Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

SCIENZE STATISTICHE

Ciclo XXIX

Settore Concorsuale di afferenza: 13/D2 – Statistica Economica

Settore Scientifico disciplinare: SECS-S/03 – Statistica Economica

Statistical Matching Imputation among different farm data sources

Presentata da: Riccardo D'Alberto

Coordinatore Dottorato

Prof.ssa Alessandra Luati

Relatore

Dott.ssa Meri Raggi

Co-Relatore

Prof. Davide Viaggi

Esame finale anno 2017

Alma Mater Studiorum
Università di Bologna

DIPARTIMENTO DI SCIENZE STATISTICHE "PAOLO FORTUNATI"
DOTTORATO DI RICERCA IN
ECONOMIA E STATISTICA AGROALIMENTARE

Statistical Matching Imputation among different farm data sources

Riccardo D'Alberto

Relatore:
Dott.ssa Meri Raggi
Co-Relatore:
Prof. Davide Viaggi

Coordinatrice di Dottorato:
Prof.ssa Alessandra Luati

XXIX ciclo
Anno Accademico 2016/2017

Alma Mater Studiorum
University of Bologna

DEPARTMENT OF STATISTICAL SCIENCES "PAOLO FORTUNATI"

PH.D. IN
AGRI-FOOD ECONOMICS AND STATISTICS

Ph.D. Coordinator:
Professor Alessandra Luati

Supervisor:
Assistant Professor Meri Raggi

Co-Supervisor:
Professor Davide Viaggi

External Examiner:
Professor Paolo Sckokai

Alma Mater Studiorum
University of Bologna

DEPARTMENT OF STATISTICAL SCIENCES “PAOLO FORTUNATI”

PH.D. IN
AGRI-FOOD ECONOMICS AND STATISTICS

Author:	Riccardo D’Alberto
Title:	Statistical Matching Imputation among different farm data sources
Department:	Statistical Sciences “Paolo Fortunati”
Degree:	Ph.D.
Convocation:	February 2017

Acknowledgements

I would like to sincerely thank Assistant Professor Meri Raggi for the kindness and the invaluable support (both human and scientific) she gave to me. She is a true example of academic commitment and dedication. Many thanks also to Professor Davide Viaggi for his supervision and his precious advises. Finally, I have to thank my father and my mother; their unconditional love and guidance are the grounds of what I am.

Abstract

This work addresses the challenge of integrating different data sources, dealing with both statistical methodology and a practical application to farm data. It reviews the existing literature on Statistical Matching (SM) imputation, focusing on non-parametric micro SM imputation "hot deck" methods, which allow to reduce the bias generated by model-based integration approaches. Implementing new combinations of these techniques with not commonly applied distance functions, we propose, through a simulation study, a robust recursive strategy for the imputation goodness validation (which is missing in the SM imputation literature) taking into account the different characteristics of the recipient and donor datasets and corroborating the few common prescriptions from the SM imputation literature. This work applies both the combinations of the "hot deck" techniques and the imputation goodness validation strategy to three different farm data sources, two official administrative datasets and one project survey, referred to the Emilia-Romagna Region farms sample. Taking into account the specificities of the different farm data sources integration issues, we propose also a reference framework for the farm data sources harmonization. Then, we firstly integrate the three different farm data sources and, secondly, on the basis of the new synthetic dataset generated through imputation, run a Propensity Score Matching (PSM) analysis. Indeed, this work also proves the usefulness of the consequent application of both the SM imputation and the PSM methodologies under the observational studies research context. The main research finding concerns the relevant (significant) evidence that the common prescription of the SM literature (i.e. that the biggest dimensionality ratio between the donor and the recipient datasets is always the best one in terms of the imputation results) can be relaxed in the case in which the matching variable(s) in the donor dataset have a "proper" variability. Indeed, even a narrower dimensionality ratio between the recipient and the donor, being the variance of the matching variable(s) in the former dataset lower than the variance of the matching variable(s) in the latter one, can produce optimal estimates of the original variable through the imputed ones (i.e. does generate good imputation results). Moreover, both the imputation goodness validation strategy and the reference framework for the farm data harmonization proposed, constitute relevant research contributions. Finally, with respect to the rigorous PSM application to an integrated dataset, we discuss the significant effect of the treatment (the farms Agri-Environmental Schemes uptake), on the land rented in, taking into account the agricultural economics literature.

Contents

List of Tables

List of Figures

Chapter 1

Preface

1.1 Introduction

Chapter 1 discusses the most relevant issues, both methodological and practical, surrounding the core research aim of the present work, i.e. the problem of different data sources integration. Issues taken into account by the present work concern three main aspects, i.e.: *i.* the practical and methodological statistical challenges behind data integration (such as the computational efficiency and the theoretical definition of new different combinations among non-parametric micro Statistical Matching imputation techniques and non-default distance functions), *ii.* the peculiar problem of different farm data sources integration and, *iii.* the agricultural economics research interest for policy impacts evaluation which has to be carried under the observational studies research context.

1.2 Data sources integration: issues in perspective

Different data sources integration is a current, debated issue, related obviously to statistical sciences but also to many other research fields. In the big data era, the opportunity of an easy and quick collection of a huge amount of data from different sources, increases the ambitious chance to easily access this kind of data and integrate/aggregate them for analysis purposes in different research field, ranging from economics to social sciences. Nevertheless, big data often prove to be hardly accessible; they are usually collected by private for strictly private purposes being, consequently, privately owned. Moreover, they often prove to be not completely reliable for research objectives. Therefore, despite the wide appeal big data do have, official administrative and survey data sources maintain a wide desirability, on the one side because of the countless possibilities of data integration/aggregation offered by the increasing amount of project surveys produced and, on the other side, because they remain the main reference data sources in order to access and use several key relevant information. Then, their desirability is still high, and it is even increased if we take into account all the theoretical issues related to these data sources, still far from being properly and completely both investigated and solved.

Nowadays there is a widespread and increasing demand for data integration/aggregation, obtained from different sources through different designs thought and realised for different research purposes. The above-mentioned increasing demand is due to the fact that new data collection requires always time, money and energies. Moreover, currently, there is an odd paradox consisting, on the one side, of a widespread production of privately owned data and, on the other side, of a sensible shortage of public, reliable, open informative data. Considering how urgent sometimes researchers' need of data can be and how difficult is both collecting new data and accessing official administrative data sources, different data sources integration/aggregation can clearly represent an optimal useful solution. Finally, taking into account the fact that the accessibility of the official administrative data sources is always conditioned to release constraints due to privacy claims which reduce data informative power, integration/aggregation procedures do acquire even more significance.

In order to integrate and/or aggregate different data sources there are several statistical methodologies. The oldest methodology for data integration is record linkage, originally implemented with the specific purpose of duplicated records identification in datasets where unique identifiers are unavailable, and progressively used for equal records matching among different datasets Winkler (2005). Record linkage is commonly divided into two different macro-approaches, i.e.: the deterministic record linkage methodology and the probabilistic record linkage one. The former is based on the exact accordance of units characteristics (usually based on alpha-numeric variables modalities), in order to match units pairs. This methodology presents the disadvantage that it does not properly work in conditions of uncertainty related to the above-mentioned units characteristics. The latter is rather based on the computed probabilities of two different units to constitute a pair, given their observed variables. Following Fellegi and Sunter (1969) then, we assign a probability of being referred to the same statistical unit, to each records pair we want to aggregate and which does belong to two different datasets. As Winkler (2005) shows, this basic method evolved until now, from being a practical "data-cleaning" procedure to being an "entity resolution" methodology. Indeed, since the first half of 70's, record linkage methods evolved providing every time a more complex theoretical background and a more efficient practical strategy to reach the purposes of merging/purging datasets, managing huge amount of records, being scalable and adaptive, visually representing connections among records through graph partitioning, optimizing likelihoods in order to speed up computational algorithms, developing generalized distance functions and the theoretical framework behind units pairs matching. In recent years, moreover, Tancredi and Liseo (2011) developed a hierarchical Bayesian approach for record linkage, focused on population sizing. It is a new original approach based on a no reduction of the available information (there is not the usual 0 to 1 comparison mechanism behind the model), and on the fact that uncertainty is used both in estimating the population size and in performing the record linkage process itself.

The second group of methodologies concerns the statistical upscaling /downscaling, commonly used to enlarge or to narrow information referred to a specific territorial and/or aggregate level. As Bloschl (2005) points out, statistical upscaling/downscaling techniques have been developed mainly in environmental and meteorological research fields, serving the principal purpose of representing and adapting, in the best possible way, data collected at different space levels and time scales, following an estimation logic. These techniques of scale changing are usually divided into two main subgroups; the former group does include stochastic-dynamic models, the latter one does involve descriptive statistics approaches.

The third group of methodologies, as the above-mentioned one, is the most recently developed and the one we mainly focus on in the present work. It consists in Statistical Matching (SM) imputation techniques which have been theoretically defined, for the first time, in a formally complete and exhaustive way, by D'Orazio et al. (2006) and further developed by Rässler (2012). SM imputation techniques represent a widespread "easy" and computationally quick solution to different data sources integration through semi-parametric and non-parametric approaches. Nevertheless, SM imputation techniques do serve different research purposes, such as: *i.* different data sources integration, *ii.* surveys missing values imputation, *iii.* new datasets building via mixed matching methods. Considering the two different SM imputation approaches, the one structured upon the non-parametric micro techniques relies on the possibility of avoiding the variables family distribution specification and/or the estimation of variables and model parameters, consequently resorting to the observed data available. Therefore, SM imputation through non-parametric micro techniques, on the one side allows researchers to work with observed (real) data and, on the other side, to avoid bias deriving from model misspecification. As Little and Rubin (2002) point out:

> "the objective of imputation is not to get the best possible predictions of the missing values, but to replace them by plausible values in order to exploit the information in the recorded variables in

the incomplete cases for inference about population parameters".

The so called "hot deck" SM imputation techniques serve the above-mentioned purposes, allowing researchers to handle the missing data issue by replacement. The core advantage of these techniques is that the replacement of an unobserved, both of a missing value and/or a variable, consists always in a substitution from an observed response of a similar unit. They are commonly called "hot deck" because they recall procedures for data storage through the use of punch cards, referring specifically to the deck of donors cards available for a non-respondent. When the deck was "hot", it meant it was being processed (D'Orazio, 2014).

We both study and apply the "hot deck" techniques with respect to three different research trajectories, i.e.:

1. we explore new combinations of not default distance functions and non-parametric micro SM imputation techniques matching algorithms;

2. we develop and implement a cohesive theoretical framework concerning the above-mentioned combinations;

3. we organise and structure a robust recursive strategy for imputation goodness validation when non-parametric micro techniques are used.

In addition to the developed combinations of different non-parametric micro SM imputation techniques and not default distance functions, the present work acquires relevance because of the lack in the existing literature, at the best of our knowledge, both of a consistent discussion concerning how to properly validate results from non-parametric micro SM imputation and how to correctly formalize the theoretical framework behind these methodologies. Indeed, despite to the fact that these methods have become extensively used and applied in the last fifteen years, there are neither a systematic strategy and/or proved tools to check the results of imputation through these techniques, nor there has been a significant improvement of their theoretical formalization. Therefore our effort is motivated by the need of both a deeper theoretical formalization of the non-parametric micro techniques and a strategy for the imputation goodness validation which is coherent with the non-parametric micro nature of the applied techniques.

1.2.1 Farm data integration

Data integration is a currently debated research issue which acquires even more relevance with respect to data specifically related to agricultural holdings (farms). Indeed, only in the most recent years, few SM imputation applications have concerned farm data which have been consequently used for different research purposes, such as: *i.* the evaluation of farms competitiveness improvement fostered by farm-investment support (Kirchweger and Kantelhardt, 2012), *ii.* the evaluation of Agri-Environmental Schemes windfall effects in specific case studies in France (Chabé-Ferret and Subervie, 2013) and, *iii.* the evaluation of farm-investment support effects on agricultural modernisation in Czech Republic (Ratinger et al., 2013). On the contrary, in others research fields, there have been several applications concerning different kinds of data and more specifically related to the data integration itself, such as: *i.* data integration concerning Italian families incomes and consumptions (Coli et al., 2005), *ii.* the integration of data related to different US electoral population samples (Vavreck and Rivers, 2008), *iii.* the integration of different statistical surveys referred to the Italian families consumptions collected by *Banca d'Italia* (Sisto, 2006), *iv.* the integration between the Italian Population and Housing Census and others official administrative statistical surveys (D'Orazio, 2008), *v.* the integration of different macroeconomics data (Kum and Masterson, 2008) and, *vi.* the integration between *ad hoc* statistical surveys carried out both on Italian families and playtime (Donatiello et al., 2016).

The lack of a widespread application of the SM imputation methodologies to farm data seems to be surprising if we consider the research needs of the agricultural economics and the relevant shortage of available, complete and reliable data on agricultural holdings referred to EU and specifically Italian farms. This is firstly due to the fact that these data are usually collected for public purposes only by few institutions whereas few are the privately owned farm data (for example, project surveys). Secondly, with specific reference to the Italian case, farm data are hardly accessible and the few accessible data sources are usually released in an incongruous time span. Thirdly, farm data present a wide heterogeneity, not only if we take into account the differences among the project surveys that have been increasingly produced within the research projects financed by the EU, but also, surprisingly, with respect to the official administrative data produced by the different level institutions structured in a hierarchic and synergistic frame. Indeed, if we take into account the Italian case, we can notice that there is a strict link among regional statistical offices, the Italian Institute of Statistics (ISTAT) and the European one (Eurostat). Nevertheless, these institutions do use heterogeneous set of questionnaires, survey methods, sampling designs, variables codes and descriptions, sometimes collecting even different kind of information (and consequently different variables and variables values/modalities), operating in different accounting years (which usually do not overlap), adopting different bureaucratic procedures and standardized data manipulation criteria for farm data release. Since both this heterogeneity among farm data and their shortage do often undermine researchers work, integrating different farm data sources can constitute a optimal research strategy to have at disposal complete and reliable data.

Three farm data sources constitute the relevant reference point for researchers who want to analyse Italian farm data, i.e.: the Farm Accountancy Data Network (FADN), upgraded annually and managed by Eurostat, the General Census on the Italian Agriculture made every 10 years by ISTAT, and the statistical survey on farms structure and productivity, the so-called *"Indagine sulla Struttura e sulla Produzione delle Aziende Agricole"* - SPA, which is carried out every 2 years by the same above-mentioned institution. Nevertheless, these farm data sources often present the availability, heterogeneity, unreliability and incompleteness issues discussed previously. For example, it is extremely difficulty to access these data sources and/or completely dispose, for research purposes, their contents (observed units and variables but also detailed sample design description and records references). Moreover, it happens that they do have information on farms collected by different questionnaires, for different accounting years, with respect to different variables which do not properly overlap. Finally, these data sources present huge differences with respect to their dimensionality, their farm samples, their designs and the procedures of pre-release data manipulation.

The quality of the data at disposal is obviously one of the most determinant factor for the goodness of the research results. In the specific context of agricultural economics, the quality of farm data is fundamental when researches approach policy impacts evaluation and causal effects analysis which are complex analysis, anyway, not only because of the shortage of reliable and complete farm data sources, but also for the peculiar context of agricultural economics research whose target subjects can hardly commit to an experimental design framework analysis. Indeed, agricultural holdings are assigned or uptake policies (i.e. “treatment”), whose impacts and causal effects are not valuable through experiments but merely observable. Farms are business units which have to adopt compulsory and/or voluntary policy measures which can not be merely randomly assigned, leading researchers into the observational studies theoretical framework where causal effects can be analysed following the theory of potential outcomes proposed by Rubin (2005).

The above-mentioned data issues, the specific observational studies research context and the EU call for a robust standardized policy impacts evaluation procedure, all these elements increase the straightforward need, operating in the agricultural economics research context, of complete, homogeneous and recurrently collected farm data. Therefore, integrating farm data from different data sources can be an optimal solution in order to face several issues, i.e.:

- the shortage of complete official administrative farm data collections made up on regular basis by national and regional institutions at different territorial levels;
- the excessively long time interval between the collection of data and their availability and/or release;
- the fact that official administrative farm data sources are hardly accessible;
- the constraints deriving from privacy claims which force researchers to deal with the loss of key-information and with the reduction of the variables informative power;
- the characteristics of the hugest official administrative farm data source available, FADN one, which presents the peculiar structure of an unbalanced data panel (see paragraph 3.1 for further details).

Taking into account these issues and considering that, nowadays: *i.* an increasing amount of data are produced and owned regularly by private for private purposes, *ii.* official administrative data tend to be diminished with respect to big data produced, despite both the key information they hold and their publicity nature, *iii.* an increasing number of *ad hoc* surveys are generated within the agricultural economics research projects financed by the EU, *iv.* Horizon 2020 (H2020) objectives actually stress the characteristics of availability and accessibility of survey data produced within these financed projects and, *v.* survey data are often highly heterogeneous and undoubtedly highly expensive to set up, this work aims also at using the implemented methodology for the integration of both primary and secondary farm data sources. Farm data integration through non-parametric micro SM imputation techniques combined with different not default distance functions, allows the preservation of various observed (real) information, building a new generated dataset which fulfil conditions of availability, completeness and homogeneity.

1.2.2 Our application

Our application concerns three different types of farm data (both from primary and secondary data sources), i.e.:

- FADN data;
- the SPA statistical survey made by ISTAT;
- the *ad hoc* survey CAP-IRE produced in the context of a financed (FP7 2008-2010) EU project.

The application to these farm data is structured upon the three following key-step, i.e.:

1. the different datasets harmonization procedure;
2. the data integration through different combinations of non-parametric micro SM imputation techniques and distance functions;
3. the policy impacts evaluation analysis through Propensity Score Matching (PSM) methods.

The harmonization procedure is a crucial step both for SM imputation and PSM applications; indeed, it provides the essential conditions for the set up of homogeneous datasets which have to be integrated, fitted out with the variables useful for the research purposes, properly re-coded in the same language, with homogeneous codes, similar descriptions and equivalent characteristics. In our application it constitutes a fundamental complex step, proving how heterogeneous different datasets can be, even if they belong to data sources produced by synergistic institutions.

The integration procedure instead, shows the several issues, relevant for the statistical methodology point of view, we have to face applying SM imputation to farm data, i.e.:

1. the problem of different farm samples representativeness;
2. the fact that FADN constitutes an unbalanced data panel since observed units change every year (but not on a regular basis) and farm samples overlap differently over time;
3. the wide variables heterogeneity among the official administrative data source and the survey data;
4. the remarkable presence of outliers, missing records, variables and values, both in official administrative data and surveys
5. the not exact correspondence among codes and characteristics of the (few) existing common variables even in the two official administrative datasets.

Finally, taking into account the policy impacts evaluation application, we stress that its main goal (considering that the original CAP-IRE 2009 data were not expressly collected for evaluation purposes), is to present a rigorous application of the PSM methodology, which is coherent with the observational studies research context, to farm data previously integrated by SM imputation. In others words, the PSM application, despite of its binding data-driven nature, represents a rigorous attempt to demonstrate how potentially useful the integration of different farm data sources can be for further policy impacts evaluation analysis. Even though the literature on Agri-Environmental Schemes (AES) shows a clear bent to not consider these policies as a massive affecting determinant of farms structural changes, job and employment dynamics swing and farm activities diversification, in our application of the PSM we choose to consider the farms uptake of AES as the treatment variable, and possibly evaluate AES impacts on farms structures, land tenure, job and activity diversification. Therefore, policy impacts evaluation analysis acquires relevance more for the application itself than for the economics findings, being constrained by the characteristics of data at disposal.

1.2.3 Agri-Environmental Schemes

The European Union, as the prime supranational organisation involved in the planning and implementation of agricultural policies of its Member States, is also the most important actor involved in policy impacts evaluation procedures. This is due to the fact that EU, through the Common Agricultural

Policy (CAP), is responsible of the main policy intervention on agriculture and rural areas in general. The CAP is structured upon two distinct Pillars, the 1^{st} and the 2^{nd} ones; it provides both for direct payments, market support and/or regulation measures, direct subsidies to EU producers (1^{st} Pillar), and Rural Development Policy (RDP -2^{nd} Pillar-), in all the EU Member States.

Under 2^{nd} Pillar, as reported on the European Commission website (Website, 2016):

> "RDP is a complement of the system of direct payments to farmers and to measures related to agricultural markets management, based on the specific needs of EU territories and focused on the three thematic axes of the competitiveness of the agricultural and forestry sector improvement, the environment and the countryside improvement, the quality of life in rural areas improvement and the encouragement of the diversification of the rural economy".

A key component of 2007-2013 RDP were AES, incentive-based instruments that pay off farmers who voluntarily commit to preserve and enhance the environment and to maintain landscapes and the socio-cultural rural context. Introduced into the CAP during the late 80's as an option to be eventually applied by the EU Member States, in 1992 AES became more extensively part of the CAP, in particular with regulation 2078/92. Since 2000, instead, AES become a compulsory part of RDP for EU member states, increasing their weight both in terms of total expenditure for rural development and attention given by the EU regulation.

AES have been studied since the late 90's by authors who attempted various methods sprang from different disciplines and fields of study. The complete literature review on AES written by Uthes and Matzdorf (2013), points out that AES have been analysed according to four main focuses: *i.* the ecological and environmental AES effects analysis, conducted through field experiments and quasi-experimental survey data, *ii.* the identification of the multiple factors influencing farms decisions to adopt AES, characterizing the way decisions are taken under different socio-economic and environmental circumstances, *iii.* the *ex ante-ex post* qualitative evaluations of AES focused on the existing differences among national and regional schemes and, *iv.* the model-based approaches used either for evaluating farmers willingness to adopt AES or for the estimation of their economic and environmental success under different CAP scenarios.

In the most recent years there was an increasingly use of PSM and others statistical methodologies in order to run causal effects analysis and policy impacts evaluation concerning AES, taking into account different measures, different case studies in several EU Member States and also various PSM estimators, such as Pufahl and Weiss (2009), Jaraitė and Kažukauskas (2012), Chabé-Ferret and Subervie (2013), Udagawa et al. (2014), and Arata and Sckokai (2016).

The present work applies non-parametric micro SM imputation techniques (differently combined with not default distance functions), in order to integrate different farm data sources and use the new generated dataset for policy impacts evaluation through PSM; the core idea is then to consequentially join these two distinct methodologies taking into account the observational studies research context nature. In this basic PSM application to the new generated dataset, we use farms AES uptake as the "treatment" variable, evaluating AES impacts on farms of the Emilia-Romagna Region during the 2007-2013 RDP. We try to identify whereas AES produced any effects on farms structures, farms employment and farms activities diversification even if agricultural economics literature does not consider them their massive affecting determinant. We do know that several more important factors affect farms transformation process, nevertheless we have to deal with data at disposal. Since the application of non-parametric micro SM imputation techniques newly combined with not default distance functions, for the generation of a complete and homogeneous dataset consequently used to run causal effects analysis using PSM methods, constitute the most relevant application effort of the present work, we give less relevance to agricultural economics literature and to the interpretation of the PSM results.

Chapter 2

Methodology

2.1 Introduction

Chapter 2 discusses the statistical methodologies we apply for data integration and causal effects analysis, respectively the Statistical Matching (SM) imputation and the Propensity Score Matching (PSM). With respect to SM imputation we take into account non-parametric micro techniques, combining within their matching algorithms not default distance functions. Since these techniques application has increased in the most recent years in spite of both their proper theoretical formalization and the lack of a robust procedure for imputation results validation, we discuss the new combinations of techniques and distance functions, develop the theoretical formalization of these techniques but also run a simulation study in order to propose a robust recursive strategy for imputation goodness validation.

2.2 Statistical Matching imputation

SM imputation is a statistical methodology for data integration commonly used for several purposes, raging from missing values imputation to different datasets integration. It works imputing elements (values, variables and/or records), between two different datasets, commonly defined as the recipient and the donor one. SM imputation techniques are commonly divided into two categories, macro and micro techniques. The former consist in parameters estimation related to the existing relations between jointly unobserved variables; the latter take into account the possibility of generating a new synthetic dataset filled in with variables originally present in different separated datasets. The present work takes into account the second category of non-parametric micro techniques which associate records identifying pairs of donor and recipient units between a donor and a recipient dataset, and consequently imputing elements from the former to the latter. Units pairs are generated differently according to the different techniques and the matching algorithm definition within them.

Non-parametric micro SM imputation techniques are commonly defined as “hot deck” techniques. They offer several advantages with respect to parametric ones since they do not require either any specification for model parameters nor any estimate of the variables family distribution. “Hot deck” techniques so, fit the purpose of generating a complete synthetic dataset with simple and computationally quick complete-data methods, not requiring model specifications and avoiding potential problems deriving from model misspecification. Furthermore, they allow researchers to use only plausible observed elements for the imputation, then they work with observed data rather than model-based estimations.

Due to their non-parametric nature, SM imputation techniques do not require a complex theoretical framework. Nevertheless, this has determined, with their lately increasing application, a slow and inappropriate development of the theoretical formalization of both the different techniques and their matching algorithms and the distance functions applicable within them.

Saying A and B two different datasets, the former defined as the recipient and the latter defined as the donor one; saying i and j two different units with $i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B$; saying $\mathbf{X} = \{X_1, \ldots, X_l, \ldots, X_L\}$ the set of common variables between datasets A and B such that:

$$\underset{n_A \times L}{\boldsymbol{X}^A} = \left\{X_1^A, ..., X_l^A, ..., X_L^A\right\} = \begin{bmatrix} x_{11}^A & ... & x_{1l}^A & ... & x_{1L}^A \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1}^A & ... & x_{il}^A & ... & x_{iL}^A \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_A 1}^A & ... & x_{n_A l}^A & ... & x_{n_A L}^A \end{bmatrix}$$

and

$$\underset{n_B \times L}{\boldsymbol{X}^B} = \left\{X_1^B, ..., X_l^B, ..., X_L^B\right\} = \begin{bmatrix} x_{11}^B & ... & x_{1l}^B & ... & x_{1L}^B \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1}^B & ... & x_{il}^B & ... & x_{iL}^B \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_B 1}^B & ... & x_{n_B l}^B & ... & x_{n_B L}^B \end{bmatrix}$$

where ${X_l}^A$ is a vector of dimension $(n_A \times 1)$ and ${X_l}^B$ is a vector of dimension $(n_B \times 1)$.

Saying then the set of the following variables exclusively present in dataset A, i.e.:

- $\underset{n_A \times P}{\mathbf{Z}} = \{Z_1{}^A, \dots, Z_p{}^A, \dots, Z_P{}^A\}$, where $Z_p{}^A$ is a vector of dimension $(n_A \times 1)$;
- $\underset{n_A \times Q}{\mathbf{Y}} = \{Y_1{}^A, \dots, Y_q{}^A, \dots, Y_Q{}^A\}$, where $Y_q{}^A$ is a vector of dimension $(n_A \times 1)$;
- $\underset{n_A \times S}{\mathbf{T}} = \{T_1{}^A, \dots, T_s{}^A, \dots, T_S{}^A\}$, where $T_s{}^A$ is a vector of dimension $(n_A \times 1)$.

