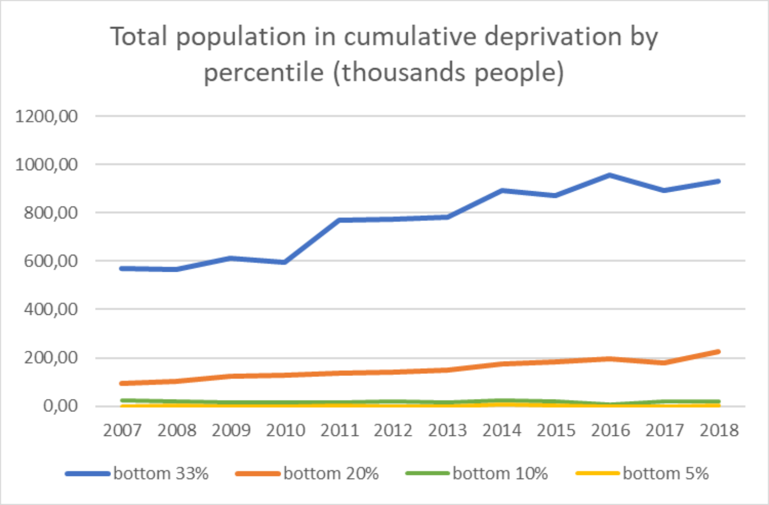


Figure 3.5: Cumulative deprivation from a time series perspective

With Figure 3.6, the headcount of cumulatively deprived is shown in proportion with the headcount of people in all other possible positional combinations in the sample. In average, the probability of falling into cumulative deprivation for those who belongs to the bottom tercile of the sample is around 3%. While, if we look at the people who

[12] Survey data are useful to have access to a multidimensional set of information on individuals and households. However, due to a limited sample size, too small quantiles of the sample distributions are not well represented. Therefore, for describing the cumulative deprivation case, the focus is on the terciles and quintiles.

are in the bottom 20% in all the dimensions, it is 0.65%. In both cases, the average percentage of the population falling in cumulative deprivation is respectively 2.4 and 4 times higher than the probability of the hypothetical case of no dependence. Figure 3.6 shows that the pattern of cumulative deprivation intensity has a peak in the years 2014 and 2015.

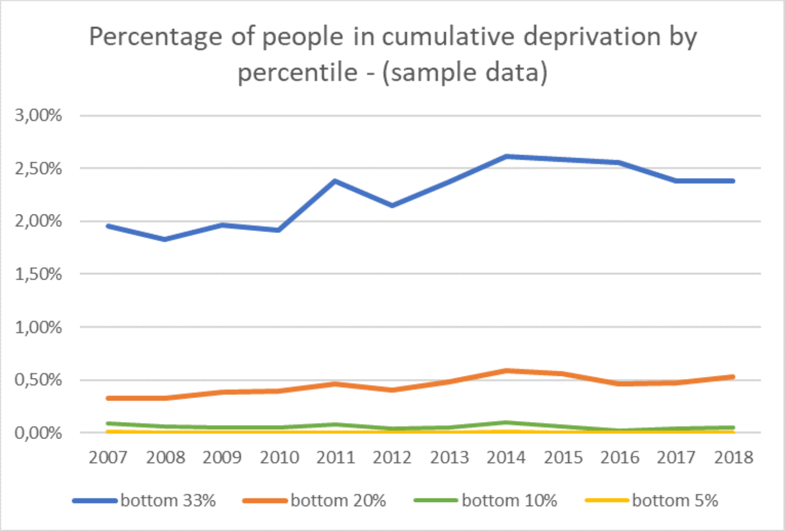


Figure 3.6: Cumulative deprivation from a time series perspective

The evidence is that, while the total population in cumulative deprivation is growing, it is possible to identify the years 2014 and 2015 as the years of the higher increase of the coexistence of the deprivation in income, education, working condition, housing quality and health status among people in the sample. This episode is undoubtedly non-random, and it reflect the lagged response of multidimensional poverty to the crisis disrupted in previous years. It could be explaining how the austerity measures cutting public expenditure may exacerbate the inability of welfare policies to avoid the vicious cycle of poverty. In order to give more solidity to this ascertainment, a description of the cumulatively deprived people's characteristics is necessary.

Who is cumulatively deprived?

It is hereby illustrated with more detail the composition of the population belonging to the bottom tercile of the population distribution in each dimension, namely, the cumulatively deprived people. For practical reasons, the proposed timeline for this observation is for only one year, and the chosen year is 2014, as it is the peak of the observed trend. In this year there were precisely 869,75 thousands people, the 2.9% of total population represented by our sample. Table 3.5 illustrates the proportions of who is cumulatively deprived by certain socio-demographic groups: gender, migration status and activity status. It is noticeable that there is a higher proportion of females than males within the cumulatively deprived population. Not surprisingly, the cumulatively deprived people are mostly inactive or unemployed. Last, the $10,5\%$ of the cumulatively deprived is coming from an Extra-EU country. [13]

Table 3.5: Population in Cumulative Deprivation by Socio-Demographic Characteristics

		Cumulatively deprived population (Thousands)	**Cumulatively deprived population**
Gender	Male	381,76	43,9%
	Female	487,99	56,1%
Country	Local	767,06	88,2%
	EU	10,85	1,2%
	Extra-EU	91,28	10,5%
Activity status	Employed or self-employed	182,00	20,9%
	Unemployed	242,37	27,9%
	Inactive	423,92	48,7%

Figure 3.7 shows the proportions of the deprived males and females by activity status.

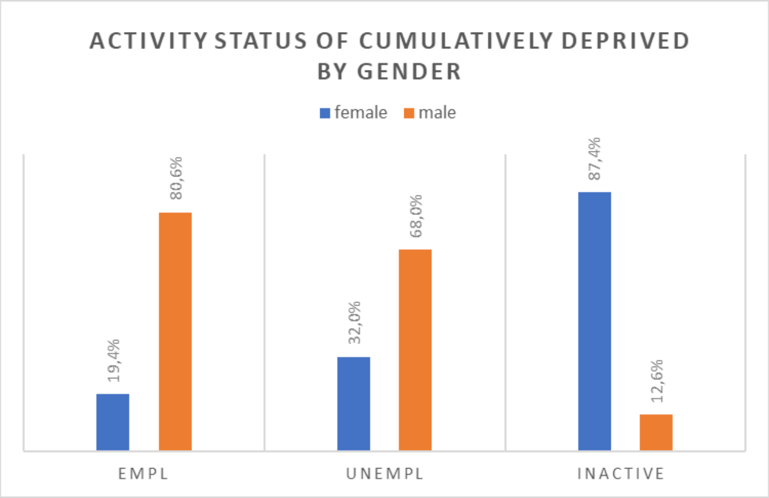


Figure 3.7: Cumulatively deprived males and females by activity status

This picture depicts how the traditionally weaker elements of the society in terms of poverty risks are those who live in cumulative deprivation. For example, it is not

[13] The Extra-EU population amounts to the $7,8\%$ of the 25-60 years old population.

surprisingly emerging the strong gender gap with respect to both poverty and the participation to the workforce. Being the population of male more present among the employed population, 80% of the cumulatively deprived among the employed people are men. The females progressively "catch-up" with the men when observing the unemployed population, up to gain the primacy in the case of inactivity. Recalling the proportions in table 3.5, it is shown that the inactive population represents almost half of the total cumulatively deprived population.

Among the inactive population in 2014, the 81% are females.[14] Thus, even if at this stage of the study it is not provided a parametric estimation, it is possible to expect that being a female could represent a higher probability to fall in cumulative deprivation sharpened by the higher probability for females of being inactive.

The estimated Downward Diagonal Dependence Curve

In the following section, the diagonal sections of the copula functions for each year are presented and illustrated. The copula diagonal section is a slice of the copula density taken precisely at the points in which the positions are equal in all the dimensions (figures 3.4 and 3.3 already provided an intuitive illustration). The study of the diagonal section of the yearly copula allows us to compare the estimates with the cases of co-monotonicity and independence. The co-monotonic case coincides with the upper bound of the copula as defined by Fréchet-Hoeffding Bounds and introduced in the methodological section. The co-monotonic copula function describes a society in which the dependence between the dimensions involved is at its maximum level. In this society, each position quantile represents a “cast” (Decancq, 2020), and each individual who belongs a given quantile of ranking for one dimension will belong to the same quantile of ranking for the other dimensions. By the contrary, when considering the independence case, there is no statistical connection between the dimensions (that are, in our case, income, job and housing conditions, educational attainment and health status). Given the multidimensionality of the copula function, the diagonal copula section is derived taking into consideration the population outranked by a specific set of positional combinations such that: $(p_{income} = p_{job} = p_{housing} = p_{education} = p_{health})$. The higher the distance between the proportion of people falling in each specific combination, from the upper bound curve is, the less dependent the dimensions involved are. Figure 3.8 shows what Decancq (2020) defines as the downward diagonal dependence curve for each year. This curve coincides with the yearly copula diagonal sections plotted together with the yearly upper F-H bound and with the diagonal section of a

[14] Personal elaboration, numbers refer to the table shown the Appendix B, table B.2

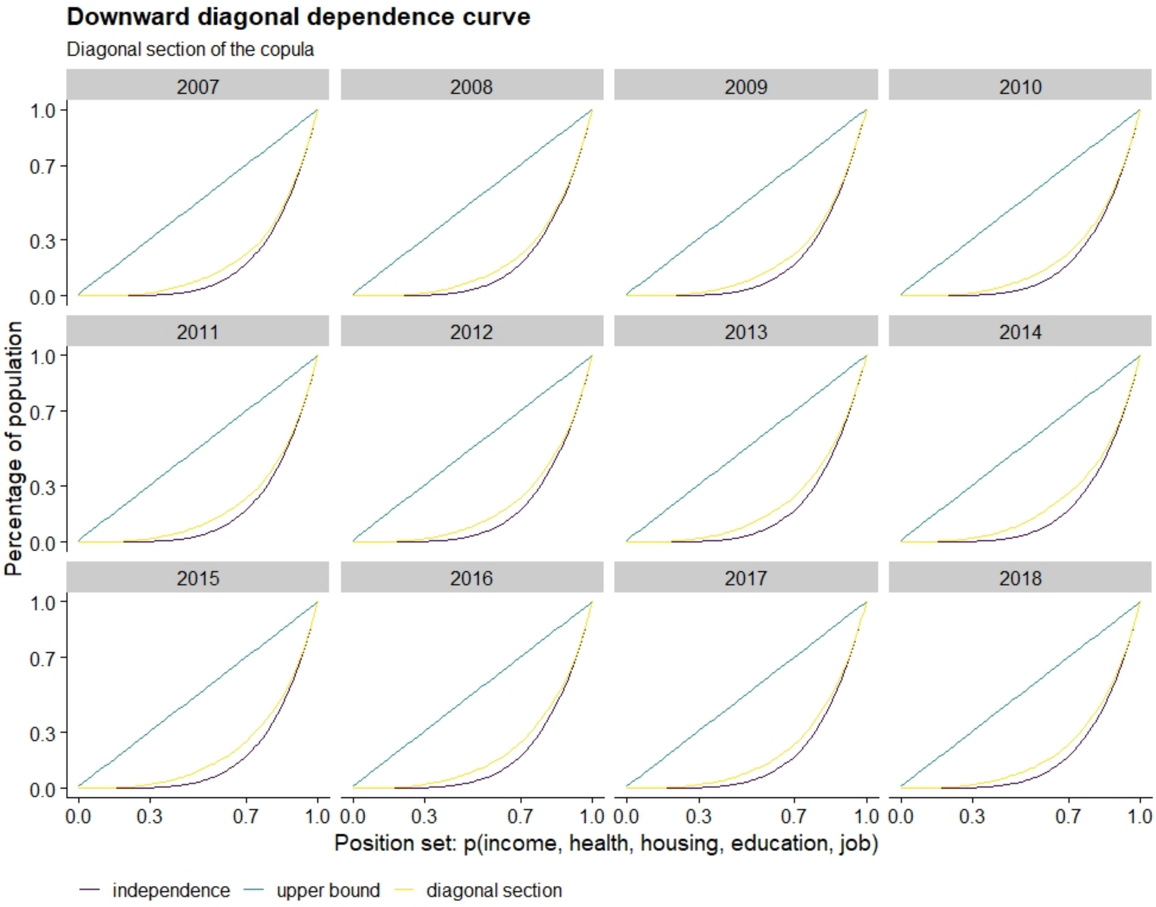


Figure 3.8: Yearly downward diagonal dependence curves

five-dimensional independence copula.[15]

Despite at a first sight the co-monotonic case (the upper bound) may look far from the empirical diagonal copula section, a more in-depth observation across the years will show some time variations. Indeed, the proximity of the diagonal sections with both the independence and the upper bound cases varies across time. In the years 2014 to 2016, it is possible to see that proximity between the diagonal sections with their upper-bound is higher. Those years coincide as well with the years of maximal distance between the diagonal copula section and the diagonal independence section. From a time-series perspective, it is not possible to say yet which year corresponds to the highest diagonal dependence level for all $p \in [0, 1]$. However, it appears to be a clear dominance of the years from 2014 to 2016 for what concerns the copula diagonal section.

[15] A zoom on the diagonal section of the 2014 copula is provided in the Appendix B with figure B.3

Partial dependence orderings

A more in-depth investigation of the inter-dimensional association is to analyse the pair-wise comparisons of each yearly diagonal copula section. With the pair-wise comparison it is possible to provide a partial ordering of dependence among the years. Table 3.6 indicates when the row-year is dominating the column-year for the downward diagonal dependence ordering (D), and the upward diagonal dependence ordering (U). The two results consist respectively of a comparison among the two diagonal sections of the yearly copulas and survivals. The presence of the zero indicates the indecisiveness of the result. The indecisive cases take place when, comparing one by one the percentiles of two different yearly copula sections, we do not find a strict dominance of one copula section against the other one. [16]

Table 3.6: Partial dominance comparisons

	2007	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**
2007	-	0	0	0	0	0	0	0	0	0	0	0
2008	0	-	0	0	0	0	0	0	0	0	0	0
2009	U	0	-	0	0	0	0	0	0	0	0	0
2010	0	0	0	-	0	0	0	0	0	0	0	0
2011	U	U	U	U	-	0	0	0	0	0	0	0
2012	0	0	0	0	0	-	0	0	0	0	0	0
2013	U	U	0	D,U	0	D,U	-	0	0	0	D	0
2014	D,U	D,U	D,U	D,U	D,U	D,U	U	-	0	0	0	0
2015	U	U	D	D,U	0	D,U	D,U	0	-	0	D	D
2016	0	0	0	U	0	U	0	0	0	-	0	0
2017	0	0	0	0	0	0	0	0	0	0	-	0
2018	0	0	0	D	0	0	0	0	0	0	0	-

This table shows that the years 2013 to 2015 are dominating the past ones in terms of dependence comparisons. At the same time, more inconclusive is the comparison with the most recent years, that only in few cases appear to be downward dominated by the years 2013 to 2015. This result provides useful insights on when the multidimensional dependence within our sample is mostly concentrated. Unsurprisingly, it emerges that the years of higher dependence coincides with the years of the increase in the sample people in cumulative deprivation.

[16] The partial dominance analysis is done by using the data on the copula and diagonal survival sections of the ten percentiles case.

The Diagonal Dependence Index

The Diagonal Dependence index as it is presented in equations 3.14, 3.15 and 3.16, is shown in figure 3.9. Its computation is provided over different combinations of dimensions. The case in which all the five dimensions are present and the cases in which one dimension at a time has been removed. This approach represents a way to assess the sensitivity of the overall diagonal dependence to the single dimension contribution.

As it is possible to see from figure 3.9, the level of dependence has raised in the last ten years. Its trend results smoother in comparison with the trend of people in cumulative deprivation, but it clearly shows a growing phenomenon. It is not easy to draw conclusions on the nature of this increase, but a natural consequence of an increased diagonal dependence is represented by an increase of the risk of falling into the poverty trap. The higher the dependence observed is, the more likely we expect to observe cumulative deprivation as a growing phenomenon.

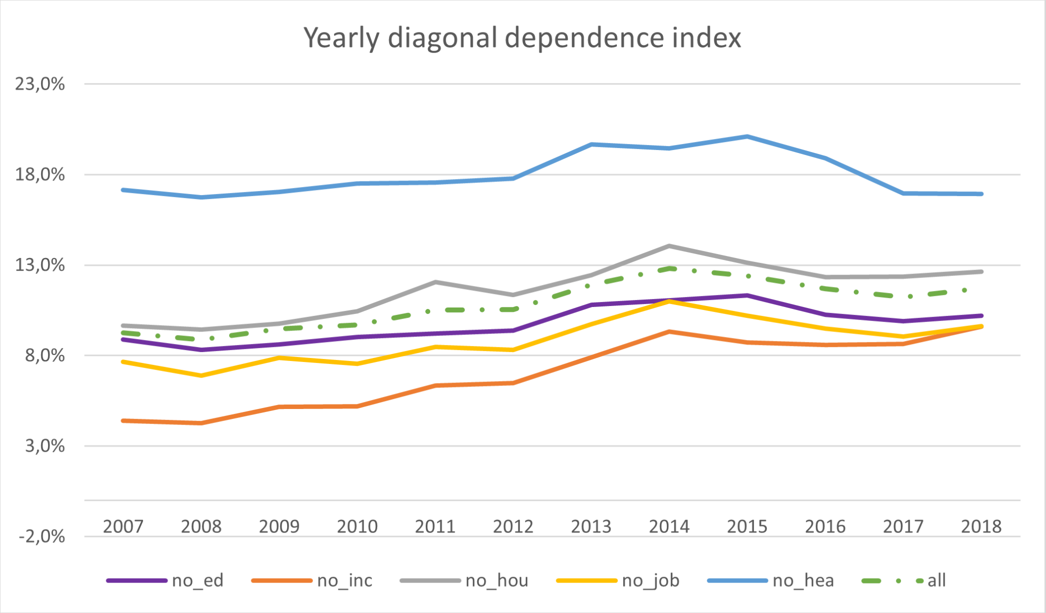


Figure 3.9: Downward diagonal dependence indices: comparing the full 5-dimensional set with 4-dimensional sets

The yearly diagonal dependence index shows a positive trend among all the presented dimensional combinations. Focusing on the 5-dimensional case, the positive trend slowed down after its peak in 2014 to remain stable at around 11%. As expected, the removal of health status causes an increase of observed dependence. Given that health, more than the other dimensions, is influenced by demographic and genetic conditions, especially when considering a vast sample age interval such the one of the workforce. The removal of the income dimension reduces the magnitude of the multidimensional dependence phenomenon especially in past years. Nevertheless, this specific dependence case rapidly catches-up other dimension combinations in more recent years. The increase of dependence among all the other socioeconomic dimensions is getting frighteningly stronger.

