Working Paper Series, N. 7, May 2020

Matching on poset-based Average Rank for Multiple Treatments (MARMoT) to compare many unbalanced groups

Margherita Silan
Department of Statistical Sciences
University of Padua
Italy

Giovanna Boccuzzo
Department of Statistical Sciences
University of Padua
Italy

Bruno Arpino
Department of Statistics, Computer Science, Applications
University of Florence
Italy

Abstract: In this article, we propose an original matching procedure for multiple treatment frameworks based on partially ordered set theory (poset). In our proposal, called Matching on poset-based Average Rank for Multiple Treatments (MARMoT), poset theory is used to summarize individuals' confounders and the relative average rank is used to balance confounders and match individuals in different treatment groups. This approach proves particularly useful for balancing confounders, even in situations in which the number of treatments considered is high. We apply our approach to the estimation of neighbourhood effect on the fractures among older people in Turin (a city in northern Italy).

Keywords: Matching, Multi-treatment, Neighbourhood effect, Poset.

Final version (2020-02-27)

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Margherita Silan
silan@stat.unipd.it

Matching on poset-based Average Rank for Multiple Treatments (MARMoT) to compare many unbalanced groups

Margherita Silan
Department of Statistical Sciences
University of Padua
Italy

Giovanna Boccuzzo
Department of Statistical Sciences
University of Padua
Italy

Bruno Arpino
Department of Statistics, Computer Science, Applications
University of Florence
Italy

Abstract: In this article, we propose an original matching procedure for multiple treatment frameworks based on partially ordered set theory (poset). In our proposal, called Matching on poset-based Average Rank for Multiple Treatments (MARMoT), poset theory is used to summarize individuals' confounders and the relative average rank is used to balance confounders and match individuals in different treatment groups. This approach proves particularly useful for balancing confounders, even in situations in which the number of treatments considered is high. We apply our approach to the estimation of neighbourhood effect on the fractures among older people in Turin (a city in northern Italy).

Keywords: Matching, Multi-treatment, Neighbourhood effect, Poset.

1 Introduction

The ideal situation in order to make inference about treatment effect is with randomized trials, however, quite often in many fields, randomization of subjects into different treatment groups is unfeasible (Oakes, 2004). Thus, the use of observational data represents a challenge because of selection bias, that happens when the treated group of subjects differs systematically from the control group, according to covariates that may also affect the occurrence of the outcome. Indeed, the distribution of covariates among the treatment groups may differ considerably, creating what is called an unbalanced situation. Thus, in these cases the crucial question is whether differences with respect to the outcome between treatment groups can be attributed to the treatment itself, rather than to differences between subejts' characteristics in the groups (Austin, 2011). This is why methodological techniques

usually applied to solve this issue are focussed on reaching the balance of covariates' distribution in the treatment groups. A widely used method to balance the distributions of observed characteristics among treatment groups consists in the use of propensity score techniques.

The propensity score is the probability to be treated conditional on a set of independent variables. Thanks to this method, individuals in different treatment groups are matched by their propensity score (Rosenbaum, 1987). The propensity score matching technique enables to adjust for confounders, and contrast only comparable groups of subjects, with similar observed characteristics. Moreover, it presents also some additional advantages. Assumptions about the functional form of the relationship between the covariates and the outcome are not necessary with propensity score matching, so there are few risks of model misspecification. Collinearity among confounders is not a problem (Harding, 2003) because the adjustment for observable confounders is managed separately from the estimation of the treatment effect on the outcome.

Several studies and scientific papers deal with the use of propensity score techniques in presence of a dichotomous treatment, where only two groups of subjects need to be balanced with respect to covariates. The use of propensity score techniques in multiple treatment frameworks is less straightforward. In the literature there are few works that estimate the effect of non-dichotomous treatments (McCaffrey et al., 2013; Lopez and Gutman, 2017; Yoshida and Franklin, 2017; Rose and Normand, 2019), comparing three or four treatment groups. However, using propensity score matching is complicated when the number of treatment groups to consider is huge, because of methodological, interpretative and computational issues. Indeed, several issues raises in dealing with a multivariate treatment with propensity score techniques. For instance, it is more difficult to identify a common support for all the treatment groups, that safeguards the comparisons between them. Also the computation of the propensity score is less trivial, especially the model specification. Moreover, the matching algorithm increases its complexity together with the number of treatment group considered.

The method we propose consists in a multiple matching based on partially ordered sets (poset). We test the Matching on poset-based Average Rank for Multiple Treatments (MARMoT) with simulations, and prove its utility in balancing the observed covariates among groups. Our simulation study takes into account 23 treatment groups and shows more than satisfying results, indeed, the MARMoT approach highly improves the balance of covariates, even starting from strongly unbalanced situation. As far as the empirical application is concerned, we observed MARMoT performance with real data considering 10, 23 and 70 treatment groups. Even with real data the results are satisfactory, especially with 10 and 23 treatment groups, while in the last case there is still room for improvement.

As a case study, we estimate the neighbourhood effect on hospitalized fractures among elderly residents in Turin, a city in the north of Italy. The focus of this paper is on comparison of neighbourhoods with different compositions, indeed individuals with different fracture risk factors may live in different areas. Our matching technique enable us to adjust for confounders using a poset-based average rank in a multiple treatments framework, even when the number of treatment groups (neigh-

bourhoods) is very high. The main methodological contributions of this paper thus consist in the proposal of this new matching approach based on poset theory, its validation in a simulation study and its application to estimate the neighbourhood effect on hospitalized fractures among individuals aged sixty or more based on three different geographical partitions.

In the second section we describe the case study, the data analysed, the three different geographical partitions considered, and the confounders observed. The methods to balance the observed confounders in multiple treatments frameworks are described in section 3, followed by a brief introduction to poset theory, and an in-depth explanation of our original methodological proposal. In the fourth section we describe the design structure and the results of a simulation that we performed to test the reliability of our original proposal. Section 5 illustrates the empirical application with real data, comparing different geographical partitions.

2 Case study

During the last 20 years, there has been growing interest in the effects of context on individuals' lives (Arcaya et al., 2016), prompting important new research in social epidemiology. Such effects are usually called "neighbourhood effects" and were defined by Oakes (2004) as the independent causal effects of neighbourhoods on a given health or social outcome. In the literature, the term neighbourhood is often used to delineate individuals' immediate residential environments and the material and social characteristics of these environments that presumably have an impact on personal outcomes (Diez Roux, 2001). Various types of outcomes are considered, such as life course events (Rabe and Taylor, 2010), educational achievement (Leckie, 2009) or health outcomes (Cubbin et al., 2000; Pickett and Pearl, 2001). The last of these are the most often analysed, and concern mental health (Mair et al., 2008; Truong and Ma, 2006), early childhood health (Christian et al., 2015), all-cause mortality (Meijer et al., 2012) and older people's health (Roux et al., 2004; Yen et al., 2009). Risk factors of health attributable to neighbourhood include deprivation, walkability, food environment, air pollution, crime and social cohesion (Arcaya et al., 2016).

Interest in the neighbourhood effect on hospitalized fractures among over-60-year-olds stems from a real need expressed by Turin's Epidemiological Service. Neighbourhoods may affect elderly fracture rates in two main ways: they may be difficult to walk around, or have inadequate street lighting, and thus increase the risk of falls; and/or people living in the area may be discouraged from engaging in physical activity, and their muscle tone and bone structure consequently deteriorate (Ambrose et al., 2013; Barnett et al., 2017; Sànchez-Riera et al., 2010). The focus here is on people over sixty, partly because of their greater exposure of hospitalized fracture, and also because they are assumed to be a more stable resident population. Indeed, some researchers have found older people more susceptible to neighbourhood effects because they spend more time in their neighbourhoods than younger people (Melis et al., 2015; Turrell et al., 2014). Older people are also less likely to move house (the annual rate for the observed population was only around 1%).

2.1 Turin Longitudinal Study

The data used in our analysis come from a longitudinal study conducted in Turin, that gave rise to an integrated database, which combines administrative data flows on residents drawn from censuses and population registry with health data flows (hospital discharge records, prescription charges and exemptions, and territorial drug prescriptions). The hospital discharge records contain information on the patient's diagnosis, admission modality (emergency, compulsory, voluntary), and dates of admission and discharge. The prescription charges database lists all exemptions from payment of health services to which some patients are entitled due to chronic conditions or low income. The territorial drug prescriptions database contains details of prescribed drugs, the quantities involved, and their classification (based on their therapeutic, pharmacological and chemical properties). The census data includes not only basic demographic details, such as age, sex, and place of birth, but also some important information about individuals' socio-economic status, such as their occupation, education, home ownership, and family composition.

All these different data sources have been pooled together over time. Starting with the censuses and population registries available in 1971, Turin's residents have been registered and tracked as a historical migration dataset, considering all movements of individuals living in Turin for at least one day from 1971 onwards (Costa et al., 2017). Several other data sources were added over time, such as the cause of death archives in 1971, the cancer registry in 1985, the hospital discharge records in 1995, drug prescriptions data in 1997, and so on.