Saying the set of the following variables exclusively present in dataset B, i.e.:

- $\underset{n_B \times M}{\mathbf{K}} = \{K_1{}^B, \dots, K_m{}^B, \dots, K_M{}^B\}$, where $K_m{}^B$ is a vector of dimension $(n_B \times 1)$.

We have two datasets A and B such that: $\left\{ \underset{n_A \times L}{\mathbf{X}^A}, \underset{n_A \times P}{\mathbf{Z}}{}^A, \underset{n_A \times Q}{\mathbf{Y}}{}^A, \underset{n_A \times S}{\mathbf{T}}{}^A \right\}$ is the recipient dataset and $\left\{ \underset{n_B \times L}{\mathbf{X}^B}, \underset{n_B \times M}{\mathbf{K}}{}^B \right\}$ is the donor one.

For sake of simplicity, we assume that S=*1*, then T is a vector of dimension $(n_A \times 1)$. Moreover we choose to consider here the simplest case in which Q=*1* so that Y is a vector of dimension $(n_A \times 1)$.

Following D'Orazio et al. (2006), having two matching samples (i.e. datasets) A and B, we assume that:

- **Assumption 1.** $A \cup B$ can be considered as a unique sample of the $n_A + n_B$ i.i.d. observations from the joint distribution of ($\mathbf{X}$, $\mathbf{Z}$, $\mathbf{K}$).

- **Assumption 2.** The recipient dataset A, with the dimensionality n_A and the donor dataset B, with the dimensionality n_B, are always chosen such that $n_A \leq n_B$.

This latter assumption is motivated by the core idea that:

> "the larger is the donor file, the more accurate is the estimated distribution of $\mathbf{Z}$ given $\mathbf{X}$ if consistent estimators are used. This reason always justifies the strategy of choosing as recipient file the one with the smaller sample size" (D'Orazio et al., 2006).

The above-mentioned key assumptions are at the basis of the SM imputation through non-parametric micro techniques, which are the following ones, i.e.:

- Nearest Neighbour Distance Hot Deck (nnd)
- Constrained Nearest Neighbour Hot Deck (nndc)
- Random hot deck (rnd)
- Rank hot deck (rnk)

We re-organize the non-parametric micro SM imputation techniques application in the following consecutive steps:

1. a descriptive analysis of data samples and observed variables;
2. the choice of the donor and recipient datasets;
3. the harmonization of the recipient and donor datasets;

4. the choice of the matching and the imputed variables;

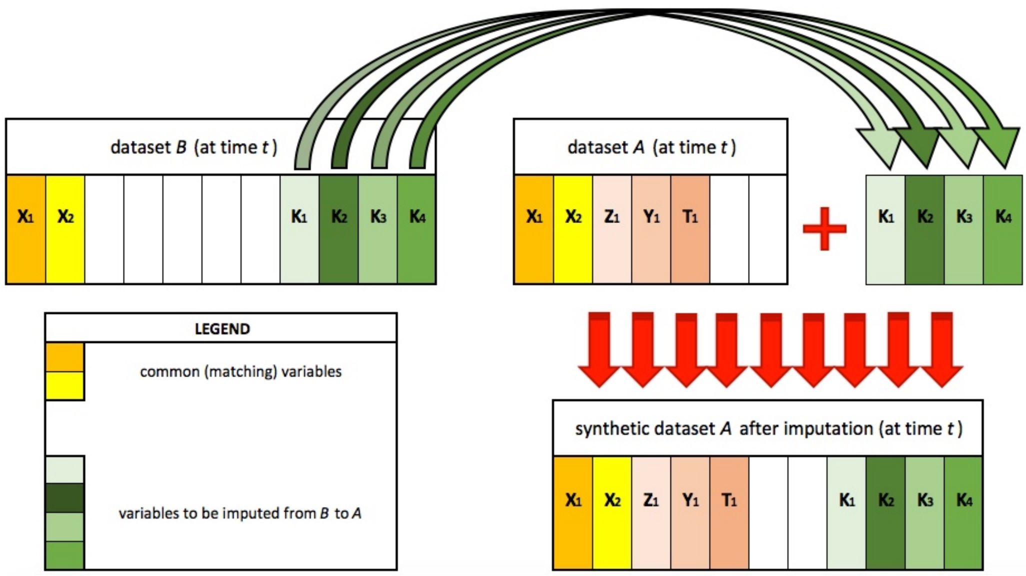
5. the choice of the matching technique and the distance function combined within its matching algorithm;

6. the imputation running and the generation of the synthetic dataset;

7. the imputation goodness validation.

Figure 2.1 shows schematically how the above-mentioned techniques do function. Having two different datasets A and B, referred to the same time (year) t, we choose among the set of common variables between A and B which ones we want to use as matching variables (i.e. the orange and yellow ones), and which ones we want to impute (i.e. the ones in green shades). Therefore, we create the synthetic dataset which is complete and homogeneous with respect to the two above-mentioned datasets of interest.

Figure 2.1: SM imputation scheme



2.2.1 Nearest Neighbour Distance Hot Deck

For sake of simplicity, we assume *L=1* so that X is a single (continuous) variable. Saying i the recipient unit in dataset A and j^* the donor unit in dataset B chosen to be matched, i.e. chosen to constitute a pair with the unit i, Nearest Neighbour Distance Hot Deck associates pairs of units in the way that the following equation holds, as suggested by D'Orazio et al. (2006):

$$d_{ij^*} = \mid x_i^A - x_{j^*}^B \mid = \min_{j=1,\dots,n_B} \mid x_i^A - x_j^B \mid ,$$

where d is the absolute value of the difference between the two units i and j (j^*). The minimum value of difference d is always computed such that $1 \leq j \leq n_B$.

Nearest Neighbour Distance Hot Deck technique is a frequently used SM imputation technique, since its logic is quite intuitive and it usually performs the best imputation fit. Indeed, by default nnd identifies in the donor dataset which units are to be considered the "nearest" to the unit in the recipient dataset which have the closest values of the variable or variables to be imputed. Basically, it always chooses the nearest donor unit to the recipient one, as the one eligible for the imputation. In order to determine the proximity between donor and recipient units, nnd algorithm computes the differences (distances) among units in terms of the chosen matching variable X which is in common between the two datasets. Obviously, matching variables can be even more than one; rather, more relevant variables we take into account, the better is the imputation fit. This SM imputation technique allows the choice of the nearest unit to be imputed always by solving the so-called "travelling salesperson problem" (Ballin et al., 2009).

It is also possible to sharpen this technique by creating the so-called "imputation donation classes", defined using existing common categorical variables (the minimum required number of common categorical variable is four), between the two datasets. Donation classes are useful in order to create homogeneous groups of units within which it is possible to choose donor and recipient units to be matched. Indeed, when donation classes hold, distances are always computed only among units belonging to the same donation class. Imputation does benefit from the donation classes building both in terms of matching precision increasing and computational matching effort lightening.

By default, nnd uses each available donor unit for the recipient one, more than once if it adequately matches it. However, a "constrained " version of this technique does exist.

2.2.2 Constrained Nearest Neighbour Hot Deck

For sake of simplicity, we assume *L=1*, so that X is a single (continuous) variable. Saying i the recipient unit in dataset A and j the donor unit in dataset B, Constrained Nearest Neighbour Hot Deck associates pairs of units, as suggested by D'Orazio et al. (2006), taking into account the following difference:

$$d_{ij} = \mid x_i^A - x_j^B \mid .$$

Imposing constraints to the nnd technique consists in minimizing the following function:

$$\sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (d_{ij}\ \omega_{ij})\ ,$$

with $\omega_{ij} \in \{0, 1\}$ representing the matched pair of units i and j. ω_{ij} is equal to 0 if the pair of units i and j are matched and equal to 1 otherwise.

nndC technique needs that the following set of constraints do hold:

$$\begin{aligned} \sum_{j=1}^{n_B} \omega_{ij} &= 1\ , \\ \sum_{i=1}^{n_A} \omega_{ij} &\leq 1\ . \end{aligned}$$

These two constraints basically mean that one donor unit j can be selected by the matching algorithm in order to be matched with the recipient unit i just once, while it could be the possibility that no recipient units i are founded for the donor unit j.

For both nndC and nnd techniques it happens that when two or more donor units are selected because they are at the same distance from a recipient unit, the matching algorithm always select the donor unit randomly.

2.2.3 Random Hot Deck

Random Hot Deck technique constitutes the most naïve SM imputation technique among the four hot deck techniques (D'Orazio et al., 2006). Indeed, rnd picks basically at random the donor unit to be matched with the recipient one. This technique represents then the most uncertain one among the four above-mentioned since it does not properly guarantee the correspondence among values of the observed variables for donor and recipient units (when not only a variable X is the common one but it is rather possible to use a set of common variables $\mathbf{X}$).

Nevertheless, this technique can be sharpen considering a proper threshold in the way that donor units, whose distances from the recipient unit is less than the set up threshold, and only those ones, are taken into account by the matching algorithm. Besides, it is possible to set up different ways to pick donor units to be matched with the recipient ones. For example it is possible to set a certain exact distance between donor units and recipient ones which has to be respected by the matching algorithm, it is possible to take into account only donors at the available minimum distance from the recipient, it is also possible to select among donor units whose proportion with respect to the recipient unit lies between 0 and a set up threshold t, and it is finally possible to reduce the chosen donor units at the squared root of the closest recipient one.

rnd technique usually disposes the possible subset of donor and recipient units pairs as defined by:

$$ {n_B}^{n_A} \ . $$

This is true if no donation classes are built. Whereas, saying X_1 and X_2 two existing common variables between the dataset A and the dataset B which constitute a donation class, rnd reduces the subset of units such that:

$$ (n^B_{X_1})^{n^A_{X_1}} \ + \ (n^B_{X_2})^{n^A_{X_2}} \ . $$

2.2.4 Rank Hot Deck

For sake of simplicity, we assume *L=1*, so that X is a single (continuous) variable. Saying i the recipient unit in dataset A and j the donor unit in dataset B, Rank Hot Deck associates pairs of units considering the empirical cumulative distribution function of the variable X (D'Orazio et al., 2006). rnk is composed by two key steps; indeed, rnk first ranks donor and recipient units, i.e.:

$$F_{X^A}(x^A) \;=\; \frac{1}{n_A}\sum_{i=1}^{n_A} I\ (x_i \leq x)\ ,$$

for the recipient dataset A, being I the set of indices of $x_i \leq x$, and:

$$F_{X^B}(x^B) \;=\; \frac{1}{n_B}\sum_{j=1}^{n_B} I\ (x_j \leq x)\ ,$$

for the donor dataset B, being I the set of indices of $x_j \leq x$.
Second, rnk matching algorithm associates to each recipient unit a donor unit in the way that the following equation holds:

$$|F_{X^A}(x_i^A) \;-\; F_{X^B}(x_{j*}^B)| \;=\; \min_{j=1,\ldots,n_B} |F_{X^A}(x_i^A) \;-\; F_{X^B}(x_j^B)|\ ,$$

where the minimum of the distance between $F_{X^A}(x^A)$ and $F_{X^B}(x^B)$ is computed such as $1 \leq j \leq n_B$.

2.2.5 Distance functions

SM imputation techniques use matching algorithms in order to compute distances between donor and recipient units. These algorithms work differently also according to the distance function set. By default, "hot deck" techniques use the Manhattan distance function whereas in the present work we discuss different combinations of techniques and not default distance functions changing the matching algorithm association process with respect to the different recipient-donor datasets characteristics (dimensionality ratio, variables at disposal, variables values/modalities, variability of the matching variable(s) used).

For sake of simplicity, we assume that L=*1*, so that X is a single (continuous) variable. Saying i the recipient unit in dataset A, j the donor unit in dataset B and h another unit from a third dataset C, with h = *1*, $\ldots, n_C$, we define the distance function δ as a distance function, if and only if, as suggested by D'Orazio et al. (2006), the three following prescriptions are verified, i.e.:

- $\delta_{ij} = \delta_{ji}$, which means that there is always symmetry between the two distance functions;
- $\delta_{ij} \geq 0$, which means that the distance function is always a non-negative function;
- $\delta_{ij} = 0$, which means that identity property does hold.

Given the δ distance function, we define Δ as a metric if and only if these two assumptions hold (Mardia and Jupp, 1979), i.e.:

- **Assumption 1.** $\Delta_{ij} = 0$, if and only if $i = j$, which means that there is an identity of the equals;
- **Assumption 2.** $\Delta_{ij} \leq \Delta_{ih} + \Delta_{hj}$, which represents a triangle inequality.

Considering that for each unit i we observe the set of variables $\mathbf{X} = \{X_1, \ldots, X_l, \ldots, X_L\}$ defined as continuous variables, where X_l is a vector of dimension $(n \times 1)$, D is the class of distance functions defined by the use of the so-called "Minkowski-Ruum" metric as suggested by Mardia and Jupp (1979), such that:

$$D_{ij} = \left[\sum_{l=1}^{L} {c_l}^{\theta} \, |x_{li} - x_{lj}|^{\theta} \right]^{\frac{1}{\theta}} ,$$

where c_l is a factor of scale for the l-th variable and θ is an index defined as $\theta = 1, \ldots, +\infty$, representing for each value of θ a different kind of metric.

Saying $\theta = 1$ then, the Manhattan metric function is defined such that the following equation holds:

$$\Delta_{ij}{}^{Mn} = \sum_{l=1}^{L} |x_{li} - x_{lj}| \, . \tag{2.1}$$

The Manhattan metric function calculates the distance, or "proximity", between two units always computing the absolute value of the sum of the differences between donor and recipient units in terms of the values of their observed variables.

The Mahalanobis metric function is defined, instead, in the following way:

$$\Delta_{ij}{}^{Ms} \;=\; \left(\mathbf{X}_i{}^{A} \;-\; \mathbf{X}_j{}^{B}\right)' \;\Sigma^{-1}_{\mathbf{X}^A\mathbf{X}^B}\;\left(\mathbf{X}_i{}^{A} \;-\; \mathbf{X}_j{}^{B}\right) \tag{2.2}$$

where Σ is the covariance matrix of the $\mathbf{X}$ variables and the above-mentioned distance function defines the "proximity" of units taking into account the statistical relationship among the observed covariates $\mathbf{X}$.

Slightly different from the previous two, the Gower distance function (which works on the basis of the Gower's dissimilarity coefficient), takes into account the different modalities of the chosen discrete variables. The distance is then computed by averaging the suitable distances for each donor and recipient unit in terms of the values of their observed variables, in the way that the following equation holds (Gower, 1971):

$$\Delta_{ij}{}^{Gw} \;=\; \frac{1}{L}\sum_{l=1}^{L} c_{l\Delta_{ijl}} \;,$$

where $\frac{1}{Rp}$ is the standardization of the chosen variables, made out either by using the standard deviation or using the above-mentioned range $Rp = max(x_{il}) - min(x_{jl})$; maximum and minimum are always considered with respect to i and c_l is a factor of scale for the l-th variable, equal to 1 for binary variables and equal to $\frac{1}{R_p}$ for continuous and ordinal categorical ones.

Therefore, the Gower distance function can be used in the way that the following equation holds:

$$\Delta_{ij}{}^{Gw} \;=\; |(x_{il}) - min(x_{jl})| \;.$$

From the above-mentioned distance function, the Exact distance one can be developed to be used within SM imputation techniques matching algorithms, taking into account eventually present categorical variables. Exact distance function works like to the so-called "Sørensen-Dice SS" logical similarity index (Gallagher, 1999). Nevertheless, due to the fact that this distance function does not satisfy the triangle inequality assumption, it cannot be considered a proper metric distance function and it should be considered rather as a "dissimilarity index". It ranges from 0 to 1, always converting the recipient and the donor units into categorical variables, then setting the distance between them to 0 if a units pair has the same response category and to 1 otherwise.

These distance functions can be combined with the SM imputation techniques (with the exception of the Rank Hot Deck), according to the existing different characteristics among the matching variables and between the donor and recipient datasets.

2.3 Simulation study

The different combinations of the distance functions within the matching algorithms of the "hot deck" techniques, generate different synthetic dataset. Taking into account the subject of imputation, the specific and peculiar characteristics of recipient and donor datasets, the objectives of the imputation process itself, we analyse the different combinations performances. We run then a simulation study in order to both analysing how the different combinations perform and proposing a structured method for the imputation goodness validation.

When non-parametric micro SM imputation techniques are used, indeed, researchers do not have at disposal a systematic methodology for the checking of the imputation results. In others words, there are no formalized tools in order to check how different combinations of SM imputation techniques and distance functions do perform together and how much good their combined application is (i.e. which is the best synthetic dataset generated). Since the "hot deck" techniques have a peculiar non-parametric nature, in order to validate their application, certainly, it is not possible to merely apply the checking procedures commonly in use within parametric SM imputation techniques. Therefore, the main goal of the simulation study is to verify how these different combinations do perform taking into account the different recipient and donor datasets characteristics. Moreover, we are interested in developing a systematic strategy useful for SM imputation goodness validation and suitable for choosing the best synthetic dataset generated by the imputation process.

Therefore, we analyse the imputation results of the Rank Hot Deck technique and the plausible combinations of distance functions within the matching algorithms of the other "hot deck" SM imputation techniques reported in table 2.1.

Table 2.1: Plausible combinations of SM imputation techniques and distance functions

Technique	Distance function	Combination
Nearest Neighbour Distance Hot Deck (nnd)	Manhattan (mn) Mahalanobis (ms) Exact (e)	nnd.mn nnd.ms nnd.e
Constrained Nearest Neighbour Hot Deck (nndc)	Manhattan Mahalanobis Exact	nndc.mn nndc.ms nndc.e
Random Hot Deck (rnd)	Manhattan Mahalanobis Exact	rnd.mn rnd.ms rnd.e
Rank Hot Deck (rnk)		

The simulation study is based on two consequent steps, a previous recipient and donor datasets variables simulation and a consequent SM imputation running. This latter step follows the above mentioned scheme of non-parametric micro SM imputation techniques and distance functions combination and the methodological steps described in paragraph 2.2.

We focus on two simulated datasets, a recipient and a donor one, which we characterise differently with respect to three main aspects, i.e.:

- the different dimensionality ratio between recipient and donor datasets;
- the different variability of matching variable(s);
- the possibility of running SM imputation either having previously built matching donation classes or not having built them.

We simulate the recipient dataset R and the donor dataset D; R and D are always simulated such that $n_R < n_D$, as prescribed by the SM imputation literature (Singh et al., 1993). For both R and D we do simulate a set of common variables $\mathbf{X} = \{X_1, X_2, X_3\}$ and a set of common variables $\mathbf{K} =$

$\{K_1, K_2\}$. Indeed, saying i and j two different units with $i = 1, \ldots, n_R$ and $j = 1, \ldots, n_D$, datasets R and D share two sets of common variables, such that:

$$\underset{n_R \times 3}{\boldsymbol{X}^R} = \left\{X_1^R, X_2^R, X_3^R\right\} = \begin{bmatrix} x_{11}^R & \ldots & x_{12}^R & \ldots & x_{13}^R \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1}^R & \ldots & x_{i2}^R & \ldots & x_{i3}^R \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_R1}^R & \ldots & x_{n_R2}^R & \ldots & x_{n_R3}^R \end{bmatrix}$$

and,

$$\underset{n_D \times 3}{\boldsymbol{X}^D} = \left\{X_1^D, X_2^D, X_3^D\right\} = \begin{bmatrix} x_{11}^D & \ldots & x_{12}^D & \ldots & x_{13}^D \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1}^D & \ldots & x_{i2}^D & \ldots & x_{i3}^D \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_D1}^D & \ldots & x_{n_D2}^D & \ldots & x_{n_D3}^D \end{bmatrix}$$

and,

$$\underset{n_R \times 2}{\boldsymbol{K}^R} = \left\{K_1^R, K_2^R\right\} = \begin{bmatrix} k_{11}^R & \ldots & k_{12}^R \\ \vdots & \ddots & \vdots \\ k_{i1}^R & \ldots & x_{i2}^R \\ \vdots & \ddots & \vdots \\ k_{n_R1}^R & \ldots & k_{n_R2}^R \end{bmatrix}$$

and,

$$\underset{n_D \times 2}{\boldsymbol{K}^D} = \left\{K_1^D, K_2^D\right\} = \begin{bmatrix} k_{11}^D & \ldots & k_{12}^D \\ \vdots & \ddots & \vdots \\ k_{i1}^D & \ldots & x_{i2}^D \\ \vdots & \ddots & \vdots \\ k_{n_D1}^D & \ldots & k_{n_D2}^D \end{bmatrix}$$

We use X_3, K_1, K_2 for referring to the matching and the imputation variables present indiscriminately both in datasets R and D. We use instead

X_3^R, K_1^R and K_2^R for referring to the matching and the imputation variables originally "observed" in the recipient dataset R and X_3^D, K_1^D and K_2^D for referring to the matching variable in the donor dataset D and the variables to be imputed from D to R.

Therefore, the core idea is to simulate two different datasets, a recipient and a donor one, which share three potential matching and two imputation variables. Variables we want to impute from the donor to the recipient are simulated also in the latter one; this is due to the imputation goodness validation purposes. Indeed, we choose to simulate R and D datasets as if the imputation variables were originally present (i.e. "observed") also in the recipient one in order to analyse the differences among the variables originally present in the recipient dataset and the imputed ones, following a pre-post imputation logic.

Both the variable K_1 and the variable K_2 are simulated as the realization of a log-Normal(μ, σ^2) multiplied for a Bernoulli(θ), with $\theta = 1/2$. The variable X_1 is simulated as the realization of a Bernoulli(θ) with $\theta = 1/2$. The variable X_2 is a categorical variable indicating the main variable value between K_1 and K_2. The variable X_3 is simulated as the sum of the realizations of the variables K_1 and K_2.

We simulate two different conditions of recipient-donor datasets dimensionality ratio; one dimensionality ratio is 1 to 10, i.e. $n_R = 1000$ and $n_D = 10000$, the other is 1 to 3, i.e. $n_R = 1000$ and $n_D = 3000$. For each of these two conditions we then simulate two different cases of matching variable(s) variability. Choosing, for sake of simplicity, the solely variable X_3 as the matching variable between datasets R and D, we simulate the case in which var(X_3^R) > var(X_3^D) and the case in which var(X_3^R) < var(X_3^D). For sake of simplicity, from now on we will refer to var(X_3^R) as var(R) and to var(X_3^D) as var(D). Finally, for each one of the possible combinations of these two different conditions, we run SM imputation both with the building of donation classes (using variables X_1 and X_2) and without building them. These different conditions are motivated by our expectations with respect to the imputation goodness results which we discuss in details in paragraph 2.3.5.

Therefore, the resulting simulation study is based upon four different simulated pairs of recipient and donor datasets. We then choose to run eight SM imputations (applying the different combinations), both with and without the building of donation classes, as summarized in table 2.2.

Table 2.2: Simulation study and imputation scheme

Simulation Nr.	1		2		3		4	
Ratio	1 to 10		1 to 10		1 to 3		1 to 3	
Variability	var(R) > var(D)		var(R) < var(D)		var(R) > var(D)		var(R) < var(D)	
Imputation Nr.	1	2	3	4	5	6	7	8
Donation classes	with	without	with	without	with	without	with	without

In order to find the best combination of SM imputation technique and distance function, we propose an imputation goodness validation using three combined tools, i.e.:

- we check the distributions of the variables originally present in the recipient dataset and the variables imputed from the donor one in a pre-post imputation logic;
- we check the distributions of the differences between the values of variables K_1^R and K_1^D and K_2^R and K_2^D in the synthetic dataset generated (we define these differences "z");

- we evaluate the MSE of the above-mentioned differences.

For sake of clarity, figures in paragraphs 2.3.1-2.3.5 show distributions of the variables X_3, K_1 and K_2 in the recipient dataset R and the donor dataset D, and distributions of variables K_1^R, K_1^D and K_2^R, K_2^D in the various synthetic datasets generated by the different combinations of SM techniques and distance functions (plus the Rank Hot Deck technique itself). Sometimes distributions are cut up to the class value 200; this is done when distributions exceed the "suitable" needs of representation. The eventually presence of outliers, anyway, is always discussed with respect to the figures representing the distributions of the differences z.

For each simulation, first, we discuss the imputation with donation classes for a specific technique combined with the three distance functions, taking into account the variable K_1 (showing and commenting both the pre-post distributions and the distributions of the differences z); then we take into account the variable K_2. Second, we discuss the imputation with donation classes with the same above-mentioned cases. We replicate the scheme for the four simulations and the eight imputations. When the imputation results for different combinations of the distance functions are far too similar, they are omitted.

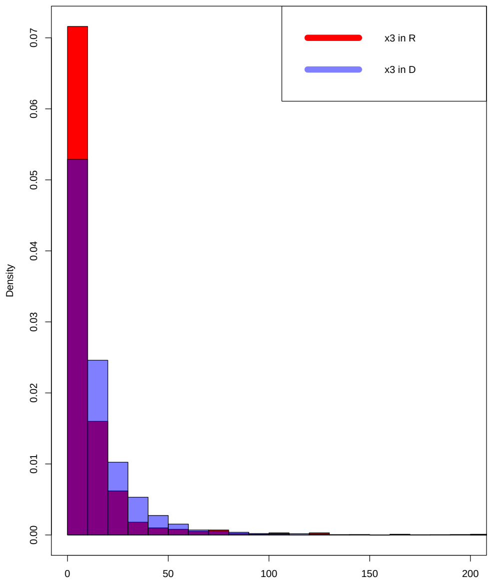
2.3.1 Results from simulation 1

Figure 2.2 shows that from simulation 1 we have the recipient dataset R and the donor dataset D characterised, with respect to the matching variable X_3, by a higher variance and a noteworthy presence of outliers in R (recipient).

We notice that with the sensible exception of the class 0-10, variable X_3^D values always overcome variable X_3^R values due to the bigger dimensionality of the donor dataset D.

Figure 2.2: Simulation 1, variable X_3 in R and D

X_3		
	R	D
mean	11.574	15.284
var	1731.413	476.384
min	0.056	0.179
max	1172.981	874.083



Taking into account the imputation variable K_1 in datasets R and D, beyond the difference in the maximum values of the variables K_1^R and K_1^D (K_1^R has a higher upper value), figure 2.3 shows that there is a slightly higher frequency of variable K_1^R in class 0-10, whereas there is a tendency of the variable K_1^D to overcome the variable K_1^R (with the exception of class 120-130 for which there is no coverage at all, i.e. there are not such values of the variable K_1 in the donor dataset D). With respect to the imputation variable K_2, figure 2.3 shows that, with the exception of the higher frequency of variable K_2^R in class 0-5, there is always a complete over-correspondence for the other values of the variable K_2 between datasets R and D.