Table B.1 form the Appendix B contains the p-values resulting from the t-test of the mean difference applied to every couple of indices considered. Making this comparison is a way to verify that the contribution of each dimension is relevant to describing the puzzle of the multidimensional dependence among socioeconomic facets of life. As it emerges from Table B.1, the average yearly DDI in the case of all dimensions is not statistically different from the average yearly DDI in the case of not including education. The removal of this dimension does not represent a loss in terms of the final multidimensional dependence, meaning that the education dimension may not add the most substantial information to the story because it is already proxied by the other dimensions. Despite that, the other comparisons do not show the total "in-utility" of considering it as a dimension of cumulative deprivation. Thus, we proceed using the five dimensional case as the most complete case in terms of information considered and reported.

Cumulative deprivation and other poverty measures - Part 1

In order to contextualise the explanatory value of the diagonal dependence index and the cumulative deprivation condition, a comparison with other socioeconomic indicators is provided.

Figure 3.10 represents a comparison between the cumulatively deprived and people counted by the AROPE rate as being at risk of poverty and social exclusion. The AROPE is an index number indicating the presence of the following three phenomena: *i)* being at risk of poverty in terms of low income (ARP), *ii)* being severely materially deprived (SMD), *iii)* being in a low work intensity condition (LWI). The presented comparison is for 2014, being the year with the highest diagonal dependence.

Given that part of the conditions accounted in the AROPE rate are in a different way included in the cumulative deprivation counting, we observe that a considerable portion of cumulatively deprived people are as well accounted within the AROPE index. Furthermore, we can notice that approximately the 83% of cumulatively deprived people are counted in at least one of the AROPE dimensions combinations.

Of course, cumulative deprivation is a rarer phenomenon because it accounts for low educational attainment and bad health status, which are entirely neglected by the

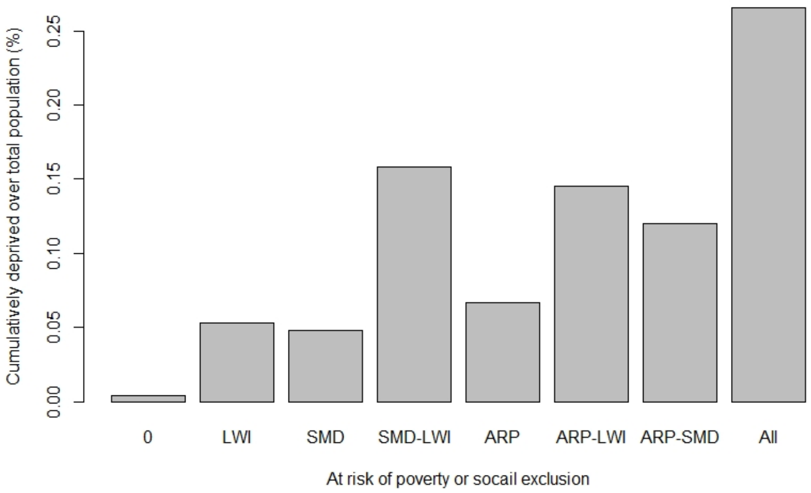


Figure 3.10: Percentage of cumulatively deprived who are counted in the AROPE index

European poverty index. In this way, the population involved in the count is sensibly lower than the total amount of people at risk of poverty and social exclusion.

Figure 3.11 provides a yearly comparison of the proportions of cumulatively deprived people with two relevant estimates of the incidence of absolute and relative poverty: the AROPE rate and the incidence of absolute poverty computed by the Italian National Statistical Institute (Istat). The two indicators represent two very different methods of explaining a social phenomenon that is currently intensely debated. The absolute poverty is expressing the exposure to poverty not only with respect to income, but also with respect to the purchasing power of income.

From figure 3.11 it emerges that, while the trends are all increasing, the magnitudes vary in a puzzling way. Despite the cumulative deprivation is conceptually closer to a relative poverty measure, it captures the interaction between a poor income and poverty in participating to society through the acquisition of the means of living. Notwithstanding, such a measure can be placed half-way between the two extreme relative and absolute poverty indicators. Further investigation of this evidence is provided in the following paragraph.

Cumulative deprivation and other poverty measures - Part 2

There are few studies dealing with comparisons between the relative and absolute poverty thresholds, among the most relevant examples is the study of Goedemé et al. (2017). Following the intention of Goedemé et al. (2017) to contextualise the poverty

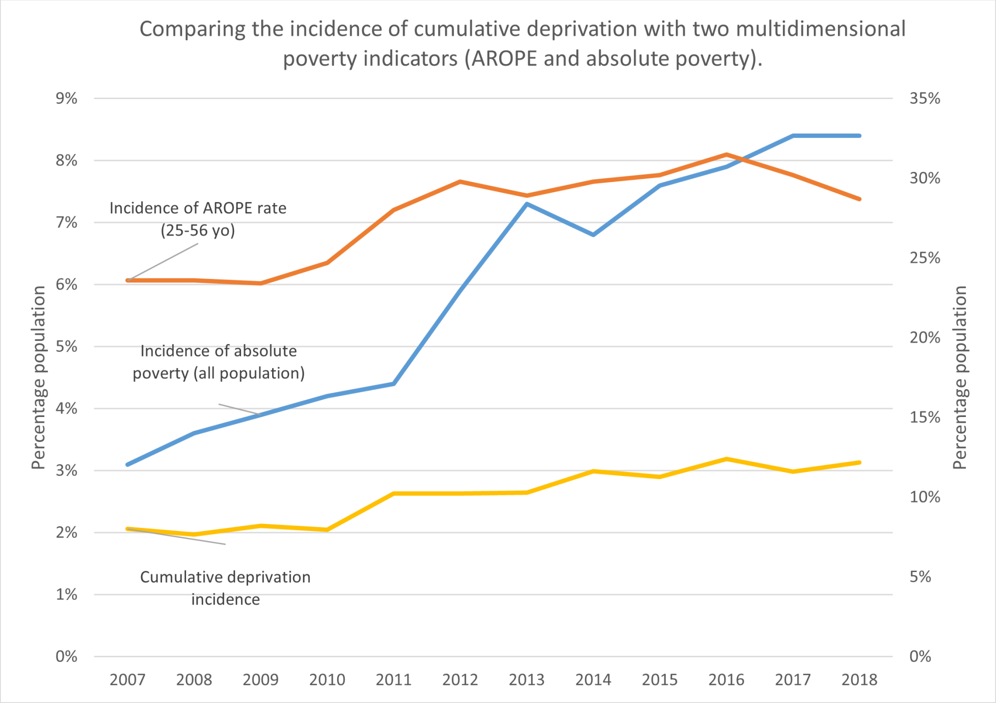


Figure 3.11: Comparison between cumulative deprivation incidence (left-axis) and other socioeconomic indicators: AROPE rate (right-axis) and Absolute poverty incidence (left-axis).

thresholds within the peculiarities and needs related with different household types and geographic areas, it is hereby presented a graphical comparison between the two mostly representative estimates of poverty income thresholds with the cumulative deprivation. In this matter the income thresholds for the relative and absolute poverty estimates are related with the maximum income observable within the cumulatively deprived sample. More specifically, the "cumulatively deprived income threshold" is identified with the maximum level of equivalent household disposable income observed within the cumulatively deprived sample, distinguishing for the different household type and geographic location (aggregating NACE-2 into three areas: North, Centre and South; and distinguishing by urbanisation level).

Given the many differences in data availability between the three different indicators considered, the yearly descriptive comparison is provided for selected observed thresholds:

- two adults and two minors (between 0 and 17 years old) for 2007 and 2011
- single person household for 2010 and 2013

The Istat estimates available to the public on absolute poverty thresholds are missing after 2013. The relative poverty threshold estimates come from the Eurostat database and are provided aggregates for the overall Italian territory but they represent the chosen household type. The following figures show the comparison for selected years between the time interval under study. As it emerges from figures 3.12 and 3.13, the maximum observed income within the cumulatively deprived group of people belonging to the two adults and two children household type is always systematically lower than the relative poverty threshold. This implies that these people are counted as well in the absolute and relative poverty indices.

The observed income levels vary across the Italian territory and most of the times they imitate the absolute poverty threshold trends, even if being constantly below it. This implies that the cumulatively deprived population belonging to this specific household type is not considered, as far as the Istat estimates show, to have enough adequate material means to live with.

A slightly different picture is provided by the single person household type, figures 3.14 and 3.15. In this matter, the cumulative deprivation income threshold may happen to be higher than the two poverty thresholds, meaning that, this type of deprivation may not be counted within the standard poverty measurement.

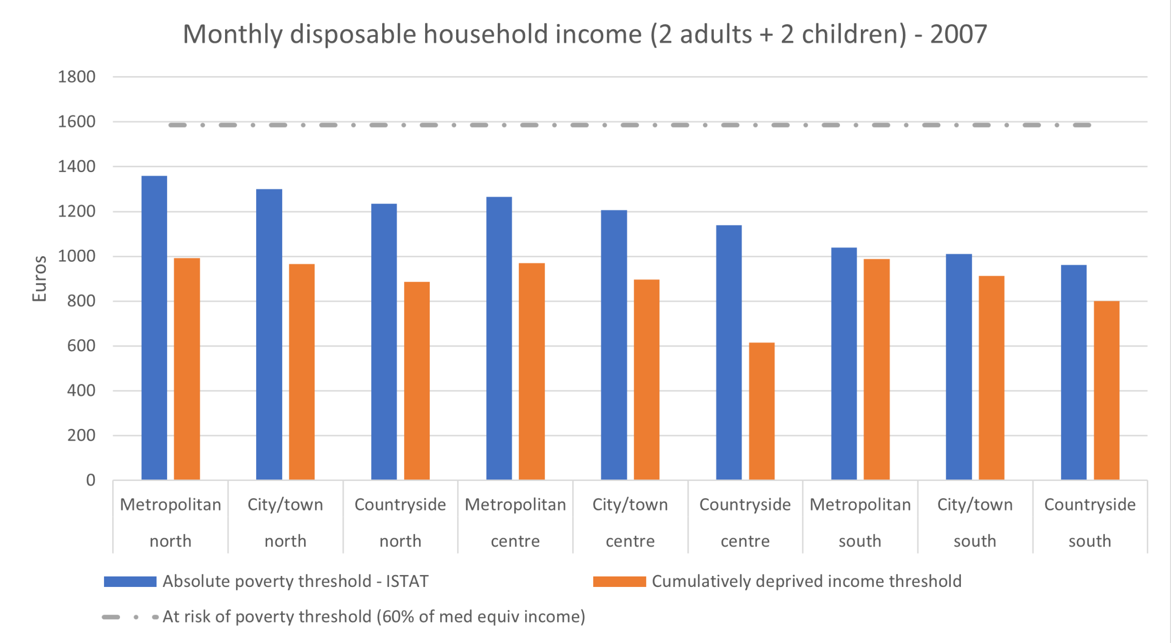


Figure 3.12: Poverty thresholds comparisons

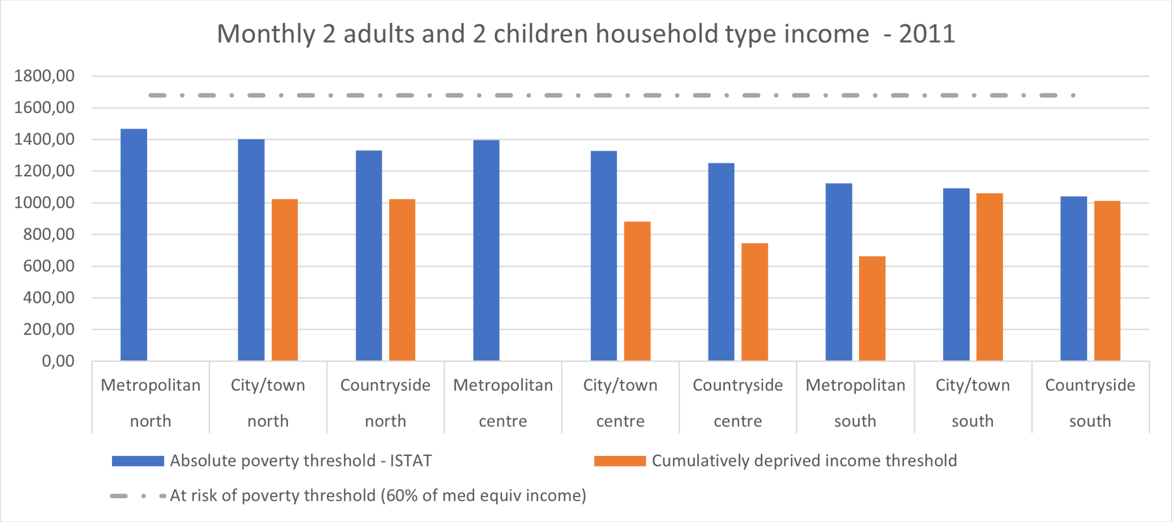


Figure 3.13: Poverty thresholds comparisons

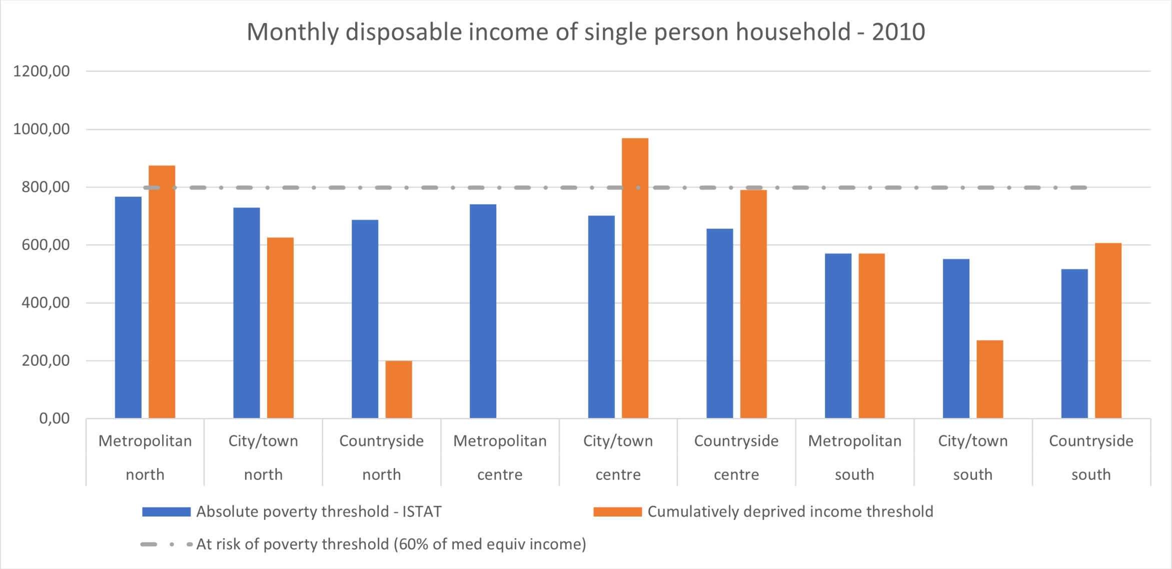


Figure 3.14: Poverty thresholds comparisons

3.5 Conclusions

This paper represents a contribution to the studies on social inequalities from a multidimensional perspective. This analysis proposes an applied measurement of the statistical dependence whose aim is to underline a group of well-being 'dimensions'.

Coming from the proposal of Decancq (2020), the tool adopted for this study is the copula function evaluated for the multidimensional set of well-being dimensions. The within-dimensional perspective has been inspected from the point of view of the population with similar outcomes in all the dimensions, to observe the phenomenon of

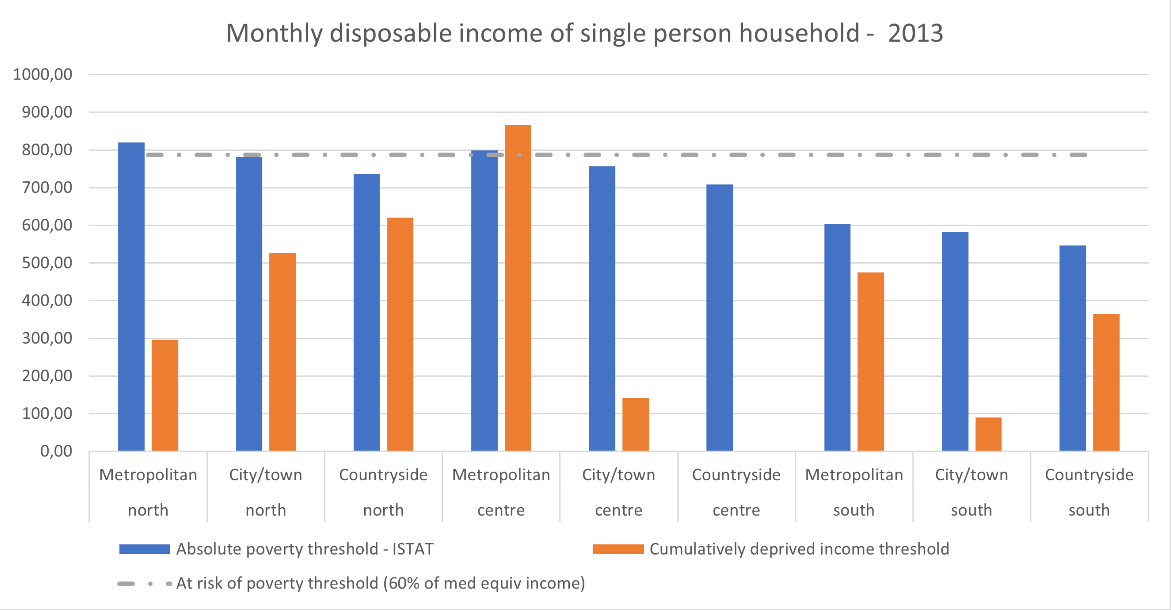


Figure 3.15: Poverty thresholds comparisons

cumulative deprivation. When applied to data on socioeconomic deprivation, this tool can provide useful insights for the integration of the standard poverty indicators. This is due to its capacity to detect the degree of statistical interrelation across multiple dimensions when observing a specific part of their distribution.

The cumulative deprivation condition is computed through the observation of the relative positions of individuals within the sample distribution. Therefore, the process to define the belonging to a certain status has a relativistic perspective. As it has been already argued within poverty studies, poverty cannot be solely measured in relative terms since its definition requires to capture what are the basic needs of people and households within different contexts of life.