2.2 Examined population

The study population consists of all individuals considered in the 2001 population census, aged 60 or more as at 31^{st} December 2001. In order to be able to collect information on possible confounders represented by past health-related information, we focus on individuals living in Turin between 1^{st} January 1997 and 31^{st} December 2001. We measure the outcome, i.e. hospitalized fractures, during the year following the census (i.e. 2002). We therefore limit our analyses to individuals who lived in Turin throughout the year 2002. Our study design enables us to measure the time-varying confounders prior to the treatment, which is measured before the outcome is observed. In this application, we focus on assessing the differences in the proportion of individuals experiencing at least 1 hospitalized fracture in 2002 among populations living in different neighbourhoods at the time of the 2001 census.

2.3 Neighbourhoods

The city of Turin can be split into 10 districts, 23 areas, or 94 zones, considering neighbourhoods that might affect health (Arcaya et al., 2016). The three partitions may relate to different living conditions (deprivation, walkability, crime, and social cohesion) and population characteristics, but the three geographical layers are only partially hierarchical. For instance, the same zone may belong to two or more areas, or districts.

Table 1: Distribution of the population by geographical partition.

Partition	Minimum	1*st* Q.	Median	Mean	3*rd* Q.	Maximum
10 Districts	10608	18777	21897	22583	29107	33072
23 Areas	3584	7976	9609	9819	12606	18089
94 Zones	3	625	1870	2402	3876	7758

Table 1 shows some summary statistics of the sizes of the populations in each geographical partition. The ten districts have an average population of 22,583, with the least populated accounting for 10,608 individuals, and the most populated for 33,072. The populations of the areas range between 3,584 and 18,089, with a mean area population of 9,819. The number of individuals living in each zone varies even more.

In our empirical analysis, we compare proportions of hospitalized fractures among neighbourhoods considering the three geographical partitions. In the case of the 94 zones, however, we needed to reduce the neighbourhoods considered because some of them were too small, as shown in the last row of Table 1. We therefore excluded zones with a population of less than 625 (corresponding to the first quartile of the distribution of zone populations). The number of individuals living in the zones thus discarded accounts for only 3% of the whole population, and the final number of zones considered is 70. For the sake of brevity, in the simulations we focus on the intermediate partition, i.e. the city divided into 23 areas.

2.4 Variables

Based on the literature on neighbourhood effects on older people's health (Roux et al., 2004; Yen et al., 2009), we consider the following variables as possible confounders: gender, age (considering five-year age brackets: 60-64, 65-69, 70-74, 75-79, 80 and over), region of birth, family composition, educational attainment, last known occupational condition, and home ownership. The region of birth is coded, distinguishing between individuals born: in Piedmont (the region to which Turin belongs); in other regions of northern Italy; central Italy; southern Italy or islands; or outside Italy. The variable representing family composition combines marital status with the number of components: living alone; married and living only with partner (family of two); unmarried and not living alone (family of two or more); married and living in a family of more than two people. The last known occupational situation is a variable obtained from the census data from 1971 to 2001, and aims to capture the last type of occupation prior to retirement. This was not possible for some individuals because they were already retired in 1971 (or in all the censuses concerning them), or they were not working for other reasons. The occupation variable distinguishes between the above-mentioned case and home-makers, entrepreneurs, white-collar workers, and manual workers.

The percentage of hospitalized fractures in 2002 is quite low, at 0.9% of Turin residents over 60 years old, with some differences between neighbourhoods. The percentages of the outcome considered vary between 0.67% and 1.18% among the

different areas.

3 Methods

Propensity score techniques are used to approximate a randomized trial with observational data. Unlike other techniques in which the analyst models the outcome given all measured confounders and treatments, the propensity score approach focuses on modelling the treatment allocation process (Williamson et al., 2014). However, the treatment allocation model needs to be well specified and should include all confounders.

3.1 Propensity score techniques

Before describing how propensity score methods are used in the multi-treatment case, we consider the simple case of a binary treatment, i.e. a situation with only two neighbourhoods, that we call 0 and 1.

Two fundamental variables are associated with each individual: a binary variable T that represents the dichotomous treatment and takes a value of 1 if individuals receive treatment 1 (lives in neighbourhood 1), or 0 if they receive treatment 0; and the outcome variable Y. Each individual i also has a pair of possible outcomes, i.e. Y_{0i} and Y_{1i}, which are respectively the outcomes under the treatments $T = 0$ and $T = 1$. Each individual receives only one of the treatments (the control treatment or the active treatment) (Austin, 2011). The effect of living in neighbourhood 1 for the individual i is $\tau_i = Y_{1i} - Y_{0i}$, i.e. the difference between the outcome for individual i who lives in neighbourhood 1 and the outcome for the same individual if he/she were living in neighbourhood 0 (Holland, 1986).

In practice, it is impossible for the same individual to live in two different neighbourhoods at the same time, so we can only observe one potential outcome, which corresponds to the allocated treatment for each individual. This is called the "fundamental problem of causal inference".

In order to use propensity score methods, some assumptions to estimate a causal effect are needed, as regards: *temporality* (the selected treatment T must occur before the outcome); and the *strong ignorability*, which is composed of two assumptions, unconfoundedness and positivity; and the *stable unit treatment value assumption (SUTVA)*. Based on the assumption of unconfoundedness, the potential outcomes (Y_1, Y_0) are independent from the allocated treatment (T), given a set of observable variables X, which are unaffected by the treatment, $Y_1, Y_0 \perp\!\!\!\perp T|X$. This assumption is also known as "selection on observables" because it amounts to assuming that there are no unmeasured confounders, since all the variables involved in the selection process have been observed, measured, and included in the propensity score computation. The positivity (or overlap) assumption requires that any individual have a positive probability of being included in the treatment or control group, $0 < P(T_i = 1|X_i) < 1$. The SUTVA includes two assumptions: the *no interference*, and the *stable unit treatment value assumption*. According to the SUTVA, the potential outcomes for any unit do not vary with the treatments allocated to other

units, and, for each unit, there are no different forms or versions of treatment level leading to different potential outcomes (Imbens and Rubin, 2015).

However, if the interest of the research is just focused on a controlled descriptive comparison, the only assumption that is fundamental is the *overlap* assumption that reassures about the comparability of groups. In this framework it is possible to define a more general estimand (Li et al., 2013) that we are considering in our analysis. As an adaptation of the estimand proposed by Li et al. (2013), we define the Average Controlled Difference among Groups on the treated (ACDG) as the expected difference in the outcome among the two neighbourhood for those who live in neighbourhood 1:

$$ACDG = E(Y_1 - Y_0|T = 1) \tag{1}$$

In words, this estimand can be interpreted as the difference between the average outcome for those who live in neighbourhood 1 and the average outcome we would observed for these people had lived in neighbourhood 0.

To be able to include all the observable confounders, we may have to deal with a large number of covariates. This problem is called the "curse of dimensionality", and it can be solved by using a so-called "balancing score" (Caliendo and Kopeinig, 2008). A balancing score, $b(X)$, is a function of the observed covariates X such that the conditional distribution of X, given $b(X)$, is the same for treated ($T = 1$) and control ($T = 0$) units; in other words, $X \perp\!\!\!\perp T|b(X)$ (Rosenbaum and Rubin, 1983). Rosenbaum and Rubin (1983) demonstrated that the propensity score e_i, the probability that each individual has to receive the treatment, $e_i = P(T_i = 1|X_i)$, is the coarsest balancing score. Propensity scores are generally estimated using parametric models such as logistic regression. If these models are misspecified, the balance of covariates may not be satisfactory. That is why several different methods for estimating the propensity score have recently been implemented and compared (Setoguchi et al., 2008; Li et al., 2013), including some CART-based methods (Lee et al., 2010) (for instance pruned, bagged and boosted (McCaffrey et al., 2004)), neural networks and random forests. Using data mining techniques in this field has been shown to achieve a better balance and a lower bias of causal estimators based on propensity scores. Indeed, these flexible data-driven algorithms also allow researchers to fit complex relations, overcoming variable selection and model building processes automatically (Cannas and Arpino, 2018). If confounders and treatments have non-linear or non-additive relations, machine learning techniques are able to gather and handle them automatically in the estimation process. Although these techniques provide models that are difficult to interpret, they are an important resource for estimating propensity scores because the interpretation aspect is not fundamental at this step in the analysis, and interest focuses mainly on the balance that can be reached with propensity score adjustments.

The Absolute Standardized Bias (ABS) measure is usually employed to measure the balance of each confounder X between treatment groups:

$$ASB = \frac{|\bar{X}_0 - \bar{X}_1|}{\sqrt{\frac{S_0^2}{2} + \frac{S_1^2}{2}}} \tag{2}$$

where $\bar{X}_0$ and $\bar{X}_1$ are the means of the variable X of individuals living respectively in neighbourhoods $T = 0$ and $T = 1$; and S_0 and S_1 are the standard deviations of the variable X for individuals living in neighbourhoods $T = 0$ and $T = 1$, respectively.