Figure 2.3: Simulation 1, variables K_1 and K_2 in R and D

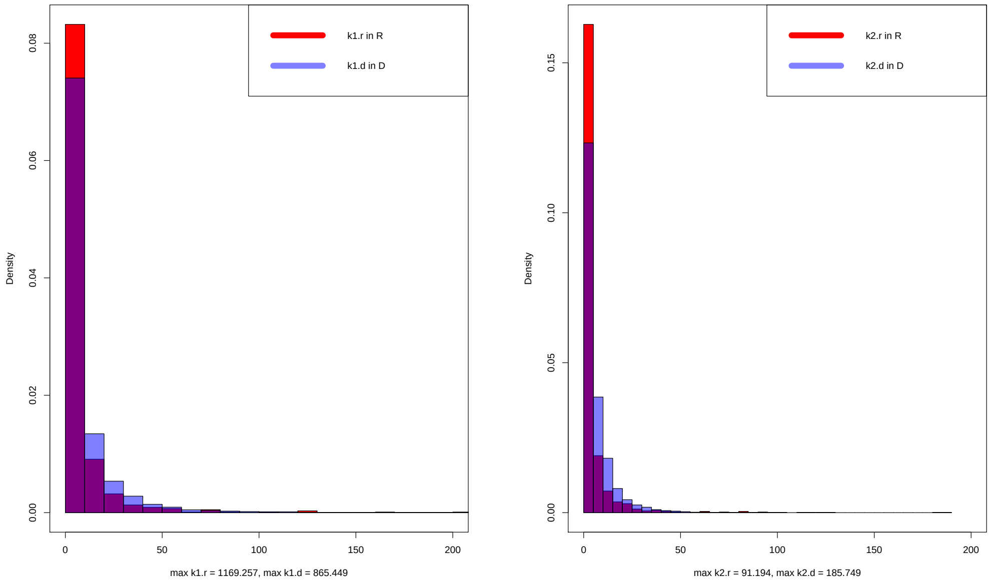
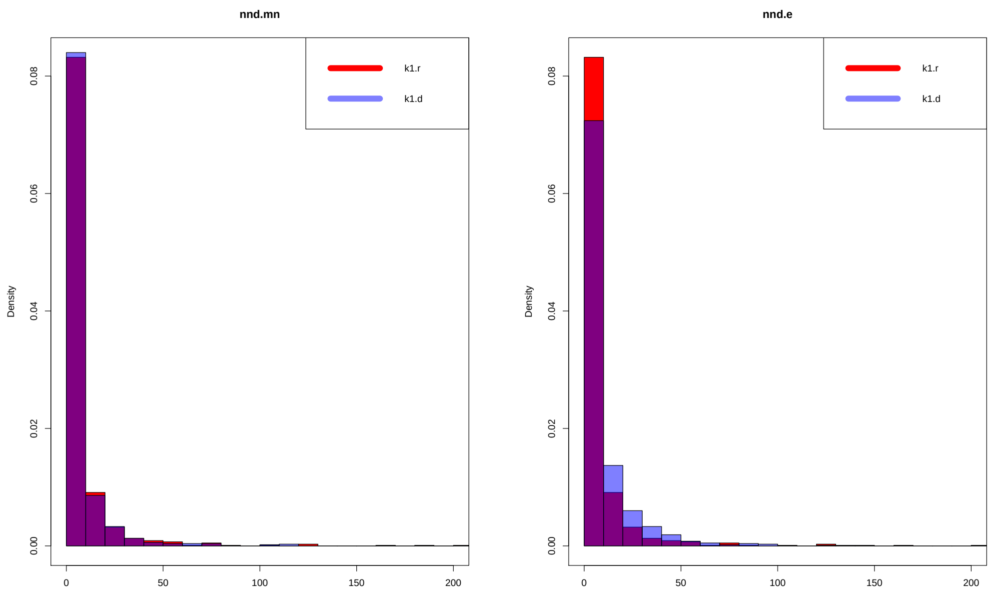


Figure 2.4 shows imputation (with donation classes) results, in terms of the different distributions of the original variable K_1^R and the imputed variable K_1^D in the synthetic datasets generated by combinations nnd.mn and nnd.e. Results of the combination nnd.ms are omitted since they are very similar to the combination nnd.mn. We can see that nnd.mn (and nnd.ms), generate a good synthetic dataset in terms of the overlap between variables K_1^R and K_1^D. Indeed, there is a not significant overestimate of variable K_1^R in classes 10-20, 60-70, 110-120, and a small not significant underestimate of variable K_1^R in classes 10-20, 40-60, 120-130. Anyway, the overall tendency of these combinations is to well represent the variable values observed in the recipient dataset. The combination nnd.e instead, generates a synthetic dataset in which the variable K_1^R in class 0-10 are slightly underestimated

whereas there is an evident tendency to overestimate (and almost doubling, for example for the class 20-30), the recipient variable K_1^R up to value 50.

Figure 2.4: Simulation 1, distributions of K_1^R, K_1^D in nnd imputation (with don. cl.)



Taking into account the distributions of differences z (i.e. the differences between the values of the original K_1^R, K_2^R variables and the imputed K_1^D, K_2^D variables), figure 2.5 shows that the combination nnd.mn (and nnd.ms), perform far better than the combination nnd.e, allowing also a really better control of the outliers. Indeed, the right tail of the z_{K1} distribution for nnd.mn is due only to the difference in the upper maximum values of the variable K_1 in R (recipient) and D (donor), whereas the right tail of the z_{K1} distribution for nnd.e reveals the presence of bad matching units pairs.

Figure 2.5: Simulation 1, distributions of z_{K1} in nnd imputation (with don. cl.)

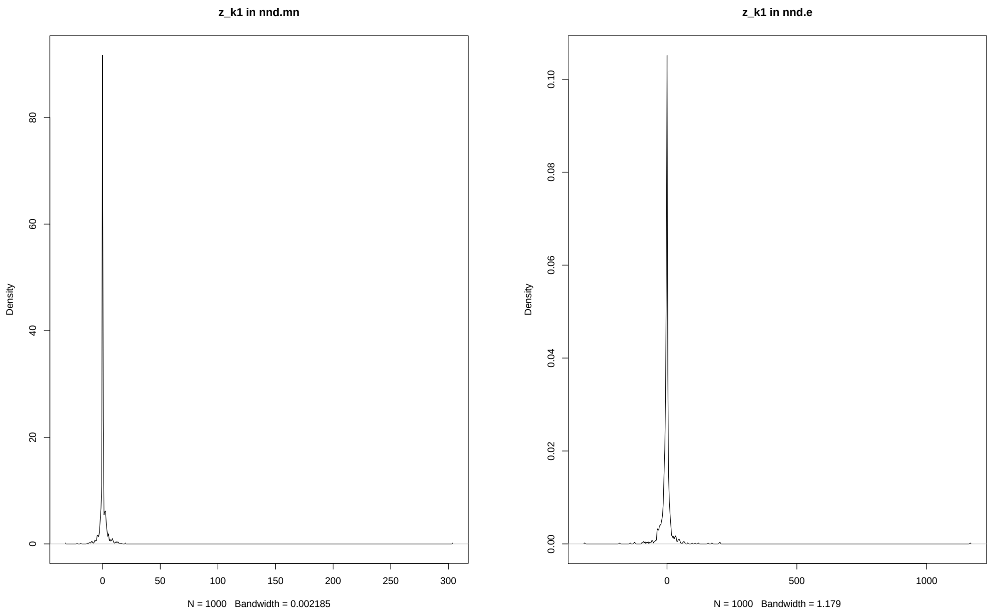
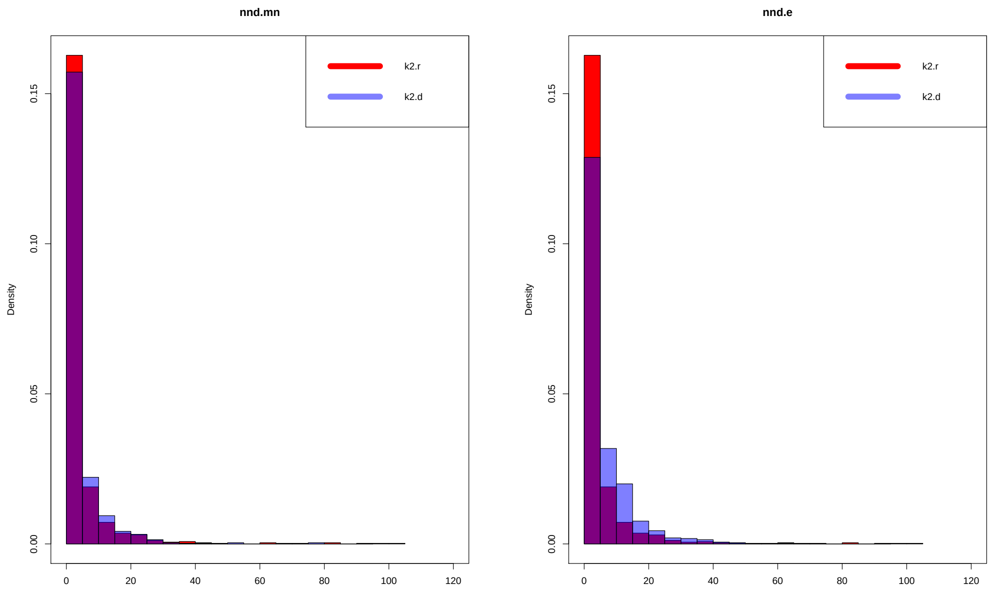


Figure 2.6 shows imputation (with donation classes) results, in terms of the different distributions of the original variable K_2^R and the imputed variable K_2^D in the synthetic datasets generated using the same above-mentioned combinations. We can see, again, a better performance of combination nnd.mn (and nnd.ms), which generate a good synthetic dataset with a small not significant underestimate of the class 0-5 and a small overestimate of variable K_2^R in classes 5-10 and 10-15, but an overall good representation. The nnd.e combination instead, generates a synthetic dataset in which the variable K_2^R in the class 0-5 is underestimated and there is an evident tendency to evidently overestimate the other values up to value 40.

Figure 2.6: Simulation 1, distributions of K_2^R, K_2^D in nnd imputation (with don. cl.)



Taking into account the distributions of differences z, figure 2.7 shows how both the combinations nnd.mn (and nnd.ms), and nnd.e for the variable K_2^R perform better than the above-mentioned ones for K_1^R. This is probably due to the smaller variance of the variable K_2^R with respect to K_1^R, so that matching units pairs are better associated, differences among them are closer and the z_{K2} distributions are almost 0-centred.

Figure 2.7: Simulation 1, distributions of z_{K2} in nnd imputation (with don. cl.)

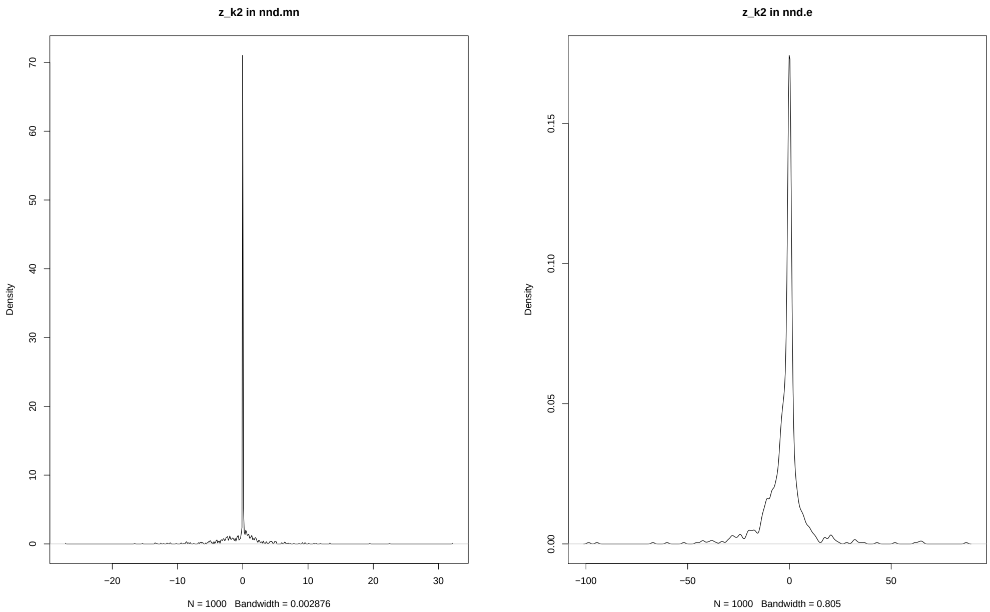
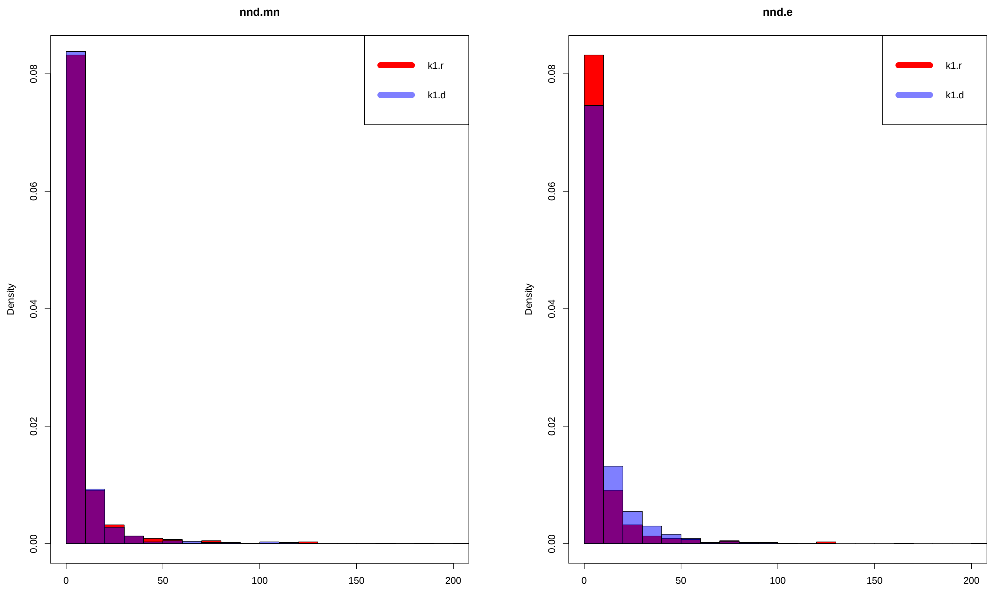


Figure 2.8 shows imputation (without donation classes) results, in terms of the different distributions of the original variable K_1^R and the imputed variable K_1^D in the synthetic datasets generated by combinations nnd.mn and nnd.e. Results of the combination nnd.ms are omitted. We can see that nnd.mn (and nnd.ms), generate a good synthetic dataset in terms of the overlap between variables K_1^R and K_1^D (there is a clear underestimate of the variable K_1^R in the class 40-50 and slightly overestimates of its high values). The combination nnd.e instead, generates a synthetic dataset in which the variable K_1^R in the class 0-10 are slightly underestimated whereas there is a tendency to overestimate the recipient variable K_1^R up to value 50.

Figure 2.8: Simulation 1, distributions of K_1^R, K_1^D in nnd imputation (without don. cl.)



Taking into account the distributions of differences z, figure 2.9 shows that the combination nnd.mn (and nnd.ms), perform far better than the combination nnd.e, allowing also a far better control of the outliers. It is also evident, anyway, that the quality of the matching units pairs and the control of the outliers are not as good as with respect to the imputation with the donation classes building.

Figure 2.9: Simulation 1, distributions of z_{K1} in nnd imputation (without don. cl.)

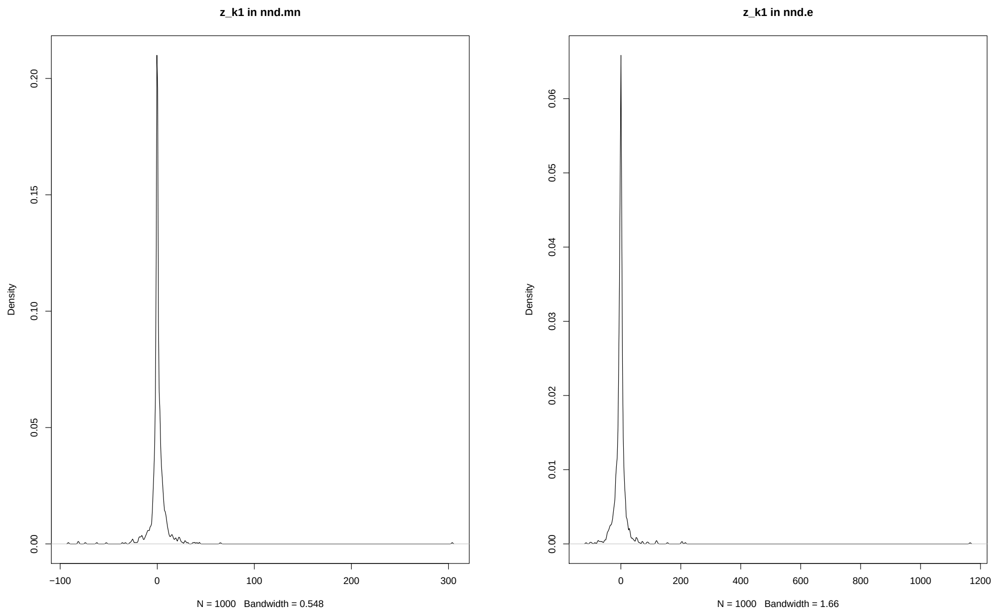
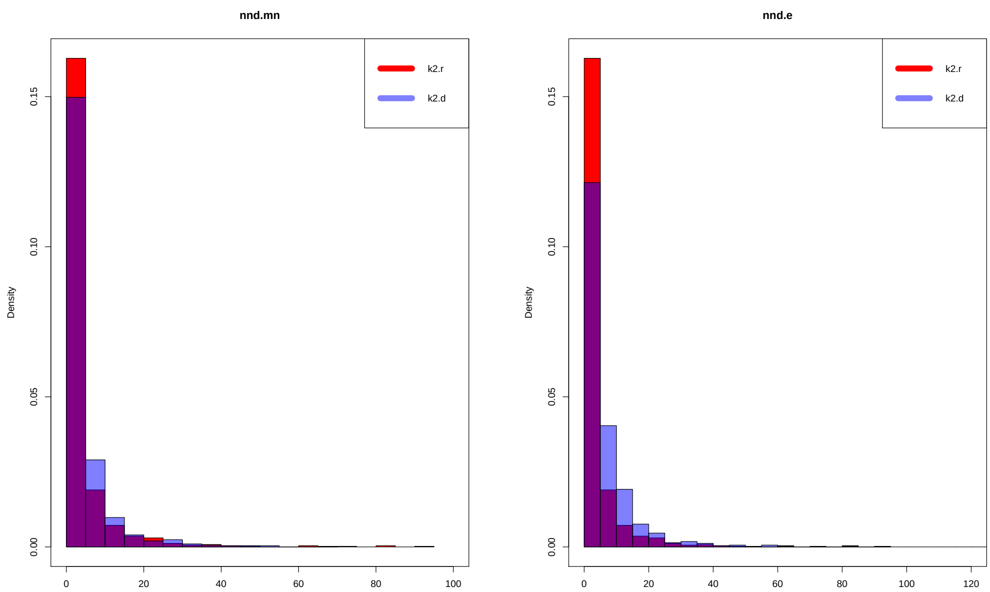


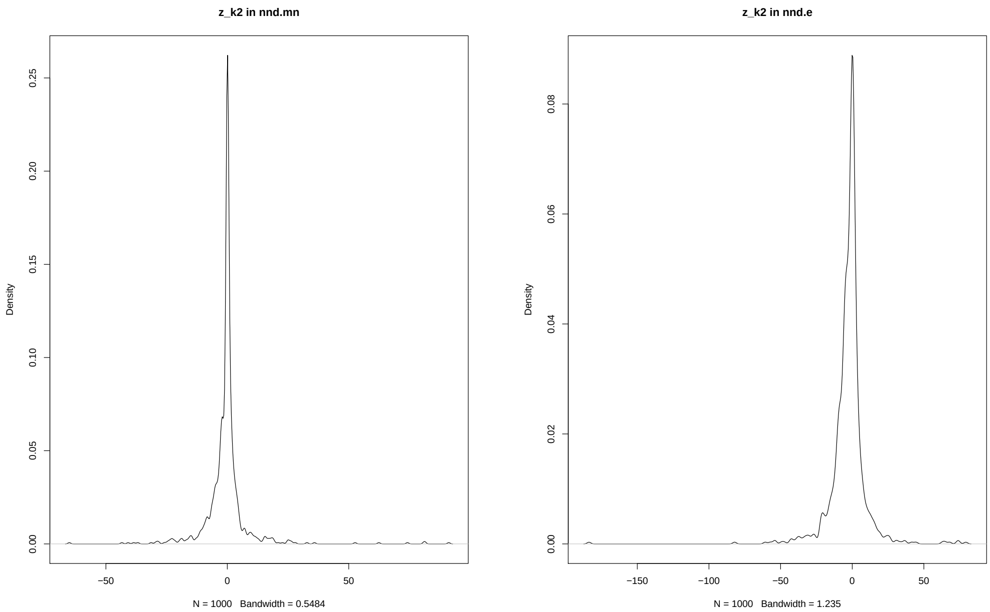
Figure 2.10 shows imputation results (without donation classes), in terms of the different distributions of the original variable K_2^R and the imputed variable K_2^D for combinations nnd.mn (and the omitted nnd.ms), and nnd.e. The latter two combinations generate a synthetic dataset with a small not significant underestimate of class 0-5 but an overestimate of variable K_2^R in class 5-10. More significant is the nnd.e combination overestimate of K_2^R which is doubled in the classes 5-10, 10-15, 15-20.

Figure 2.10: Simulation 1, distributions of K_2^R, K_2^D in nnd imputation (without don. cl.)



Taking into account the distributions of differences z, figure 2.11 shows how both the combinations nnd.mn (and nnd.ms), and nnd.e for the variable K_2^R perform not so good with respect to the matching units pairs, with a clearer tendency of the combination nnd.e to not even properly control for the outliers.

Figure 2.11: Simulation 1, distributions of z_{K2} in nnd imputation (without don. cl.)



For sake of brevity, distributions of K_1^R, K_1^D and K_2^R, K_2^D in the synthetic datasets generated by combinations nndc.mn, nndc.ms and nndc.e, and the respective differences z_{K1}, z_{K2} distributions, are omitted (both the imputations with and without donation classes), because they generate results which are highly similar to the combinations with the unconstrained SM imputation technique (i.e. the Nearest Neighbour Distance Hot Deck one). Anyway, we stress that combinations within nndc, in the case of donation classes building, show an overall tendency to slightly reduce the overestimates of both the variables K_1^R and K_2^R.

Figure 2.12 shows imputation (with donation classes) results, in terms of the different distributions of the original variable K_1^R and the imputed variable K_1^D in the synthetic datasets generated by combinations rnd.mn, rnd.ms, rnd.e. We can see that both combinations rnd.mn and rnd.ms generate a good synthetic dataset in terms of the overlap between variables K_1^R and K_1^D with an overall tendency to not exceed in the (under)overestimates of the variable K_1^R, almost by the rnd.ms combination. The combination rnd.e instead, generates a synthetic dataset with a clear presence of overestimates of K_1^R (for example in the class 10-20 which is doubled).

Figure 2.12: Simulation 1, distributions of K_1^R, K_1^D in rnd imputation (with don. cl.)

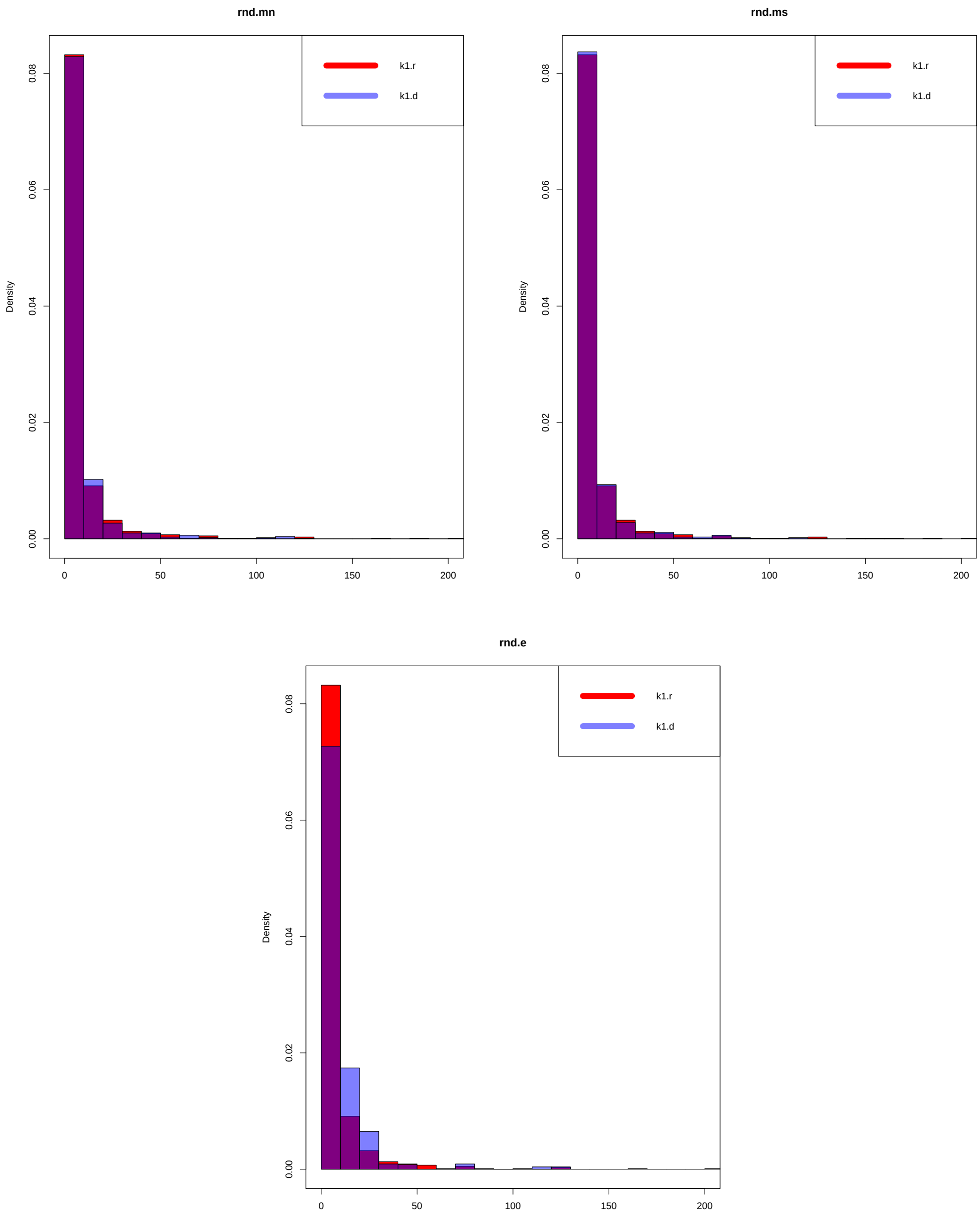


Figure 2.13 instead, shows how the above-mentioned combinations do not perform well in controlling the outliers with respect to the variable K_1^R.