Despite that, keeping the multidimensional perspective to the measurement of poverty constitutes a pillar for quantifying it in more absolute terms.

Such an approach is thus thought to be essential for investigating the effects of socioeconomic inequalities and for exploring an aspect that has been frequently neglected by the standard approaches studying inequalities: the interaction among different socioeconomic deprivations. The index of multidimensional dependence obtained within the copula-based technique, provides a statistically solid technique and a valid support to the general definition of poverty conditions.

The study has been carried out with EU-SILC data on Italy between 2007 and 2018, the dimensions selected for the experiment are: the income, the educational attainment, the labour, health and housing conditions. The research outcome shows an increasing trend of the cumulatively deprived population in Italy together with an increased diagonal dependence. This evidence is considered a symptom of a strong connection between income and other life dimensions. Unsurprisingly, the weaker agents of society (females, migrants, unemployed people), turned out to be more frequently counted among the cumulatively deprived. The presented multidimensional dependence study has been contextualised within the wider poverty indicators framework through a comparison of the income levels of cumulatively deprived people with the income thresholds of absolute and relative poverty indicators for Italy. From this analysis it emerges that the income of the cumulatively deprived households is constantly lower than the relative poverty threshold (the 60% of the average income per capita) for all household types considered, whilst it shows to be fluctuating (from above and from below) around the absolute poverty threshold. The dependence among dimensions of well-being can enable investigating the ability of welfare systems to eliminate barriers characterising the unequal distribution of access to public services. Future extensions of the study can go in several directions. For instance, it can be functional to investigate through a parametric estimation the predictive power of several individual characteristics in determining the cumulative deprivation incidence. Additionally, the determinants of cumulative deprivation could be analysed form a pseudo-panel perspective accounting for the time dimension role. Undoubtedly, an extension of the diagonal dependence index measurement to other countries can be beneficial for its interpretation. The use of both time and cross-country perspective would undoubtedly provide further understanding in order to compare different welfare systems.

3.6 Theoretical Appendix

3.6.1 The positional outcomes are the margins of the cumulative distribution function F_X

The marginal distribution function of the j^{th} dimension of F_X, is denoted as F_j and defined as:

$$F_j = \frac{\partial F_X(x_1, ..., x_d)}{\partial x_j} \qquad for\ j \in (1, ..., d) \tag{3.17}$$

In probability theory, the marginal distribution of a multidimensional *cdf* expresses the probability of X_j to have outcome of at least x_j for every possible value of all the other dimensions involved. The marginal distribution of our original multivariate *cdf* for dimension j can also be defined as $F_j(x_j) = F(\infty, ..., \infty, x_j, \infty, ..., \infty) = Pr[X_j \leq x_j]$, $x_j \in \mathbb{R}$.

The marginal distributions for each component can be easily derived using their ranking distributions. When computing the rank ordering of each of the sample realisations of the continuous random variable X_j, it is possible to observe the probability of realisation of such outcome for any given value of all the other $i \neq j$ dimensions. In other words, the rank ordering allows us to derive a marginal distribution for each dimension.

The marginal distribution for every outcome x_j is exactly expressing the proportion of population who has weakly less than x_j in dimension j. By looking at the individuals who show such an outcome, we can derive their rank position expressed as a real number scale between 0 and 1.

The position vector $p = (p_1, ..., p_d)$ is describing the positions assumed by a single person in all the considered dimensions of well-being. The re-scaling procedure on the rank values of each entry of the position vector lead to observe the transformed entries taking values in the space $[0, 1]^d$. When the position vector of individual i is equal to (0, ..., 0), it implies that the person is outranked in all the dimensions. Otherwise, in case the positional vector of i is (1, ..., 1) this implies that this person is top-ranked in all dimensions. Concluding, it is possible to say that $P_j = F_j(X_j)$, so the position of the individual i in dimension j with respect to the others, can represent the marginal distribution function of the random variable X_j.

3.6.2 The probability Integral Transform Theorem

The *probability integral transform* theorem is useful for our case because it allows to say that all the marginal distributions that can be derived from the positional random vector follow a standard uniform distribution. In more technical terms, if X is a continuous random variable with cumulative distribution function $F_X(x)$ and if $Y = F_X(X)$, then Y is a standard uniform random variable.

Demostration: Let X be a continuous random variable with *cdf* $F_X = Prob(X \leq x)$; let Y being another continuous random variable defined as $Y = g(X)$. Let g be strictly increasing and differentiable, thus g^{-1} uniquely exists; let $g = F_X$. The distribution of Y is obtainable as follows.

$$\begin{aligned} F_Y(y) &= Prob(Y \leq y) \\ &= Prob(F_X(X) \leq y) \\ &= Prob(X \leq F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y \end{aligned} \tag{3.18}$$

According to the properties of the uniform distribution with margins $a = 0$ and $b = 1$, $F_Y = y$ means that $F_Y \sim U(0, 1)$.
In our specific case, Y equals the position random variable defined before as $P_j = F_j(x_j)$. Following the probability integral transform theorem, P_j will be uniformly distributed on the interval $[0, 1]$.

Bibliography

R. Aaberge, A. B. Atkinson, and S. Königs. From classes to copulas: Wages, capital, and top incomes. *The Journal of Economic Inequality*, 16(2):295–320, June 2018. ISSN 1569-1721, 1573-8701. doi: 10.1007/s10888-018-9386-x.

S. Alkire. *Multidimensional Poverty Measures as Relevant Policy Tools*. Number 118. OPHI, 2018.

S. Alkire and J. Foster. Understandings and misunderstandings of multidimensional poverty measurement. *The Journal of Economic Inequality*, 9(2):289–314, 2011.

A. B. Atkinson and F. Bourguignon. The comparison of multi-dimensioned distributions of economic status. *The Review of Economic Studies*, 49(2):183–201, 1982.

F. Bourguignon and S. R. Chakravarty. The measurement of multidimensional poverty. In *Poverty, Social Exclusion and Stochastic Dominance*, pages 83–107. Springer, 2019.

A. Charpentier, J.-D. Fermanian, and O. Scaillet. The estimation of copulas: Theory and practice. *Copulas: From Theory to Application in Finance*, pages 35–64, 2007.

K. Decancq. Copula-based measurement of dependence between dimensions of well-being. *Oxford Economic Papers*, 66(3):681–701, 2014.

K. Decancq. Measuring cumulative deprivation and affluence based on the diagonal dependence diagram. *METRON International Journal of Statistics*, June 2020.

K. Decancq and E. Schokkaert. Beyond gdp: Using equivalent incomes to measure well-being in europe. *Social Indicators Research*, 126(1):21–55, 2016.

J. Deutsch and J. Silber. Measuring multidimensional poverty: An empirical comparison of various approaches. *Review of Income and Wealth*, 51(1):145–174, 2005.

C. García-Gómez, A. Pérez, and M. Prieto-Alaiz. Copula-based analysis of multivariate dependence patterns between dimensions of poverty in europe. *Review of Income and Wealth*, 2020.

T. Goedemé, T. Penne, T. Hufkens, A. Karakitsios, A. Bern, B. Simonovits, E. C. Alvarez, E. Kanavitsa, I. C. Parcerisas, J. R. RomanÃ, et al. What does it mean to live on the poverty threshold? lessons from reference budgets. Technical report, 2017.

M. Hofert, I. Kojadinovic, M. Mächler, and J. Yan. *Elements of Copula Modeling with R*. Springer, Jan. 2019. ISBN 978-3-319-89635-9.

E. L. Idler and Y. Benyamini. Self-rated health and mortality: a review of twenty-seven community studies. *Journal of Health and Social Behavior*, pages 21–37, 1997.

R. B. Nelsen. *An Introduction to Copulas*. Springer Science & Business Media, June 2007. ISBN 978-0-387-28678-5.

A. Pérez. Measuring The Dependence Between Dimensions of Welfare: a Study Based on Spearman's Footrule and Gini's Gamma. page 20, 2015.

A. Pérez and M. Prieto-Alaiz. Measuring the dependence among dimensions of welfare: A study based on spearman's footrule and gini's gamma. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 24(Suppl. 1):87–105, 2016.

C. Quinn. Using copulas to measure association between ordinal measures of health and income. Technical report, HEDG, c/o Department of Economics, University of York, 2007.

A. Sklar. Random variables, distribution functions, and copulas: a personal look backward and forward. *Lecture Notes-Monograph Series*, pages 1–14, 1996.

M. Sklar. Fonctions de repartition an dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Univ. de Paris*, 8:229–231, 1959.

J. E. Stiglitz, A. Sen, J.-P. Fitoussi, et al. Report by the commission on the measurement of economic performance and social progress, 2009.

K. Tkach and C. Gigliarano. Multidimensional poverty index with dependence-based weights. *Social Indicators Research*, pages 1–30, 2020.

E. Van Doorslaer and A. M. Jones. Inequalities in self-reported health: Validation of a new approach to measurement. *Journal of Health Economics*, 22(1):61–87, 2003.

Chapter 4

Model-based Recursive Partitioning to estimate Unfair Health Inequalities

4.1 Introduction

> *"All animals are equal, but some animals are more equal than others."*
> Animal Farm, G. Orwell (1989)

It is a fact that the health conditions vary greatly across people both within and between countries (WHO, 2013)[1]. The variability of health conditions across the population is not only and exclusively due to biomedical or genetic factors, but also to socioeconomic factors. Health inequalities, as income inequalities, strongly depend on the context in which people are born, live, work and age. The debate on health inequalities owes to the main contributions, among all, of Marmot (2005); Sen et al. (2004), the Commission on Social Determinants of Health and the World Health Organization (2008). The COVID-19 pandemic had a violent impact on the world-wide economies, bringing the discussion of health on the top of the governments' agenda. Unsurprisingly, the distribution of health conditions and health-care access across the population has gained an increasing attention among researches and policy makers.

There can be a different political and philosophical point of view regarding the social judgements towards the determinants of socioeconomic inequalities. Despite that, the general idea about them in modern economies is that the market is not able to lead to an equal distribution of the outcomes in society, and there exists a set of characteristics beyond the individual's control that impact on the unequal societal outcomes.

[1] The World Health Organization (2013) defined health inequalities as: *"avoidable inequalities in health between groups of people within countries and between countries"*.

For what concerns health itself, there are structural difficulties to empirically determine how health inequalities are affected by socioeconomic conditions. Indeed, a person's socioeconomic background can have a direct influence on the state of their health and indirectly affect it through their actions and lifestyle habits. In the empirical studies on health inequalities, the capability to disentangle the exact contribution of each source of inequality depends on the researcher's normative assumptions on the individual responsibilities.

The Equality of Opportunity theory shifts the attention of the traditional welfare theory from studying only the outcomes, to studying the inputs and provides a valid theoretical background to model the unequal distribution of health outcomes.

In societies increasingly sensitive to issues of equal distribution and inclusive economic policies, the theory of equality of opportunity has gained its place in the debate. Equality of Opportunity theory is based on the fact that individual advantages observable as income, education or health are determined by attributes for which it is morally correct to hold individuals accountable (fair inequalities), and by those circumstances that are beyond individual's control, and for which the individuals should not be held accountable (unfair inequalities). These circumstances are generally represented by demographic factors, such as gender and age, and parental socioeconomic background. The models of Inequality of Opportunity (IOP) address the analysis of the contribution of each factor to the outcome (advantage or disadvantage) formation.

The main precursors of the Equality of Opportunity theory in political philosophy are Rawls (1971), Dworkin (1981) and Cohen (1989). Equality and freedom are two values emphasised in this theory. Ideally, equality of opportunity is achieved when the life lottery effect on life plan choices is abolished and all individuals are free to choose from the same set of opportunities. However, these two principles are not always compatible. In the empirical studies, any definition of Equality of Opportunity requires a balancing process between equality and freedom.

The principles of equality and freedom have been translated into the Inequality of Opportunity models as two types of policy actions against unequal distribution: the compensation principle and the reward principle. As they are defined in this theoretical framework, the two principles may not be always compatible with each other. Meaning that the prevalence of one does not guarantee the realisation of the other. The first one states that any inequality due to circumstances beyond the individual's control is unfair and should be eliminated. The second specifies how well-being should relate to responsibility characteristics.

The recognition of the existence of "unfair" sources of inequality would thus raise the question: *to what extent hold people responsible for their preferences and choices?*

For answering this question, Fleurbaey and Schokkaert (2009) refer to the debate on the equality of *use* and equality of *access*. On one side, the equality of access can be intended as a non-discriminatory principle, requiring that all individuals have the same *means* while they can make different choices regarding their outcome generating process. On the other hand, the proponents of equality of use require that all individuals have the same chances to obtain valuable outcomes no matter what their means and choices are. We can identify, in the literature on Inequality of Opportunity (IOP), two main schools of thought regarding the role of the policy makers in correcting these distortions. On one side, there is the model defined by Roemer (1998) which conceives inequalities in society as a materialistic phenomenon, where the societal distribution of outcome is a result of the distribution of the circumstances. On the other side, Fleurbaey and Maniquet (2012) state that there are some individual choices and preferences which can be observed and should be respected comparing the outcomes. While the former conceives a policy intervention to *compensate* for the inequalities originated in the society by sources beyond individual control, the latter includes also the objective of ensuring that arbitrary causes of inequality are *rewarded*.

The empirical non-parametric studies measuring IOP basically refer to the Roemerian model framework which requires to divide the initial population in groups defined by socioeconomic characteristics, also named population *types*. The effort is assumed not to be observable in absolute terms. Most of the Roemerian IOP empirical applications estimate a counterfactual distribution of the group-specific outcome variable - a modified distribution of the outcome variable which accounts only for the unfair sources of inequalities -, in order to isolate the "socially-defined" unfair components from the original outcome distribution. Up to the IOP theory, if the types precisely describe the full contribution of non-arbitrary circumstances, the between-types inequality coincide with the unfair inequality. Alternatively, Fleurbaey and Schokkaert (2009) propose to take into account for the reward principle a sort of "responsibility-sensitive" component in the analysis of the health distribution. Departing from the traditional partition á lá Roemer (1998), F&S propose a model to isolate the exact (direct and indirect) contribution of the non-arbitrary factors on the health inequalities and evaluate the direct unfairness and the fairness reward with respect to the efforts distribution and the population types in the society.

From the vast literature on IOP, we can distinguish the various empirical applications on the base of technical and theoretical characteristics, e.g. *i)* the way to group people in socioeconomic groups which define the non-arbitrary sources of inequality, *ii)* the transformation of the dependent-variable distribution to derive the modified distribution, *iii)* the assumptions concerning the observation of individual responsibilities.

While all the three decisions have been traditionally addressed through a normative choice of the researchers, the recent outbreak of data-driven techniques in estimating economic models lead to a change of direction. More specifically, in IOP models, the recourse to data-driven techniques has characterised the process of population types partition. Generally, the recourse to data-driven techniques has been justified by the possibility to optimise the estimation processes. Yet it has been highly debated across a number of scholars afraid that data-driven methods risk to replace the researchers' choices in the definition of what ought to be - which is normatively stated - with what is - which is empirically observed -.[2] By contrary, the choice of using data-driven techniques mainly wishes to optimise estimation procedures, relaxing some assumptions which are mostly related with the empirical technique rather than with a normative position, but also want to *test* some model assumptions' limitations by letting the data provide information on underlying phenomena. Therefore, it is believed that the adoption of a data-driven method to the specific task of performing the population type-partition could help in finding the *salient types* among all the possible existing population types. Whilst the traditional type partitioning computes inequality in the outcomes across some population types which may turn out not to be really socially different, and, therefore, lead to statistically biased Inequality of Opportunity estimation, the data-driven technique helps in maximising the heterogeneity observable across population groups which are still depending on the normatively defined socioeconomic characteristics. For this reason the adoption of data-driven type-partitioning of the population is notably bringing improvements to the study of IOP (Brunori et al., 2018).

Studies focused on income inequality of opportunity, have mostly used a measurement approach which is agnostic with respect to the effort characterisation. Notwithstanding, the main concern of the researchers of IOP in income have been the measurement of the relative portion of inequality originated from the family background, as a ratio of total inequalities. On the side of health unfair inequalities, many empirical contributions had the ambition to identify the indirect effect of socioeconomic background on health outcomes, bringing the lifestyle into the picture as a proxy of the individual effort. Jusot et al. (2013) worked on disentangling the direct and indirect contribution of the circumstances on the health status by estimating parametrically the effort response to circumstances and using this prediction to estimate the health response to effort. One pitfall of performing a single regression for the whole population is that it is

[2] The "Is-Ought" problem raised by Hume with the *Hume's guillotine* implies that a reasoner cannot logically decide whether a moral statement is true if he has only access to descriptive and non-moral statement.

imposed a constant response of effort to all socioeconomic categories. Not considering the heterogeneous effect of lifestyle behaviours on outcome may represent a weakness for this type of analysis. Hence, Carrieri and Jones (2018) presented the estimation of the direct effect of effort, allowing it to vary in its parameters across types. Although their attempt to describe the interaction of circumstances and effort through heterogeneous slopes is getting closer to the theoretical intention of Fleurbaey and Schokkaert (2009), the authors provide a self-defined population type-partition. As a consequence, they encountered the traditional problem affecting all the standard non-parametric applications of the Roemerian partitioning: the course of dimensionality. Addressing the limitations affecting this literature on the definition of types, Carrieri et al. (2020) are presenting a data-driven partitioning approach to derive the population types and to estimate the type-specific lifestyle-to-effort relations. More precisely, they are using Finite Mixture Models (FMM) to find the latent population subgroups as a class mixing the given circumstances and to model the dependent-to-effort relation.

One of the major problems emerging from the health IOP studies is the recourse to self-reported health status to proxy individual health. An example of a more objective health definition is the allostatic load measure, which has been recently constructed and used in the IOP literature Carrieri et al. (2020); Davillas and Jones (2020). Allostatic load is a cardinal biological measure aggregating nurse-recorded health information dimensions, its observation is available in two waves of the UK Household Longitudinal Survey (UKHLS).

With this paper, the aim is to extend the literature on IOP in health, by bringing into the picture the interaction between the different socioeconomic contexts and lifestyle behaviours. This issue is addressed using the Model-Based recursive Partitioning (MOB). MOB is a tree-based supervised learning algorithm developed by Zeileis et al. (2010). The algorithm fits a tree based regression on predetermined partitioning variables and estimates a statistical model in each terminal node. Intuitively, the algorithm initially fits the full-sample model and recursively searches for a partition in two sub-samples that would allow the model to better fit the data. MOB stops splitting the sample when no further split would result statistically significant.