Propensity scores may be involved in the balancing procedures in four main ways (Austin, 2011): matching, stratification, covariate adjustment, and inverse probability of treatment weighting. According to the propensity score matching approach, treated and untreated individuals with the same propensity scores are matched and their outcomes are compared. The above-described theoretical framework focuses on a binary treatment, while propensity score matching (PSM) is less straightforward to implement if the number of treatments increases, as explained in the following section.

3.2 Propensity score techniques in a multi-treatment framework

The set of multiple treatments can be represented by a series of dummies, $D_{it}(T_i)$ (Linden et al., 2016), where T_i is a categorical treatment variable that takes values from 1 to K:

$$D_{it}(T_i) = \begin{cases} 1 & if \quad T_i = t \\ 0 & otherwise. \end{cases} \quad for \quad t = 1, \ldots, K \tag{3}$$

We will consequently have a set of potential outcomes, $\mathbf{Y} = (Y_{1i}, ..., Y_{Ki})$ for individual i, considering all different treatments, and only one of them is observed.

In a multi-treatment framework, the definition of the Average Controlled Difference among Groups for each treatment t will be

$$ACDG_{t,t^c} = E[Y_t - Y_{t^c}|T = t]. \tag{4}$$

This estimand compares every treatment group (neighbourhood) t with the rest of the population t^c (the rest of the city). However other comparisons maybe more meaningful, indeed in other cases, the most informative comparison may be between two neighbourhoods, or between each neighbourhood and a common reference (e.g. the neighbourhood with lowest rate of hospitalized fractures). These other estimands can be obtained with minimal variations to definition 4.

In order to perform a controlled descriptive comparison, just the *overlap* assumption is needed in the multi-treatment case, as in the dichotomous treatment framework, under the circumstance that there are K treatments, and not just two (Lopez and Gutman, 2017).

Finally, for the measure of balance or ASB, there is more than one possible expression available in the literature, depending on the treatment comparisons of interest. In this work, we define the ASB as

$$ASB = \frac{|\bar{X}_t - \bar{X}|}{\sqrt{\frac{S_t^2}{2} + \frac{S^2}{2}}} \tag{5}$$

where $\bar{X}$ and $\bar{X}_t$ are the means of the variable X of individuals living respectively in the whole city, and in the neighbourhoods t; and S and S_t are the standard deviations of the variable X vis-á-vis individuals living respectively in the whole city, and in the neighbourhood t.

In a multi-treatment framework, the propensity score also needs a different specification. Imbens (2000) proposed a modified definition of the propensity score. The generalized propensity score (GPS) is the conditional probability of receiving a particular level of the treatment, given the pretreatment variables. Generalized propensity score applications remain largely scattered in the literature, however, with few applications in regimes involving three (or four) treatments (Lopez and Gutman, 2017). Some of these involve binomial comparisons (Lechner, 2001, 2002) that may pose problems in terms of common overlap and computational effort when the number of treatments increases. Other attempts have focused on forming triplets to compare subjects in a three-treatment framework using matching algorithms (Hade, 2012; Rassen et al., 2011), or larger numbers with vector matching (Lopez and Gutman, 2017). The application of IPTW approaches has been explored by combining different techniques (McCaffrey et al., 2013; Linden and Yarnold, 2016). Other methods that have been tested and compared (Linden et al., 2016) include: regression adjustment (Spreeuwenberg et al., 2010); marginal mean weighting through stratification (Hong, 2010, 2012); and doubly robust methods like the Inverse Probability of Treatment Weighting (IPTW) regression adjustment (Uysal, 2015).

None of these methods are practical, however, if the number of treatments greatly increases. Some important assumptions (such as the overlap) become difficult to satisfy, and estimating the propensity score becomes computationally demanding. The most common model for estimating a GPS is the multinomial logistic regression (Lopez and Gutman, 2017): using this model, K propensity scores e_{it} with $t = 1, ...K$ are estimated, one for each treatment, and they sum to 1. The dependent variable of such a model in a framework with many treatments is therefore categorical with many levels. The result of such a model in a multi-treatment framework would be an estimation of many small probabilities, with small differences between them (generally speaking, with 23 treatments we would expect a mean of the predicted values of around 0.04 for each individual).

An alternative approach, to solve the curse of dimensionality without needing to estimate the probability of receiving each treatment, is template matching. This method can handle the balance of many treatments, and it has been used to compare the performance of hospitals, for instance, reducing the bias due to their different case-mix of patients (Silber et al., 2014). Taking this approach, a sample of individuals represented in all the treatment groups is selected so as to make the individuals in all the treatment groups included in the analysis comparable. This sample becomes the template. Then the matching algorithm matches individuals from all treatment groups with the template, and all other individuals are discarded. The analysis is thus restricted to individuals belonging to the common support of covariates across all the treatment groups. The matching procedure remains similar to the binary case, focusing only on the template and its selected variables. The final dataset will comprise a sample of individuals for each treatment group that resembles the template as much as possible. This simplification enables a huge number of treatments to be managed, but limits the analysis to the individuals comprising the template, and to the choice of template. This means that the target population experiencing the estimated effects may differ considerably from the whole sample population, even though it will be relevant with respect to the chosen template.

We propose an original alternative approach to deal with covariate balance when comparing many treatments. Our method involves matching on a score (average rank) obtained using partially ordered sets (poset) theory. This approach, that we label MARMoT (Matching on poset-based Average Rank for Multiple Treatments) allows us to make the distribution of confounders similar across many treatments.

3.3 Matching on poset-based Average Rank for Multiple Treatments (MARMoT)

3.3.1 Introduction to poset theory

A partially ordered set (poset) is, in mathematics, a set of elements where a binary relation that indicates an order can be traced, the word "partially" refers to the fact that not every pair of elements needs to be comparable. Poset theory is a theoretical field between graph theory and discrete mathematics that quickly developed after the 1970s thanks to technological advances that made greater computational efforts manageable (Brüggemann and Patil, 2011). The main concepts needed to understand why this method is useful to overcome the curse of dimensionality without using a parametric model or introducing some subjective criteria are explained with a toy example.

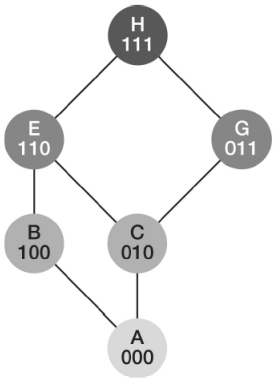
When dealing with a population, the people comprising it can be ranked and ordered using a single variable: level of education, for instance, enables two different individuals to be arranged in an order. From the mathematical standpoint, an order is a binary relation between the elements in a set that respects specific properties. Let P be a set, an order on P is a relation $(\leq)$ between two elements in the set P such that, for all $x, y, z \in P$, the following properties hold:

- Reflexivity: $x \leq x$
- Antisymmetry: $x \leq y$ and $x \geq y$ implies $x = y$
- Transitivity: $x \leq y$ and $y \leq z$ implies $x \leq z$.

A set equipped with such a relation is said to be ordered. If the comparison is drawn using several variables, it may be that some elements are neither equal nor ordered, in which case they are defined as incomparable (Davey and Priestley, 2002). The word "partially" is added to "ordered set" when some of its elements are incomparable, so the order relation has to be changed to a partial order relation, which takes the incomparability (indicated with $||$) of the elements into account:

Incomparability: $x||y \leftrightarrow x \not\leq y \quad and \quad y \not\leq x, \quad x, y \in P$.

Comparing the individuals in a population gives rise to a list of comparabilities and incomparabilities, which can be represented in a graphic form called a Hasse diagram. This diagram represents the elements in a poset: each node is an element, two or more equal elements still form one node, and every line segment is an order relation between comparable objects. Let us suppose that we have a population comprising six individuals characterized by three dichotomous variables, as represented in Table 2: age (which takes a value of 0 for individuals who are between 60 and 70 years old, and 1 if they are older); education (which takes a value of 0 if they



(a)

H	H	H	H	H
E	E	E	G	G
G	G	B	E	E
C	B	G	B	C
B	C	C	C	B
A	A	A	A	A

(b)

H =	6
E =	4.6
G =	4.2
B =	2.8
C =	2.4
A =	1

(c)

Figure 1: A poset and its linear extensions: part (a) represents the Hasse diagram for the individuals in Table 2; part (b) lists all the linear extensions for these individuals; and part (c) their exact average rank.

have a higher education, and 1 otherwise); and homeowner (which takes a value of 0 if they own the house in which they live, and 1 otherwise). The set of observed characteristics of each individual is called "profile". These variables are ordered according to the risk of experiencing the outcome, where a value of 1 corresponds to the highest risk of hospitalized fracture.