Figure 2.13: Simulation 1, distributions of z_{K1} in rnd imputation (with don. cl.)

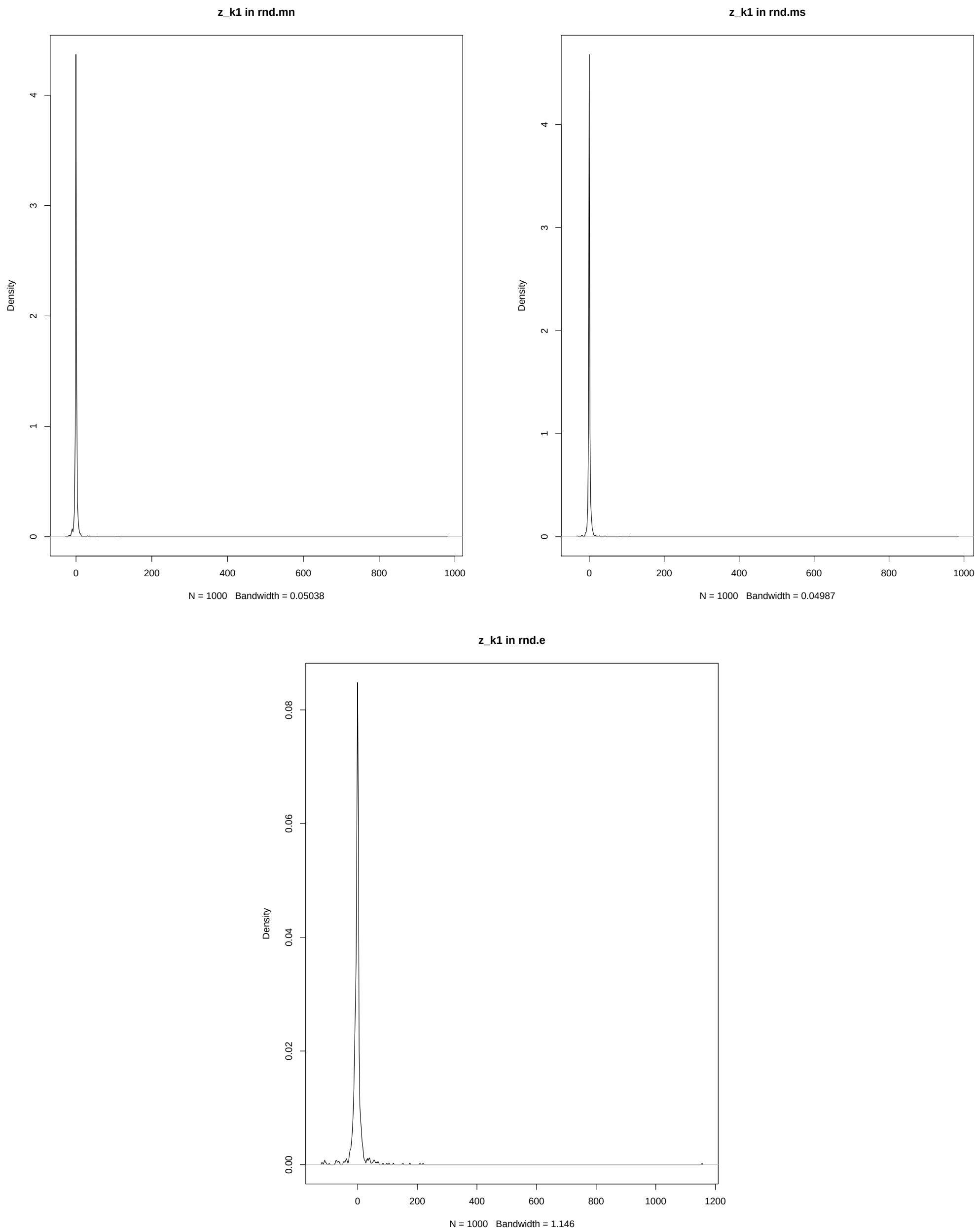


Figure 2.14 shows imputation (with donation classes) results for the variable K_2^R in the synthetic datasets generated using the above-mentioned com-

binations. The synthetic dataset generated presents a good overlap tendency between the variables K_2^R and K_2^D (probably even due to the far lower variance of the variable K_2^R with respect to K_2^D, than the variance for K_1^R, K_1^D). Nevertheless, it is true with respect to combinations rnd.mn (and the omitted rnd.ms); combination rnd.e indeed, generates a synthetic dataset in which K_2^R is clearly overestimated (except for the class 0-5 which show a not significant underestimate).

Figure 2.14: Simulation 1, distributions of K_2^R, K_2^D in rnd imputation (with don. cl.)

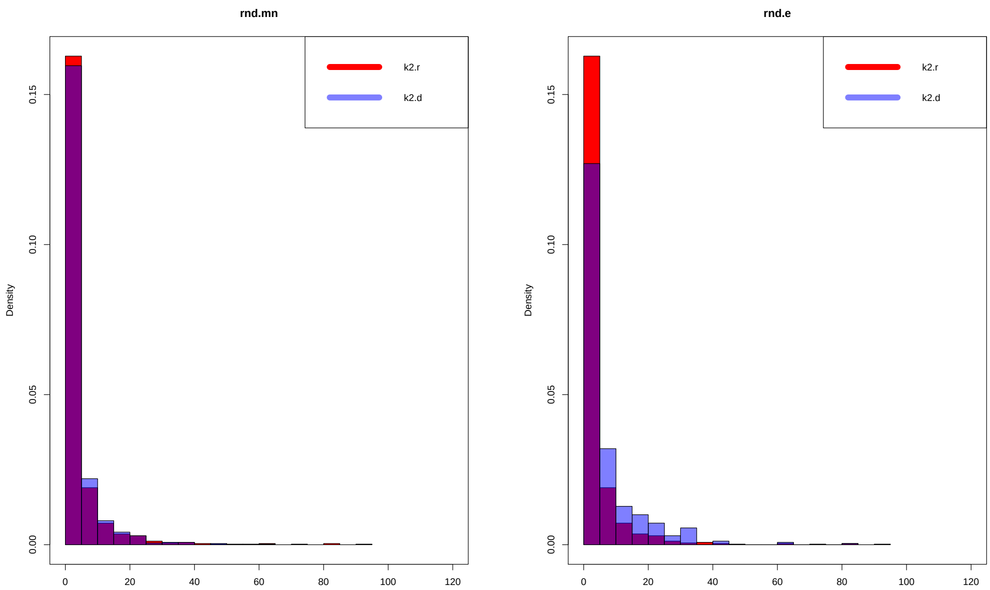


Figure 2.15 shows how differences z_{K2} are clearly better than the distributions of z for K_1^R; anyway, taking into account the far lower difference between the variances of the variables K_2^R, K_2^D the matching units pairs are not sufficiently closer and the differences between K_2^R and K_2^D tend not to be perfectly 0-centred, indicating a not good control of the outliers.

Figure 2.15: Simulation 1, distributions of z_{K2} in rnd imputation (with don. cl.)

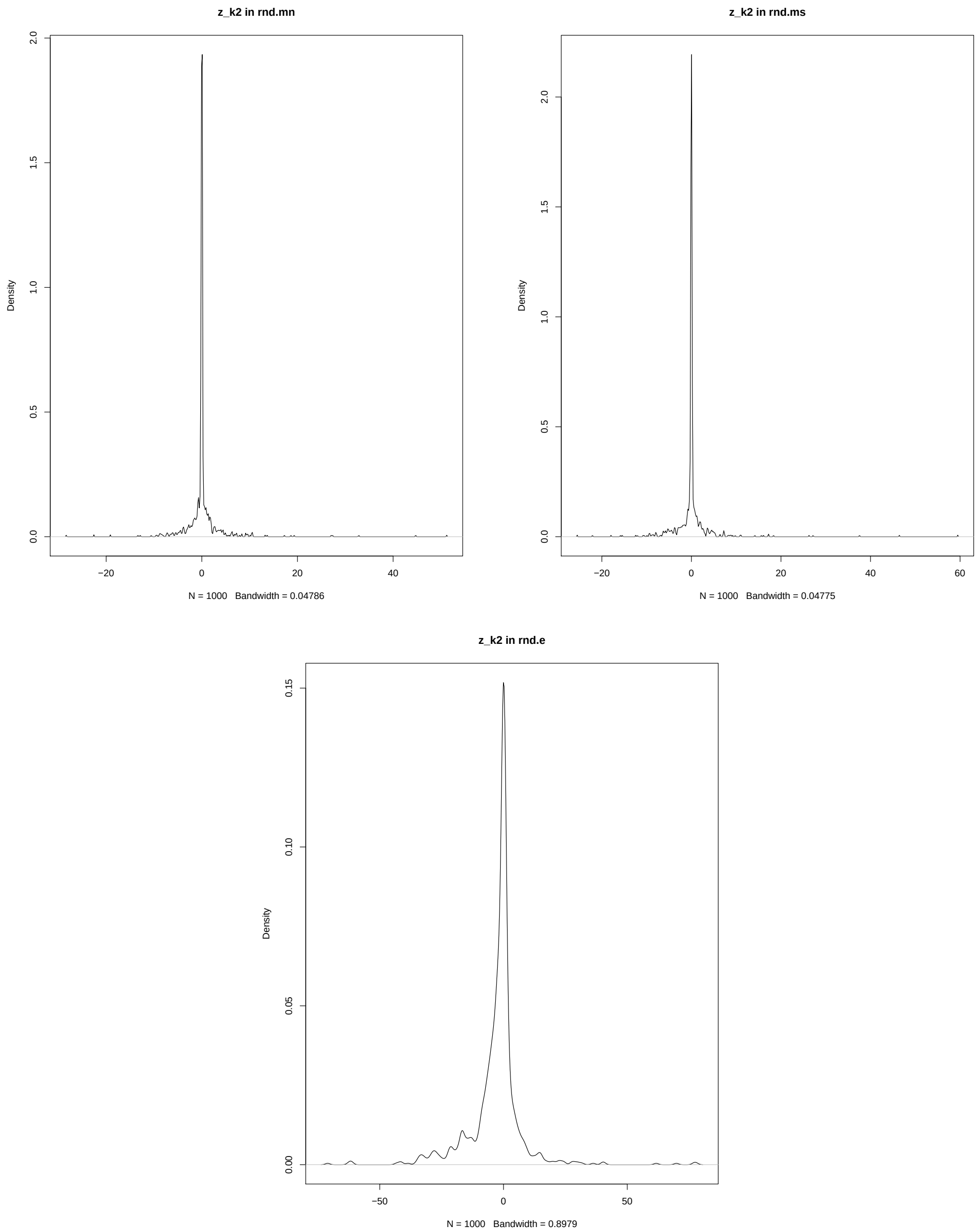


Figure 2.16 shows imputation (without donation classes) results, in terms of the different distributions of the original variable K_1^R and the imputed

variable K_1^D in the synthetic datasets generated by combinations rnd.mn, rnd.ms, rnd.e. We can see that both combinations rnd.mn and rnd.ms generate a synthetic dataset with under(over)estimates of K_1^R (for example, the former slightly underestimates the variable K_1^R in the classes 10-20 and 40-50, slightly overestimating the classes 60-70, 80-90 and 90-100 whereas the latter slightly underestimates the classes 10-20, 20-30, slightly overestimating the classes 60-70, and from the value 80 to value 110). Really bad results are generated by the combination rnd.e which generates a synthetic dataset with a clear presence of overestimates of the variable K_1^R in the classes 10-20 and 20-30 (for which K_1^R is doubled), and relevant underestimates in the classes ranging from the value 30 up to value 70.

Figure 2.16: Simulation 1, distributions of K_1^R, K_1^D in rnd imputation (without don. cl.)

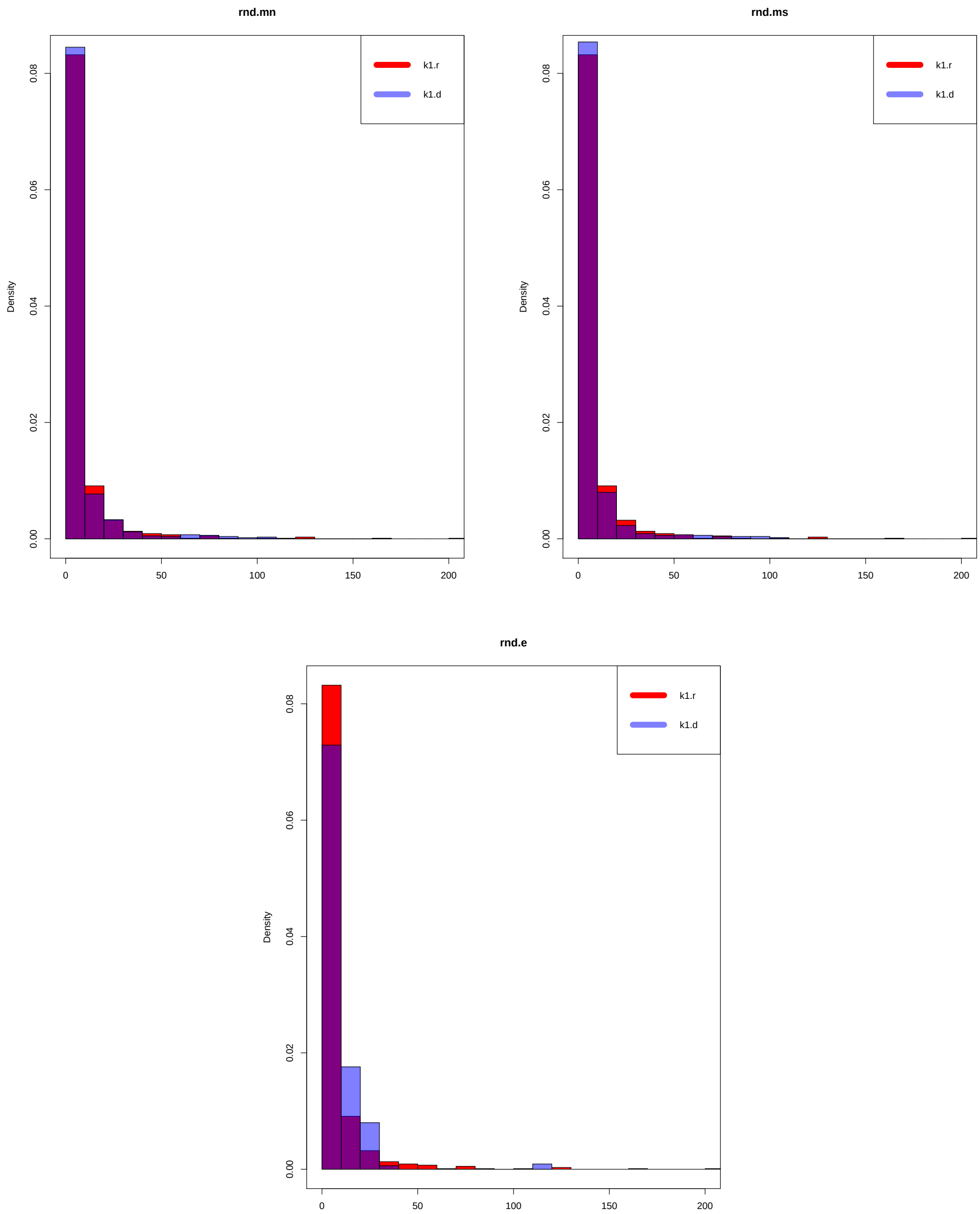
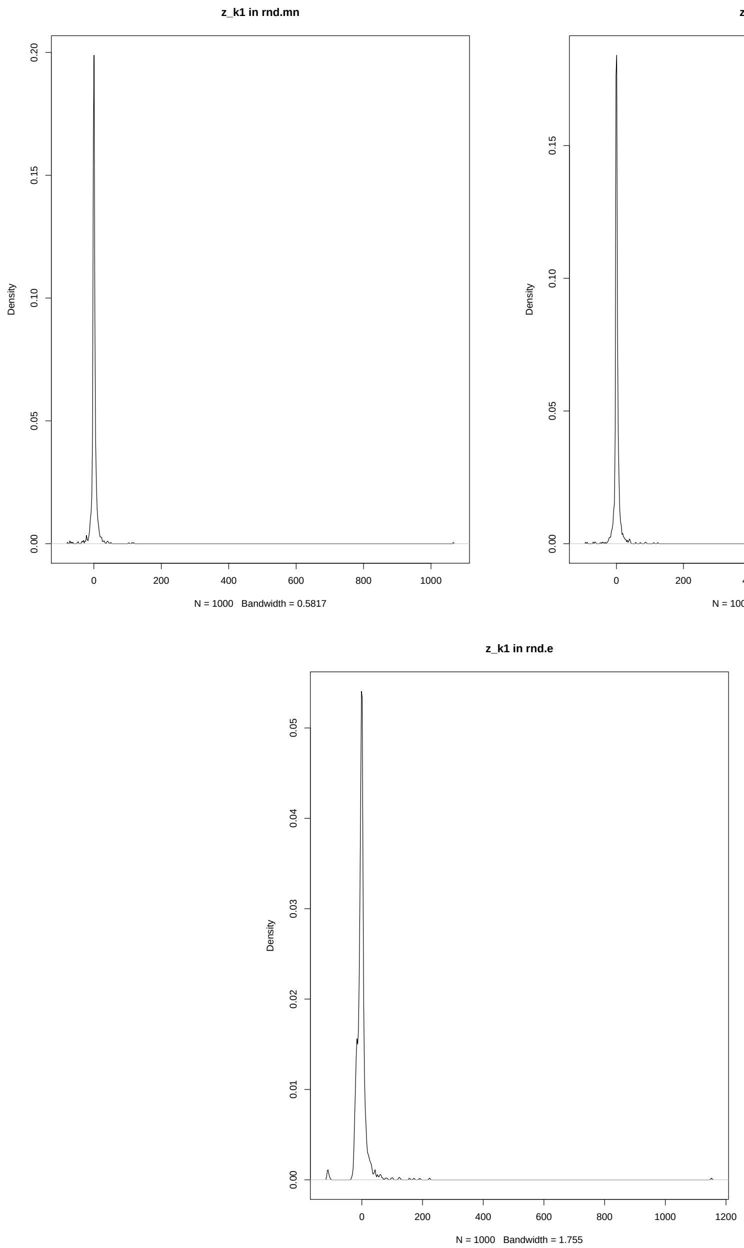


Figure 2.17 shows how the above-mentioned combinations clearly do not perform at all well in the control of the outliers with respect to K_1^R.

Figure 2.17: Simulation 1, distributions of z_{K1} in rnd imputation (without don. cl.)



Imputation results from the above-mentioned combinations do not even get better in the synthetic dataset generated with respect to K_2^R and K_2^D.

Indeed, figure 2.18 shows that combination rnd.mn (and rnd.ms which is really similar, then omitted), generate a synthetic dataset in which K_2^R is slightly overestimated but the overall tendency shows a good overlap between K_2^R and K_2^D. Nevertheless, combination rnd.e overestimates almost all the values of K_2^R.

Figure 2.18: Simulation 1, distributions of K_2^R, K_2^D in rnd imputation (without don. cl.)

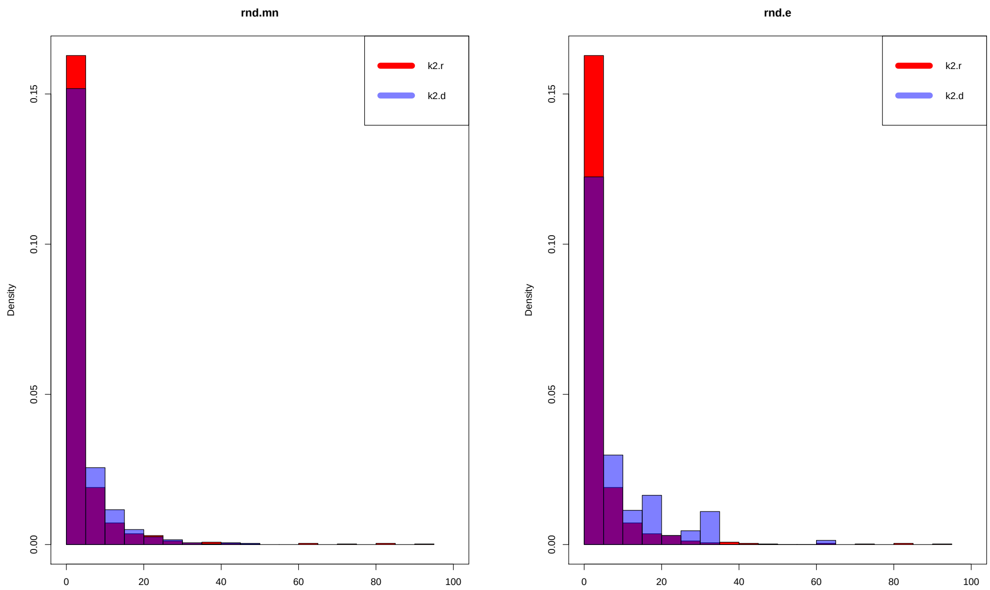


Figure 2.19 shows differences z for the variable K_2^R with an evident performance decrease with respect to the same combinations applied with the donation classes building.

Figure 2.19: Simulation 1, distributions of z_{K2} in rnd imputation (without don. cl.)

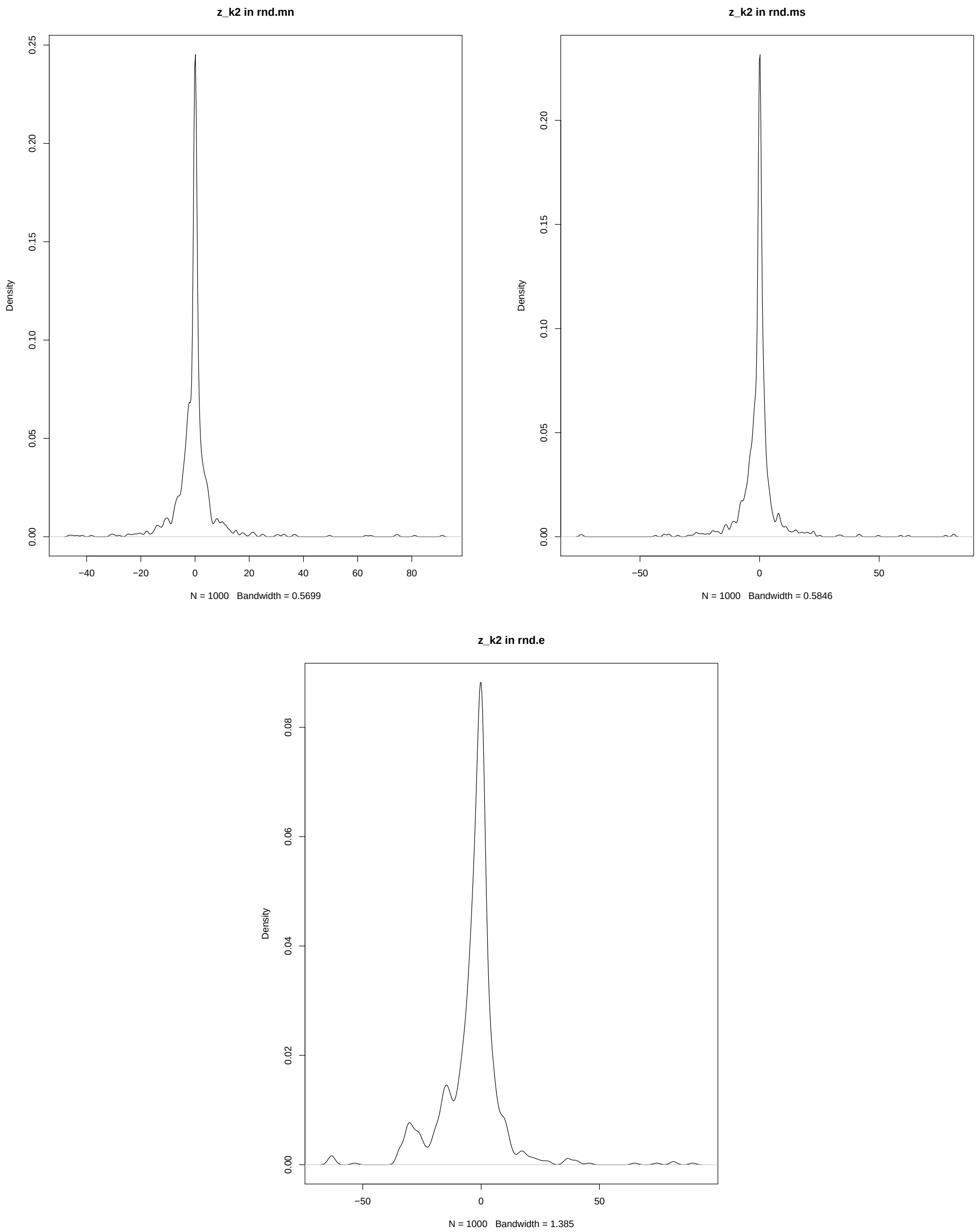


Figure 2.20 shows imputation (with donation classes) results in terms of the different distributions of the original variable K_1^R and the imputed

variable K_1^D in the synthetic datasets generated using rnk technique, both for the variable k_1^R and the variable k_2^R. This technique generates not really good synthetic datasets, in which there is a clear tendency to overestimate k_1^R and even more k_2^R.

Figure 2.20: Simulation 1, distributions of K_1^R, K_1^D in rnk imputation (with don. cl.)