The innovative contribution of this approach states in the fact that the sole relation between health and effort proxied by the lifestyle behaviour is estimated, allowing the effort to vary in its intercept and slope according to different relevant circumstantial realisations. Thus it is possible to identify the indirect effect of circumstances through the different parameters relating the health and lifestyle. With this paper, estimates of the Direct Unfariness (DU) and Fairness Gap (FG), following the methodology of Fleurbaey and Schokkaert (2009) are presented and discussed.

Such an application, as far as it is concerned, is new to the literature of IOP. The empirical application is done using the same data source of Carrieri et al. (2020). This exercise focuses on both providing insights on the IOP in health and in assessing the validity of new possible data-driven techniques to be adopted for IOP empirical studies.

In Section 4.2, an extensive illustration of the IOP theoretical framework is presented; discussing as well the main partitioning and estimation techniques adopted in the literature. In Section 4.3, the empirical characterisation the approach is displaced. The data are illustrated in Section 4.4. Section 4.5 contains the results and a related discussion. Section 4.6 conclusion remarks are provided.

4.2 Inequality of Opportunity: from theory to practice

The Equality of Opportunity theory provides a framework to practically measure IOP based on different definitions of it. All the definitions provided by the models illustrated in the following section share an important consideration: there is a portion of inequality in society that can be defined unproblematic, being due to responsibility matters; and an unfair portion of inequality which is, instead, determined by the socioeconomic circumstances. The models propose different ways to disentangle the contributions of fair and unfair sources but they allow some normative space for attributing responsibilities to individuals. Therefore, the decision regarding which factors are to be classified as circumstances beyond individual control and which are to be considered as personal choices, is left in the hands of the society (and the policy maker).

The literature on Inequality of Opportunity has its roots in economic theory with the seminal contributions by Roemer (1998) and Fleurbaey and Schokkaert (2009). The different definitions of inequality of opportunity provided by these models led to different measures of IOP. What follows in the next section is an illustration of both models from a theoretical and methodological point of view.

4.2.1 Model frameworks

Roemerian model

Let the population be finite and indexed as $i \in \{1, ..., N\}$, where N is large. Each individual i has three attributes $\{y_i, C_i, e_i\}$, respectively, the individual advantage (e.g. income or health status), the circumstances (e.g. demographics, family background), and the effort (e.g. hours worked, lifestyle).

It is possible to partition the population into K types according to the set of circumstances. The types are, thus, based on personal non-arbitrary characteristics. Given that the elements included in the circumstances are finite and each one has a discrete domain, the partition of the whole population onto finite groups is given by $K : \Pi = \{T_1, ..., T_k, ..., T_K\}$. This partition is homogeneous, hence, each group is non overlapping $T_l \cap T_k = \oslash$, $\forall l \neq k$.

With the type partitioning we can observe population grouped in different opportunity sets. Within each group, people are facing the same non-arbitrary circumstances of life. While the circumstances are assumed to be known by the policy maker, the effort exerted by the individual is not necessarily observable. This characteristic is not a limitation for its identification because the outcome is assumed to be a function of the effort, $y_i = f(e_i)$. Furthermore, the outcome is assumed to be monotonically increasing in the effort, so the following statement is always true, for every $k \in K$ groups:

$$y_i^k(e_i) \geq y_j^k(e_j) \iff e_i^k \geq e_j^k,\ T_k \in \Pi,\ i \neq j,\ \forall e_i, e_j \in \Re^+ \tag{4.1}$$

Furthermore, the effort is not independent from the circumstances and, as a consequence, its type-specific distribution should be accounted as a characteristic of the type. Accordingly, the absolute value of effort, when observable, is not an accountable information due to its type-specificity.

In the case in which the effort is not observable, the Roemer Identification Axiom identifies the *rank* position of the person in advantage distribution within a type-group with the relative effort exerted $e^k(\pi)$. In a world in which opportunities are equally distributed, the income is uniformly distributed across all types for a given amount of effort exerted.

$$y_i^k(\pi) = y_j^l(\pi),\ \ \forall \pi \in [0, 1]\,;\ i \neq j,\ \forall T_k, T_l \in \Pi. \tag{4.2}$$

The effort rank and the advantage rank coincide by assumption of the model, $e_i^t(\pi) = y_i^t(\pi)$. It implies that a society achieves equal opportunity when the different type-specific advantage distributions are the same:

$$F^k(y) = F^l(y),\ \ \forall l, k | T_k\ \in \Pi,\ T_l \in \Pi \tag{4.3}$$

Summarised in equations 4.2-4.3 are the strong equality of opportunity assumptions of Roemer. A weaker IOP assumption is the equalisation of the type-specific income across different types, μ^k.

$$\mu^k = \mu^l, \ \ \forall l, k | T_k \ \in \Pi, \ T_l \in \Pi \tag{4.4}$$

Considering the effort degrees as the quantiles of the effort distribution, a further population partition can be realised. This partition is called *tranches*, and allows to compare the effort degree across the different types. The partition of the population into tranches is defined as $Q : \Theta = \{S_1, ..., S_Q\}$.

In the literature of IOP, the within-types (between-tranches) inequality is acceptable, while the between-types (within-tranches) inequality is the target of the compensatory policies. The policy that equalises the opportunity tries, can be of different strength and complexity with respect to the type of redistributive transfers.

The *ex-ante* compensation states that "people should face the same opportunity set, independently from the effort they exert". The policy would imply to transfer and redistribute across people belonging to different types, in order to equalise their average circumstance-related income.

The *ex-post* compensation proposes that "people exerting the same amount of effort should have the same income, regardless of their circumstances". The ex-post compensatory policy requires a more complex organisation of transfers among the people in order to completely equate the type-specific income distributions conditional to effort.

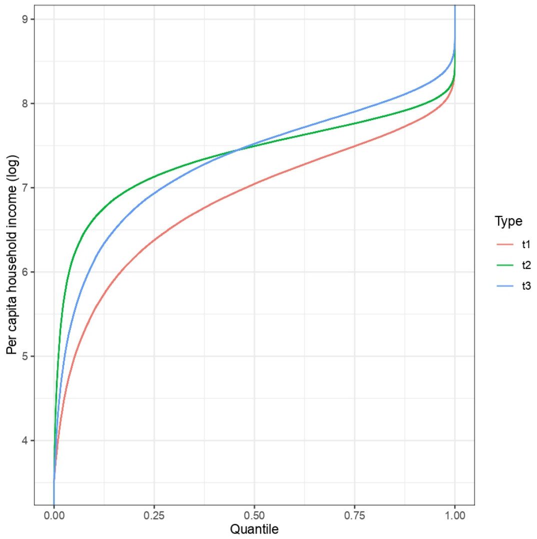


Figure 4.1: Type-specific income distributions

If we look at the type-specific income distributions at figure 4.1, we can see that, following the ex-ante criterion, no IOP emerges from the comparison of the mean income of type 2 and type 3, while type 1 will represent the policy recipient. The ex-ante IOP will be measured looking at the differences among the mean income of people belonging to type 1 with respect to the mean income of people in type 2 and 3. Conversely, if we measure IOP with an ex-post criterion, a re-distributive transfer policy would be necessary to equate all the three distributions.

Fleurbaey-Schokkaert model

This model is rooted from the "responsibility-sensitive egalitarianism" proposed by Fleurbaey (1995), Fleurbaey (2008), Fleurbaey and Schokkaert (2009) and Fleurbaey and Maniquet (2012). In this framework, the outcomes generated in the society are determined by factors beyond the individual's control. However, people are held responsible, to some extent, for their achievements.

Recalling the duality between the principle of reward and compensation, the model that is hereby illustrated is proposed by Fleurbaey and Schokkaert (2009) and presents two measures of inequality of opportunity which respectively refers to the two principles. While Roemer mainly associates its models to the IOP in income, Fleurbaey and Schokkaert, (2009) (F&S) focus their attention in the inequalities of opportunity in health consumption and outcomes.

As in the Roemerian model, the approach of F&S requires the original distribution of health outcomes to be replaced by a counterfactual distribution reflecting all and only unfair health inequalities $\tilde{H}$.

In order to be fully consistent with both the reward and compensation principle, $\tilde{H}$ would require to:

- preserve the outcome inequality between individuals with the same effort degree (within-tranche inequality)
- do not preserve any outcome inequality between individuals characterised by the same circumstances (within-type inequality)

Fleurbaey (2008) have discussed the impossibility of computing a counterfactual distribution which is consistent with both conditions.[3] What follows this consideration is the formulation of two distinct measures of inequality of opportunity which are respectively consistent with one principle, while preserving consistency for the other principle at least for a reference type or effort degree.

Let us assume that the policy maker can only observe two factors influencing individual health status, income and lifestyle. The latter represents, to some extent, a source of inequalities for which the people can be hold responsible for, namely the

[3] See also Brunori (2016) for further discussion.

effort. Furthermore, let us assume that the effects of the income and lifestyle factors (l) on health are additive and separable on the health status. The health status of i can be represented as a function

$$h_i = \boldsymbol{h}(y_i, l_i);\ \forall\, i \in (1, ..., n). \tag{4.5}$$

A different health status between two individuals $h_i(y_i, l_i) \neq h_j(y_j, l_j)$ can be acceptable, if, and only if, such inequality does not reflect a illegitimate variation in the incomes, i.e. if $y_i = y_j$.
It is possible to individuate the illegitimate sources of inequality in the health status by knowing the exact relation each individual shows between her health, income and lifestyle. If we observe people's health at a given reference lifestyle $\tilde{l}$, we can evaluate the *Direct Unfariness* between two individuals by measuring the difference:

$$\tilde{h}_i(y_i, \tilde{l}) - \tilde{h}_j(y_j, \tilde{l}) \tag{4.6}$$

The inequality within types is preserved while the between types heterogeneity is neglected in the measure of direct unfairness. Formally, it respects the principle of compensation, and it preserves consistency with the reward principle at least for each reference effort degree addressed.

When $h_i(y_i, l_i) = h_j(y_j, l_j)$, such equality is considered being fair if and only if $l_i = l_j$. If we observe people's health at a given reference income y^*, we can measure the *Fairness Gap* by measuring the difference:

$$h_i^*(y^*, l_i) - h_j^*(y^*, l_j) \tag{4.7}$$

In this case, the inequality within tranches is preserved by the counterfactual distribution, and the inequality between types is observed with respect to a reference type. The fairness gap is consistent with the principle of reward for the reference circumstance, it being insensitive to changes in inequality within the reference circumstance. Moreover, it is fully consistent with the principle of compensation, and because of being obtained with respect to a reference type, it preserves the differences across types.

As already underlined, the model itself is based on the assumption that the health status can be described by an additive and separable function of income and lifestyle. Thus, applying the appropriate identification of the structural model, it is possible to decompose each effect as a function of two distinct phenomena. An empirical application would translate this assumption in the researcher's normative belief with regard to the individual health condition formation and the identification of the different "legitimate" and "illegitimate" actions in determining individual health.

Wrapping up the presented models, the outlined theory of IOP admits that society may have distorted rewards due to some context factors that should be considered morally not acceptable. Given the strong ideological implications in assigning responsibilities to the people, the IOP models considered the effort as being a result of both arbitrary and non-arbitrary factors. More specifically, the IOP models provide a framework to analyse the mutual influence of circumstances on the outcomes through the efforts. Such modelling exercise aims to guarantee a meritocratic reward within societies, providing an equalisation policy on the non-arbitrary sources of inequality. Although meritocracy can be ideologically questionable in the real societies, we believe that it is worth investigating the deep relations between the person's conditions and the socioeconomic context.

4.2.2 Estimation methods and IOP measurement

The models of Inequality of Opportunity are basically looking at outcome inequalities through a multidimensional perspective which is given by the observation of different opportunity sets. Different opportunities imply different problems and needs and, as a consequence, heterogeneous interactions between effort, choices and final outcome. For this reason, Roemer and Trannoy (2015) are defining the equality of opportunity as a process which varies in time and across societies. The empirical approaches to measure the Inequality of Opportunity Index have partly addressed the issue of disentangling the direct and indirect effect of circumstances. Most of these examples instead focus on measuring the effect of the observable circumstances, without really considering their indirect effect through the effort. In practice, the inequality of opportunity index is computed on a transformed distribution of the outcome which captures the sole influence of the opportunity set on the outcome (also named *counterfactual* distribution). The IOP index generally measures a reduced portion of the whole inequality, capturing only the between-types component. The measurement approach varies in the literature, respectively for measuring the Roemerian *ex-ante* or *ex-post* inequality.

The empirical contributions providing the measurement of the ex-ante IOP index is using the within-type outcome means as the counterfactual distribution coordinate.

The non-parametric estimation approaches are initially operating a population partitioning into types.[4] Van de Gaer (1993) proposed to compute the smoothed income distribution $\tilde{Y}^{EP}$ by replacing the individual income of each component of the group-

[4] This operation shows several potential problems which will be discussed in Section 4.2.3.

type with the type-specific mean income:

$$\mu_i^k = \frac{1}{N_k} \sum_{i \in T_k} y_i^k, \tag{4.8}$$

where T_k is the k^{th} type and N_k is its size. μ_i^k represents a benchmark income associated with the type, it represents the opportunity set accessible to people. The absolute level of IOP is the measure of inequality between all the mean incomes for each type: $\tilde{Y}^{EA} = I(\mu_i^k)$, $\forall i = (1, ..., N)$ individuals and $\forall k = (1, ..., K)$ types. The relative amount of unfair inequality over the total inequality has been generally presented as the ratio of the absolute ex-ante IOP index over the total inequality index.[5]

Non-parametric or semi-parametric approaches require to define the population partitioning into the Roemerian types. At the cost of some discretionality in the functional form structure, with the parametric approach the counterfactual distribution is estimated using the regression coefficients derived form the outcome generating function. The parametric approach does not require the realisation of a population partitioning á lá Roemer. Bourguignon et al. (2007) and Ferreira and Gignoux (2011) propose to estimate the outcome counterfactual distribution by training a reduced form of the outcome generating function. The ex-ante parametric approach defines the outcome generating function as follows:

$$y = f(C, E(C, \nu), u) \tag{4.9}$$

Assuming linearity and additive separability of the functions f and E, it is possible to estimate the reduced form of the outcome y_i of individual i in terms of her characteristics C_i with a simple log-linear regression: ln $y_i = C_i\boldsymbol{\beta} + \epsilon_i$.

The coefficients $\boldsymbol{\beta}$ are representing the overall effect of the circumstantial characteristics of i to the outcome, without distinguishing their direct and their indirect effects as function of efforts $E(C, \nu)$. According to Roemer and Trannoy (2015) this approach can be associated with the estimation of a reduced form of direct unfairness as defined by Fleurbaey and Schokkaert (2009). The counterfactual distribution in this case coincides with the predicted income obtained with the estimate $\hat{\boldsymbol{\beta}}$ (Ferreira and Gignoux, 2011), p.634.

$$\hat{\boldsymbol{\mu}}_{Y_{EA}} = \exp[C_i\hat{\boldsymbol{\beta}}] \tag{4.10}$$

[5] Alternatively, another transformation adopted is the standardised advantage distribution: $\nu_i^k = y_i^k \frac{\mu}{\mu^k}$. In this case, the between-groups inequality is eliminated through a re-scaling operation of the subgroup-means used for the index calculation.

The *smoothed* income of individual i is obtained by ignoring the residuals and counting only the role of the observed circumstances in determining the income. The vector $\hat{\boldsymbol{\mu}}_{Y_{EA}}$ is a parametric analogue of the smoothed distribution shown in the equation (4.8).[6] The ex-ante approach has been very popular in the IOP measurement due to its very simple computation. By contrary, this approach has been criticised because it literally ignores the interplay between the circumstances and the effort providing an incomplete measure of unfair inequalities (Fleurbaey et al., 2017). We can observe two different ways to deal with the analysis of the effect of circumstances on the outcome through the effort. On one side, non-parametric approaches for ex-post IOP assume that the effort distribution is a characteristic of the type itself, while its relative distribution is informative about the indirect effect of circumstances. On the other side, there are parametric approaches which propose a modelling framework to reduce the variance explained by the error term in equation 4.10 through the inclusion of the effort.

The ex-post non-parametric IOP measures generally refer to the Roemer Identification Axiom for identifying the unobserved relative effort and deriving the counterfactual distribution to which the inequality measure is applied. Checchi and Peragine (2010) propose to weight the income of individual i with the average income divided by the average income over the same tranche of i.

$$\tilde{y}_i = y_i^k(\pi)\frac{\mu}{\mu^\pi}, \tag{4.11}$$

where $\mu^\pi = \frac{1}{N_q}\sum_{i\in\Pi_q} y_i$, Π_q is the q^{th} tranche, or quantile of effort, and N_q is the size of tranche-group q.

With the ex-post approach it is necessary to estimate the whole distribution of the advantage within each type, not only its first moment as in the case of ex-ante approach. The ex-post parametric IOP measurement, instead, brought in a two-stage estimation which accounts to the indirect effect of the socioeconomic background on the outcome through the effort.

An interesting example is provided by Jusot et al. (2013), who proposed a linear parametric estimation of a structural model of health inequality by using individual health-related behaviours as proxies of effort. Let, for the sake of simplicity, the health status to be determined by a single circumstance and a single effort variable. The estimation procedure takes place in two stages. First, the effort is estimated as a

[6] Ferreira and Gignoux (2011) p.634 present the parametric analogue to the standardised distribution as the predicted standardised income distribution: $\hat{\boldsymbol{\nu}}_{Y_{EP}} = \exp[\bar{C}_i\ \hat{\boldsymbol{\beta}} + \hat{\epsilon}]$. Where $\hat{\nu}_i$ represents the retaining of the within-type variation.

function of the circumstance:

$$e_i = \gamma_0 + \gamma_1 c_i + \nu_i \tag{4.12}$$

Second, the general health equation is estimated by plugging the predicted effort and the regression residual of the equation (4.12):

$$h_i = \beta_0 + \beta_1 c_i + \beta_2(\hat{\gamma_0} + \hat{\gamma_1} c_i + \nu_i) + u_i \tag{4.13}$$

which can be expressed as well in the following way:

$$h_i = \alpha_0 + \alpha_1 c_i + \epsilon_i \tag{4.14}$$

where $\alpha_0 = \beta_0 + \beta_2\gamma_0$ and $\alpha_1 = \beta_1 + \beta_2\gamma_1$ and represents the total contribution of the circumstances (direct and indirect). The residual term $\epsilon_i = \beta_2\nu_i + u_i$ contains the direct effect of effort and a zero-mean error term. In this way, it is possible to derive the indirect effect of circumstances through the effort observed with the variation in the type-specific estimated effort, $\hat{e_c}$, and the direct contribution of effort with the difference between the observed effort and the type-specific effort $e_i - \hat{e_c}$.