Table 2: Toy example for a group of observations.

Subject	Age	Education	Homeowner
A	0	0	0
B	1	0	0
C	0	1	0
E	1	1	0
G	0	1	1
H	1	1	1

In this example, for the sake of simplicity, we included only dichotomous variables, but categorical and discrete variables may be also considered in a poset. However, in order to contain the entropy of the poset, it is recommended to reduce each discrete variable in few meaningful classes.

A Hasse diagram can be used to visualize the order relations between the elements in a poset, and it is based entirely on the order of the elements, disregarding any quantitative information.

In Figure 1(a), the six individuals are represented by their profile in the Hasse diagram, where each node stands for a profile. When two individuals are comparable, they are connected by line segments in the diagram, like A and B or B and E, whereas there is no ascending or descending path between incomparable elements, like B and C.

The list of all the ranks that each individual may occupy is shown in part (b) of Figure 1, where all the linear extensions of the poset are listed. Linear extensions are all the possible rankings of elements in the poset that respect its comparabilities (the connections in the Hasse diagram) and incomparabilities (Brüggemann and Patil, 2011; Davey and Priestley, 2002). The average rank (AR) of a node represents the mean of all the ranks that the element occupies in all possible linear extensions, starting from the known order relations, as listed in Figure 1 part (c).

The AR is a single value for each element in the set that describes the relative position of a given element with respect to the rest of the population. It can be normalized in the interval [0;1].

AR's involvement in the MARMoT approach is just as a balancing tool: its purpose is to reduce data dimensionality and balance on observable individuals' characteristics. There is no need in finding a substantial interpretation to AR values, for our purposes.

3.3.2 Approximating the average rank

If the number of individuals and variables increases, the linear extensions become too many to be examined thoroughly, and it becomes computationally almost impossible to find the exact AR as in the example in Table 2. That said, satisfactory approximations of the number of linear extensions of a poset can be found in works by Dyer et al. (1991), and De Loof (2009).

Researchers have used two main approaches to obtain a computationally efficient calculation of the AR, by sampling linear extensions (Fattore, 2016; Lerche and Sorensen, 2003), or defining an approximation formula. Different approximation formulas have been proposed in the literature, such as the Local Partial Order Model (Brüggemann and Carlsen, 2011), or the one based on Mutual Probabilities (De Loof, 2009). The present work is based on De Loof's approach (2009) because it provides better results than other methods in terms of accuracy with a large sample size (De Loof et al., 2011).

Two concepts help us to understand this approximation, for a sample P with $|P|$ elements:

The rank probability $P(rank(x) = i)$ is the fraction of linear extensions in which an element's rank equals i, where i assumes the value of all possible ranks in the sample of size $|P|$, so $i = 1, \ldots, |P|$.

The mutual rank probability $P(x > y)$ of two elements $x, y \in P$ is the fraction of linear extensions in which the element x is ranked higher than element y.

Now we can establish a relation between the last-mentioned two concepts and the real AR of elements x, $\bar{h}(x)$, starting from a sample P with $|P|$ elements, including x and y:

$$\bar{h}(x) = \sum_{i=1}^{|P|} i \cdot P(rank(x) = i) = 1 + \sum_{y=1}^{|P|} P(x > y). \tag{6}$$

In other words, the first part of formula 6 describes the real AR value, $\bar{h}(x)$, as the expected value, multiplying each possible rank value i by the fraction of linear

extensions in which the element's rank equals i. The second part of formula 6 expresses the real AR value as the sum of all the mutual rank probabilities that involve the element x. Starting from this formula, we need to find an approximation for the mutual rank probability. To do so, we have to define three subsets of the poset P, given a generic element $x \in P$:

Downset: $O(x) = \{y \in P : y \leq x\}$;

Upset: $F(x) = \{y \in P : y \geq x\}$;

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

If $y \in O(x)$, then $P(rank(x) > rank(y))$ equals 1, and if $y \in F(x)$, then $P(rank(x) > rank(y))$ equals 0, so the mutual rank probabilities only need to be approximated with respect to the reciprocal ranks of the incomparable elements. The following approximation was proposed by Brüggemann et al. (2004)

$$P^*(x > y) = \frac{[o(x) + 1][f(y) + 1]}{[o(x) + 1][f(y) + 1] + [o(y) + 1][f(x) + 1]}, \tag{7}$$

where $o(x) = |O(x) \setminus \{x\}|$ and $f(x) = |F(x) \setminus \{x\}|$ are respectively the number of elements in the downset and the upset of x without $\{x\}$. Two more quantities are needed to approximate the AR according to the De Loof (2009) formula, $\tilde{o}(x)$ and $\tilde{f}(x)$:

$$\tilde{o}(x) = o(x) + \sum_{y \in U(x)} P^*(x > y) \quad and \tag{8}$$

$$\tilde{f}(x) = f(x) + \sum_{y \in U(x)} P^*(x < y), \tag{9}$$

and the AR approximation proposed by De Loof (2009) is

$$AR(x) = o(x) + 1 + \sum_{y \in U(x)} \frac{[\tilde{o}(x) + 1][\tilde{f}(y) + 1]}{[\tilde{o}(x) + 1][\tilde{f}(y) + 1] + [\tilde{o}(y) + 1][\tilde{f}(x) + 1]}. \tag{10}$$

That is to say that using formula 10, the AR of x is given by the number of elements in its downset and the sum of probabilities of being a part of x's downset for all incomparable elements with respect to x, using the approximation of the *mutual rank probabilities*. Following the toy example in Table 2, the steps needed to approximate the AR with the De Loof (2009) approach are solved in Table 3, including the estimation of the AR.

In the present work, the approximated AR was computed using the R software, with an R function proposed by Caperna (2019, 2016) that can cope with large datasets (Boccuzzo and Caperna, 2017; Caperna and Boccuzzo, 2018).

3.4 The Matching

We use our MARMoT technique to address the so-called curse of dimensionality, the need to summarize confounders, applying a poset-based AR of the individuals.

Table 3: A numerical example of the approximation of the average rank according to De LoofDe Loof (2009) approach.

x	$o(x)$	$f(x)$	U(x)	$Pr^*(x > y)$						$\tilde{o}(x)$	$\tilde{f}(x)$	AR(x)
				$y = A$	$y = B$	$y = C$	$y = E$	$y = G$	$y = H$			
A	0	5	0	.	0.20	0.25	0.08	0.10	0.03	0.00	5.00	1.00
B	1	2	C, G	0.80	.	0.57	0.25	0.31	0.10	1.88	3.12	2.90
C	1	3	B	0.75	0.43	.	0.20	0.25	0.08	1.43	3.57	2.43
E	3	1	G	0.92	0.75	0.80	.	0.57	0.25	3.57	1.43	4.57
G	2	1	B, E	0.90	0.69	0.75	0.43	.	0.20	3.12	1.88	4.10
H	5	0	0	0.97	0.90	0.92	0.75	0.80	.	5.00	0.00	6.00

The individuals' characteristics are summarized by unique numbers, and individuals who have a similar AR have comparable profiles. AR enables us to proceed with a matching whereby each individual in a given neighbourhood is allocated an individual with a similar AR in all the other neighbourhoods, and those who cannot be matched are discarded in order to respect the overlap condition and make all neighbourhoods comparable simultaneously.

Once the AR has been computed, the first step is to build a frequency table, as the table 4, where each row corresponds to one observed value of the AR (AR_r, $r = 1, \ldots, R$), and each column represents a treatment group (t, $t = 1, \ldots, K$).

Table 4: An example of the frequency table involved in the matching of the MAR-MoT approach.

AR	t_1	t_2	$\ldots$	t_k	$\ldots$	t_K
AR_1	$f_{1,1}$	$f_{1,2}$	$\ldots$	$f_{1,k}$	$\ldots$	$f_{1,K}$
AR_2	$f_{2,1}$	$f_{2,2}$	$\ldots$	$f_{2,k}$	$\ldots$	$f_{2,K}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
AR_j	$f_{j,1}$	$f_{j,2}$	$\ldots$	$f_{j,k}$	$\ldots$	$f_{j,K}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
AR_J	$f_{J,1}$	$f_{J,2}$	$\ldots$	$f_{J,k}$	$\ldots$	$f_{J,K}$

In order for each value of the AR to be represented equally in all the treatment groups, the desired result would be a table where $f_{r,1} = f_{r,2} = \cdots = f_{r,t} = \cdots = f_{r,K} = f_r$, $\forall t = 1, \ldots, K$ in every row r.