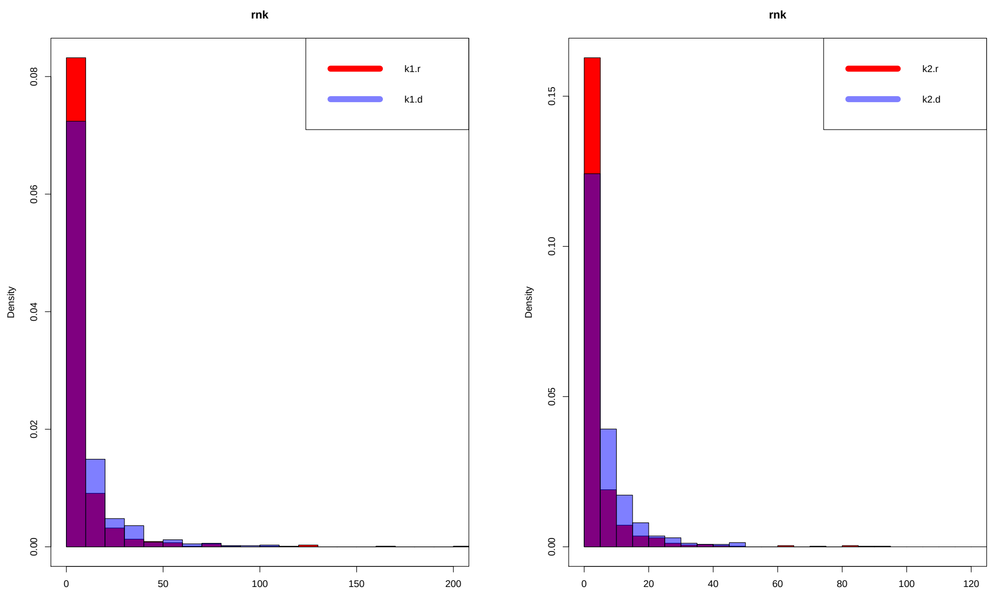
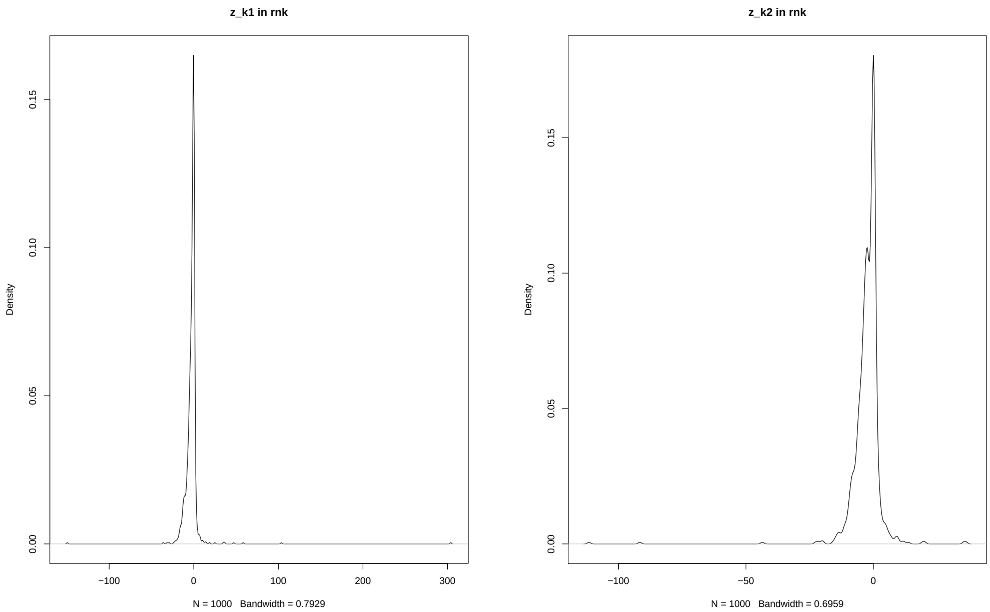


Figure 2.21 confirms that the rnk technique does not control the outliers and does not guarantee a good matching pair for units, both considering the variable K_1^R and the variable K_2^R.

Figure 2.21: Simulation 1, distributions of z_{K1}, z_{K2} in rnk imputation (with don. cl.)



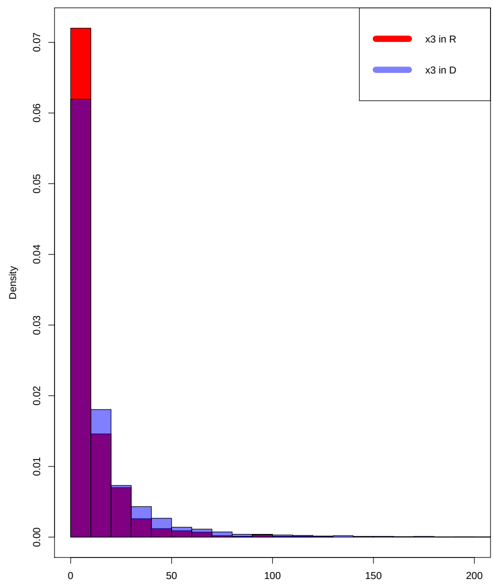
For sake of brevity, distributions of K_1^R, K_1^D and K_2^R, K_2^D in the synthetic datasets generated by rnk, and the respective differences z_{K1}, z_{K2} distributions, are omitted since they basically show results similar to the above-mentioned ones, even with a relevant decrease of imputation goodness.

2.3.2 Results from simulation 2

Figure 2.22 shows that from simulation 2 we have the recipient dataset R and the donor dataset D characterised, with respect to the matching variable X_3, by a higher variance and a noteworthy presence of outliers in the donor dataset D. With respect to simulation 1, simulation 2 characterises the recipient dataset R and the donor dataset D also by the significant difference in means of the matching variable X_3. We notice that with the sensible exception of the class 0-10, variable X_3^D always overcome variable X_3^R due to the bigger dimensionality of the donor dataset D.

Figure 2.22: Simulation 2, variable X_3 in R and D

	X_3	
	R	D
mean	9.552	15.183
var	196.919	1074.066
min	0.061	0.032
max	124.824	1760.312



Taking into account the imputation variable K_1 in datasets R and D, beyond the difference between the upper values of K_1^R and K_1^D due to the much lower maximum value of the variable K_1^R, figure 2.23 shows that there is an overall almost equally correspondence between R (recipient) and D (donor). With respect to the imputation variable K_2 in datasets R and D, figure 2.23 shows that, with the exception of the higher frequency of the variable K_2^R in the class 0-5, there is always a complete over-correspondence for the other values of K_2.

Figure 2.23: Simulation 2, K_1 and K_2 in R and D

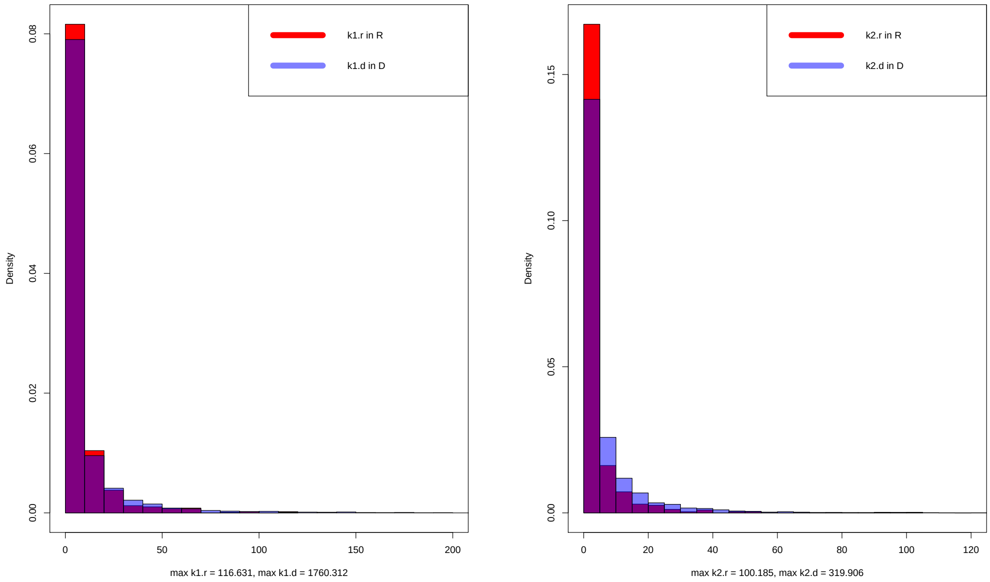
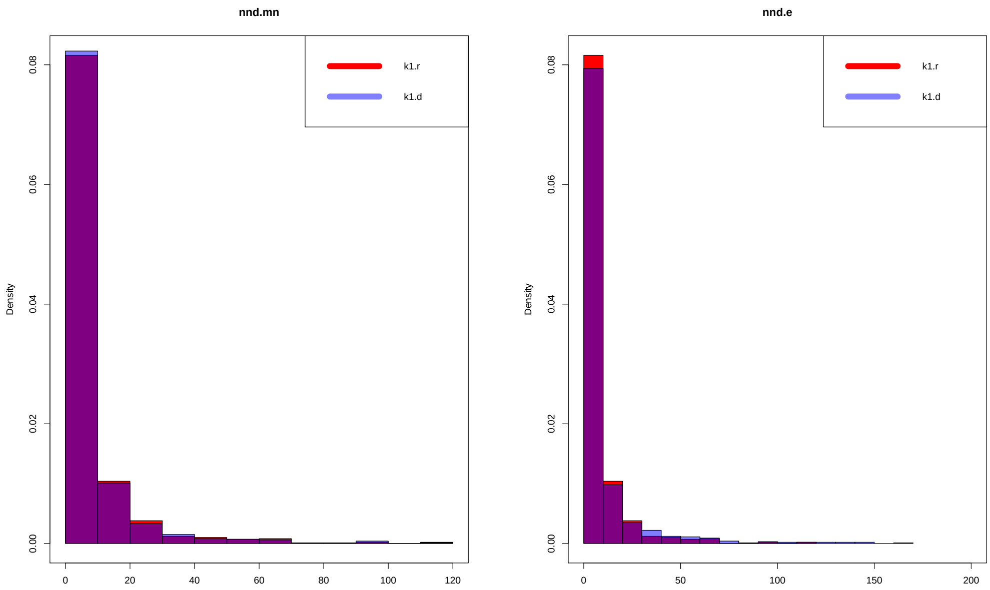


Figure 2.24 shows imputation (with donation classes) results for the combinations nnd.mn (nnd.ms is really similar, then omitted), and nnd.e. Both the combinations generate really good synthetic datasets in terms of the overlap between the variables K_1^R and K_1^D.

Figure 2.24: Simulation 2, distributions of K_1^R, K_1^D in nnd imputation (with don. cl.)



Taking into account the distributions of differences z, figure 2.25 shows how the combinations nnd.mn (and nnd.ms), and nnd.e perform well in controlling the outliers values, being both almost 0-centred. Also the matching units pairs present an association of really close K_1^R and K_1^D variables values.

Figure 2.25: Simulation 2, distributions of z_{K1} in nnd imputation (with don. cl.)

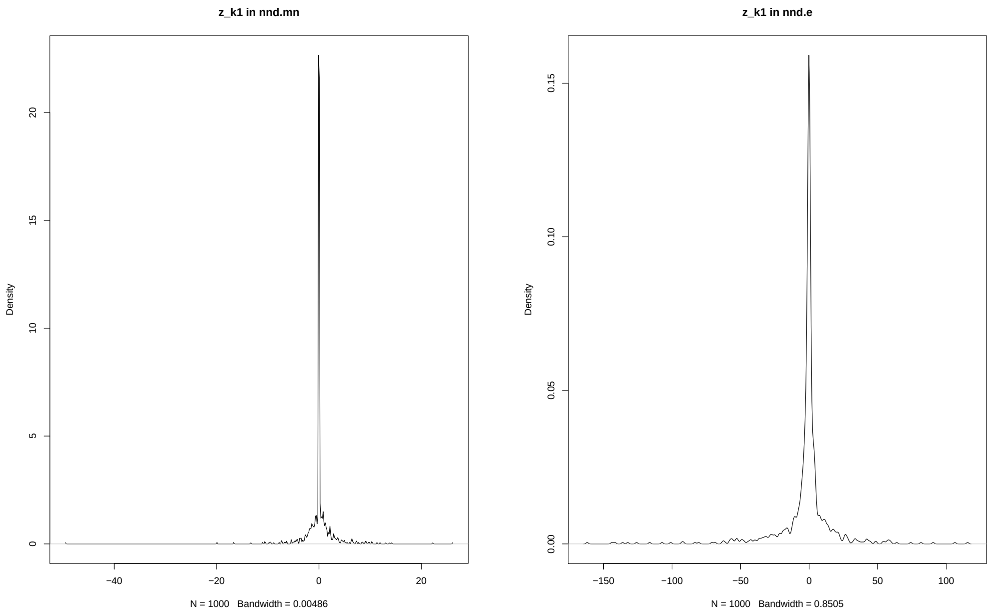


Figure 2.26 shows imputation (with donation classes) results for the variable K_2^R with the above-mentioned combinations. With respect to K_2^R, these combinations show a trend really more similar to the ones showed by the same combinations in simulation 1. Indeed, there is a better performance of nnd.mn (and the really similar nnd.ms), combinations (which slightly (under)overestimate the variable K_1^R), than the combination nnd.e. This one generates a synthetic dataset in which overestimates are clearly more significant.

Figure 2.26: Simulation 2, distributions of K_2^R, K_2^D in nnd imputation (with don. cl.)

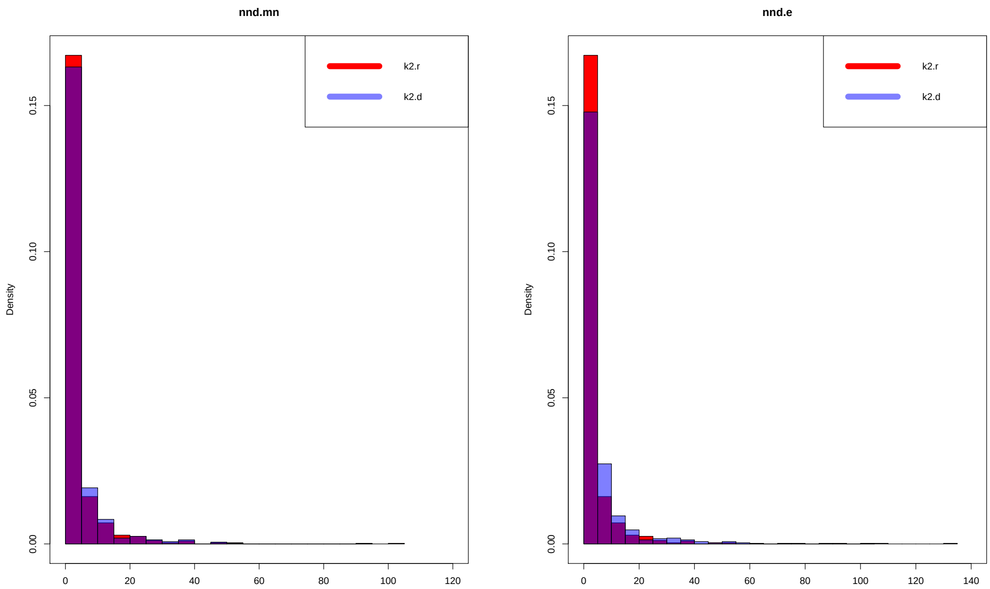


Figure 2.27 shows a slightly better performance of the combinations nnd.mn (and nnd.ms), and nnd.e with respect to the variable K_2^R if we take into account results for the distributions of the differences z referred to the variable K_1^R (slightly 0-centred, i.e. they control well the outliers).

Figure 2.27: Simulation 2, distributions of z_{K2} in nnd imputation (with don. cl.)

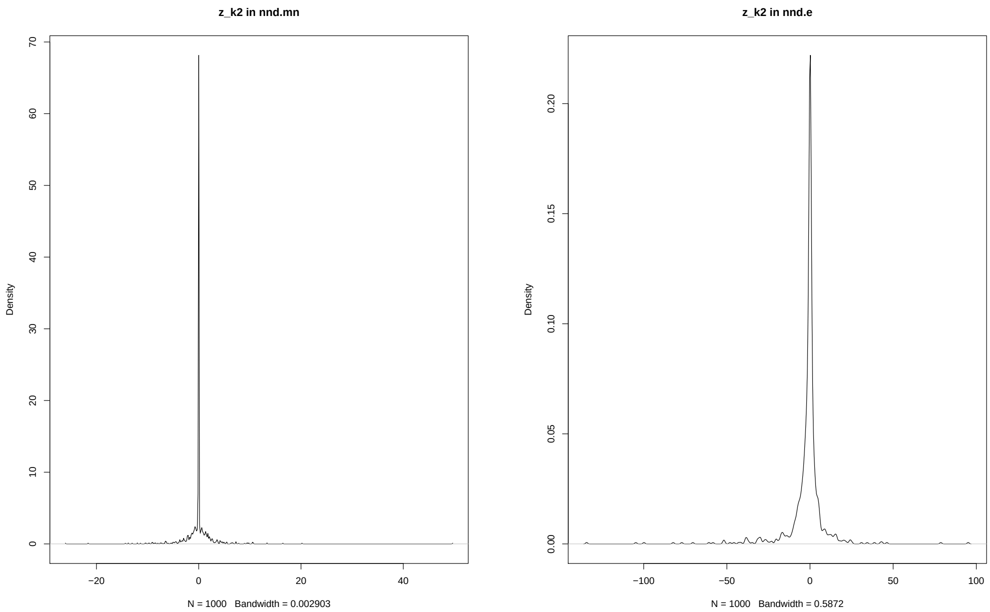
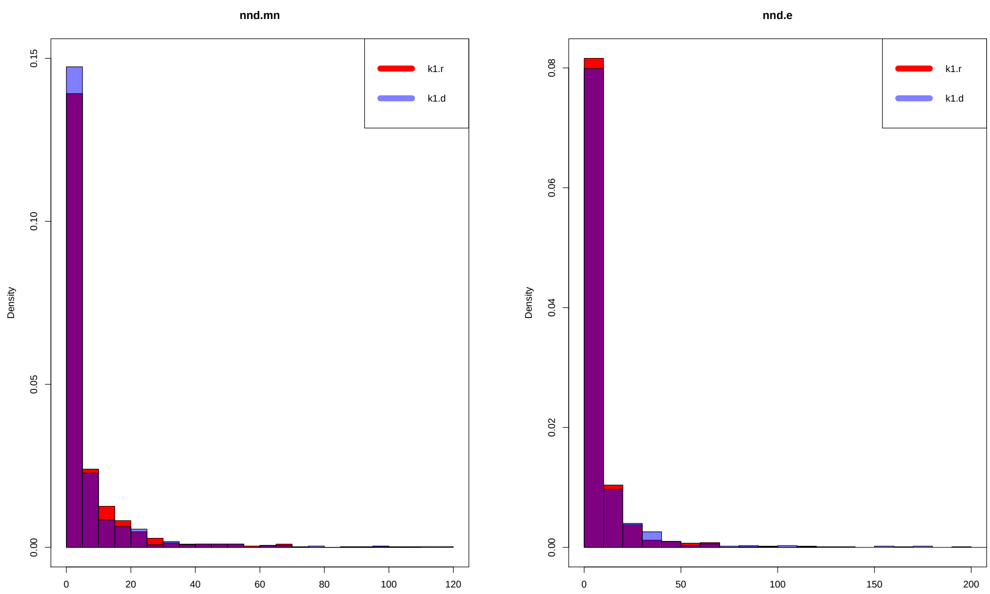


Figure 2.28 shows imputation (without donation classes) results, in the synthetic datasets generated from combinations nnd.mn (and nnd.ms), and nnd.e. They generate slightly worse synthetic datasets if we take into account the same combinations applied building donation classes. Indeed, almost for the combinations nnd.mn (and nnd.ms) there is a not significant but still present tendency to underestimate K_1^R.

Figure 2.28: Simulation 2, distributions of K_1^R, K_1^D in nnd imputation (without don. cl.)



Consequently, even distributions of the differences z, as figure 2.29 shows, present a not so good association of matching units pairs but a discrete control of the outliers.

Figure 2.29: Simulation 2, distributions of z_{K1} in nnd imputation (without don. cl.)

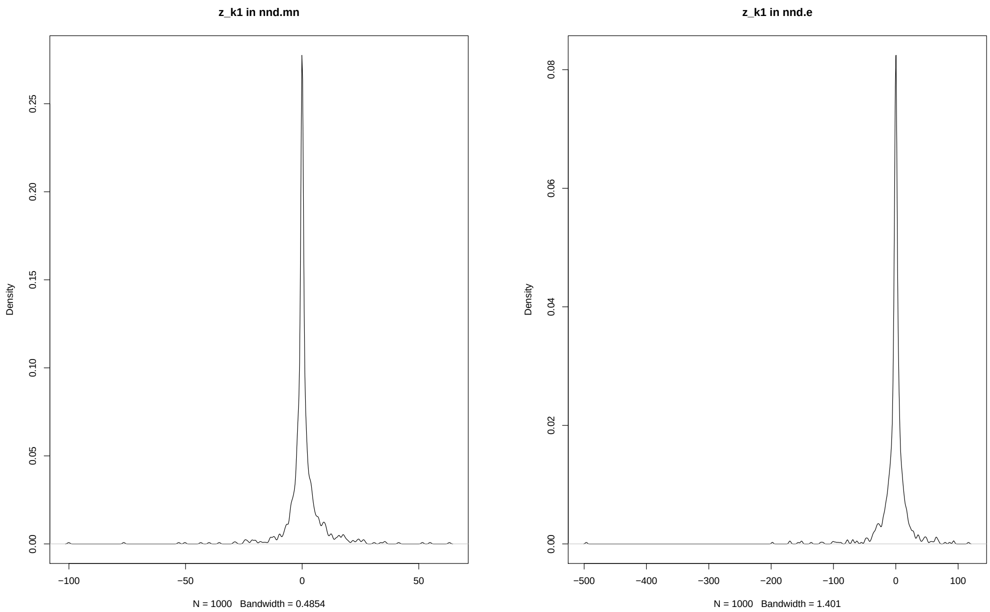


Figure 2.30 shows imputation (without donation classes) results, for variable K_2^R with the above-mentioned combinations. There is, again, a better performance of nnd.mn (and nnd.ms), which generate a good synthetic dataset with small not significant overestimates (for examples in the class 0-20). Surprisingly, even the combination nnd.e generates a synthetic dataset in which K_2^R is more overestimated but not significantly.

Figure 2.30: Simulation 2, distributions of K_2^R, K_2^D in nnd imputation (without don. cl.)

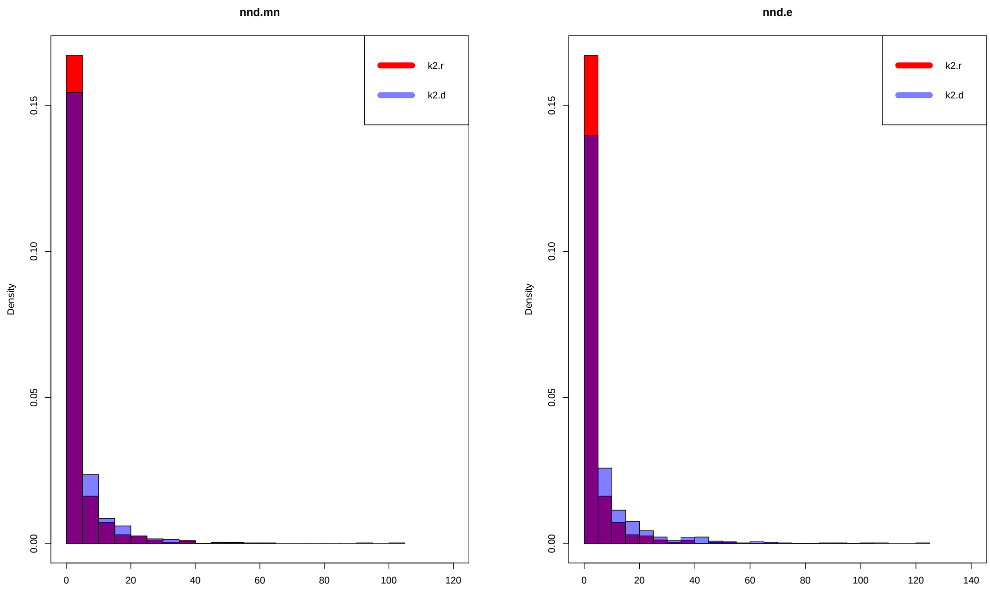
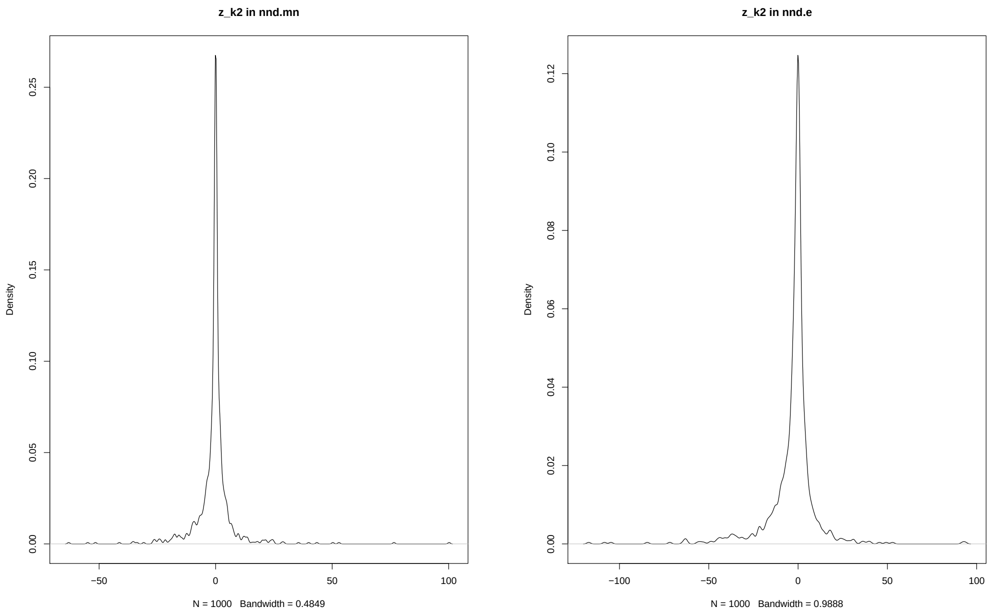


Figure 2.31 shows how the combinations nnd.mn (and nnd.ms) and nnd.e perform well with respect to the variable K_2^R (the differences distributions are both almost 0-centred).

Figure 2.31: Simulation 2, distributions of variables z_{K2} in nnd imputation (without don. cl.)