4.2.3 Criticism around the traditional empirical applications

The variety of estimation approaches clearly shows how complex is the process of translating the theory to the practice. Despite the richness of the data sources, there are several potential biases emerging with the estimation of inequality of opportunity in the real world.

First, the circumstance set provided by data is incomplete. We will always lack of some unobserved socio-demographic circumstances that matter in the process of defining the outcome inequality across people. As a consequence, the IOP estimators could be lower-bounded in representing the real inequality of opportunity (Ferreira and Gignoux, 2011). Second, in non-parametric estimations, the course of dimensionality due to an excessive number of partition units is very likely to take place. Hence, the too small sample size of certain types leads to biases in the estimated counterfactual distributions. Third, in order to deal with the first two issues, researchers arbitrarily refer to *ad hoc* definitions of types, which may ignore unobserved non-trivial social groups (Brunori et al., 2018; Donni et al., 2015). Fourth, Brunori et al. (2019) discuss the risk of IOP overestimation given by the dependence of the IOP estimates on the type-specific sampling distributions. Fifth, in parametric estimations, circumstances may be wrongly identified as fixed and additive. Notwithstanding, in many cases, they interact with each other and have a different effect according to other characteristics (Hufe and Peichl, 2015). Last, as a vicious cycle, Brunori et al. (2018) stressed that too accurate model specification in parametric estimations may lead to model overfitting and as a consequence, to upward-biases in the IOP estimates.

4.2.4 A new generation of empirical applications: a detailed illustration

There are two main issues related with the presented IOP estimation techniques. First, traditional parametric estimations imply the assumption of a single statistical model to be holding for the entire sample, and we estimate the coefficients indicating a constant relation among the dependent and the explanatory variable. However, it can happen that different parameters hold for different subgroups (types). Second, non-parametric approaches applied to the hand-generated Roemerian types allow for deriving the type-specific counterfacutal distribution across different groups, but suffer over the course of dimensionality.

The new methodologies to construct IOP indices can be classified according to the solutions introduced in either directions. On one side, there is the adoption of data-driven techniques for dealing with the population partitioning avoiding the curse of dimensionality (Brunori and Neidhöfer, 2020; Brunori et al., 2018; Davillas and Jones, 2020; Donni et al., 2015). On the other side, there has been a data-driven technique dealing with both partitioning and type-specific model estimation (Carrieri et al., 2020).

Among data-driven techniques adopted to perform population partitioning, we can distinguish two main approaches: Latent-Class Analysis (LCA) and tree-based methods. Both approaches use the information provided by the data to reduce the total amount of groups in a sample, and maximising the between-groups variance explained. The following paragraphs briefly illustrate all the adopted data-driven methodologies and the proposed new methodology, the Model-based Recursive Partitioning.

Tree-based methods derive from *decision trees* which, in statistics, can be used to visually represent the "decisions", or if-then rules, that are used to generate predictions. There are essentially two key components to building a decision tree: determining which *features* to split on the prediction sample and setting a rule to stop splitting. Tree-based methods aim at obtaining predictions for an outcome variable h as a function of a set of input variables $C = (C^1, ..., C^p, ..., C^P)$. Specifically, they use the set of input variables to partition the population into a set of non-overlapping groups (terminal nodes, or *leaves* of the tree). The partition is performed as far as there is a significant difference in the distribution of the dependent variable conditional on a specific realisation of the input variables. In other words, they test the null hypothesis of independence between the outcome and the input variables:

$$H_0^{C^P} : D(H|C^P) = D(H) \tag{4.15}$$

for each realisation of every input variable, and obtain a p-value with each test. Afterwards, the rule to stop the splitting will determine whether the p-value is statistically significant.

Brunori et al. (2018) propose an application of tree-based methods adopting the conditional inference trees training algorithm and testing whether the socioeconomic circumstances of a person significantly cause the income conditional distribution variation. This approach has been associated to a test for the existence of ex-ante inequality of opportunity (Brunori et al., 2018), where the presence of a split implies that the expected income is statistically varying according to different realisations of individual circumstances.

Latent-Class Analysis is a data-driven approach which takes as inputs all the possible discrete grouping variables and finds K latent groups in the population which contain a combination of those variables, each characterised by different conditional probability distribution.

While both techniques efficiently reduce the number of groups, the LCA provides less easily interpretative results requiring a successive decomposition of the latent classes in order to obtain the non-overlapping groups as defined by Roemer. Notwithstanding, the tree-based approach turned out to be more handy than LCA in the definition of the Roemerian types and in the investigation of the *structure* of the opportunity sets.

Clustering models which include a non-partitioning model estimation

The very recent contribution of Carrieri et al. (2020) presents an extension of the use of data-driven techniques to the more complex task of disentangling the direct and indirect effects of circumstances, considering as well some proxies of health-related effort. In other words, they provide an estimation of the overall relation between health, lifestyle and circumstances that is peculiar for each latent class identified as a Roemerian type.

The authors adopt Finite Mixture Models (FMM) to perform the population partitioning and to exploit the unobserved type-specific heterogeneity that characterises the outcome-to-effort relation. FMM is a model-based clustering algorithm,[7] that treats the distribution of the data as a mixture of K distributions, each appearing with mixing proportion where the class assignments (clusters) are unknown and learned from the data. In their paper Carrieri et al. (2020) use FMM to explore the information on the heterogeneous relationship between outcomes (health conditions) and regressors (health-related behaviours) within the various socioeconomic circumstances that the individuals are facing. The Model-Based recursive Partitioning (MOB), which is used for this paper's empirical application, is an alternative data-driven methodology with respect to FMM, to perform the estimation of the health-to-lifestyle relation and derive the socioeconomic population subgroups.

Zeileis et al. (2010) technically formalised and implemented the MOB. Differently form the FMMs, the MOB is a tree-based technique. In the next paragraphs, both methodologies are illustrated and compared within the context of IOP in health estimation. Let the individual health conditions h_i be described by two observable factors, lifestyle and socioeconomic background, respectively the effort e_i and the circumstances c_i of the IOP model. Let the following function be describing the health generating process for a given lifestyle and socioeconomic background:

$$f(h_i|\boldsymbol{c_i}, \boldsymbol{e_i}) \tag{4.16}$$

The IOP framework requires the estimation of a linear model such as

$$h_i = x_i^T \beta + \epsilon_i \tag{4.17}$$

where $x^T = [c_i, e_i]$. If there are social groups for which this linear relation systematically changes, the coefficients vector β would not efficiently hold for all n observations. Therefore, there would be a certain amount of population subgroups for which the β parameters are better defined.

The aim of the IOP framework is to individuate the K different subgroups (types) of the initial population, based on c_i, for which a linear model estimating the response of health to effort fits best. Hence, we can represent the full model as a weighted sum of the $k = 1, ..., K$ models associated with each subgroup parameters $\beta_{(k)}$:

$$f(h_i|c_i, e_i, \boldsymbol{\beta_{(1)}}, ..., \boldsymbol{\beta_{(K)}}) = \sum_{k=1}^{K} \pi_k(c_i) \cdot f(h_i|\boldsymbol{e_i}, \boldsymbol{\beta_{(k)}}) \tag{4.18}$$

depending on the adopted technique between FMM and MOB, the weight $\pi_k(c_i)$

[7] FMM represent a generalisation of the LCA because they do not necessarily need discrete partitioning variables.

and the subgroup models will be identified with a different process.

Finite Mixture Models

In the FMM, the conditional density of the health outcome, eq. 4.16, is assumed to be drawn from a population which is characterised by a finite additive mixture of K distinct clusters, eq. 4.18, each one assigned with proportions that are function of the circumstances $\pi_k(c_i)$. Each subgroup corresponds to a *component*, and for each component a linear regression of h_i on e_i is estimated. Thus, the estimated coefficients for each parameter vary across the components. The overall mixture of models is obtained aggregating each component model with the corresponding circumstance-related weight. Generally, the weight represents a smooth and monotonic transition from one component to another (Frick et al., 2014). The estimation of all possible K models is performed with Maximum Likelihood obtaining $\boldsymbol{\beta_k}$ subgroup-specific parameter estimates. The selection of the optimal K number of latent classes is usually subject to a statistical information criterion. The parameters are necessary to estimate the posterior distributions for the membership of each individual in the K classes. The posterior density will be used to form the population partitioning. Carrieri et al. (2020) performed the partitioning following the technique of the *modal assignment* by placing each individual into the type with the highest posterior distribution.

The final types are not clearly identifiable in terms of circumstances, indeed, they will be described by a mix of the characteristics not easily associated with a social group. The identification is done as well following the modal assignment rule, specifically through the estimation of the posterior distribution conditional to specific realisation of each circumstance. Therefore, this process could be somewhat convoluted and not easily adaptable to the identification of social classes as IOP aims at. This may happen due to the fact that some circumstances may be almost equally identify more than one latent types in terms of posterior probabilities, i.e. if there are two latent types and the posterior for being a migrant is respectively 0.51 for type A and 0.49 for type B, with the modal assignment rule, we would say that the type A represents migrants without anymore considering that type B is representing migrants the 49% of times.

Model-Based Partitioning

The Model-Based recursive Partitioning is an algorithm which estimates a full model as shown in eq. 4.17, and assess the parameters' *instability* for all specified covariates (c_i). The estimation of the model can take place either with OLS or ML techniques. In either cases, given the observation of the dependent variable, the parameter coefficients of the model are derived by optimising the objective function. When a model fits well, we should see that the sum of the deviations of the estimated $\hat{h}_i$ from the observed h_i should approximate zero. However, if the parameters change along a specific partitioning variable c_{ik} we would observe *"systematic deviations from zero"* (Frick et al., 2014). As a consequence, the full model might not be the best solution though we should account for the c orderings.

By means of Generalised M-fluctuation tests (Zeileis and Hornik, 2007), it is possible to compute a statistic summing all the deviations along a categorical partitioning variable.[8] If the fluctuation test statistic turns out to be highly significant with respect to a certain threshold α, we are rejecting the hypothesis of no-instability and we prefer to split the sample and estimate two distinct models for two distinct realisations of $\boldsymbol{c}$.

Schematically, Zeileis et al. (2010) illustrate the steps of the algorithm as follows:

1. Fit the model $h_i = \alpha + \boldsymbol{e_i}\,\beta$ given the set of all the potential partitioning variables $c_1, ..., c_k$, i.e. estimate the model in the entire sample.

2. Check whether there is any partitioning variable causing parameter estimates for the model to be unstable. If there is overall instability, select the variable associated with the highest instability source, otherwise stop.

3. Compute the exact split point which optimises the objective function of the estimation (OLS or ML type) according to the selected unstable partitioning variable.

4. Split the node into child nodes and restart the procedure, i.e. estimate the same models into different subgroups.

The output of the model-based recursive partitioning algorithm depends on a tuning parameter $\boldsymbol{\alpha}$ which determines the p-value threshold for performing each split. More specifically, the $\boldsymbol{\alpha}$ parameter is used as a threshold measure for assessing the statistical significance of the instability test that the MOB performs before splitting the sample. Thus, this parameter determines the depth of the output tree. The tuning of the algorithm can be performed with a machine-learning technique ensuring that MOB stops splitting the sample when no further split would result in a better out-of-sample fit of the data. In this paper, the tuning of the algorithm is performed by 5-fold cross validation. The critical p-value selected is 0.105. This means that the out-of-sample mean-squared-error in predicting individual health is minimised when MOB is allowed

[8] This statistic is distributed as a χ^2 and we can compute the Bonferroni-adjusted p-value for testing its significance, (Zeileis et al., 2010).

to split the sample until it is possible to reject the null hypothesis of a same $\hat{\alpha}$,$\hat{\beta}$ obtaining a p-value ≤ 0.105.

The final model can be represented as a tree for which all the branches are associated with the split of the sample determined by the instability test, and the terminal nodes are the sub-samples used for the final fitted models. In the IOP framework, each terminal node describes a population type according to which splitting path brings to it. The full model weights describe the path to the terminal node as the product of all the splits leading up to the node. Differently form the case of FMMs, the weights represent abrupt variations in the model behaviour and, when we have multiple splits for the same covariate, the tree is representing a non-monotonic transition from one subgroup to another (Frick et al., 2014).

A contextualised comparison between FMM and MOB

Following the discussion of Frick et al. (2014), the two models have different abilities in detecting the correct subgroups given certain characteristics of the model: *i)* how strongly the relationship between the dependent and the regressor differ across groups; *ii)* the level of association between the partitioning variables and the groups.

In the IOP framework, the former can be interpreted with the following questions: to what extent the effort is indirectly determined by the circumstances? Do we observe a stable or variable relation between lifestyle and health across social groups? Is there a strong association between a given circumstance realisation and the definition of a population type? From the perspective of inequality of opportunity in health, a more equal word with respect to opportunities would be the one in which the effort-to-health relation is stable and there are no relevant social groups. Given that the real world is far from being like that, we need the tool that better responds to the capabilities to detect these aspects in the modelling framework.

Frick et al. (2014) provide a simulation to investigate how these techniques vary according to the data characteristics. Their simulation brings the following considerations:

- if there is a strong association between the partitioning variables and the groups formed, the MOB is able to detect smaller variations in the type-specific coefficients, $\beta_{(k)}$, than FMM because they employ a significance test for the single parameter.

- by contrary, FMMs are able to detect latent subgroups even when there is no association with the covariates (as far as the latent groups reflect significant differences in the $\beta_{(k)}$).

- FMMs better perform groups when the partitioning variables simultaneously determine them.
- the MOB has a better performance in the instability checking, when both important and least important partitioning variables are included in the test
- if the association between the partitioning variables is smooth and monotonic, FMMs perform a good partitioning.
- When the groups are strictly identifiable with abrupt shift form a category to another, the MOB is more suitable technique for partitioning

The FMM is more concerned in individuating, among all possible group-specific estimated models, when parameters are varying. MOB is instead more concerned with the partitioning variables, and how they interact with the full model.

In the context of IOP in health and the choice of the more adaptable empirical methodology, we face a trade off. If the main concern is to find meaningful social groups, the model-based partitioning technique is preferred. If the aim is to individuate some tight variations in the relation between health and lifestyle and, afterwards, investigate the influence of socioeconomic conditions, the latent-class analysis is a suitable technique. In this paper, the priority is to identify meaningful social groups and perform a group specific analysis of the health-to-lifestyle relation. Therefore, a model-based recursive partitioning is specified, and the unfair inequalities are quantified through the computation of the direct unfairness and fairness gap measures á lá F&S.

4.3 Estimation Strategy

The estimation strategy comprehends the tuning of MOB algorithm and the computation of Direct Unfariness and Fairness Gap.

The health condition is modelled as a linear function of life-style, but the parameters of the function are not assumed to be the same across individuals characterised by different socioeconomic background.[9] Hence, the MOB output coincides with a tree whose k terminal nodes are populations subgroups used for estimating k linear regressions as:

$$h_i = \alpha_k + \beta_k e_i \qquad (4.19)$$

[9] The outcome of the estimation can be easily compared with the outcome of Carrieri and Jones (2018), who implemented the same linear model estimation after having performed the classical Roemerian type partition (semi-parametric approach).

for $k = (1, .., m)$ terminal nodes.

For computing the counterfactual distribution $\tilde{H}_{DU}$ and $\tilde{H}_{FG}$ the inputs are the type population partition, the effort levels and the consequent tranche partitions into type-specific effort degrees, and the type k estimated coefficients of the health-to-effort relation, $\hat{\alpha}_k$ and $\hat{\beta}_k$. As shown in section 2.1, in order to compute the $\tilde{H}_{DU}$, the health status of each individual has to be fitted by assuming everyone is exerting the same reference effort. Being the effort, or lifestyle behaviour, in absolute terms related with socioeconomic factors, the analysis presented refers to the Roemerian approach to applying the tranche partitioning. To do so, the tranche groups have been obtained by looking at the type-specific degree of effort. Consequently, the reference effort $\bar{e}_q$ is obtained by averaging the level of effort within each tranche. Given that the degree of effort is grouped into 5 quantiles q we obtain a direct unfairness measure for each quantile-specific reference effort.

$$\tilde{h}_i^{k,q} = \hat{\alpha}_k + \hat{\beta}_k \bar{e}_q, \tag{4.20}$$

The inequality measure is obtained by computing the variance of the counterfactual distribution:

$$DU_q = I(\tilde{h}^{k,q}) = var(\tilde{h}^{k,q}). \tag{4.21}$$

The fairness gap counterfactual distribution, $\tilde{Y}_{FG}$, is obtained by subtracting the health fitted with the j^{th} reference-type coefficients $\hat{\alpha}_j$ and $\hat{\beta}_j$ to the actual fitted health:

$$\tilde{h}_i^j = \hat{\alpha}_k + \hat{\beta}_k e_i - (\hat{\alpha}_j + \hat{\beta}_j e_{qj}) \tag{4.22}$$

where $\hat{h_i} = \hat{\alpha}_k + \hat{\beta}_k e_i$, and e_{qj} is the average effort computed within the j^{th} individual i specific tranche. In other words, after having imposed a between-type constant health-to-effort relation, we observe the different health outcomes achievable by each observed effort level. The inequality measure is obtained by computing the variance of the counterfactual distribution:

$$FG_j = I(\tilde{h}^j) = var(\tilde{h}^j). \tag{4.23}$$

This measure reduces the total explained variability by removing the variability associated with a specific socioeconomic condition. When, for $j \neq k$, this difference is higher than zero it means that the society is rewarding more the effort of the type k people than type j people.[10]

[10] Note also that when estimating eq. 2 and 3, one is implicitly assuming that the error term in eq 1

4.4 Data

The data adopted comes from a sub-sample of the Wave 2 (2010-2011) of UKHLS. The UKHLS is a survey conducted in the UK on a General Population Sample. Within the whole population interviewed for the wave 2, it has been considered only the sub-sample of respondents with non-missing objective health status variable and lifestyle information. The considered sample contains 5,561 observations, that is approximately the 10% of the whole sample in the wave 2. Thus, the possibility to observe an objective health measure comes at the cost of having a sub-sample which cannot fully represent the whole population as the original sample of the UKHLS data. The variable representing the health status is a composite biological measure of health condition based on parameters such as adiposity, blood pressure, inflammation, blood sugar levels, cholesterol levels. These observations are part of a special module in the UKHLS survey which have been nurse-based collected in waves 2 and 3. The allostatic load index has been computed and used as a measure of health status by Davillas and Pudney (2020), and in the IOP framework by Davillas and Jones (2020). In their paper - at p.7 - they define it as:

> (...) an overall assessment of a respondent's physiological condition. (Allostatic load is) elevated when a person's biological systems are affected by repeated stressors, resulting in persistently elevated or altered levels of a number of biomarkers associated with 'chronic stress'.