Thus, for every row, we must choose the most appropriate frequency f_r for each AR value to impose in the balanced population. In the artificial final population, the distribution of AR values will be balanced in all the treatments groups so as to balance all confounders too. At the end of the matching procedure, each AR_r value will be present in the balanced population $K * f_r$ times, with f_r individuals in each of the K treatment groups. The value for f_r may be chosen according to different criteria: for example, it may be the maximum, the mean, the median or

the minimum of the frequencies in row r. In this work, we define the reference f_r as

$$f_r = \begin{cases} 1 & if \quad median(f_{r,1}, f_{r,2}, \ldots, f_{r,K}) = 0 \\ median(f_{r,1}, f_{r,2}, \ldots, f_{r,K}) & otherwise. \end{cases} \tag{11}$$

Instead of discarding all the AR values with $median(f_{r,1}, f_{r,2}, \ldots, f_{r,K}) = 0$, we set the minimum value of f_r at 1 in order to have a matched population that includes all the profiles in the real population. The choice of the value for f_r may affect both the final dimension of the balanced dataset, and the performance of the MARMoT method in terms of balance. For instance, if we define f_r as the maximum of the frequencies in row r, the final dimension of the dataset will be more than double the dimension obtained with the previous definition and also the quality of matches will be worse. Indeed, for AR values where the frequency matrix is sparse, individuals are duplicated creating distortion and noise in the final dataset.

Having established the frequency that each value of AR should have in each treatment, the algorithm proceeds in three different ways, depending on the dimensions of $f_{r,t}$ and f_r, for every r and every t:

1. if $f_{r,t} = f_r$: all individuals with AR_r (that is the AR value in row r) in the treatment group t are copied in the final dataset.

2. if $f_{r,t} \neq f_r$ and $f_{r,t} \neq 0$: a random sample of size f_r with replacement is selected from among the individuals with AR_r in the treatment group t, and included in the final dataset.

3. if $f_{r,t} = 0$: a random sample with replacement of size f_r is selected from among the individuals with an AR close enough (with a given tolerance) to AR_r in treatment group t, and included in the final dataset. If there are no individuals with an AR close enough, then all individuals with an AR equal to AR_r have to be discarded.

While points (1) and (2) are just a matter of matching individuals with identical AR values, point (3) is the trickiest, because it involves inexact matching, and possibly excluding some individuals from the final dataset. In this work we define the tolerance interval as $[AR_r - \frac{S_{AR}}{4}; AR_r + \frac{S_{AR}}{4}]$, considering as a caliper the value $\frac{S_{AR}}{4}$, where S_{AR} is the AR's standard deviation, similar to recommendations on caliper setting in the propensity score matching literature (Cochran and Rubin, 1973; Lunt, 2013). Thus, if all frequencies $f_{.,t}$ that correspond to AR values included in the interval $[AR_r - \frac{S_{AR}}{4}; AR_r + \frac{S_{AR}}{4}]$ equal 0, subjects with AR value equal to AR_r will be discarded with respect to all the treatments groups. This criterion ensures that the overlap assumption is respected.

As a final remark, the MARMoT method is strongly influenced by five fundamental aspects:

1. the number of variables considered, which directly affects the number of AR values (the number of rows of table 4);

2. the number of the levels of ordinal and categorical variables and the inclusion of a discrete variable that may increase the entropy of the poset (and the number of rows of table 4);

3. the number of treatments, i.e. the number of columns in table 4;

4. the size of the total population, N, which corresponds to $N = \sum_{r=i}^{R} \sum_{t=i}^{K} f_{r,t}$; and

5. the choice of f_r that affects both the final dimension of the balanced dataset and the quality of matches.

An increase of one of the first three variables without a proportional growth of the population will cause an increase of not exact matching cases with a consequent slight worsening of the balancing. An interesting development would be to test limits of this method changing the first three mentioned dimensions.

Once the MARMoT algorithm has matched the individuals and balanced the confounders, any estimand can be used to calculate the effect of a treatment. In the following paragraphs, we use the ACDG on the treated as the estimand of interest representing the neighbourhood effect.

4 Simulation study

Before using the MARMoT method to estimate the neighbourhood effect on real data, we tested it with some simulations in two different scenarios for allocating individuals to 23 treatments. The R code used for the simulation study is reported in appendix A.

4.1 Simulation design

To keep our simulation close to the real situation of interest, we considered the real population of Turin and the individuals' observed characteristics. Starting from the seven confounders described in the second section, we simulated the treatment allocation according to two different scenarios. Since the computation of the AR depends only on individual variables (which come from the observed population and are not simulated artificially), and not on the treatment, AR values computed directly on the observed data could be used, meaning that they were based exclusively on the real population, not on simulated values.

In the first scenario, the treatment allocation equation is simple and close to the real situation. The treatment is generated through a multinomial logistic model, taking neighbourhood 20 (the one with the lowest crude hospitalized fractures rate) for reference. Thus, for each neighbourhood t, and each individual i, the treatment

equation is

$$
\begin{aligned}
ln\left(\frac{Pr(T_i=t)}{Pr(T_i=20)}\right) &= \beta_0^t + \beta_1^t * Gender_i + \beta_2^t * LowerSecondary_i + \\
&+ \beta_3^t * UpperSecondary_i + \beta_4^t * Age65-69_i + \\
&+ \beta_5^t * Age70-74_i + \beta_6^t * Age75-79_i + \\
&+ \beta_7^t * Age>79_i + \beta_8^t * MarriedCouple(2)_i + \\
&+ \beta_9^t * MarriedCouple(>3)_i + \\
&+ \beta_{10}^t * NoMarriedCouple(>2)_i + \\
&+ \beta_{11}^t * HomeMaker_i + \beta_{12}^t * Entrepreneur_i + \\
&+ \beta_{13}^t * WhiteCollars_i + \beta_{14}^t * Manualworkers_i + \\
&+ \beta_{15}^t * NorthofItaly_i + \beta_{16}^t * CenterofItaly_i + \\
&+ \beta_{17}^t * SouthofItaly_i + \beta_{18}^t * OutsideofItaly_i + \\
&+ \beta_{19}^t * Homeowner_i. \qquad (12)
\end{aligned}
$$

In order to choose values for the coefficients, we estimated a multinomial logistic model on the whole population. The result was a matrix with 23 rows and 20 columns containing all treatments' equations intercepts β_0^t for $t = 1, ..., 19, 21, 22, 23$, and coefficients β_v $v = 0, ...19$ for the other variables in the model. These coefficients were perturbed by adding a random value coming from a uniform distribution between -0.01 and $+0.01$, and rounded up or down to just three decimals.

The second scenario envisages a more complex treatment allocation equation, which includes all the interactions between the seven variables considered. As in the first scenario, the choice of parameters for these treatment allocation equations was based on those estimated by a multinomial logistic model, perturbed by a uniform distribution between -0.1 and $+0.1$, and rounded up or down to just three decimals.

4.2 Results

The main results of the simulations are shown in Table 5, where column T indicates the above-described treatment allocation scenarios (coded as 1 for the linear and additive, and 2 for the one with interactions). The first part of Table 5 shows the results of the simulation as described in the previous section, the differences between the scenarios and the differences in the distribution of the individuals among the neighbourhoods.

We examined the initial balance of the two scenarios in all 1000 simulations using the ASB. Having 23 neighbourhoods and seven variables (for a total of 24 levels), we chose to summarize the information by computing the minimum, the 1^{st} quartile, the mean, the median, the 3^{rd} quartile and the maximum of all the ASB, counting ASB values over 5% and 10% for each iteration. The means of these values among all 1000 simulations before and after the balancing procedure for each scenario are given in Table 6. The balance was much improved in both scenarios after the matching procedure, which fixed even extremely unbalanced situations. After matching, the mean number of ASB over 10% corresponded to one tenth of the number beforehand. The central part of Table 5 shows the means (among the simulations) of the number

Table 5: Results of simulations in the two treatment (T) scenarios: (a) mean percentage of the distribution of individuals among neighbourhoods in 1000 iterations; (b) mean number of ASB higher than 5% before adjusting for each neighbourhood; (c) mean number of ASB higher than 5% after adjusting for each neighbourhood; (d) mean number of ASB higher than 10% before adjusting for each neighbourhood; and (e) mean number of ASB higher than 10% after adjusting for each neighbourhood.