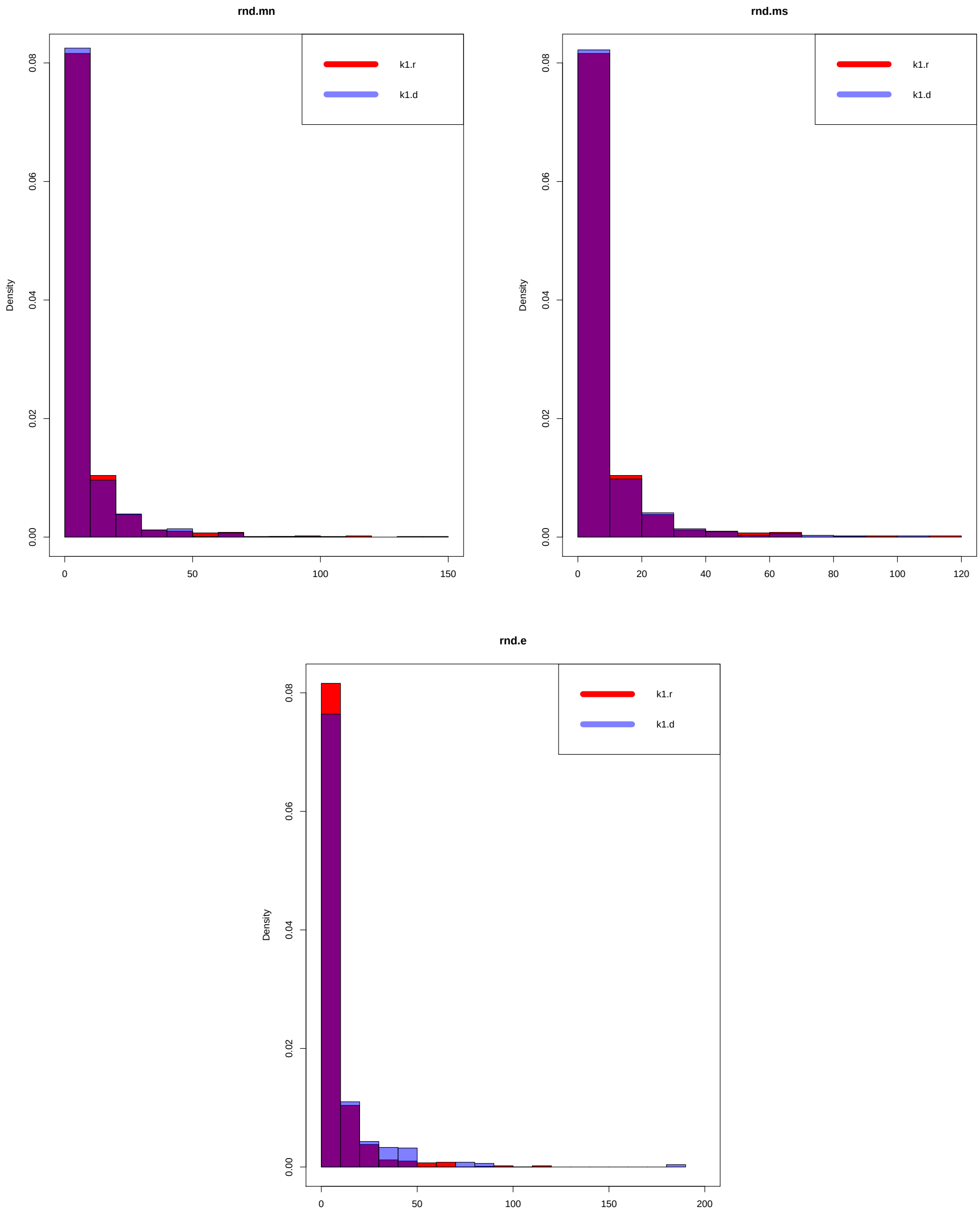


For sake of brevity, distributions of K_1^R, K_1^D and K_2^R, K_2^D in the synthetic datasets generated by combinations nndc.mn, nndc.ms and nndc.e, and the respective differences z_{K1}, z_{K2} distributions, are omitted (both the imputations with and without donation classes), because they generate results which are highly similar to the combinations with the unconstrained SM imputation technique (i.e. the Nearest Neighbour Distance Hot Deck one).

Figure 2.32 shows imputation (with donation classes) results for variable K_1^R using the combinations rnd.mn, rnd.ms, and rnd.e.These generate a good synthetic dataset in terms of the overlap between variables K_1^R and K_1^D with an overall tendency to properly estimate K_1^R. The combination rnd.e, anyway, presents a significant overestimate of K_1^R in the classes 30-40 and

40-50 (and a relevant underestimate for the classes 50-60 and 60-70).

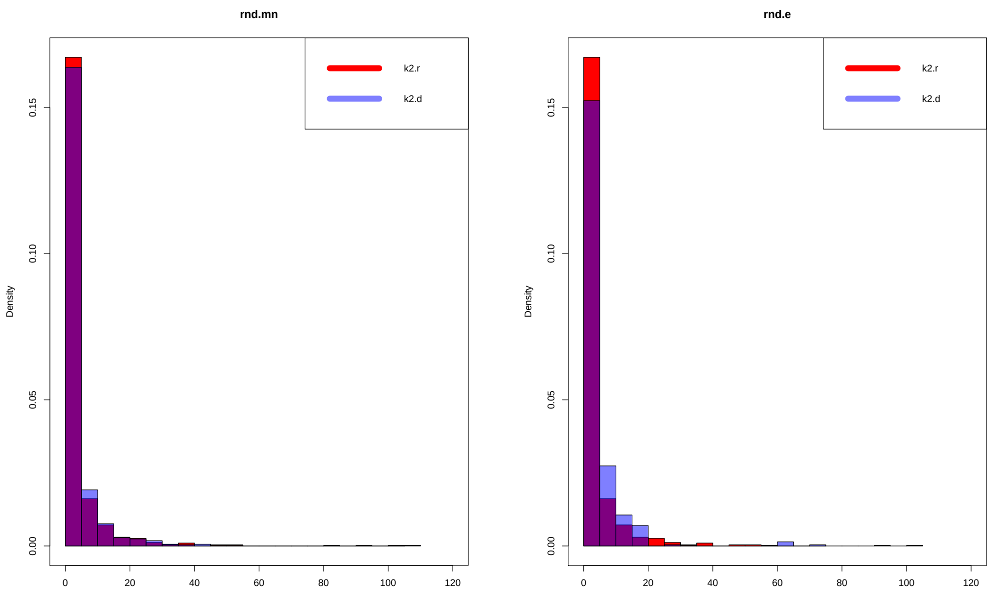
Figure 2.32: Simulation 2, distributions of K_1^R, K_1^D in rnd imputation (with don. cl.)



For sake of brevity, distribution of z_{K1} for the above-mentioned combinations is omitted since they do not represent a different tendency with respect to the one showed for the same combinations in simulation 1.

Figure 2.33 shows imputation results for combinations rnd.mn (and rnd.ms), and rnd.e with respect to the variable K_2^R. We can notice an overall good overlap between K_2^R and K_2^D for the combinations rnd.mn (and rnd.ms), while rnd.e tends both to overestimate the variable K_2^R and to clearly underestimate it in the classes 20-25, 25-30, 35-40.

Figure 2.33: Simulation 2, distributions of K_2^R, K_2^D in rnd imputation (with don. cl.)

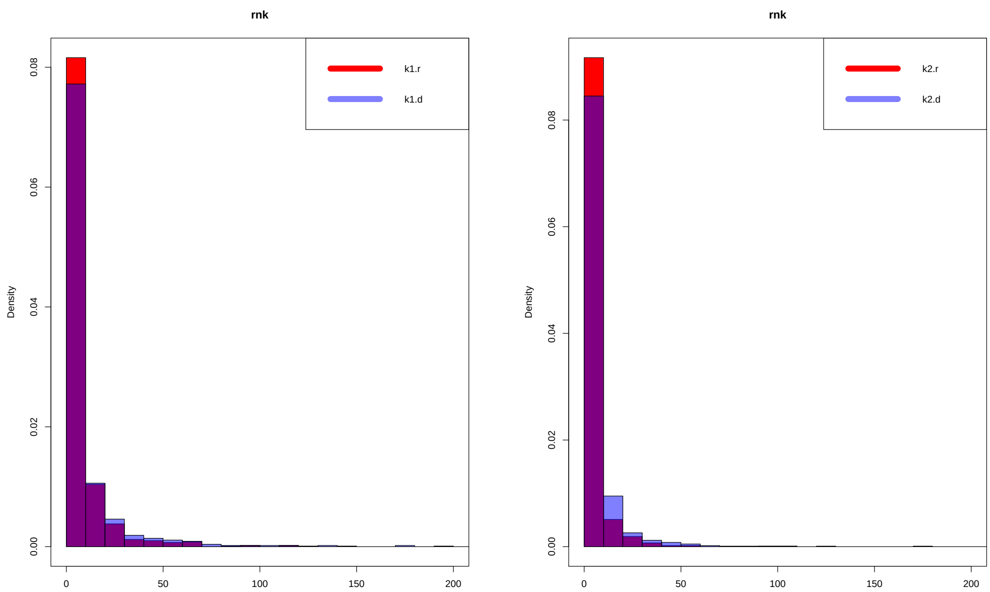


For sake of brevity, even distribution of z_{K2} for the above-mentioned combinations is omitted.

We decide to omit, for sake of brevity, the distributions of K_1^R, K_1^D and K_2^R, K_2^D for the above mentioned combinations applied without the donation classes; we omit also the distributions of the differences z_{K1} and z_{K2}. This is due to the fact that, generally, results from imputation without donation classes building related to these combinations are similar to the showed ones, just slightly worse in terms of the outliers control and of an overall tendency of overestimation.

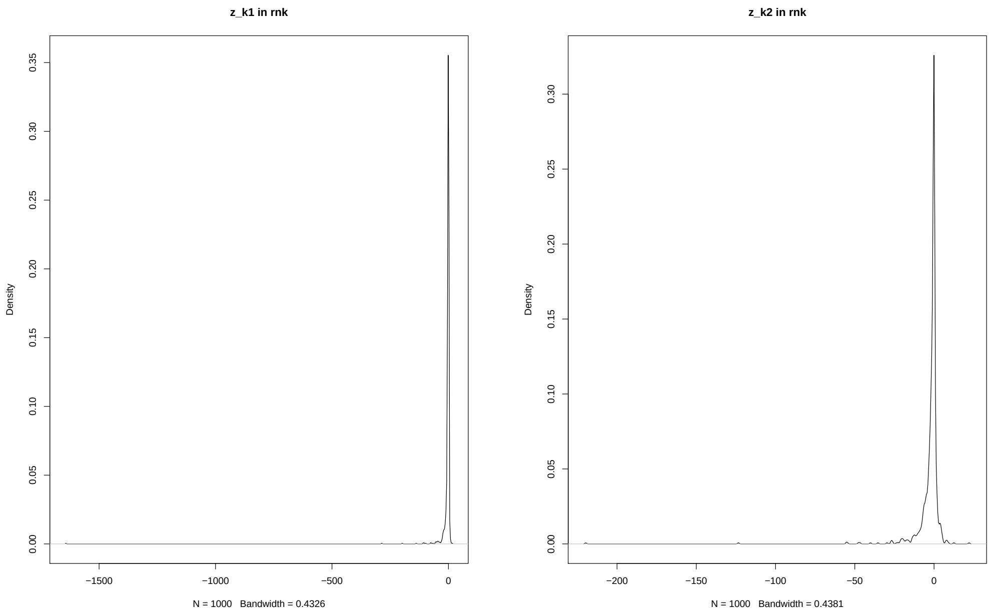
Figure 2.34 shows imputation (with donation classes) results for the variables K_1^R, K_2^R, obtained applying the rnk technique which, with the exception of the variable K_2^R (doubled just in the class 5-10), performs an overall good imputation.

Figure 2.34: Simulation 2, distributions of K_1^R, K_1^D and K_2^R, K_2^D in rnk imputation (with don. cl.)



Nevertheless, as figure 2.35 shows, rnk technique does not allow to properly control for the outliers, neither for the variable K_1^R nor for the variable K_2^R.

Figure 2.35: Simulation 2, distributions of z_{K1}, z_{K2} in rnk imputation (with don. cl.)

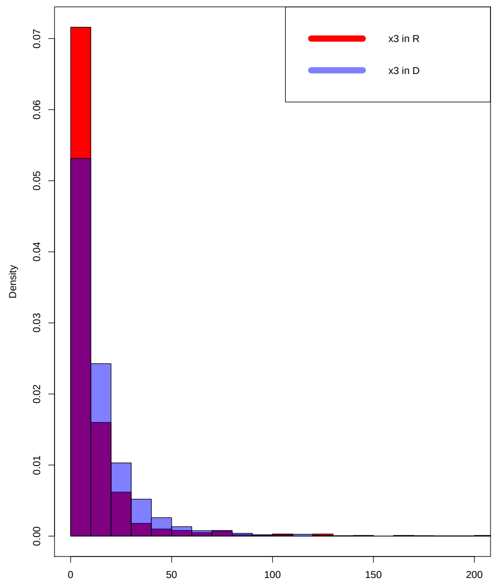


2.3.3 Results from simulation 3

Figure 2.36 shows that from simulation 3 we have the recipient dataset R and the donor dataset D characterised, with respect to the matching variable X_3, by a higher variance and a noteworthy presence of outliers in the recipient dataset R. We notice that, differently from the previous simulations, there is a significant difference between X_3^R and X_3^D with respect to the class 0-10, with a higher frequency of the variable X_3^R whereas for the other values we observe a proper over-correspondence of the variable X_3^D in the donor dataset D.

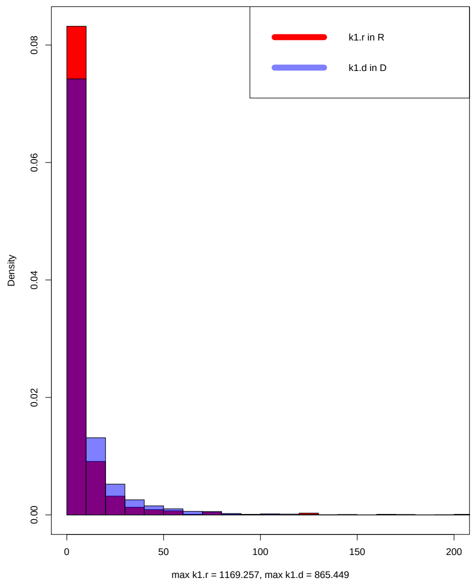
Figure 2.36: Simulation 3, variable X_3 in R and D

X_3		
	R	D
mean	11.574	15.584
var	1731.413	636.076
min	0.056	0.179
max	1172.981	874.083



Taking into account the imputation variable K_1 in datasets R and D, beside the difference between the upper values of K_1^R and K_1^D, figure 2.37 shows that there is a not significantly higher frequency of the variable K_1^R in the class 0-10 whereas there is a tendency of over-correspondence of the variable X_3^R in the donor dataset D. With respect to the imputation variable K_2 in datasets R and D, figure 2.37 shows that, with the exception of the higher frequency of the variable K_2^R in the class 0-5, there is always a complete over-correspondence for the K_2 variable between datasets R and D.

Figure 2.37: Simulation 3, K_1 and K_2 in R and D



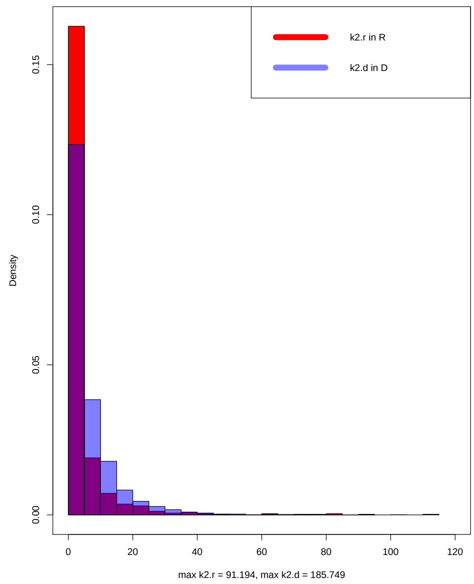
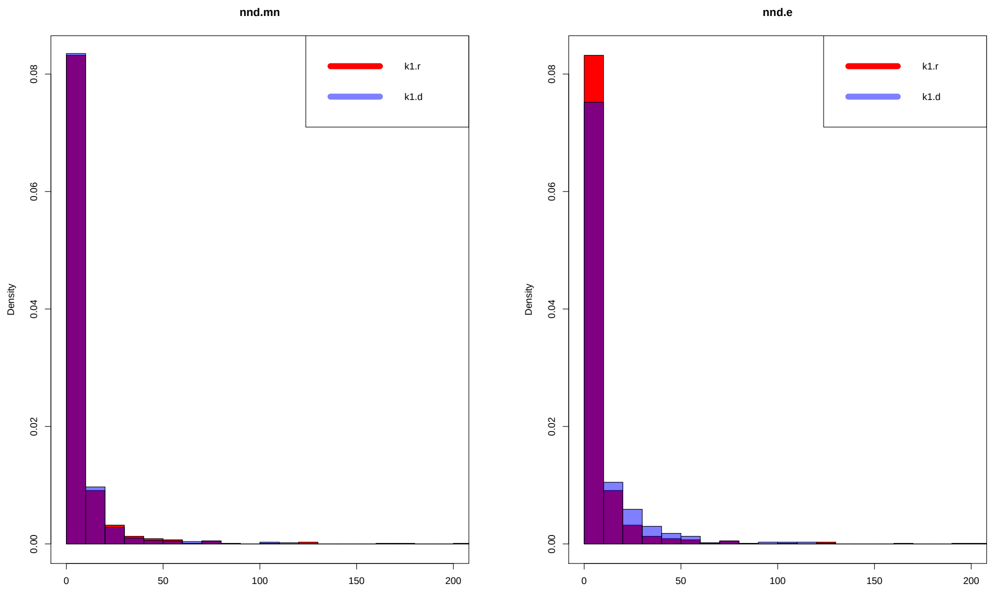


Figure 2.38 shows imputation (with donation classes) results, in terms of the different distributions of the variables K_1^R and K_1^D in the synthetic datasets generated from combinations nnd.mn and nnd.e. Results of the nnd.ms combination are omitted since they are really close to the ones generated by the nnd.mn combination. Figure 2.38 shows how both the combinations nnd.mn (and nnd.ms), and nnd.e generate good synthetic datasets, with a tendency of the latter combination to slightly overestimate K_1^R.

Figure 2.38: Simulation 3, distributions of K_1^R, K_1^D in nnd imputation (with don. cl.)



Taking into account the distributions of z_{K1} for the combinations nnd.mn (and the really similar nnd.ms), figure 2.39 shows a not so good capacity of them to properly control the outliers (which is even less good for the combination nnd.e).

Figure 2.39: Simulation 3, distributions of z_{K1} in nnd imputation (with don. cl.)

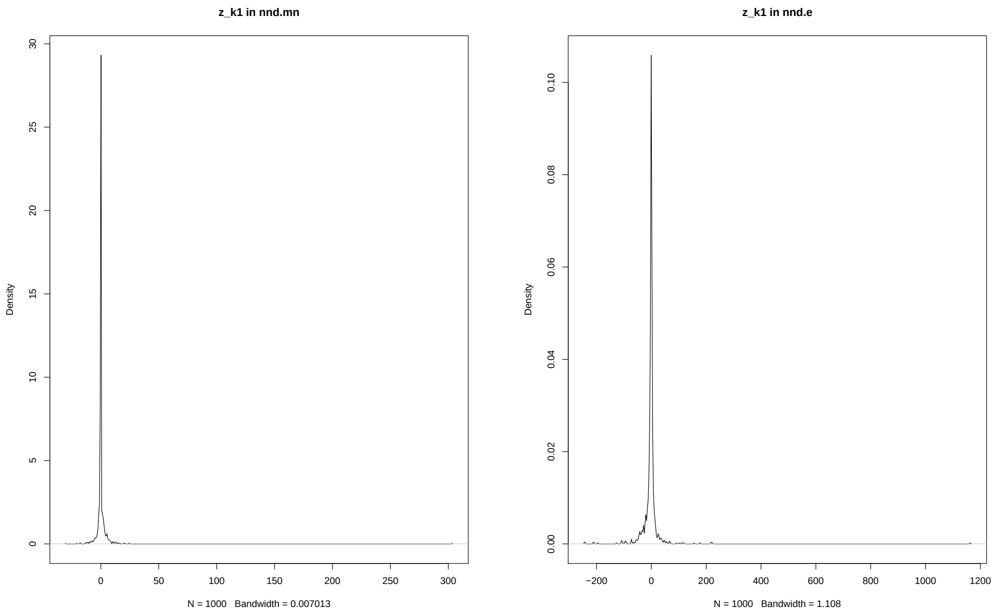
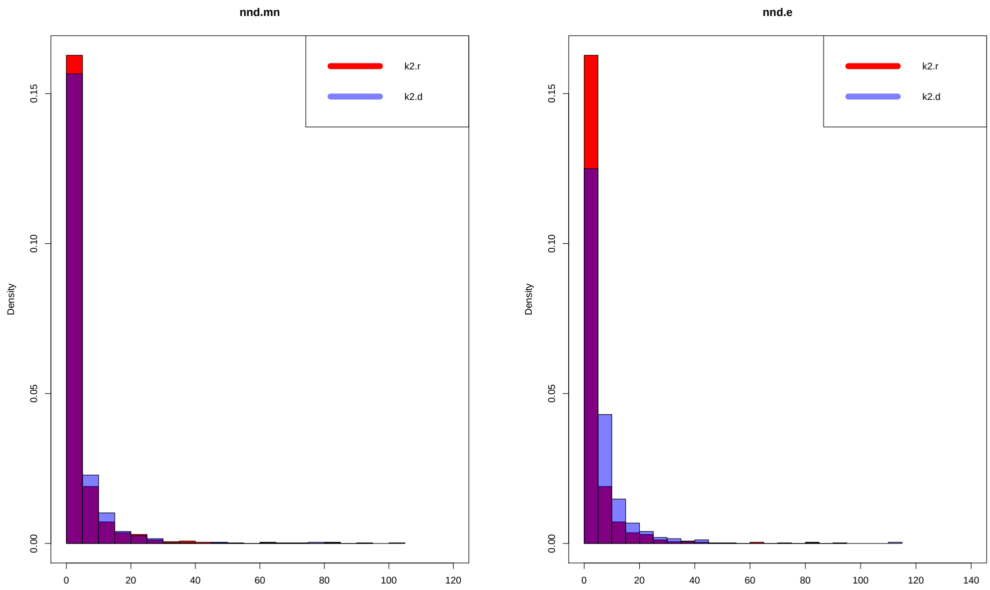


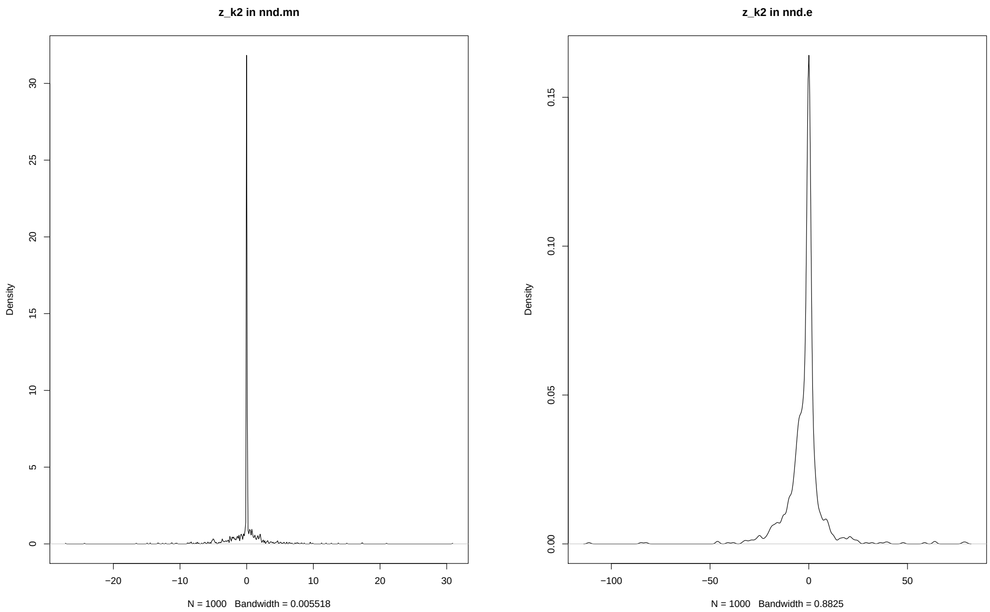
Figure 2.40 shows imputation (with donation classes) results, in terms of the different distributions of the variables K_2^R and K_2^D in the synthetic datasets generated from the above-mentioned combinations. We can notice a slightly tendency of nnd.mn (and nnd.ms) to overestimate the variable K_2^R and a clear significant tendency of the combination nnd.e to double it (for example in classes 5-10, 10-15).

Figure 2.40: Simulation 3, distributions of K_2^R, K_2^D in nnd imputation (with don. cl.)



Taking into account the distributions of z_{K2} for the combinations nnd.mn (and the really similar nnd.ms), and nnd.e, figure 2.41 shows good control of the outliers in the two former combinations but a bad matching units pairs and a lack of outliers control for the latter one.

Figure 2.41: Simulation 3, distributions of z_{K2} in nnd imputation (with don. cl.)



For sake of brevity, we omit the distributions of the variables K_1^R, K_1^D and K_2^R, K_2^D but also the distributions of differences z_{K1} and z_{K2}, resulting from the application of the above-mentioned combinations without the donation classes building. Indeed, they are similar to the above-mentioned ones with a more evident tendency for both K_1^R and K_2^R to be overestimated and a worse control of the outliers vales.

For sake of brevity, even distributions of K_1^R, K_1^D and K_2^R, K_2^D in the synthetic datasets generated by nndc.mn, nndcms and nndc.e combinations, are not showed because these combinations generate results highly similar to the ones previously discussed. Moreover, we omit even the distributions of differences z. The omitted results concern both the imputation with the donation classes and the one without them.

Figure 2.42 shows the imputation (with donation classes) results, for the rnd.mn, rnd.ms and rnd.e combinations. As we can see, the variables K_1^R and K_1^D have an overall good overlap for the former two combinations; with respect to the combination rnd.e there is instead a clear tendency to overestimate the variable K_1^R (in the classes 10-20 and 20-30 it is doubled) and to underestimate it in the upper values (for example in the classes 40-50, 50-60 and 70-80).

Figure 2.42: Simulation 3, distributions of K_1^R, K_1^D in rnd imputation (with don. cl.)