Practically, the allostatic load measure is a multidimensional index obtained through an additive aggregation of each dimension-specific z-score. In that way, the individual dimension-specific observed score is related with the overall distribution of scores of that dimension.[11]

Table 4.1 shows the descriptive statistics of the allostatic load variable and the effort variable.

Table 4.1: Summary statistics: Allostatic load (H) and Effort (E)

	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
H	6,292	−0.000	0.840	−2.728	−0.562	0.502	4.907
E	5,561	5.210	1.741	−0.473	3.964	6.598	9.114

is not to be considered part of unfair health inequality. An alternative approach could be considering the error term as part of the unfair inequality and add the term $u_i = h_i^k - \hat{\alpha}_k - \hat{\beta}_k e_i$.

[11] A more detailed explanation is provided in the Appendix C.

The allostatic load measure has been adjusted for the individual age in order to control for the age-specific variability in health. The age-adjustment is performed by regressing individual allostatic load on a polynomial of order two for the age of the person recorded by the nurse at the moment of the health dimensions collection. Controlling for the age it implies that we assume this category to be unproblematic for what concerns the health inequalities.

Effort variables considered are: smoking and drinking behaviours, dietary habits (bread/milk/fruit/vegetable), sport activity, sedentary life style. The variables of efforts are summarised in a scalar obtained by Principal Component Analysis (PCA).[12] More specifically, the outcome variable from the PCA analysis represents a measure of effort in the overall health-related behaviour, namely the lifestyle. As shown in the correlations table (table 4.2), this variable correlates with the behaviours consistently with a definition of "healthy behaviour".[13]

Health and Lifestyle	Effort
Eat fruit	0.516***
Eat vege.	0.438***
Milk habits	0.203***
Bread habits	0.193***
White bread only	-0.320***
Smoking habits	-0.396***
Ex-smoker	0.051***
Sport self-ass.	0.471***
Sport frequency	0.305***
Walking	0.769***
Drinking habits	-0.010

Table 4.2: Correlation between effort and lifestyle variables

The variables describing the socioeconomic background have been used as splitting variables in the model-based partitioning estimation. The partitioning variables are: ethnic group, place of birth, father and mother occupation, mother and father education, mother and father activity status (all information about parents are referring to the period in which the respondent was 14). Table 4.3 shows the frequencies of each circumstance category within the sub-sample.

Decision tree-based algorithms, through the practice of surrogate splitting, traditionally address the issue of imputing the missing values among the partitioning

[12] An illustration on the procedure adopted is available in Appendix C.

[13] As in Davillas and Jones (2020), some problems were encountered when drinking behaviours were present. This is partly solved using a dummy for heavy drinkers that positively correlate with allostatic load and negatively (although the correlation is not significant) with the variable of effort.

Table 4.3: Descriptive statistics - Circumstances

Variables	Frequencies (%)
Ethnic group	
uk white	91.94
irish white	0.78
other white	2.32
mixed	0.94
asian	2.69
african or arab	1.33
Mother occupational skill-level	
High skill	9.19
Medium high skill	8.60
Medium skill	27.54
Low skill	14.32
Father occupational skill-level	
High skill	16.62
Medium high skill	42.48
Medium skill	26.67
Low skill	9.79
Mother education	
not educated	1.16
primary	48.06
secondary	29.78
upp sec./tertiary	21.00
Father education	
not educated	0.95
primary	43.61
secondary	21.20
upp sec./tertiary	34.23
Mother activity status	
Working	59.11
Not working	39.02
Decreased	1.34
Unknown	0.53
Father activity status	
Working	88.74
Not working	4.12
Decreased	4.07
Unknown	3.07

variables. Due to the presence of missing values among the partitioning variables presented[14], the missing data have been imputed adopting a linear model-based imputation technique (R Hmisc package / function: *aregImpute*).[15] The imputers are all the observable circumstances, educational attainment and gender.

4.5 Estimation Results

An important step to take before presenting the estimation results is to analyse the overall relation between the objective health status and the effort variable. As it emerges from the figure 4.2, the allostatic load level and the effort exerted are negatively and significantly related. This outcome is coherent with the interpretation that a good lifestyle has an impact in reducing the physiological effects of chronic stress. Although their correlation is negative and statistically significant, the two variables do not show, from the scatter plot, to have a clear fit; which means that effort is far from being sufficient on its own to explain health status. Thus, notable improvements are expected from the implementation of the MOB which will consider whether this relation fits better for different socioeconomic background variables.

The model estimation led to four significant splits and five terminal nodes defining the relevant socioeconomic background characteristics for the health-to-effort relation determination.

The terminal nodes identifying the population types are:

- Node 2: father with high skill occupation
- Node 5: father unemployed, UK white or other white
- Node 7: father with low/medium/medium-high skill occupation, UK/other white, mother with low education (primary)
- Node 8: father with low/medium/medium-high skill occupation, UK/other white, mother with up to tertiary education (secondary with or without certification, tertiary)
- Node 9: father with up to medium/high skill occupation, African, Asian, Caribbean or mixed ethnicity

[14] The missing cases linked with the respondent's parent not being alive or living in the household at respondent's age of 14, have been excluded from the imputation. A general summary on the missing categories is available at table C.1 in Appendix C.

[15] Harrell Jr (2019)

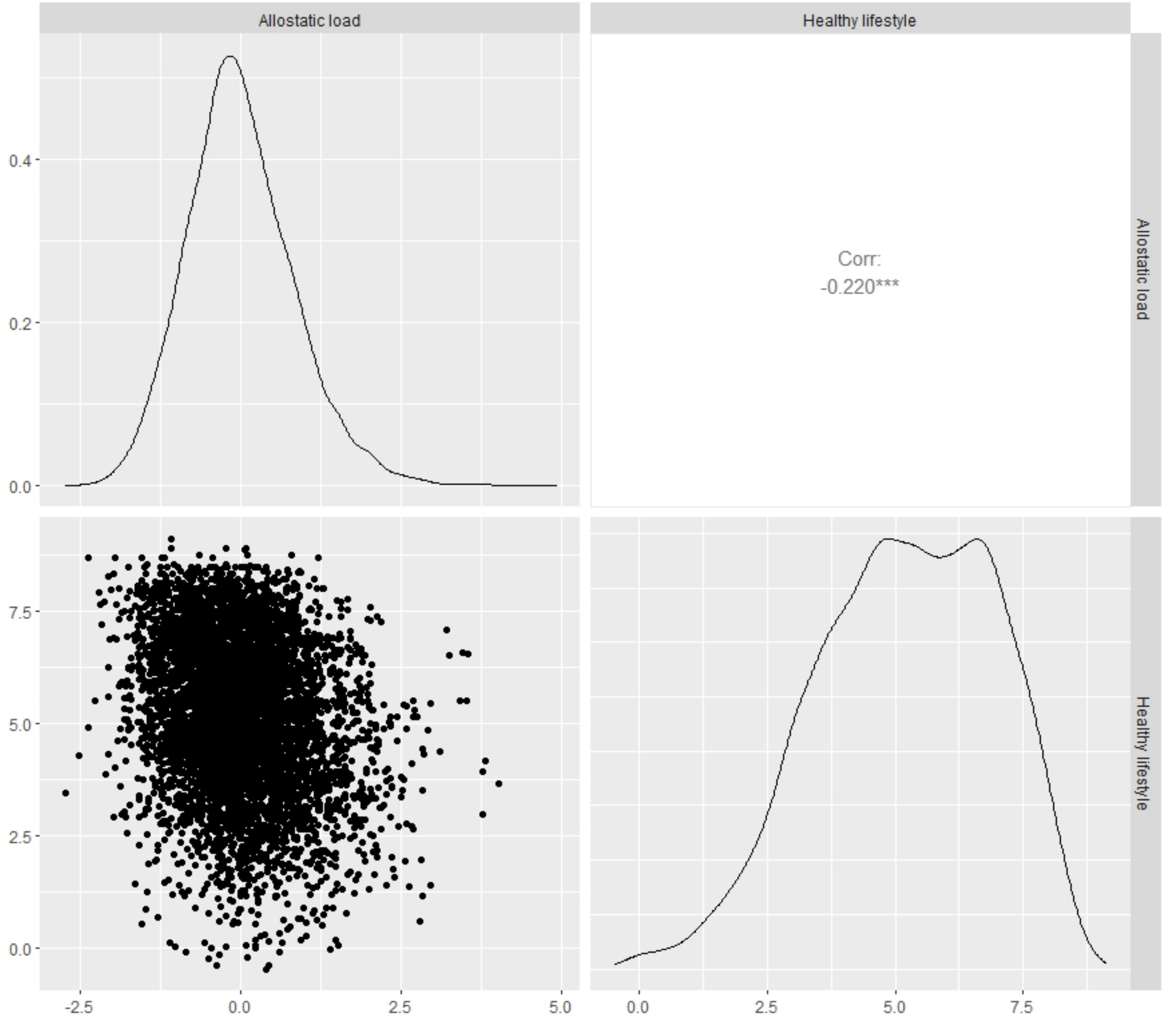


Figure 4.2: Relation between allostatic load and effort

Figure 4.3 shows the MOB tree outcome and table 4.4 shows the type- specific coefficients of the linear allostatic load-to-effort regression.

For simplicity, whenever reference is made to terminal nodes, the terminology 'population types' will be used, with each type being associated with the number of the terminal node it refers to.

As it emerges from the estimation results shown in table 4.4, the relation between effort and allostatic load is always negative. For type 5 and 9 the response of health to effort is lower in absolute term and not significant. Meaning that, for type 5 and 9, no actual return to effort is visible in the health status.[16]

Figure 4.4 serves as graphical support to the estimation results shown in table 4.4. However, this figure is showing the response of health to hypothetical effort levels thus

[16] This may be at least in part explained by the smaller sample size of the types.

Table 4.4: Regression coefficients for each terminal node (population type)

Predictors	Type 2	Type 5	Type 7	Type 8	Type 9
Intercept	0.382^{***}	0.543^{**}	0.555^{***}	0.542^{***}	0.4487^{**}
CI 95%	$[0.199; 0.572]$	$[0.248; 0.855]$	$[0.449; 0.651]$	$[0.433; 0.662]$	$[0.116; 0.815]$
Effort	-0.103^{***}	$-0.058^{\cdot}$	-0.097^{***}	-0.1150^{***}	-0.039
CI 95%	$[-0.134; -0.071]$	$[-0.119; -0.001]$	$[-0.114; -0.078]$	$[-0.136; -0.096]$	$[-0.104; -0.016]$
R^2	0.049	0.015	0.041	0.056	0.002
Adj. R^2	0.048	0.011	0.041	0.055	-0.0023
Num. obs.	925	219	2223	1886	223
RMSE	0.743	0.826	0.831	0.807	0.8404

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

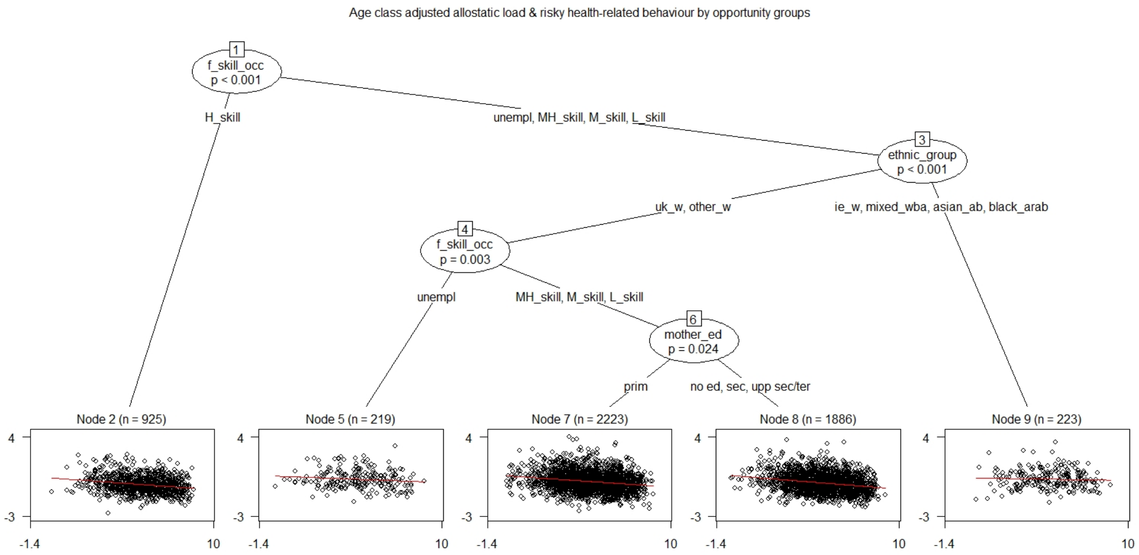


Figure 4.3: Model-based partitioning: Health to effort relation by circumstances.

it is not representative of the true effort and health levels observed. Table 4.5 provides the mean and variances of allostatic load and effort within each specific type.

It emerges that allostatic load is generally higher for who has a poor socioeconomic background. Moreover, although poorer groups show a drastically lower return on effort, this relationship does not spill over into a tendency towards riskier behaviour. Indeed, while the type-specific health shown in table 4.5 significantly changes across types both in its mean ad variance - even showing higher variability within worse-off types -, the type-specific effort varies less. This is visible too when looking at the effort density distributions for the different types, figure 4.5. This evidence leads us to the conclusion that a different return to effort in the health outcomes does not determine different behavioural choices among the observed individuals.

Type	Mean(e)	var(e)	mean(h)	var(h)	weight (%)
2	5.5534	2.6548	-0.1778	0.6420	16.62
5	4.8233	3.4202	0.2640	0.7451	3.91
7	5.0352	3.1474	0.0719	0.7044	40.03
8	5.2873	2.9148	-0.0739	0.6872	34.28
9	5.2185	3.1499	0.3063	0.7024	5.15

Table 4.5: Within type descriptive statistics

Table 4.6 provides the allostatic load distribution across types and effort quantiles (tranches). The value in the cells represent the average allostatic load within all the group belonging to a specific type and tranche. The table shows, unsurprisingly how

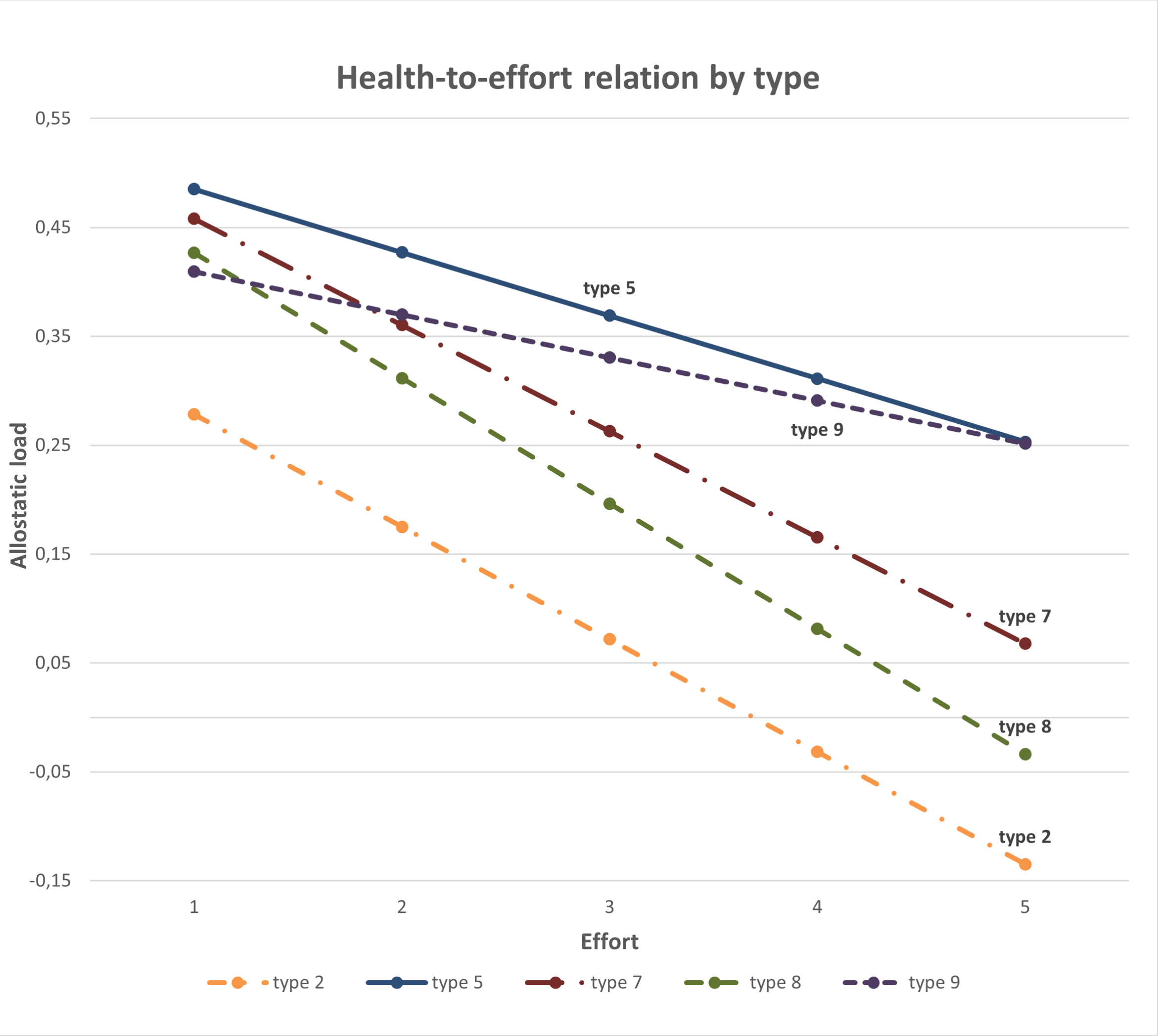


Figure 4.4: Health-to-effort relation by type

the allostatic load diminishes for higher effort quantiles but with a different magnitude across the types. Indeed, disadvantaged types (type 5 and 9) have a high allostatic load even when effort is high (q4 and q5 tranches).