	T	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
		Simulation design																						
(a)	1	3.49	3.98	4.28	2.86	4.31	4.65	3.54	3.93	3.59	5.91	5.70	7.76	7.83	5.47	5.83	5.71	1.92	4.63	4.25	2.24	1.59	2.13	4.39
(a)	2	4.11	3.58	4.03	3.54	4.18	4.84	4.02	4.44	4.08	4.54	6.04	7.44	6.15	5.61	6.35	6.30	1.83	3.82	4.63	2.18	1.93	1.86	4.50
		Balance before and after matching																						
(b)	1	15.4	12.6	15.8	10.2	15.9	11.2	11.2	5.1	9.6	6.3	9.7	11.3	5.7	3.5	14.8	10.1	10.2	7.6	16.8	10.3	10.3	11.5	16.5
(c)	1	4.3	1.5	4.0	0.2	0.8	0.6	0.8	0.0	1.1	0.1	0.1	0.3	0.0	0.0	3.2	1.4	3.7	2.3	4.5	4.5	6.7	8.6	3.8
(d)	1	9.8	8.0	10.1	1.6	5.8	3.7	5.8	0.5	3.2	0.0	5.7	2.8	0.9	1.0	9.3	6.6	6.0	6.2	10.4	8.6	7.9	7.9	10.0
(e)	1	1.2	0.0	2.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	1.6	0.0	0.7	0.8	2.3	5.0	0.5
(b)	2	15.3	13.8	14.8	14.0	16.9	13.8	12.5	7.6	16.8	10.8	9.8	14.9	8.1	7.1	15.9	10.0	13.5	12.6	14.9	10.0	16.4	11.7	16.2
(c)	2	5.1	1.9	4.6	0.1	1.1	1.0	0.5	0.0	1.4	0.0	0.8	0.0	0.0	0.0	2.2	1.2	5.1	3.9	5.1	5.1	5.8	9.2	5.5
(d)	2	12.0	9.0	10.0	7.0	9.1	7.6	6.1	1.4	8.4	2.1	5.1	8.4	1.0	0.3	11.8	6.1	8.1	8.6	11.0	8.1	11.8	8.8	12.8
(e)	2	1.5	0.0	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	2.5	0.5	1.0	1.1	2.0	5.7	0.3

of ASB higher than 5% and 10% before and after adjusting for each neighbourhood. From these results, we can see that our MARMoT method greatly improves the balance of confounders among neighbourhoods: it achieves a five-fold reduction in the number of ABS over 5%, and an almost ten-fold reduction in those over 10%, in both the treatment scenarios.

Table 6: Mean of ASB summary statistics in the first and second scenarios before and after balancing among 1000 simulations.

Scenario	Balance	Min	1^{st} Quartile	Median	3^{rd} Quartile	Max	Mean	Over 5%	Over 10%
First	Before	0.01	1.98	4.43	9.59	63.72	8.01	252	132
	After	0	0.58	1.30	2.59	16.95	2.12	52	15
Second	Before	0.02	2.42	5.63	11.89	68.41	9.10	297	175
	After	0	0.62	1.37	2.71	16.84	2.24	60	18

5 Empirical Results

In this section, we use our MARMoT technique to estimate neighbourhood effects considering 10 districts, 23 smaller areas and 94 zones. As explained in the second section, rather than considering all 94 zones, we selected 70 of them with a sufficient number of individuals (more than 625) to avoid the individuals in the excluded neighbourhoods causing problems in the balancing procedure. The computational time required by MARMoT is acceptable, as the procedure to balance the 10 districts took less than 18 minutes, the one for the 23 areas took 36 minutes, and the balancing of the 70 zones took 116 minutes. Table 7 shows that the MARMoT method substantially reduces the ASB in the three partitions, but slightly less successfully in the case of the 70 zones.

The mean of the ASB computed in the 70 zones decreases from around 10% before the MARMoT adjustment to 5.7% in the matched population. Before matching, the majority of the 70 considered zones had at least half of the computed ASB higher than 5%, while after MARMoT adjustment the number of these zones is halved and the number of zones with half of the computed ASB higher than 10% is null. The percentage of zones with a quarter of the computed ASB higher than 10% decreases from around 63% before the matching, to 21.4% after the MARMoT adjustment.

Table 7: Mean of ASB summary statistics on the empirical study in different geographical partitions before and after balance.

Partitions	Balance	Min	1^{st} Quartile	Median	3^{rd} Quartile	Max	Mean	Over 5%	Over 10%
10	Before	0.012	1.689	3.563	7.587	56.207	7.242	101	46
Districts	After	0	0.192	0.427	0.937	8.948	0.846	5	0
23	Before	0.072	1.934	4.198	9.774	63.763	7.938	248	132
Areas	After	0.003	0.482	1.169	2.330	15.802	1.973	51	11
70	Before	0.008	2.556	5.723	12.287	105.132	10.020	914	522
Zones	After	0.008	1.539	3.523	7.075	55.625	5.725	624	265

Figure 2 plots the mean of the ASB of variables in each neighbourhood, before and after the MARMoT procedure in order to visualize areas that are more difficult to balance and those that were more unbalanced in the initial situation. As a general observation, the two areas that are highly unbalanced and difficult to balance are the city center and a neighbourhood in the south of Turin, called "Mirafiori Sud". The composition of "Mirafiori Sud" is quite different from the others, indeed, in this neighbourhood there is a higher percentage of men, individuals born in the South of Italy, subjects with primary or lower education than in the rest of Turin. Moreover, the most common last occupations are home-makers and labourers.

Considering smaller areas also enabled us to identify neighbourhood effects in a greater geographical detail, even though it was more difficult to balance and it proved necessary to discard an extensive portion of the 70-zone partition (white area in Figure 2) because they are scarcely populated. Indeed, the eastern part of the map (in white) is hilly and essentially very different and scarcely comparable with the rest of Turin, the others are mainly graveyards and factories.

6 Conclusions

The aim of this paper was to develop and evaluate an original approach, based on poset theory, to deal with selection bias in a multiple-treatment framework. The main idea behind our method, that we labelled MARMoT, was to obtain a population in which each poset-based AR value that summarize the combinations of confounders, was equally represented in all the treatment groups. The MARMoT approach proved very useful in balancing for confounders and reducing biases in our estimates. The matching involved is not bound to subjective choices (of the template, for instance), and the computation time required is limited, even in the case of 70 different treatments.

Our method enabled us to estimate the neighbourhood effect on hospitalized fractures involving the elderly, considering different geographical partitions (10 districts, 23 smaller areas, and 70 more circumscribed zones) without any selection bias due to the different composition of the neighbourhoods. Indeed, once Turin residents over 60 years old residing in different Turin's neighbourhoods have a comparable composition with respect to confounders distribution, it is possible to evaluate differences among their distribution of hospitalized fractures. This information will be useful to the Piedmont Region's Epidemiological Service when implementing prevention policies for Turin's population and urban interventions focusing on the neighbourhoods at greatest risk.

The choice of geographical scale is a very important issue in neighbourhood studies, and several authors have suggested considering different scales, and examining neighbourhood effects on outcomes for individuals in more detail, in order to better discern which geographical scale is more relevant to the examined phenomena (Arcaya et al., 2016). The importance of choosing the most meaningful scale for spatial data is illustrated by a serious analytical issue known as the modifiable areal unit problem (MAUP). Using our MARMoT method, neighbourhood effects can be estimated and compared in different geographical partitions, enabling an assessment of