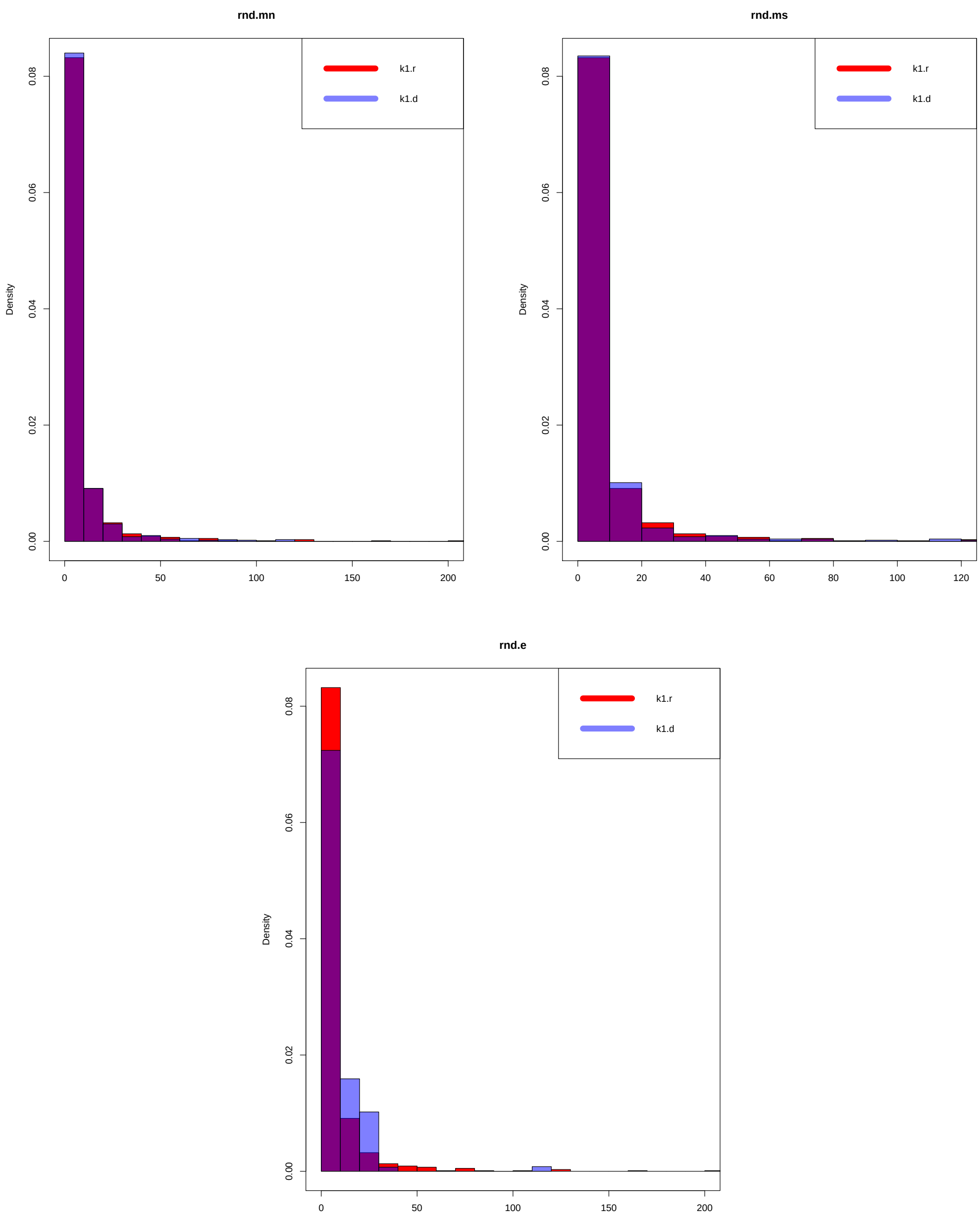


Figure 2.43 shows how the above mentioned combinations do not allow at all to properly control the outliers with respect to the variable K_1^R .

Figure 2.43: Simulation 3, distributions of z_{K1} in rnd imputation (with don. cl.)

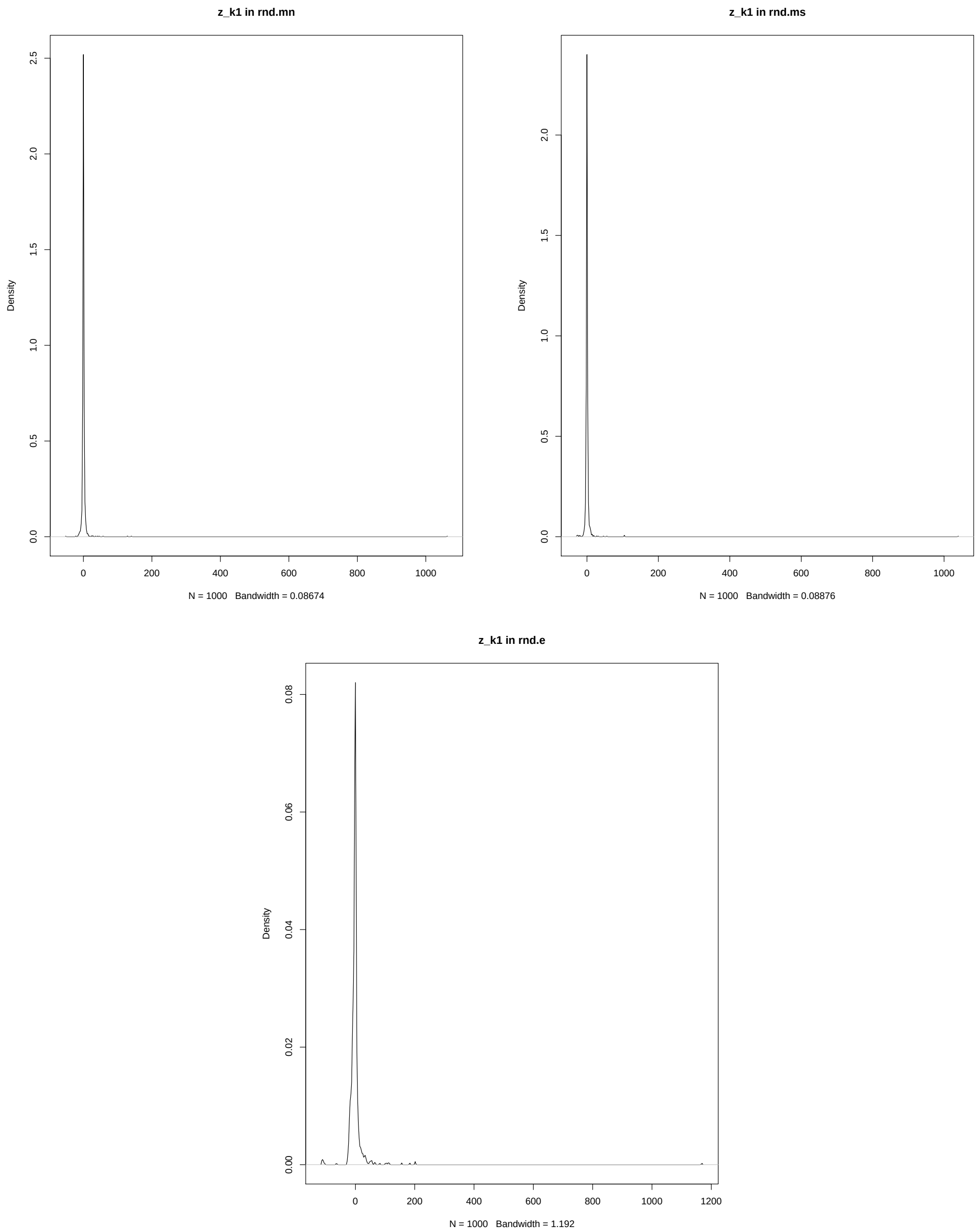


Figure 2.44 shows the imputation (with donation classes) results, for the rnd.mn, rnd.ms and rnd.e combinations concerning the variables K_2^R and

K_2^D. Whereas rnd.mn and rnd.ms do properly estimate the recipient variable K_2^R, rnd.e combination overestimates it almost doubling (for example, in the classes 15-20, 20-25, 25-30).

Figure 2.44: Simulation 3, distributions of K_2^R, K_2^D in rnd imputation (with don. cl.)

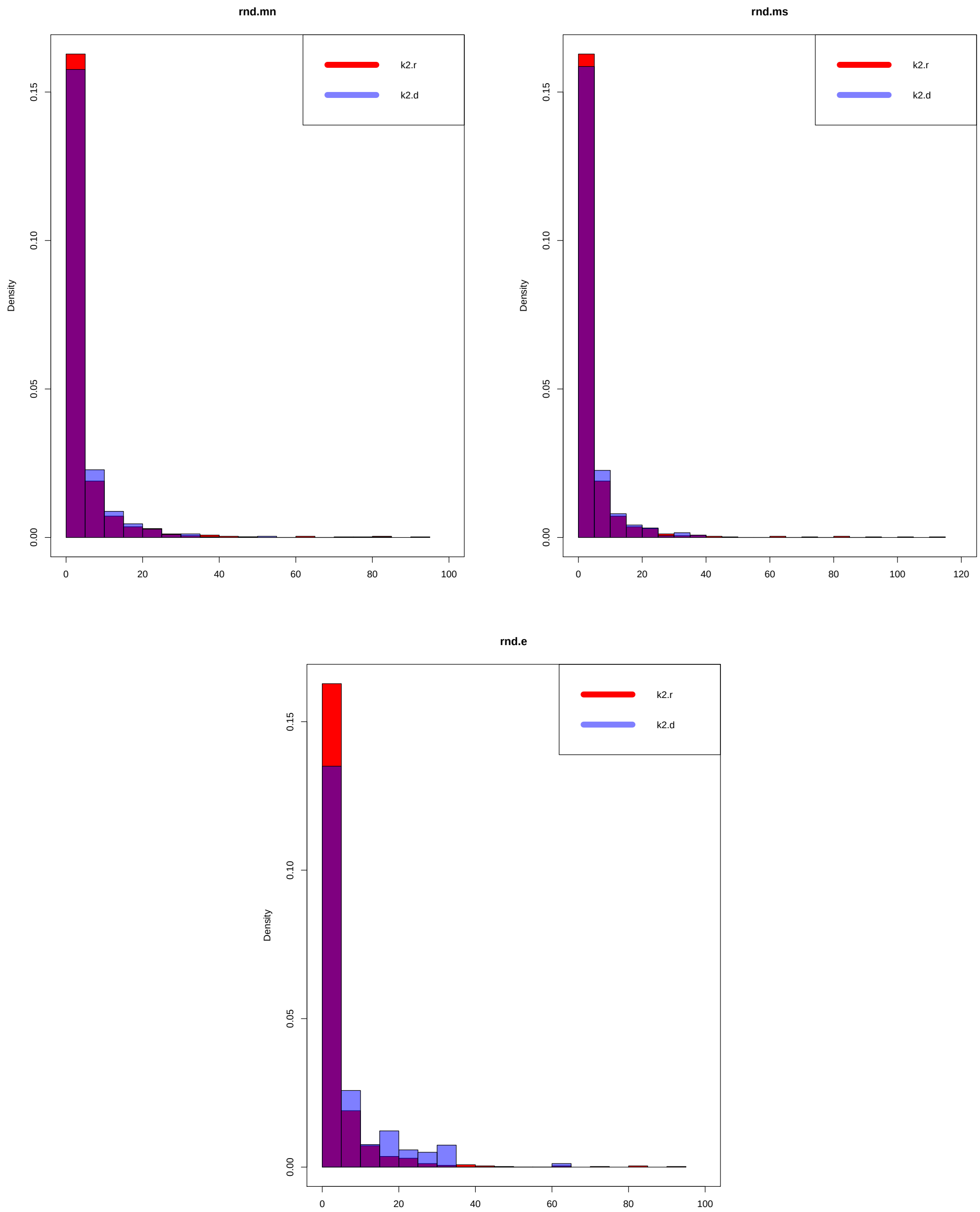


Figure 2.45 confirms the good matching units pairs associated by combinations rnd.mn and rnd.ms (which do perform even a discrete control of the outliers) whereas the combination rnd.e clearly perform a really bad imputation with respect to the variable K_2^R.

Figure 2.45: Simulation 3, distributions of z_{K2} in rnd imputation (with don. cl.)

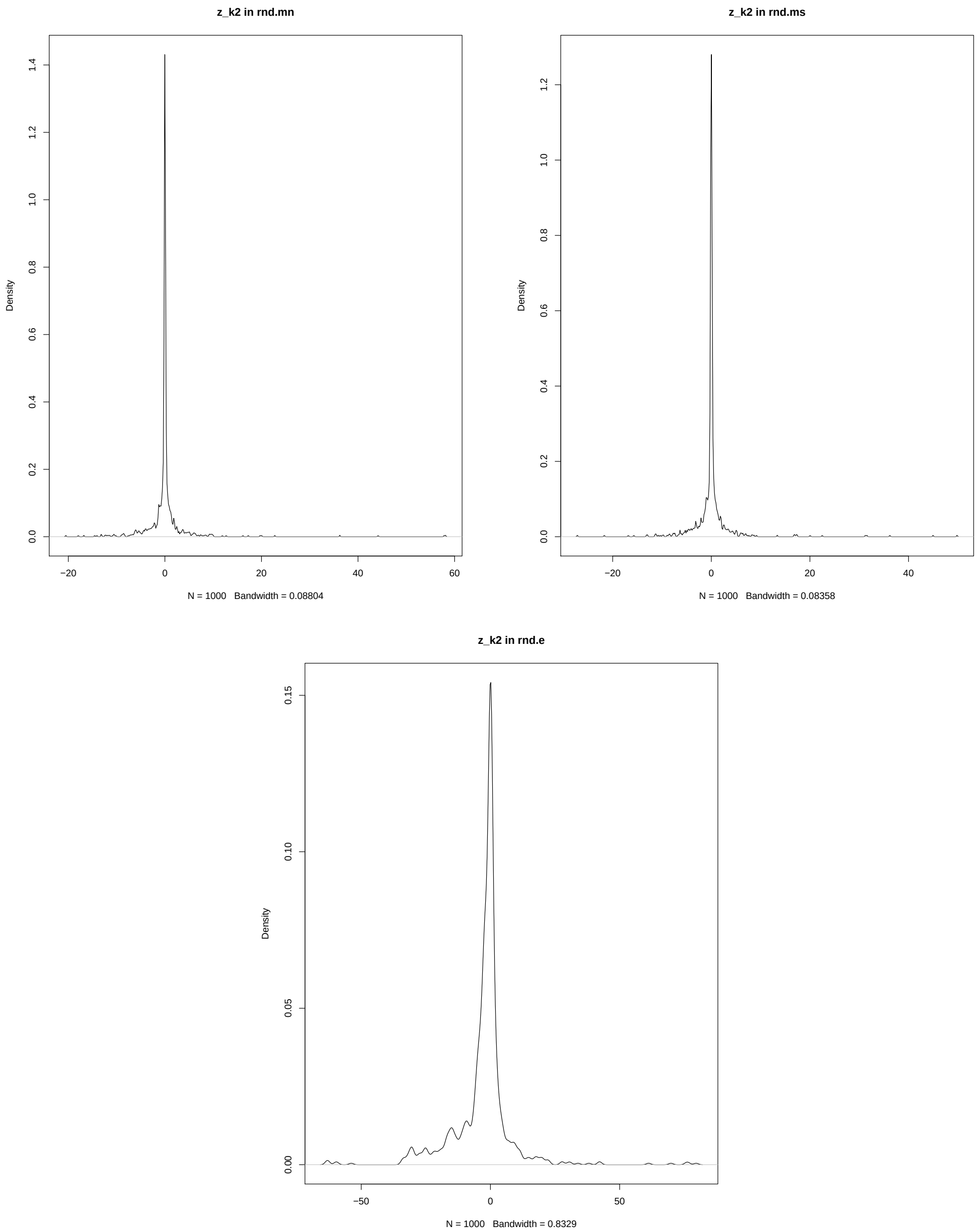


Figure 2.46 shows the imputation (without donation classes) results, for the same above-mentioned combinations; rnd.mn and rnd.ms perform again

an overall good imputation, with the rnd.ms that tends to slightly underestimate the variable K_1^R (for example in the classes 10-20 and 30-40). This tendency it nevertheless clearly evident and significant applying combination rnd.e which also overestimates K_1^R (for example in the class 10-20).

Figure 2.46: Simulation 3, distributions of K_1^R, K_1^D in rnd imputation (without don. cl.)

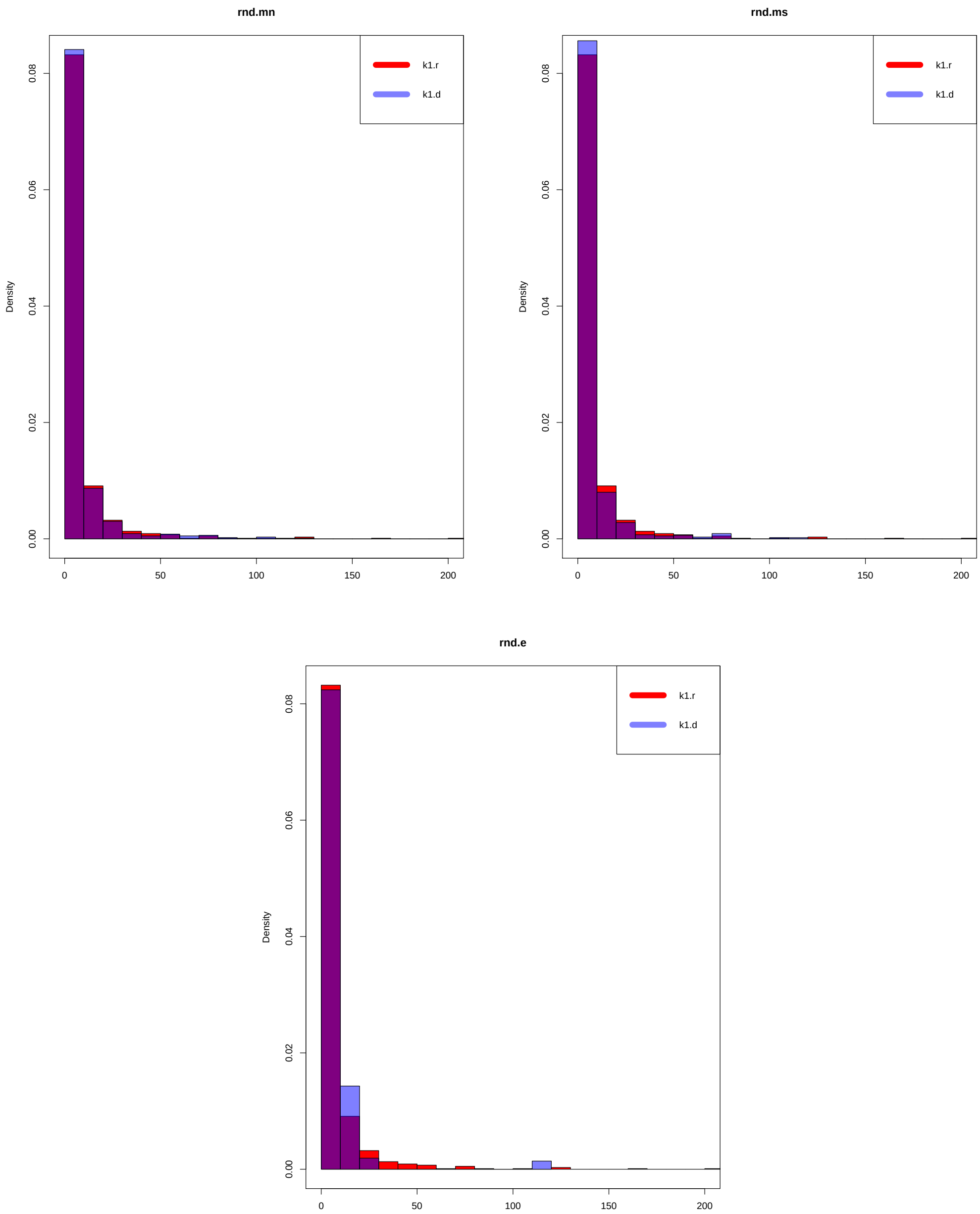


Figure 2.47 does not show different tendencies with respect to the previously discussed ones, concerning applications of rnd.mn, rnd.ms and rnd.e

which do not allow to control for outliers in spite of they perform discrete associations of matching units pairs.

Figure 2.47: Simulation 3, distributions of z_{K1} in rnd imputation (without don. cl.)

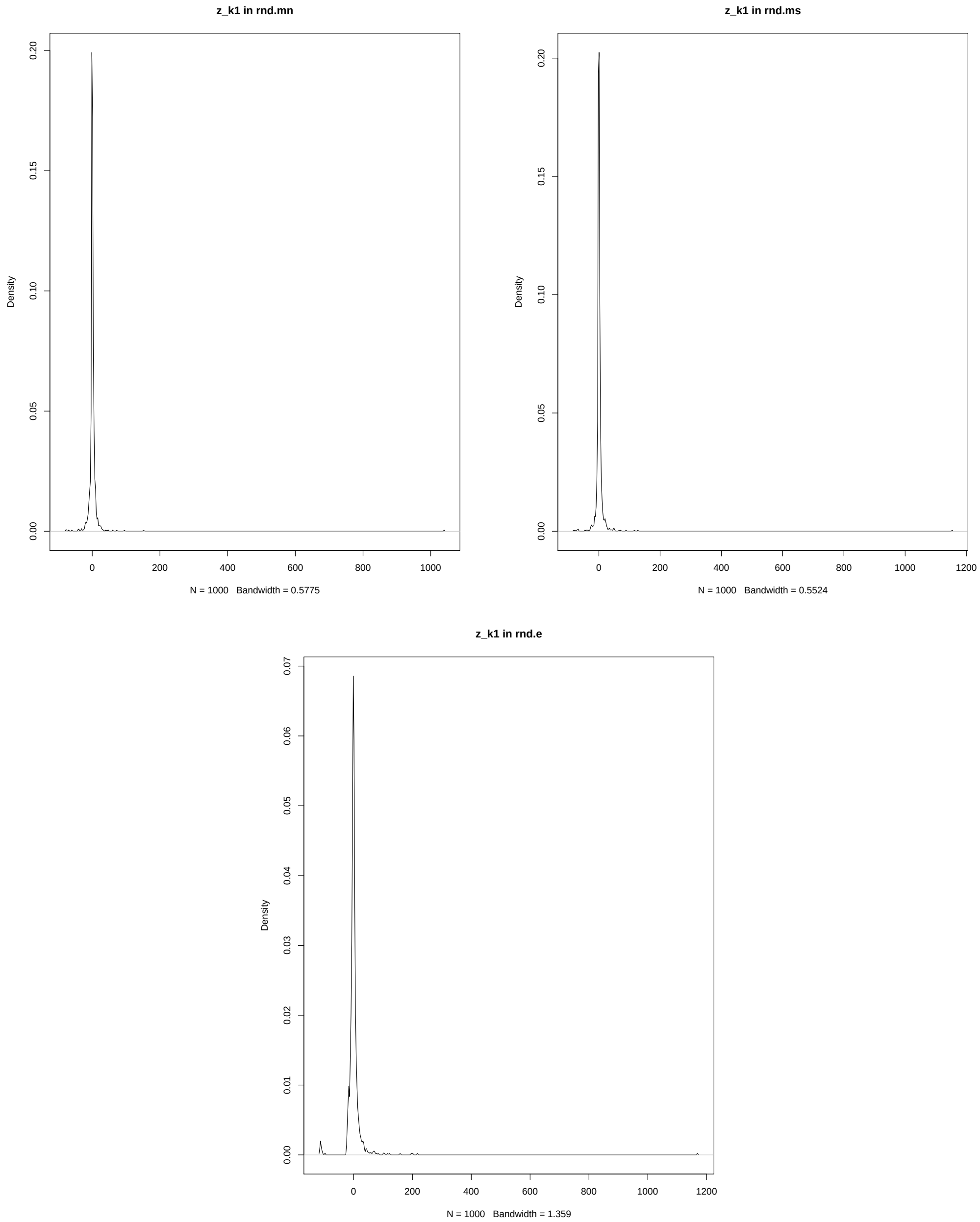


Figure 2.48 shows imputation (without donation classes) results for the above-mentioned combinations applied but referred to the variable K_2^R. The combinations rnd.mn and rnd.ms do perform similarly, with the latter one guaranteeing a less overestimate of K_2^R. The combination rnd.e instead, clearly underestimates the variable K_2^R in the class 0-5 but also overestimates (doubling it) the variable K_2^R for the upper values (for example in the class 5-10).

Figure 2.48: Simulation 3, distributions of K_2^R, K_2^D in rnd imputation (without don. cl.)

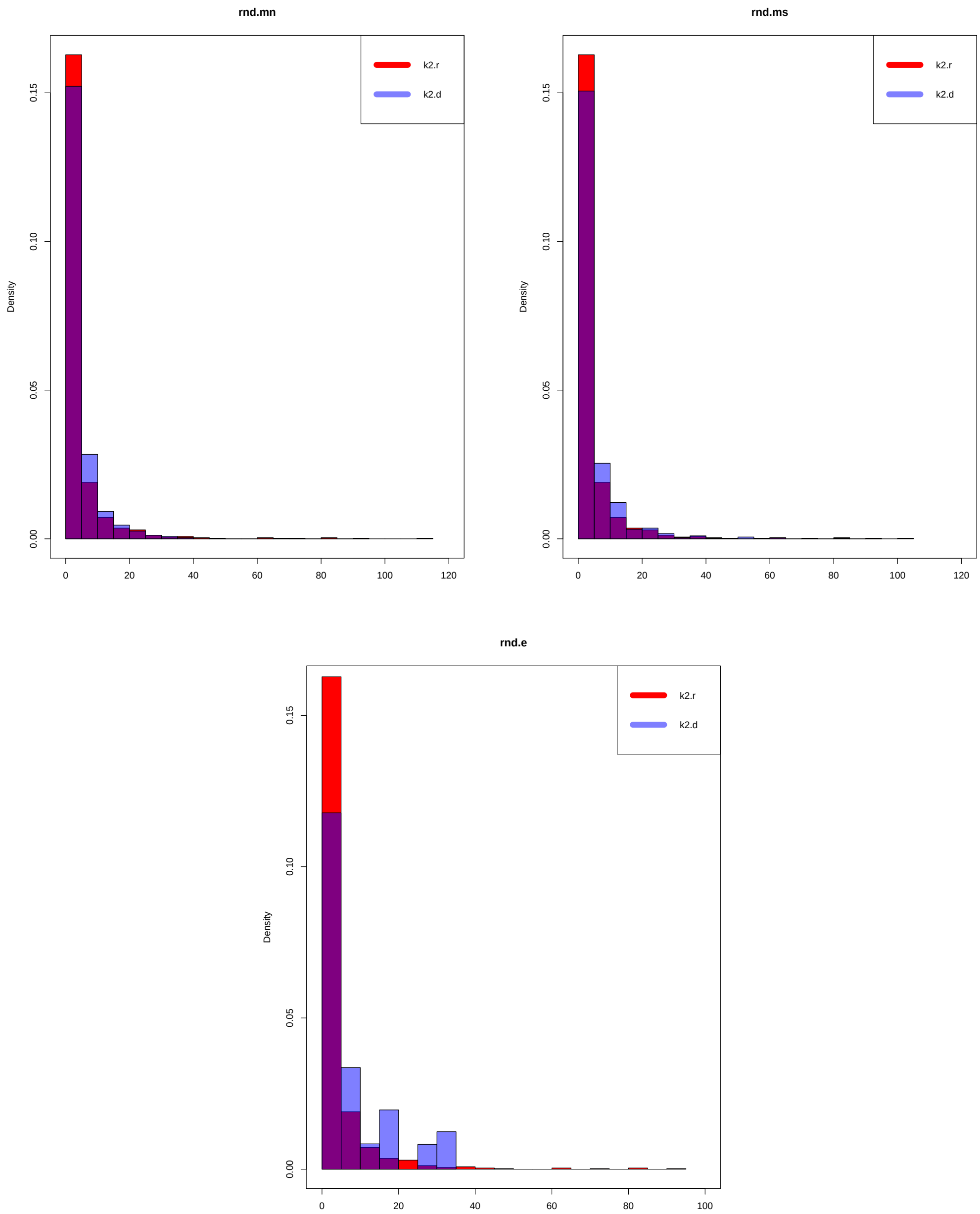
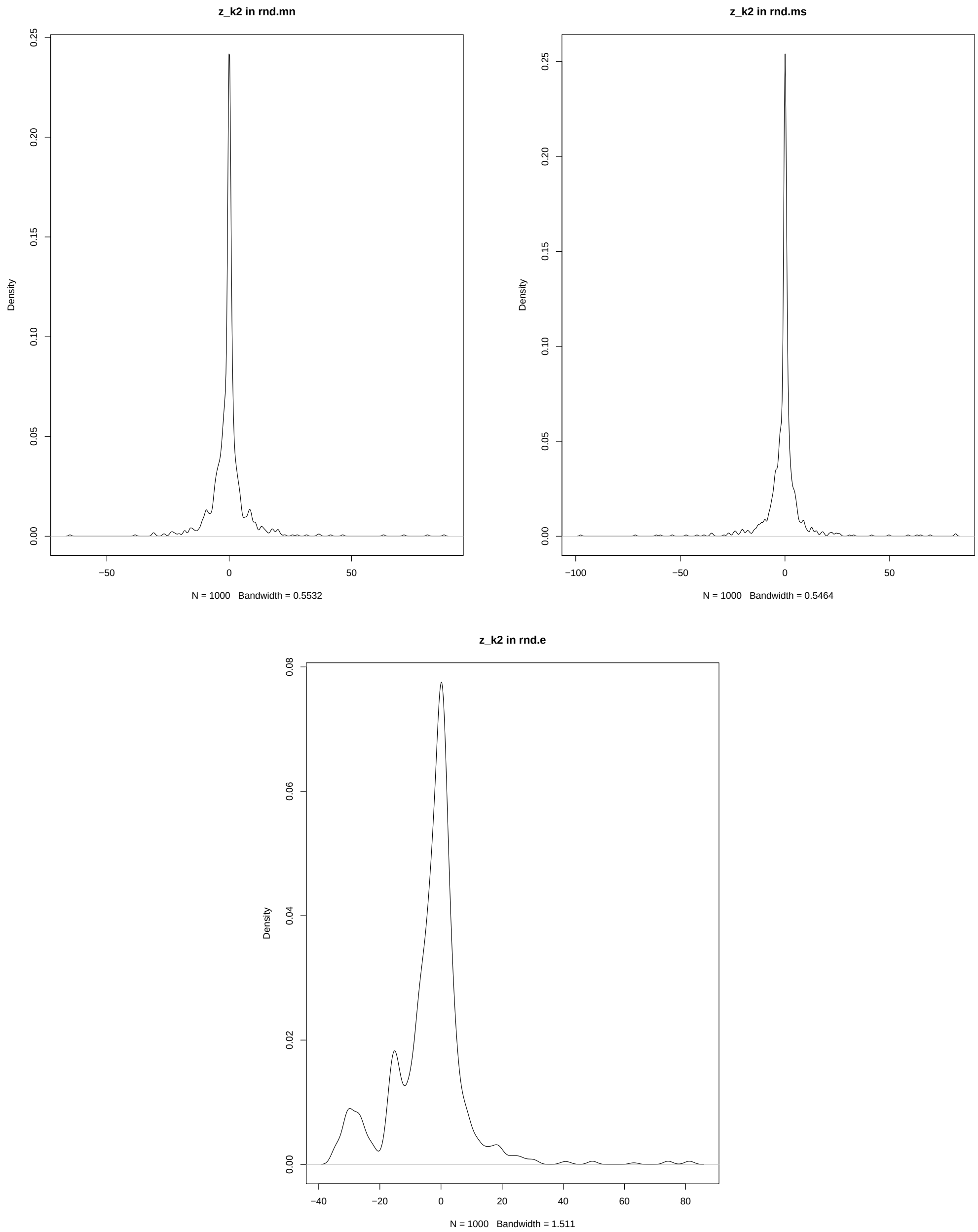


Figure 2.49 confirms that for the variable K_2^R the combinations rnd.mn and rnd.ms perform well in the control of the outliers even if they do not allow

an optimal association of the matching units pairs. The rnd.e combinations, instead, performs badly both in the association and in the outliers values control.