Finally, bootstrapped estimates of direct unfairness and fairness gap are provided. The outcome tables show five DUs, one for each effort quintile, and five FGs, one for every possible reference effort. The inequality index is the variance of the counterfactual distribution. Tables 5 and 6 report the obtained measures. [17]

Before going through the results, it is worth recalling that the estimations of DU

[17] Note that in this case, the error term is part of fair inequality because it has been excluded from the computation, this implies a low level of estimated unfair inequality with respect to the total observed inequality.

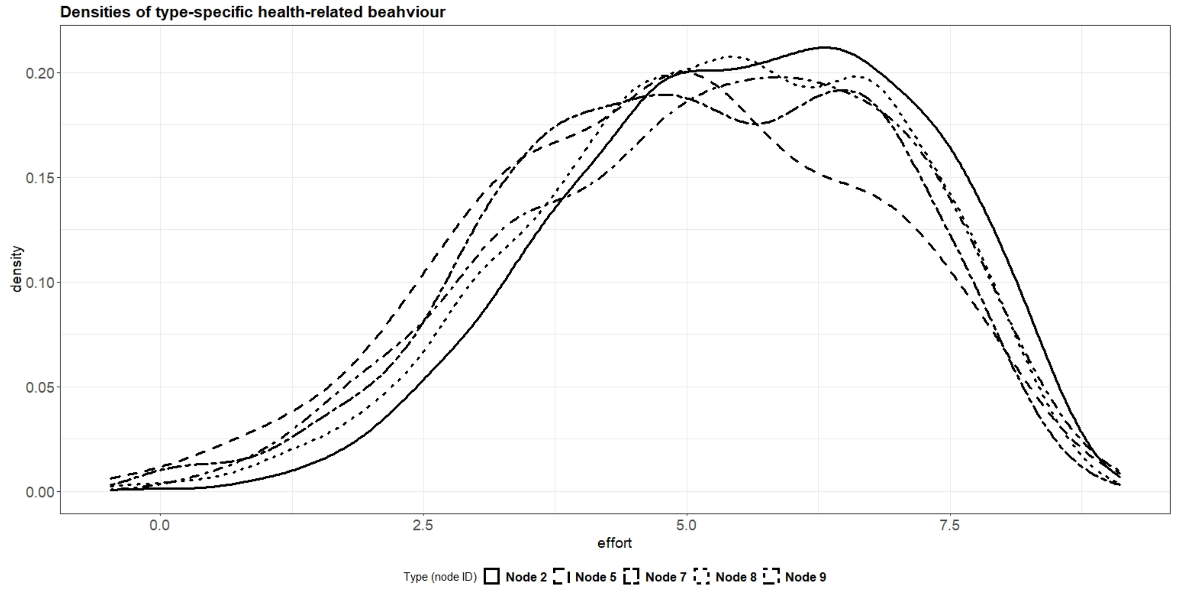


Figure 4.5: Type specific effort densities.

	Tranche				
Type	q1	q2	q3	q4	q5
2	0.1418	-0.0030	-0.3049	-0.2376	-0.3910
5	0.3770	0.2834	0.3933	0.1150	-0.0036
7	0.2901	0.1493	0.0668	-0.0706	-0.1883
8	0.2628	0.0209	-0.0953	-0.1404	-0.3365
9	0.2287	0.4278	0.2221	0.3808	0.0331

Table 4.6: Average health outcome by type and effort quantile

and FG are computed on the predicted health status. Thus, all the inequality in health that has not been explained within the model is not considered for the computation.

As it emerges from the two estimates tables, DU is monotonically increasing with effort. This is due to the interaction of intercepts - a sort of types' fixed effect - and slopes - return to healthy life-style. When using DU we are evaluating inequality considering only one reference level of effort, thus, the measure is completely insensitive to inequality in any other part of the effort distribution. This implies that, by assumption, any difference in effort is full responsibility of the individual. Therefore, the unequal health outcomes associated to different efforts are considered unproblematic. If types with lower intercepts - better health condition when exerting zero effort - have also higher returns to healthy life-style, they will tend to diverge in the DU outcomes from the worse-off types. The measure of DU is more or less representing what shown in figure 4.4 when fixing the effort levels to the reference values. Therefore, the inequality

Reference tranche	Direct Unfairness	Confidence Interval (95%)
q1	0.0065	[0.0021 ; 0.0146]
q2	0.0081	[0.0047 ; 0.0152]
q3	0.0112	[0.0074 ; 0.0173]
q4	0.0159	[0.0101 ; 0.0226]
q5	0.0231	[0.0121 ; 0.0319]

Table 4.7: Direct Unfairness for each specific effort quantile - Bootstrapped results

Reference type	Fairness Gap	Confidence Interval (95%)
Type2	0.0176	[0.0141 ; 0.0251]
Type5	0.0278	[0.0154 ; 0.0457]
Type7	0.0178	[0.0138 ; 0.0245]
Type8	0.0186	[0.0141 ; 0.0257]
Type9	0.0438	[0.0166 ; 0.0587]

Table 4.8: Fairness Gap for each specific reference type - Bootstrapped results

measure is computed on the differences in the health outcomes on the y-axis of the figure, that we observe at each corresponding effort level.

FG vary less than DU and is higher for the two types with lower return to effort. Recalling the FG formula, the counterfactual distribution on which the inequality index is computed, is the difference between the actual predicted health and the health obtainable in the case in which the return to effort was described by a certain reference society. The reference society outcomes are obtained assuming that all the people at each stage have exactly the same health-to-effort response, thus, the highest counterfactual variability would emerge from the difference between observed health and the health obtainable when the reference type is the worse-off type.

4.6 Conclusions

This study aims to provide both a methodological innovation to the IOP measurement, as well as new insights on the health inequalities. The methodological innovation is represented by the adoption of the Model-Based recursive Partitioning for estimating the health-to-lifestyle relation while considering the different socioeconomic background of people. This paper encompasses an extended contextualisation of the MOB technique in the IOP framework. Whilst, for what concerns the inequalities of opportunity in health, a normatively defined responsibility-sensitive framework has been adopted to empirically measure the Direct Unfairness and the Fairness Gap à là Fleurbaey and Schokkaert (2009).

Among the main features of the MOB's employment in IOP measurement is its ability to capture those socioeconomic characteristics which are fundamental to determine a change in the conditional distribution of the outcome health-to-lifestyle model. Hence, the MOB algorithm has been employed to estimate the type-specific relation between health and lifestyle. The empirical application was performed with data from the UKHLS data (wave 2) on a subsample for which data on objective health status - the allostatic load measure - is observable. It emerged that lifestyle plays a non-constant role in determining health outcomes. Indeed, people with a more disadvantaged family background experience worse health statuses on average. Despite that, it turns out that the differences in health outcomes are not necessarily explained by substantial variations in lifestyles; meaning that the lower return to effort is not reflecting an incentive to riskier behaviours for the disadvantaged people. Therefore, there is an "effort return gap" which have socioeconomic origins. The computed DU and FG show that the within-type inequality varies across the effort levels in a monotonic way. Poorer socioeconomic conditions lead to a lower return to efforts and a higher overall gap with respect to better-off types' achievable outcomes.

As far as it emerged from the estimation results, the objective health can be only partly described by the lifestyle, when their relation is let to vary in other socioeconomic characteristics. Indeed, there are many aspects which are not included in the model even though they are impacting the health status. Some of them are unobservable to the researcher/policymaker, e.g. the genetic endowments or other possibly relevant socioeconomic information. Some others, however, could fit in the F&S formalisation and should be taken into account, e.g. healthcare consumption and the role of public healthcare services.

Given that the predicted health is used for the inequality computations, the inequality that is actually observed in the DU and FG is not considering the individual deviations from the fitted results. These deviations contribute to the "unexplained inequality" which is, by assumption, excluded from the "unfair" inequality computation. The extension of the model including other determinants of health could partly solve this problem. Alternatively, a possible solution would be to work on the inclusion of the whole *unexplained inequality* - the regression residual - within the DU and FG computation. Due to the preliminary nature of this application, priority was given to the exploration of the new techniques' capacities rather then proceed with a more accurate structural model definition. Last, the sub-sample of the data for which objective health was observable, may not be fully representative of the society. For this reason, further implementation of this methodology should be its extension to a larger and more representative database.

Bibliography

F. Bourguignon, F. H. Ferreira, and M. Menéndez. Inequality of opportunity in brazil. *Review of Income and Wealth*, 53(4):585–618, 2007.

P. Brunori. How to measure inequality of opportunity: a hands-on guide. 2016.

P. Brunori and G. Neidhöfer. The evolution of inequality of opportunity in germany: A machine learning approach. *ZEW-Centre for European Economic Research Discussion Paper*, (20-013), 2020.

P. Brunori, P. Hufe, and D. G. Mahler. *The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees.* Policy Research Working Papers. The World Bank, Feb. 2018.

P. Brunori, V. Peragine, and L. Serlenga. Upward and downward bias when measuring inequality of opportunity. *Social Choice and Welfare*, 52(4):635–661, 2019.

V. Carrieri and A. M. Jones. Inequality of opportunity in health: A decomposition-based approach. *Health Economics*, 27(12):1981–1995, 2018.

V. Carrieri, A. Davillas, and A. M. Jones. A latent class approach to inequity in health using biomarker data. *Health Economics*, 29(7):808–826, 2020.

D. Checchi and V. Peragine. Inequality of opportunity in italy. *The Journal of Economic Inequality*, 8(4):429–450, 2010.

G. A. Cohen. On the currency of egalitarian justice. *Ethics*, 99(4):906–944, 1989.

A. Davillas and A. M. Jones. Ex ante inequality of opportunity in health, decomposition and distributional analysis of biomarkers. *Journal of Health Economics*, 69:102251, 2020. ISSN 0167-6296.

A. Davillas and S. Pudney. Using biomarkers to predict healthcare costs: Evidence from a uk household panel. *Journal of Health Economics*, 73:102356, 2020.

P. L. Donni, J. G. Rodríguez, and P. R. Dias. Empirical definition of social types in the analysis of inequality of opportunity: a latent classes approach. *Social Choice and Welfare*, 44(3):673–701, 2015.

R. Dworkin. What is equality? part 1: Equality of welfare. *Philosophy & Public Affairs*, pages 185–246, 1981.

F. H. G. Ferreira and J. Gignoux. The Measurement of Inequality of Opportunity: theory and an application to Latin America. *Review of Income and Wealth*, 57(4): 622–657, Dec. 2011.

M. Fleurbaey. Equal opportunity or equal social outcome? *Economics & Philosophy*, 11(1):25–55, 1995.

M. Fleurbaey. *Fairness, Responsibility, and Welfare.* Oxford University Press, 2008.

M. Fleurbaey and F. Maniquet. *Equality of Opportunity: the Economics of Responsability.* 2012.

M. Fleurbaey and E. Schokkaert. Unfair inequalities in health and health care. *Journal of Health Economics*, 28(1):73–90, 2009.

M. Fleurbaey, V. Peragine, and X. Ramos. Ex post inequality of opportunity comparisons. *Social Choice and Welfare*, 49(3-4):577–603, 2017.

H. Frick, C. Strobl, A. Zeileis, and M. Gilli. To split or to mix? tree vs. mixture models for detecting subgroups. 2014.

F. Harrell Jr. Package 'hmisc'. *CRAN2018*, 2019:235–236, 2019.

P. Hufe and A. Peichl. Lower bounds and the linearity assumption in parametric estimations of inequality of opportunity. 2015.

F. Jusot, S. Tubeuf, and A. Trannoy. Circumstances and efforts: how important is their correlation for the measurement of inequality of opportunity in health? *Health economics*, 22(12):1470–1495, 2013.

M. Marmot. Social determinants of health inequalities. *The lancet*, 365(9464):1099–1104, 2005.

Organization. *Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health: Commission on Social Determinants of Health Final Report.* World Health Organization, 2008.

G. Orwell. *Animal Farm.* Harlow: Longman, 1989.

J. Rawls. *A Theory of Justice.* Harvard University Press, 1971.

J. E. Roemer. *Theories of Distributive Justice.* Harvard University Press, 1998.

J. E. Roemer and A. Trannoy. Equality of Opportunity. In *Handbook of Income Distribution*, volume 2, pages 217–300. Elsevier, 2015.

A. Sen, S. Anand, and F. Peter. Why health equity. In *:*, pages 21–33. Oxford University Press, 2004.

D. Van de Gaer. *Equality of Opportunity and Investment in Human Capital.* PhD thesis, Doctoral Dissertation, Leuven: Katholieke Universiteit Leuven, 1993.

W. WHO. *The Economics of Social Determinants of Health and Health Inequalities: a Resource Book*, volume 3700. World Health Organization, 2013.

A. Zeileis and K. Hornik. Generalized m-fluctuation tests for parameter instability. *Statistica Neerlandica*, 61(4):488–508, 2007.

A. Zeileis, T. Hothorn, and K. Hornik. Party with the mpb: Model-based recursive partitioning in r. *R package version 0.9-9999*, 2010.

Appendix A

Data and Tables

Table A.1 - A.2 - A.3 show the descriptive statistics for the selected leisure activities observed respectively in years 2006, 2011 and 2016.[1]

Table A.1: Summary statistics - Year 2006

Variable	Mean	Std. Dev.	Min.	Max.	N
TV	1.332	0.748	1	5	286
Pc games	2.983	1.469	1	5	286
Internet	2.615	1.58	1	5	286
Listen to music	1.297	0.829	1	5	286
Play music	3.479	1.654	1	5	286
Sport	2.451	1.235	1	5	286
Dance/act	3.727	1.31	1	5	286
Tech activities	4.024	1.255	1	5	286
Read	2.776	1.431	1	5	286
Relax	2.297	1.216	1	5	286
Girl/boyfriend	3.007	1.757	1	5	286
Best friend	1.913	0.96	1	5	286
Clique	2.367	1.217	1	5	286
Youth centre	4.350	1.071	1	5	286
Volunteer	4.077	1.29	1	5	286
Religious	4.255	0.978	1	5	286

[1] The frequencies reported in the table are referred to the original ordinal units assigned in the SOEP data base. Precisely, 1 = Daily, 2 = Weekly, 3 = Monthly, 4 = Rarely, 5 = Never.

Table A.2: Summary statistics - Year 2011

Variable	Mean	Std. Dev.	Min.	Max.	N
TV	1.455	0.839	1	5	506
Pc games	2.779	1.605	1	5	506
Internet	1.358	0.791	1	5	506
Listen to music	1.172	0.570	1	5	506
Play music or sing	3.7	1.495	1	5	506
Sport	2.152	1.065	1	5	506
Dance or act	4.182	1.146	1	5	506
Do tech activities	4.028	1.205	1	5	506
Read	2.67	1.407	1	5	506
Relax	2.249	1.256	1	5	506
Girl/boyfriend	3.316	1.641	1	5	506
Best friend	2.006	0.976	1	5	506
Clique	2.285	1.208	1	5	506
Youth centre	4.427	0.977	1	5	506
Volunteer	4.221	1.161	1	5	506
Religious	4.213	1.02	1	5	506

Table A.3: Summary statistics - Year 2016

Variable	Mean	Std. Dev.	Min.	Max.	N
TV	1.359	0.735	1	4	493
Pc games	2.15	1.429	1	5	493
Social network	1.562	1.17	1	5	493
Internet	1.627	1.021	1	5	493
Listen to music	1.249	0.699	1	5	493
Play music	3.805	1.507	1	5	493
Sport	2.363	1.162	1	5	493
Dance/act	4.32	1.061	1	5	493
Tech activities	4.256	1.122	1	5	493
Read	2.947	1.386	1	5	493
Relax	2.266	1.246	1	5	493
Girl/boyfriend	2.771	1.888	1	5	493
Best friend	2.172	0.965	1	5	493
Clique	2.554	1.218	1	5	493
Youth centre	4.55	0.950	1	5	493
Volunteer	4.191	1.195	1	5	493
Religious	4.323	0.906	2	5	493

Table A.4 shows the estimation output of the mean difference significance test for the Complexity scores.

Table A.4: Mean Difference Test

	Mean diff.
Complexity	0.0659
	(0.72)
N	506

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The t-statistic of the test is 0.7156 with 504 degrees of freedom. The corresponding two-tailed p-value is 0.4746 so with a very high confidence we cannot reject the null hypothesis of statistically equal means.

Table A.5 shows the same mean difference significance table for all the activities' distributions.