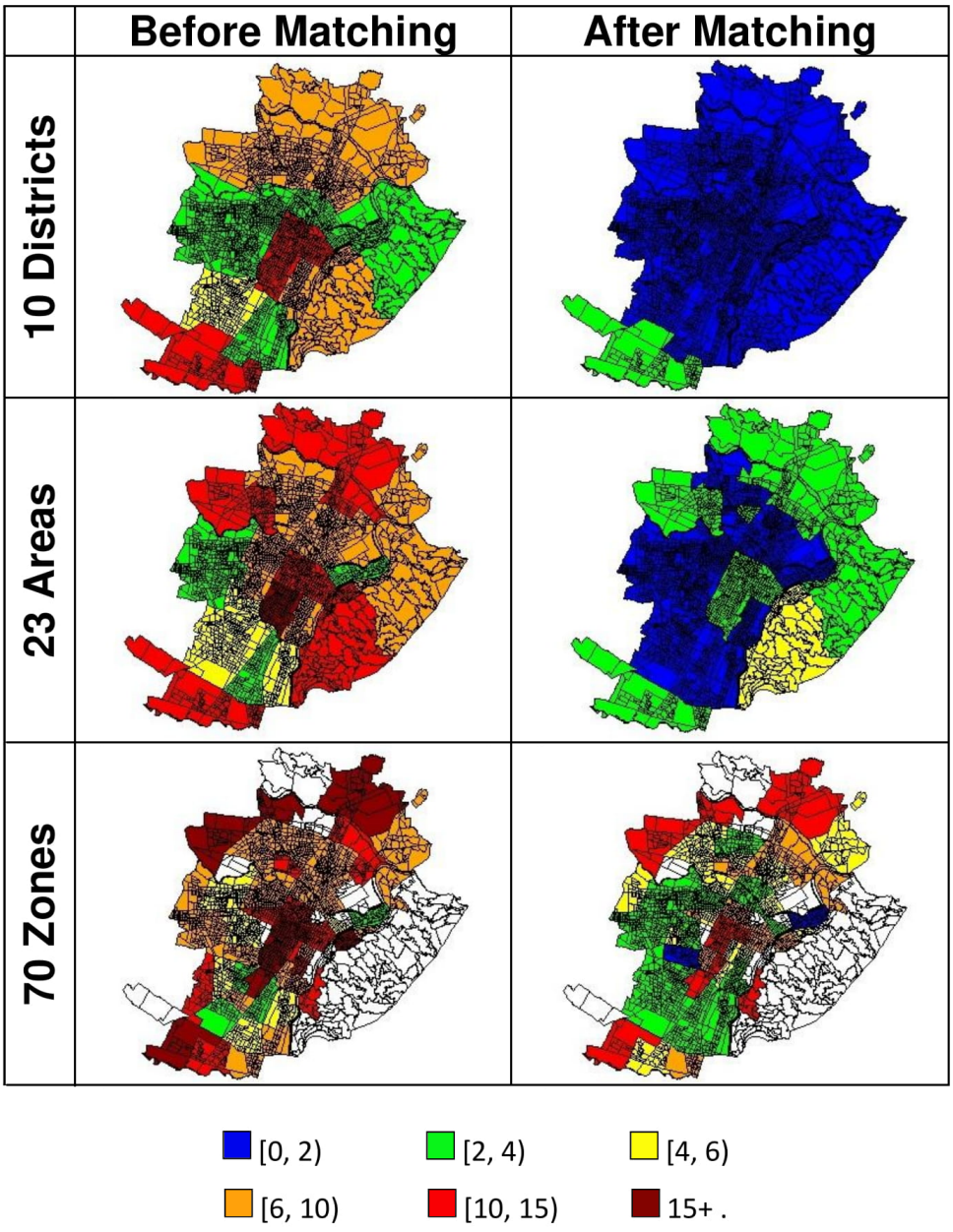


Figure 2: Mean of ASB before and after MARMoT for three geographical partitions: 10 districts, 23 areas and 70 zones.

the sensitivity of neighbourhood effect estimates to different choices of geographical scale.

Moreover, when the considered number of treatments is 70, there is still room for improvement. Several adjustments may be done tuning some choices such as the choice of frequency reference for each AR, the caliper and some additional cleaning of less frequent AR values. Further steps will be taken in these directions to improve this already promising technique.

A R code for the simulation study

```
Required( parsec, dplyr, multiwayvcov, lmtest)
### Computation of the average rank using Caperna (2009) approximation
function implemented in R, where set is a matrix that contains all
observable covariates (columns) for all the considered subjects (rows)
ar<-deloof(set)
###Normalization of the average rank value
ar_norm<-(ar-min(ar))/(max(ar)-min(ar))
data$ar<-ar_norm

### Set the number of iterations
num<-1000

### Need the data matrix that contains all individual characteristics,
personal ids and the computed average rank; and the prob.sim matrix
that contains the computed probabilities for each subject to be
assigned to each one of the 23 treatments (according to one scenario)

### Create matrices to save measure of balance and computational time
ASB_pre<-matrix(data=NA, nrow=num, ncol=8)
ASB_post<-matrix(data=NA, nrow=num, ncol=8)
time_bal<-c()

### Starting with the simulation
for (k in 1:num){
### Simulation of the treatment
ass.treat = t(apply(prob.sim, 1, rmultinom, n = 1, size = 1))
s1 = cbind.data.frame(data, treat_sim = apply(ass.treat, 1,
  function(x) which(x==1)))

### Saving balance measure relative to the population before balance
pre<-table(s1$sex, s1$treat_sim)
pre<-rbind(pre,table(s1$age, s1$treat_sim))
pre<-rbind(pre,table(s1$edu, s1$treat_sim))
pre<-rbind(pre,table(s1$fami, s1$treat_sim))
pre<-rbind(pre,table(s1$occ, s1$treat_sim))
```

```
pre<-rbind(pre,table(s1$birth_reg, s1$treat_sim))
pre<-rbind(pre,table(s1$homeowner, s1$treat_sim))
perc_treatm<-matrix(data=NA, nrow = 24, ncol = 23)
for (i in 1:23){
perc_treatm[,i]<-pre[-c(1,25),i]/tab_treats[k,i]}
var_treatm<-perc_treatm*(1-perc_treatm)
perc_tot<-apply(pre[-c(1,25),],1,sum)/(sum(tab_treats[k,]))
var_tot<-perc_tot*(1-perc_tot)
t_asb_pre<-matrix(data=NA,nrow=24, ncol = 23)
for (i in 1:24){
t_asb_pre[i,]<-(abs(perc_treatm[i,]-perc_tot[i]))/
               (sqrt((var_treatm[i,]+var_tot[i])/2))*100}
ASB_pre[k,]<-c(quantile(t_asb_pre, probs = c(0, 0.25, 0.5, 0.75, 1)),
            mean(t_asb_pre), length(which(t_asb_pre>5)),
            length(which(t_asb_pre>10)))

### Preparation of the frequency table and setting of other parameters
needed in the balancing procedure
freq<-table(s1$ar, s1$t)
ps<-sort(unique(s1$ar))
freq<-cbind(ps, freq)
### Set the caliper to define the tolerance interval
caliper<-sd(s1$ar)/4
n<-as.numeric(colnames(freq)[-1]) # Identification codes of treatment
groups considered
nT<-length(n)  # Number of treatments in the matching procedure
### Set the frequency reference for each row fr
ref<-ifelse(ceiling(apply(freq[,-1], 1, median))==0,1,
         ceiling(apply(freq[,-1], 1, median)))

### Create empty vectors to store individual identification codes that
will be included in the balanced population
new<-c()
new0<-c()
sub0<-c()
rem<-c()

### Start the balancing procedure
start<-Sys.time()
### Consider every column separately
for (i in 1:nT){
same<- freq[,i+1]==ref
zero<-freq[,i+1]==0
different<-freq[,i+1]>0 & freq[,i+1]!=ref

ok<-rep(0, dim(s1)[1])
```

```
ok<-ifelse(s1$ar %in% freq[same==TRUE,1] & s1$t==n[i], TRUE, FALSE)
new<-c(new, which(ok))

for (j in which(different)){
cond_tosample<-which(ifelse(s1$ar==freq[j,1] & s1$t==n[i], TRUE, FALSE))
if (length(cond_tosample)==1){
ok<-rep(cond_tosample, ref[j])   }
if (length(cond_tosample)>1){
ok<-sample(cond_tosample, ref[j], replace = TRUE) }
new<-c(new, ok)}

for (j in which(zero)){
diff<-abs(freq[!zero,1]-freq[j,1])
value<-ifelse(sort(diff)[1]<=caliper,as.numeric(names(sort(diff))[1]),-1)
if (value==-1){rem<-c(rem,freq[j,1])}
if (value!=-1){
cond_tosample<-which(ifelse(s1$ar==value & s1$t==n[i], TRUE, FALSE))
if (length(cond_tosample)==1){
ok<-rep(cond_tosample, ref[j])
}
if (length(cond_tosample)>1){
ok<-sample(cond_tosample, ref[j], replace = TRUE)
}
new0<-c(new0, ok)
sub0<-c(sub0, rep(freq[j,1], ref[j]))
}}}
end<-Sys.time()

###build the balanced population
balanced_pop<-s1[new,]
balanced_pop0<-s1[new0,]
balanced_pop0$ar2<-sub0
after_all<-rbind(balanced_pop, balanced_pop0)
after<-after_all[!(after_all$ar2 %in% rem), ]

########save measure of balance on the balanced population
post<-table(after$sex, after$t)
post<-rbind(post,table(after$age, after$t))
post<-rbind(post,table(after$edu, after$t))
post<-rbind(post,table(after$fami, after$t))
post<-rbind(post,table(after$occ, after$t))
post<-rbind(post,table(after$birth_reg, after$t))
post<-rbind(post,table(after$homeowner, after$t))
treatm<-as.numeric(table(after$t))[1]
perc_treatm<-post[-c(1,25),]/treatm
var_treatm<-perc_treatm*(1-perc_treatm)
```

```
perc_tot<-apply(post[-c(1,25),],1,sum)/(treatm*23)
var_tot<-perc_tot*(1-perc_tot)
t_asb<-matrix(data=NA,nrow=24, ncol = 23)
for (i in 1:24){
t_asb[i,]<-(abs(perc_treatm[i,]-perc_tot[i]))/
   (sqrt((var_treatm[i,]+var_tot[i])/2))*100  }
time_bal[k] <-end-start
ASB_post[k,]<-round(c(quantile(t_asb, probs = c(0, 0.25, 0.5, 0.75, 1)),
              mean(t_asb), length(which(t_asb>5)), length(which(t_asb>10))),
              digits=3)
```

References

Ambrose, A. F., Paul, G., and Hausdorff, J. M. (2013). Risk factors for falls among older adults: a review of the literature. *Maturitas*, 75(1):51–61.

Arcaya, M. C., Tucker-Seeley, R. D., Kim, R., Schnake-Mahl, A., So, M., and Subramanian, S. V. (2016). Research on neighborhood effects on health in the united states: A systematic review of study characteristics. *Social Science & Medicine*, 168:16–29.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424.