Table A.5: Mean difference between the two samples

	Mean diff.	
TV	0.190*	(2.47)
Pc games	0.283	(1.92)
Internet	0.087	(1.19)
Listen to music	-0.006	(-0.11)
Play music or sing	0.064	(0.46)
Sport	0.047	(0.48)
Dance/act	-0.142	(-1.35)
Tech activities	-0.083	(-0.75)
Read	0.177	(1.37)
Relax	-0.010	(-0.09)
Girl/boyfriend	-0.292	(-1.94)
Best friend	-0.101	(-1.12)
Clique	0.004	(0.04)
Youth centre	-0.099	(-1.11)
Volunteer	0.107	(1.00)
Religious	-0.007	(-0.08)
N	506	

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

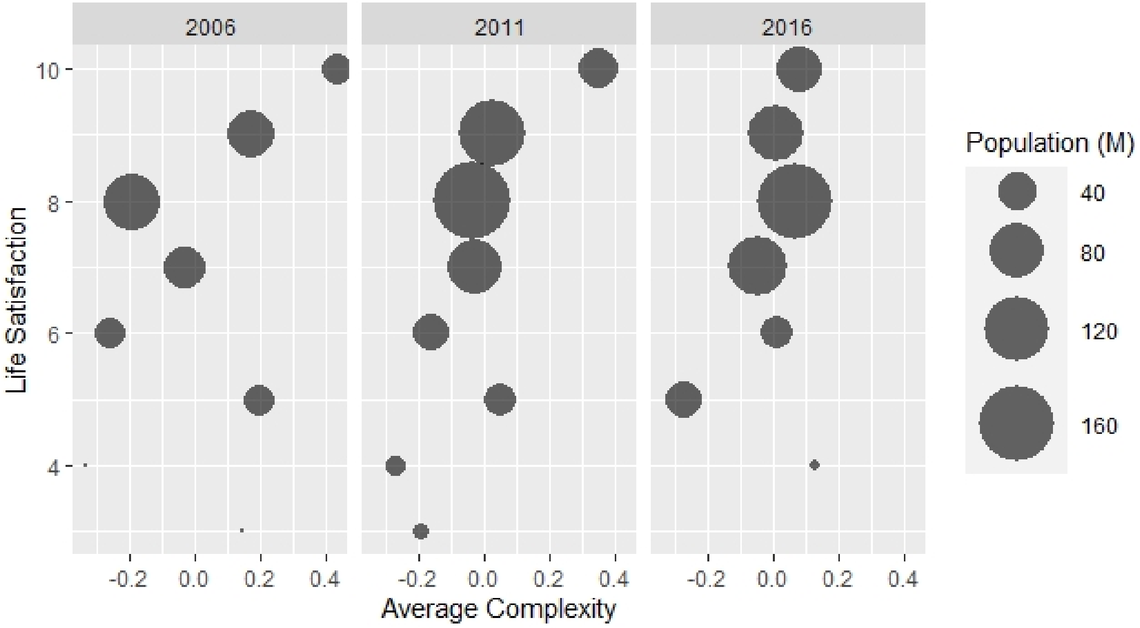


Figure A.1: Summarising the average individual complexity by levels of life satisfaction - Size of each level group

Appendix B

Data and Tables

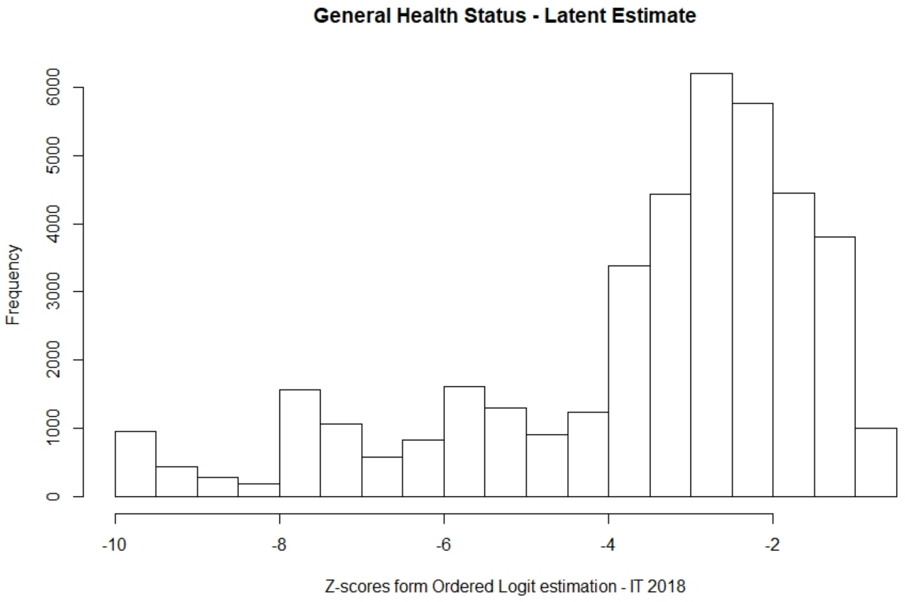


Figure B.1: Latent health estimates. Year 2018.

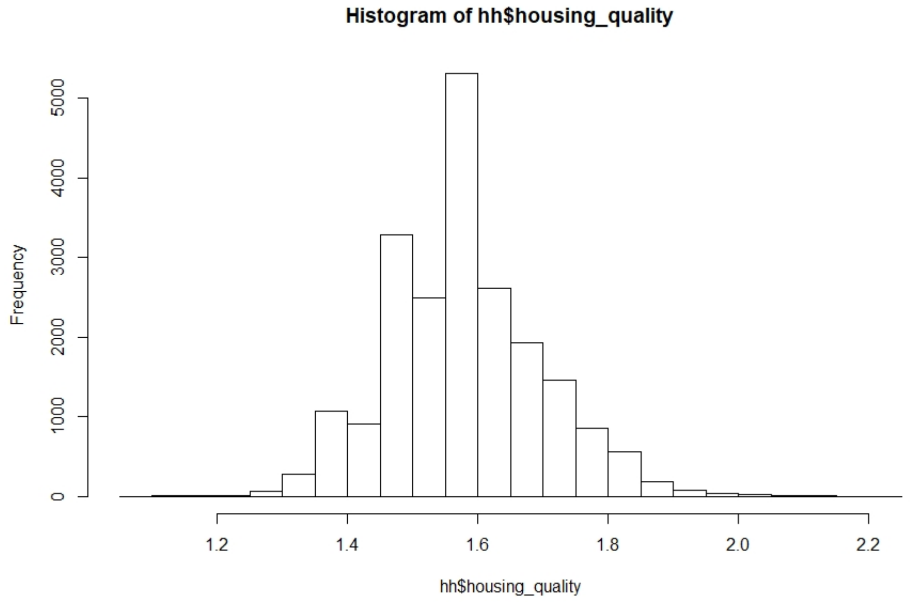


Figure B.2: Housing Quality Index

Table B.1: Average diagonal dependence indices comparisons
$H_0 : \bar{\delta}_m(X_1) = \bar{\delta}_m(X_2)$ *(Confidence levels: *** 99%, ** 95%, * 90%)*

$\bar{\delta}_m(X_1)$	$\bar{\delta}_m(X_2)$	**p_vals**
4 dim - NOJOB	4 dim - NOED	0,055*
4 dim - NOJOB	4 dim - NOHEALTH	0,000***
4 dim - NOJOB	4 dim - NOHOU	0,000***
4 dim - NOJOB	4 dim - NOINC	0,017**
4 dim - NOJOB	ALL 5 DIM	0,005***
4 dim - NOED	4 dim - NOHEALTH	0,000***
4 dim - NOED	4 dim - NOHOU	0,002***
4 dim - NOED	4 dim - NOINC	0,001***
4 dim - NOED	ALL 5 DIM	0,157
4 dim - NOHEALTH	4 dim - NOHOU	0,000***
4 dim - NOHEALTH	4 dim - NOINC	0,000***
4 dim - NOHEALTH	ALL 5 DIM	0,000***
4 dim - NOHOU	4 dim - NOINC	0,000***
4 dim - NOHOU	ALL 5 DIM	0,049**
4 dim - NOINC	ALL 5 DIM	0,000***

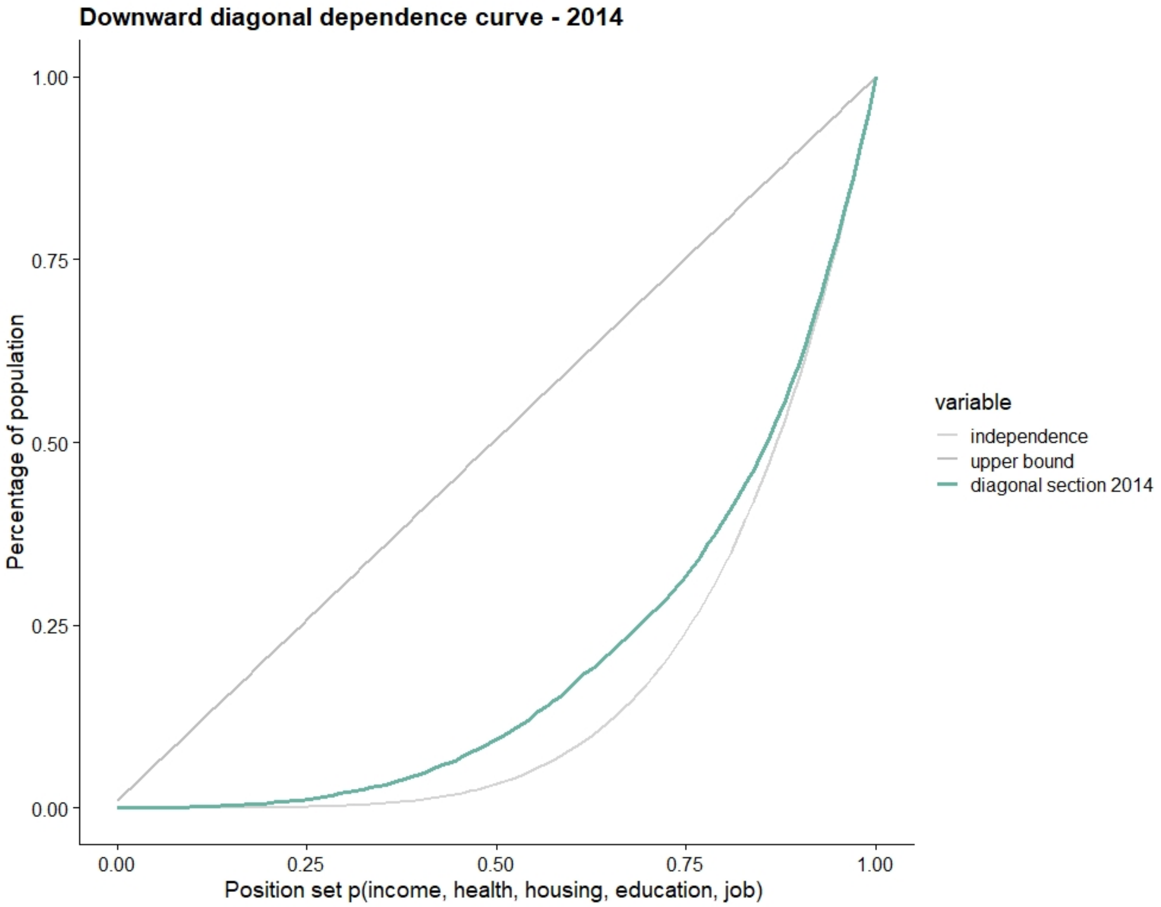


Figure B.3: Zoom of diagonal copula section

Table B.2: Total population: Cumulative deprivation, gender and activity status

cum_dep (0=No, 1= Yes)	tot (thousands)	gender	activity
0	8402,479	Male	Employed
0	2845,540	Male	Self-Employed
0	11,370	Male	Employed
0	1913,883	Male	Unemployed
0	114,474	Male	Retired
0	1110,906	Male	Inactive
0	11,474	Male	NA
0	128,219	Male	NA
0	6904,124	Female	Employed
0	1275,385	Female	Self-Employed
0	10,058	Female	Employed
0	1546,822	Female	Unemployed
0	56,475	Female	Retired
0	4551,212	Female	Inactive
0	22,878	Female	NA
0	100,370	Female	NA
1	80,664	Male	Employed
1	38,830	Male	Self-Employed
1	172,384	Male	Unemployed
1	7,641	Male	Retired
1	58,854	Male	Inactive
1	2,136	Male	NA
1	5,065	Male	NA
1	57,187	Female	Employed
1	15,788	Female	Self-Employed
1	75,590	Female	Unemployed
1	1,804	Female	Retired
1	376,702	Female	Inactive
1	0,921	Female	NA

Appendix C

Data and Tables

Missing categories

	Missing cases
Education	38
Ethinc group	40
Mother activity	83
Father activity	74
Mother skill	192
Father skill	240
Mother education	632
Father education	692
Female	0

Table C.1: Missing values among circumstance variables, net of the item missingness present on purpose

Allostatic Load measure

A detailed illustration of how the Allostatic load measure has been obtained is following from :

> We calculated a composite risk score measure to proxy allostatic load after converting HDL, Albumin and DHEAS to negative values to reflect ill-health rather than good health, and then transforming each of the nine biomarkers into a z-score and summing to produce the composite measure. The index is then scaled so that a 1-unit increase in allostatic load corresponds to an increase of one standard deviation. Higher values of allostatic load indicate worse health.

The exact dimensions included are:

> "We use waist-to-height ratio to measure adiposity and resting heart rate (HR), systolic blood pressure (SBP) and high-density lipoprotein cholesterol (HDL) to measure cardiovascular health. Lung function is measured using a spirometer as forced vital capacity (FVC), the total amount of air forcibly blown out after a full inspiration; higher FVC values indicate better lung functioning. C-reactive protein (CRP) is our inflammatory biomarker, which rises as part of the immune response to infection and is associated with general chronic or systemic inflammation (Emerging Risk Factors Collaboration, 2010). Glycated haemoglobin (HbA1c) is our blood sugar biomarker, and is a validated diagnostic test for diabetes. Albumin is used to proxy liver functioning, with low albumin levels suggesting impaired liver function. We also use dihydroepiandrosterone sulphate (DHEAS), a steroid hormone in the body, representing one of the primary mechanisms through which psychosocial stressors may affect health, with low levels associated with cardiovascular risk and all-cause mortality (Ohlsson et al., 2010)."

Principal Component Analysis and the health-related behaviours

Being the partitioning variables categorical data, the realisation of the principal component analysis,is not immediate. Indeed, in order to compute the eigenvalues of the data variance-covariance matrix, the variables are required to be numeric and to have a meaningful and equal distance between each category. Thus, the PCA has been conducted after computing the Polychoric transformation of the mixed data to obtain a meaningful covariance matrix (R polycor package / function: *hector*).

The health-related behavioural variables used are:[1]

- Habits on drinking milk: whole milk, semi-skimmed milk, skimmed milk, soya milk, any other sort of milk, don't use milk. (`usdairy`)
- Habits on bread eating: white bread only, wholemeal bread, granary - wholegrain - brown bread, no bread, other type of bread. (`usbread`)
- Eating fruit and vegetables: days per week eating a portion of fruit or vegetable. (`wkfruit, wkvege`)

[1] The original variable name from the dataset is shown within the parenthesis. The variables used for the analysis may have been modified with respect to the original.

- Smoking habits: More than 20 cig. per day, 9 to 20 cig. per day, less than 9 cig: per day.(`smoking`)
- Ex-smoker: dummy variable. (`smoking`)
- Sport activity: sport intensity in an ordinary week + self-assessed sport attitude. (`sportact, sportsfreq`)
- Walking habits: number of days walked at least 10 minutes in a month. (`daywlk`)
- Drinking alcohol habits: drinking alcohol everyday. (`scfalcdrnk`)

Estimated model (R function: *prcomp()*). Model fit measure on how well does the factor model fit the given correlation matrix: $= 0.720$. The resulting first component of the PCA accounts for almost the 40% of the total variability of the data. Given its positive correlation with the "bad behaviours", the upcoming variable has been multiplied by (-1) in order to have a meaning of "good health-related behaviour". The final outcome variable is a measure of effort in having a healthy lifestyle. Table 4.2 shows the correlations between the first component of the PCA output and the behavioural variables involved in the analysis.

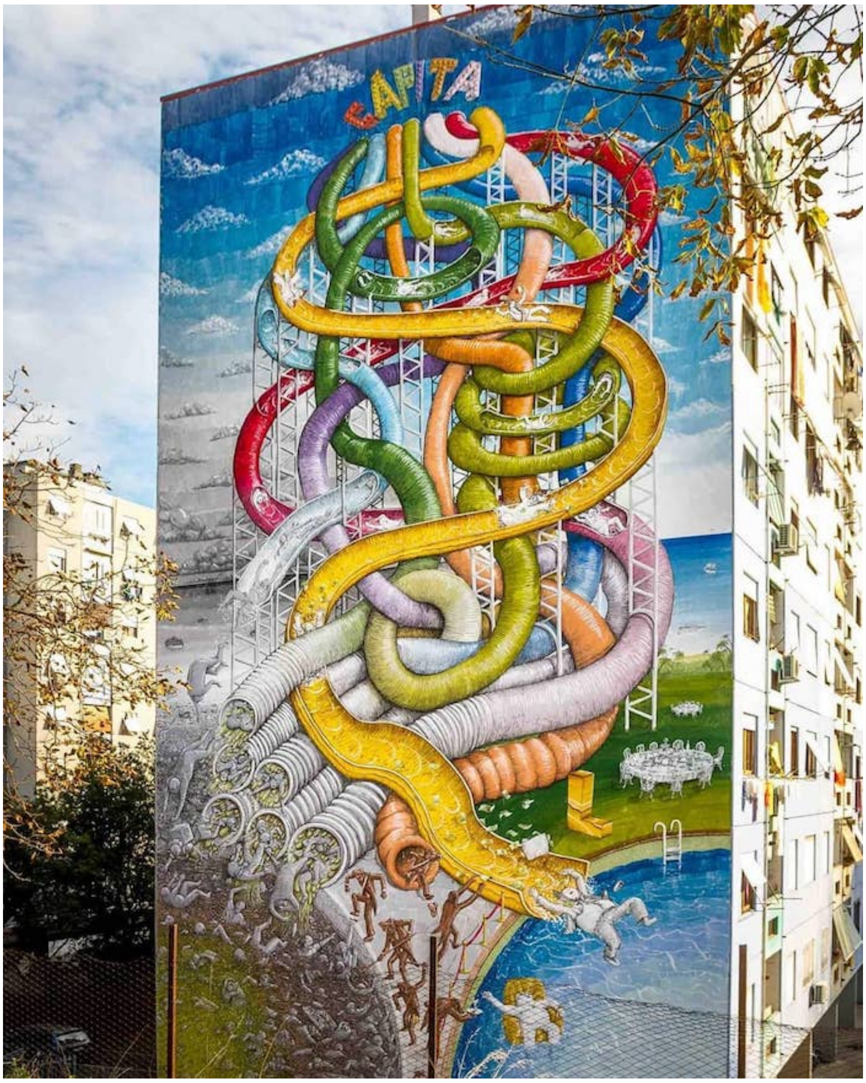
Acknowledgements

I would like to thank Prof. Paolo Brunori for his constant help and support during my whole doctoral study. He has been a mentor not only for my research experience but also for this period of life. I am grateful to Prof. Koen Decancq for his precious guidance to my research work during the visiting at the CSB in the University of Antwerp. My gratitude is also to Giuliano Resce for his useful suggestions, and to Apostolos Davillas and Andrew M. Jones for the technical support to the use of the UKHLS database.

Thanks to all the PhD students and researchers at the CSB and in the University of Antwerp for having welcomed me during my visiting period. A warm thank goes to all my PhD colleagues in the University of Siena, we have been companions of unforgettable adventures (and dis-adventures) during this PhD program and in the everyday life in Siena. I would also like to thank Francesca Fabbri, the PhD secretary, and all the other DEPS personnel. A special thank goes to those wonderful friends who read this thesis helping me with feedback and correction suggestions, Claudia, Alexandra and Margherita. Finally, I am profoundly grateful to my beloved Alessandro, my parents Antonella and Giuseppe, my sisters Letizia and Elisabetta, my grandma, and those old and new friends who believed in my choices and warmly supported me, always above any expectation, during those years.

Rome, March 2021

Giovanna Scarchilli



"It happens", Blu, 2019
Casal de' pazzi, Rome

This picture shows a coloured street art painting containing a dystopian description of the world looking like an amusement park in which, everyone plays on a kids' slide and randomly ends up in one of three possible outcomes:
the rich, the poor, the guardians.

This piece of street art has been incredibly inspiring to me and to my research. Being a metaphor of an unjustly unequal society, it is placed in a peripheral neighbourhood of Rome where is settled Rebibbia, one of the mostly crowded jails of the whole country.