Barnett, D. W., Barnett, A., Nathan, A., Van Cauwenberg, J., and Cerin, E. (2017). Built environmental correlates of older adults' total physical activity and walking: a systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity*, 14(1):103.

Boccuzzo, G. and Caperna, G. (2017). Evaluation of life satisfaction in italy: Proposal of a synthetic measure based on poset theory. In *Complexity in Society: From Indicators Construction to their Synthesis*, pages 291–321. Springer.

Brüggemann, R. and Carlsen, L. (2011). An improved estimation of averaged ranks of partial orders. *MATCH Communications in Mathematical and in Computer Chemistry*, 65:383–414.

Brüggemann, R., Lerche, D., and Sorensen, P. B. (2004). First attempts to relate structures of Hasse diagrams with mutual probabilities. *Order Theory in Environmental Sciences*.

Brüggemann, R. and Patil, G. P. (2011). *Ranking and prioritization for multi-indicator systems: Introduction to partial order applications*. Springer Science & Business Media.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72.

Cannas, M. and Arpino, B. (2018). Machine learning for propensity score matching and weighting: comparing different estimation techniques and assessing different balance diagnostics. RECSM Working Paper Number 54.

Caperna, G. (2016). *Partial Order Theory for Synthetic Indicators*. PhD thesis, University of Padova. (Available from http://paduaresearch.cab.unipd.it/9588/).

Caperna, G. (2019). Approximation of averagerank by means of a formula.

Caperna, G. and Boccuzzo, G. (2018). Use of poset theory with big datasets: A new proposal applied to the analysis of life satisfaction in Italy. *Social Indicators Research*, 136(3):1071–1088.

Christian, H., Zubrick, S., Foster, S., Giles-Corti, B., Bull, F., Wood, L., Knuiman, M., Brinkman, S., Houghton, S., and Boruff, B. (2015). The influence of the neighborhood physical environment on early child health and development: a review and call for research. *Health & Place*, 33:25–36.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446.

Costa, G., Stroscia, M., Zengarini, N., and Demaria, M. (2017). *40 anni di salute a Torino, spunti per leggere i bisogni e i risultati delle politiche*.

Cubbin, C., LeClere, F. B., and Smith, G. S. (2000). Socioeconomic status and injury mortality: individual and neighbourhood determinants. *Journal of Epidemiology & Community Health*, 54(7):517–524.

Davey, B. A. and Priestley, H. A. (2002). *Introduction to lattices and order*. Cambridge university press.

De Loof, K. (2009). *Efficient computation of rank probabilities in posets*. PhD thesis, Ghent University. (Available from https://biblio.ugent.be/publication/874495).

De Loof, K., De Baets, B., and De Meyer, H. (2011). Approximation of average ranks in posets. *MATCH Communications in Mathematical and in Computer Chemistry*, 66:219–229.

Diez Roux, A. V. (2001). Investigating neighborhood and area effects on health. *American Journal of Public Health*, 91(11):1783–1789.

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

Fattore, M. (2016). Partially ordered sets and the measurement of multidimensional ordinal deprivation. *Social Indicators Research*, 128(2):835–858.

Hade, E. M. (2012). *Propensity score adjustment in multiple group observational studies: Comparing matching and alternative methods*. PhD thesis, The Ohio State University.

Harding, D. J. (2003). Counterfactual models of neighborhood effects: The effect of neighborhood poverty on dropping out and teenage pregnancy. *American Journal of Sociology*, 109(3):676–719.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Hong, G. (2010). Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics*, 35(5):499–531.

Hong, G. (2012). Marginal mean weighting through stratification: A generalized method for evaluating multivalued and multiple treatments with nonexperimental data. *Psychological Methods*, 17(1):44.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of labour market policies*, pages 43–58. Springer.

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics*, 84(2):205–220.

Leckie, G. (2009). The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement. *J. R. Statist. Soc. A*, 172(3):537–554.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

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

Li, F., Zaslavsky, A. M., and Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19):3373–3387.

Linden, A., Uysal, S. D., Ryan, A., and Adams, J. L. (2016). Estimating causal effects for multivalued treatments: a comparison of approaches. *Statistics in Medicine*, 35(4):534–552.

Linden, A. and Yarnold, P. R. (2016). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*, 22(6):875–885.

Lopez, M. J. and Gutman, R. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, 32(3):432–454.

Lunt, M. (2013). Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *American Journal of Epidemiology*, 179(2):226–235.

Mair, C. F., Roux, A. V. D., and Galea, S. (2008). Are neighborhood characteristics associated with depressive symptoms? a critical review. *Journal of Epidemiology & Community Health*, 62:940–946.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19):3388–3414.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4).

Meijer, M., Rohl, J., Bloomfield, K., and Grittner, U. (2012). Do neighborhoods affect individual mortality? a systematic review and meta-analysis of multilevel studies. *Social science & medicine*, 74(8):1204–1212.

Melis, G., Gelormino, E., Marra, G., Ferracin, E., and Costa, G. (2015). The effects of the urban built environment on mental health: A cohort study in a large northern italian city. *International Journal of Environmental Research and Public Health*, 12(11):14898–14915.

Oakes, J. M. (2004). The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology. *Social Science & Medicine*, 58(10):1929–1952.

Pickett, K. and Pearl, M. (2001). Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. *Journal of Epidemiology & Community Health*, 55(2):111–122.

Rabe, B. and Taylor, M. (2010). Residential mobility, quality of neighbourhood and life course events. *J. R. Statist. Soc. A*, 173.(3):531–555.

Rassen, J. A., Solomon, D. H., Glynn, R. J., and Schneeweiss, S. (2011). Simultaneously assessing intended and unintended treatment effects of multiple treatment options: a pragmatic "matrix design". *Pharmacoepidemiology and Drug Safety*, 20(7):675–683.

Rose, S. and Normand, S. L. (2019). Double robust estimation for multiple unordered treatments and clustered observations: Evaluating drug-eluting coronary artery stents. *Biometrics*, 75(1):289–296.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Roux, A. V. D., Borrell, L. N., Haan, M., Jackson, S. A., and Schultz, R. (2004). Neighbourhood environments and mortality in an elderly cohort: results from the cardiovascular health study. *Journal of Epidemiology & Community Health*, 58(11):917–923.

Sànchez-Riera, L., Wilson, N., Kamalaraj, N., Nolla, J. M., Kok, C., Li, Y., Macara, M., Norman, R., Chen, J. S., Smith, E., et al. (2010). Osteoporosis and fragility fractures. *Best Practice & Research Clinical Rheumatology*, 24(6):793–810.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., and Glynn, R. J., . C. E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety*, 17(6):546–555.

Silber, J. H., Rosenbaum, P. R., Ross, R. N., Ludwig, J. M., Wang, W., Niknam, B. A., Mukherjee, N., Saynisch, P. A., Even-Shoshan, O., Kelz, R. R., et al. (2014). Template matching for auditing hospital cost and quality. *Health Services Research*, 49(5):1446–1474.

Spreeuwenberg, M. D., Bartak, A., Croon, M. A., Hagenaars, J. A., Busschbach, J. J. V., Andrea, H., Twisk, J., and Stijnen, T. (2010). The multiple propensity score as control for bias in the comparison of more than two treatment arms: an introduction from a case study in mental health. *Medical Care*, pages 166–174.

Truong, K. D. and Ma, S. (2006). A systematic review of relations between neighborhoods and mental health. *The Journal of Mental Health Policy and Economics*, 9(3):137–154.

Turrell, G., Hewitt, B., Haynes, M., Nathan, A., and Giles-Corti, B. (2014). Change in walking for transport: a longitudinal study of the influence of neighbourhood disadvantage and individual-level socioeconomic position in mid-aged adults. *International Journal of Behavioral Nutrition and Physical Activity*, 11(1):151.

Uysal, S. D. (2015). Doubly robust estimation of causal effects with multivalued treatments: an application to the returns to schooling. *Journal of Applied Econometrics*, 30(5):763–786.

Williamson, E. J., Forbes, A., and White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine*, 33(5):721–737.

Yen, I. H., Michael, Y. L., and Perdue, L. (2009). Neighborhood environment in studies of health of older adults: a systematic review. *American Journal of Preventive Medicine*, 37(5):455–463.

Yoshida, K., H.-D. S. S. D. H. J. J. W. G. J. J. G. R. J. and Franklin, J. M. (2017). Matching weights to simultaneously compare three treatment groups: comparison to three-way matching. *Epidemiology (Cambridge, Mass.)*, 28(3):387–395.

Acknowledgements

The Authors thank Prof. Giuseppe Costa and his collaborators of the Unit "SCaDU Servizio Sovrazonale di Epidemiologia" in Grugliasco (Turin, Italy) for their useful suggestions and support in the data management. The data used for the research have been managed following a formal agreement between the SCaDU Service and the Department of Statistical Science of the University of Padova.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
University of Padua
Italy