Figure 2.49: Simulation 3, distributions of z_{K2} in rnd imputation (without don. cl.)



In figure **??** we can notice how the imputation (with donation classes) results for the synthetic dataset generated by the rnk technique, with respect to both the variables K_1^R and K_2^R tends to overestimate them with a more evident tendency with respect to the variable K_2^R.

Figure 2.50: Simulation 3, distributions of K_1^R, K_1^D and K_2^R, K_2^D in rnk imputation (with don. cl.)

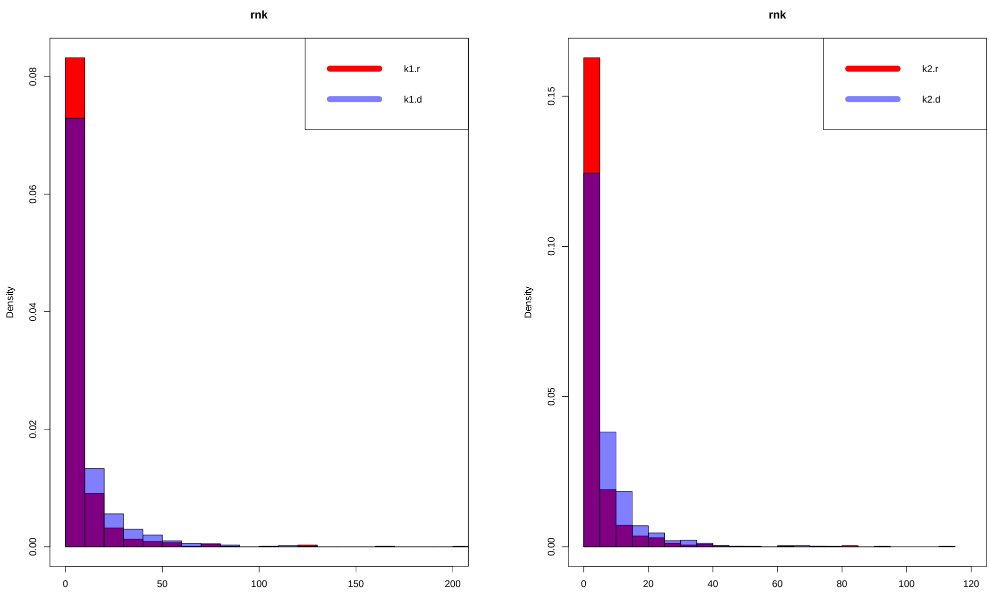
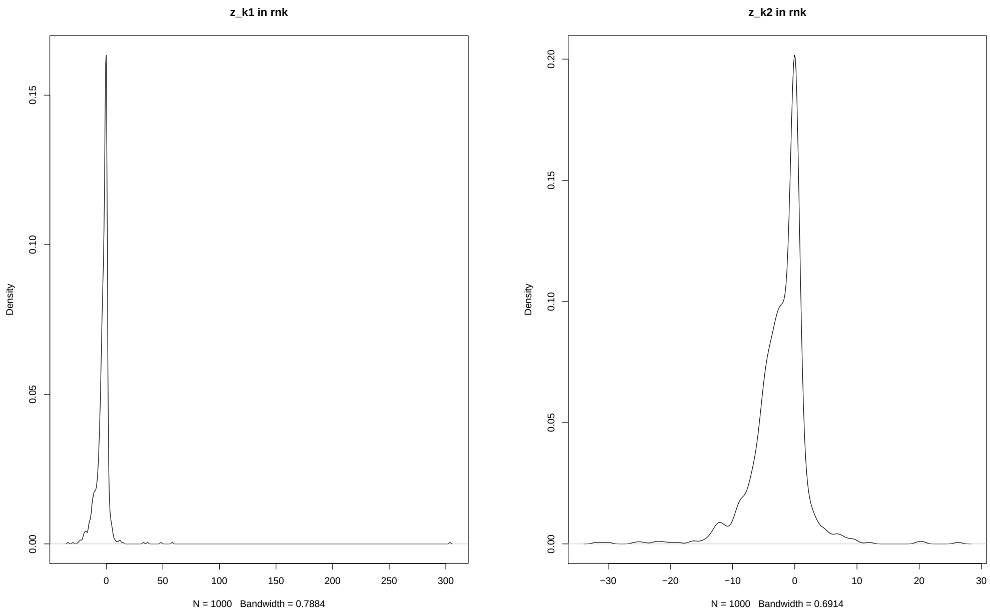


Figure 2.51 confirms what discussed previously; indeed, the rnk technique performs a better imputation with respect to the variable K_1^R (but does not properly control for the outliers), whereas it associates bad matching units pairs if we look at the variable K_2^R.

Figure 2.51: simulation 3, distributions of z_{K1}, z_{K2} in rnk imputation (with don. cl.)



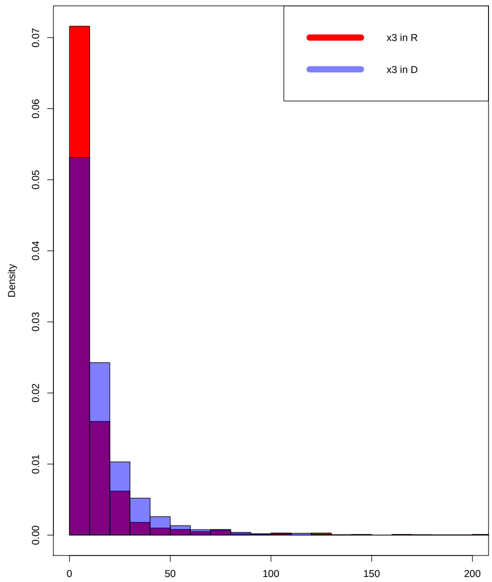
For sake of brevity, we omit the discussion of the imputation results applying the rnk technique without the donation classes building since they are similar to the above-mentioned ones, showing a slightly worse tendency to overestimate both the variables K_1^R, K_2^R and not properly controlling for the outliers.

2.3.4 Results from simulation 4

Figure 2.52 shows that from simulation 4 we have the recipient dataset R and the donor dataset D characterised, with respect to the matching variable X_3, by a higher variance and a noteworthy presence of outliers in the donor dataset D. We notice that the distributions of the matching variable X_3 both in R and D.

Figure 2.52: Simulation 4, variable X_3 in R and D

X_3		
	R	D
mean	9.552	15.487
var	196.919	1713.695
min	0.061	0.054
max	124.824	1760.312



Taking into account the imputation variable K_1 in datasets R (recipient) and D (donor), beyond the difference between the upper values of K_1^R and K_1^D due to the much lower maximum value of the variable K_1^D, figure 2.53 shows that there is an overall almost equally correspondence between R (recipient) and D (donor). With respect to the imputation variable K_2 in datasets R and D instead, figure 2.53 shows that, with the exception of the higher frequency of the variable K_2^R in the class 0-5, there is always a complete over-correspondence for the other values of K_2 in the donor dataset D.

Figure 2.53: Simulation 4, K_1 and K_2 in R and D

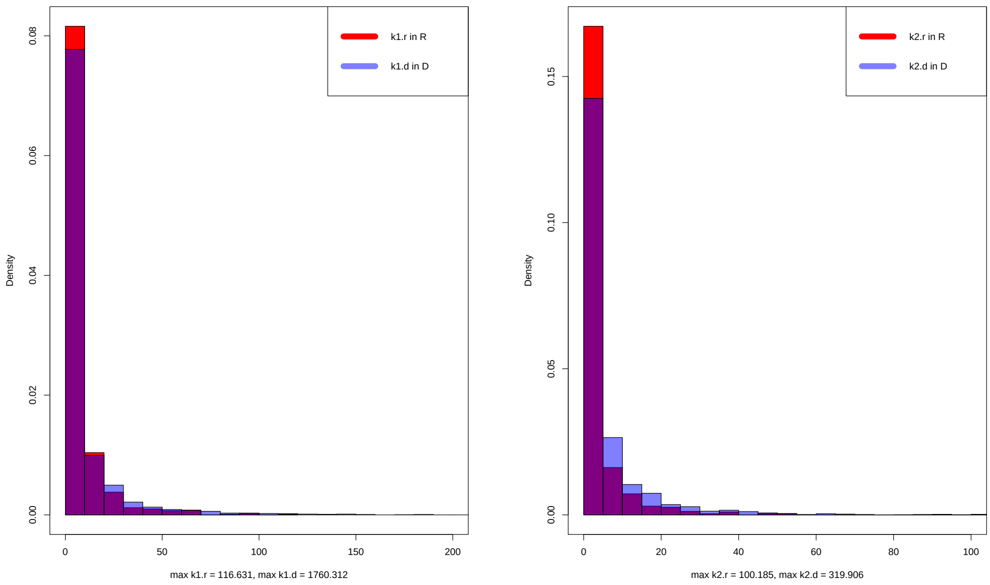
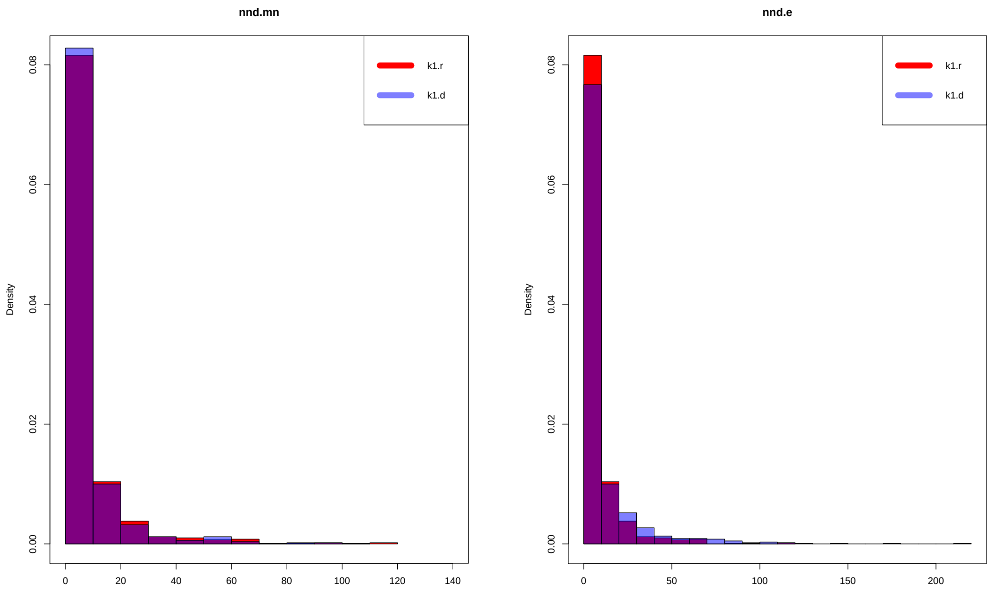


Figure 2.54 shows imputation (with donation classes) results for the combinations nnd.mn (nnd.ms is really similar, then omitted), and nnd.e. Both the combinations generate really good synthetic datasets in terms of the overlap between the variables K_1^R and K_1^D, with nnd.mn (and nnd.ms) slightly underestimating K_1^R whereas nnd.e generates a synthetic dataset in which there is a tendency to slightly overestimate the variable K_1^R.

Figure 2.54: Simulation 4, distributions of K_1^R, K_1^D in nnd imputation (with don. cl.)



Taking into account the distributions of the differences z, figure 2.55 shows how the combinations nnd.mn (and nnd.ms), and nnd.e perform well in controlling the outliers values, being both almost 0-centred; nnd.e nevertheless, tends to overestimate the variable K_1^R badly associating matching units pairs.

Figure 2.55: Simulation 4, distributions of textitz_{K1} in nnd imputation (with don. cl.)

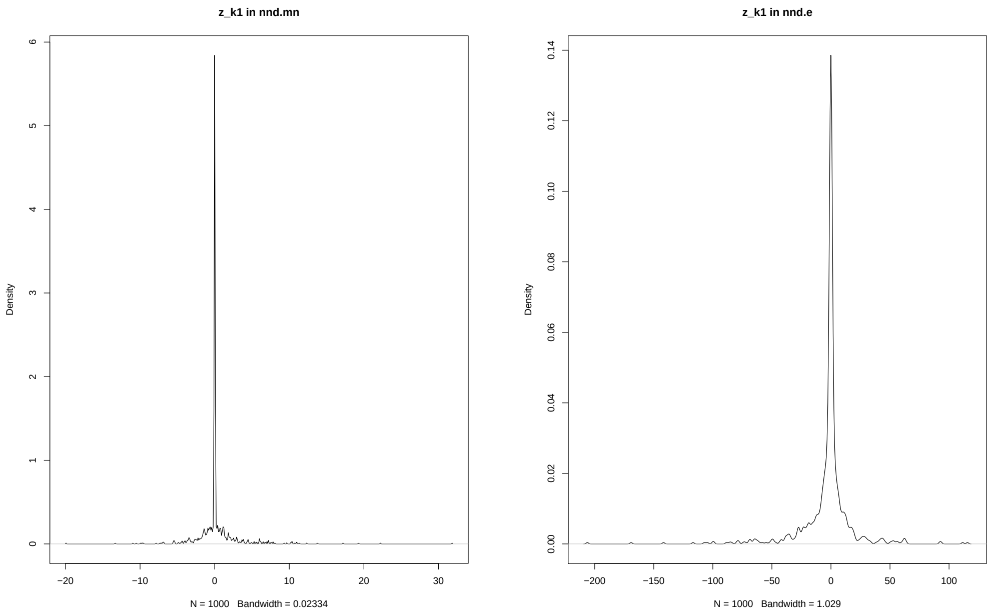


Figure 2.56 shows imputation (with donation classes) results for the variable K_2^R with the above-mentioned combinations. With respect to K_2^R, the combinations nnd.mn (and nnd.ms really similar, then omitted), perform a really good imputation in term of the overlap between the variables K_2^R and K_2^R. The combination nnd.e tends instead, to clearly overestimate the variable K_2^R (for example in the classes 5-10, 15-20 and 25-30).

Figure 2.56: Simulation 4, distributions of K_2^R, K_2^D in nnd imputation (with don. cl.)

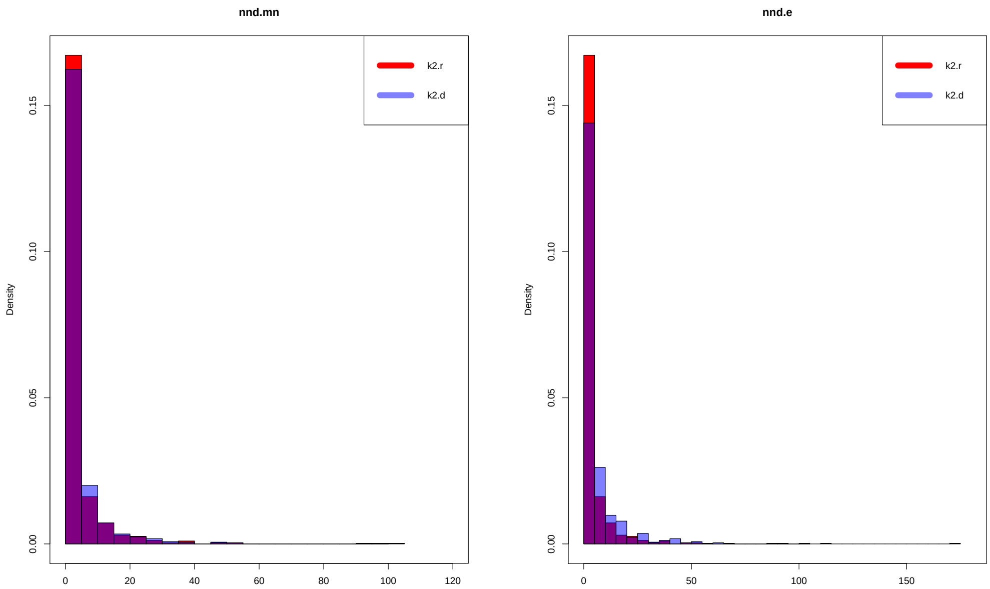


Figure 2.57 shows a slightly better performance of the combinations nnd.mn (and nnd.ms), and nnd.e with respect to the variable K_2^R if we take into account results for the distributions of the differences z referred to the variable K_1^R (slightly 0-centred, i.e. they control well the outliers).

Figure 2.57: Simulation 4, distributions of z_{K2} in nnd imputation (with don. cl.)

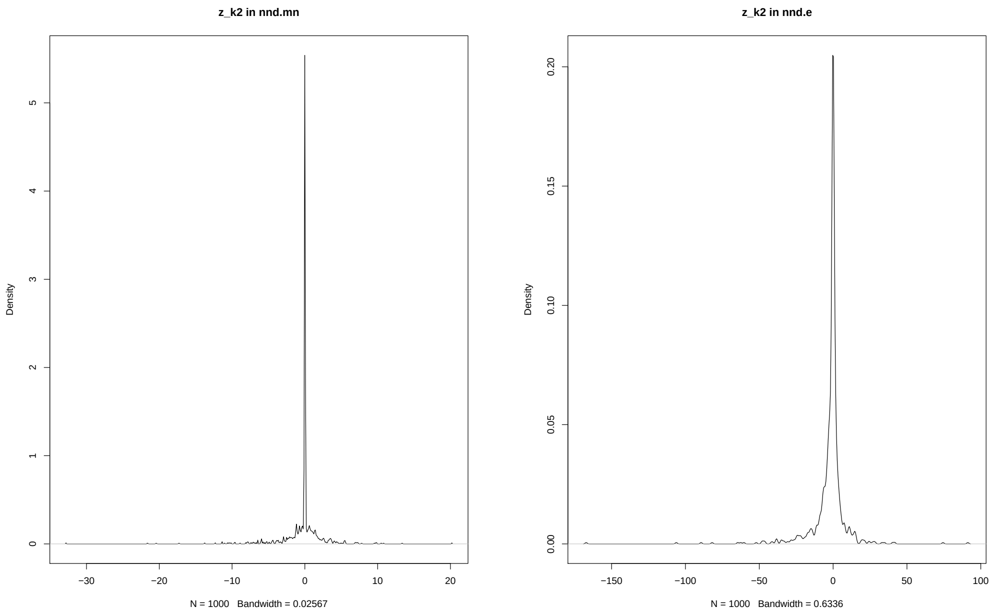
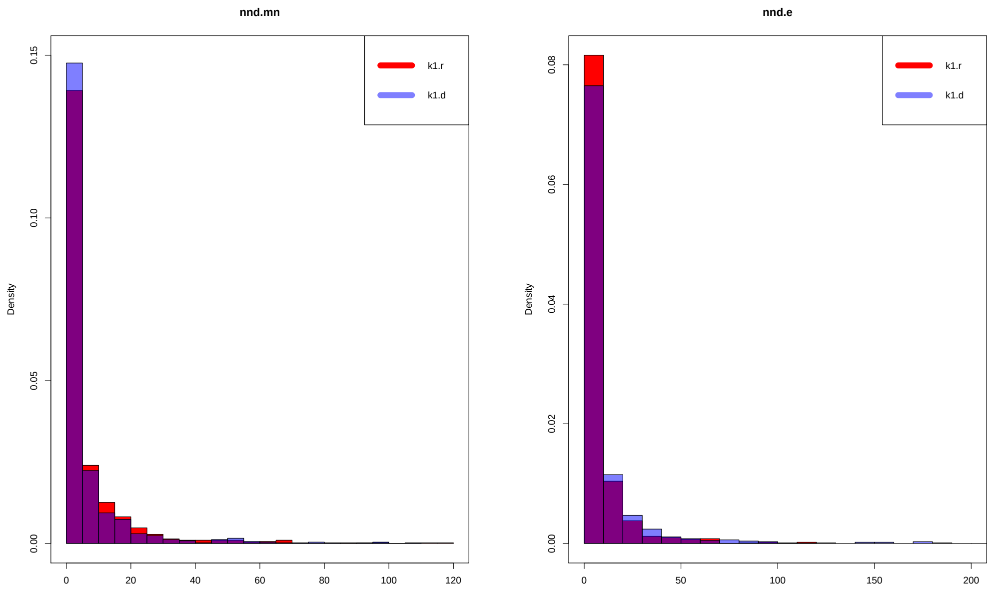


Figure 2.58 shows imputation (without donation classes) results in the synthetic datasets generated from combinations nnd.mn (and nnd.ms), and nnd.e. We can see that nnd.mn (and nnd.ms) slightly underestimate the variable K_1^R whereas the combination nnd.e tends to overestimate it. Nevertheless, neither the former tendency nor the latter are significant.

Figure 2.58: Simulation 4, distributions of K_1^R, K_1^D in nnd imputation (without don. cl.)



Taking into account the distributions of the differences z, figure 2.59 shows that the combinations nnd.mn and nnd.e do not associate good matching units pairs and have not an optimal performance with respect to the outliers control.

Figure 2.59: Simulation 4, distributions of z_{K1} in nnd imputation (without don. cl.)

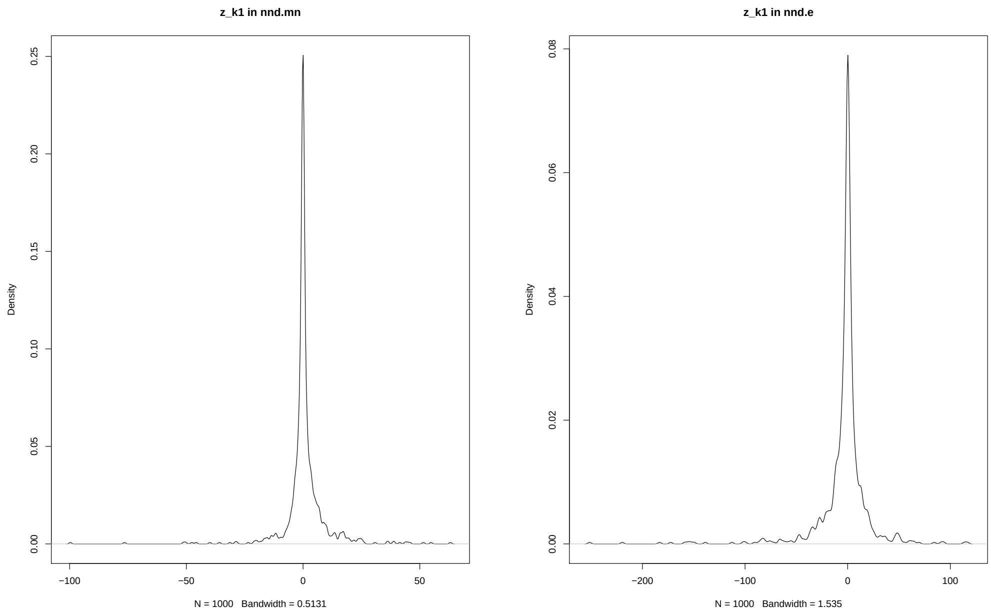


Figure 2.60 shows imputation (without donation classes) results for variable K_2^R with the above-mentioned combinations. The synthetic datasets generated with respect to the variable K_2^R present a similar tendency for both the nnd.mn (and nnd.ms) and nnd.e to overestimate K_2^R in the class 5-10, but to generally well represent the recipient variable.

Figure 2.60: Simulation 4, distributions of K_2^R, K_2^D in nnd imputation (without don. cl.)

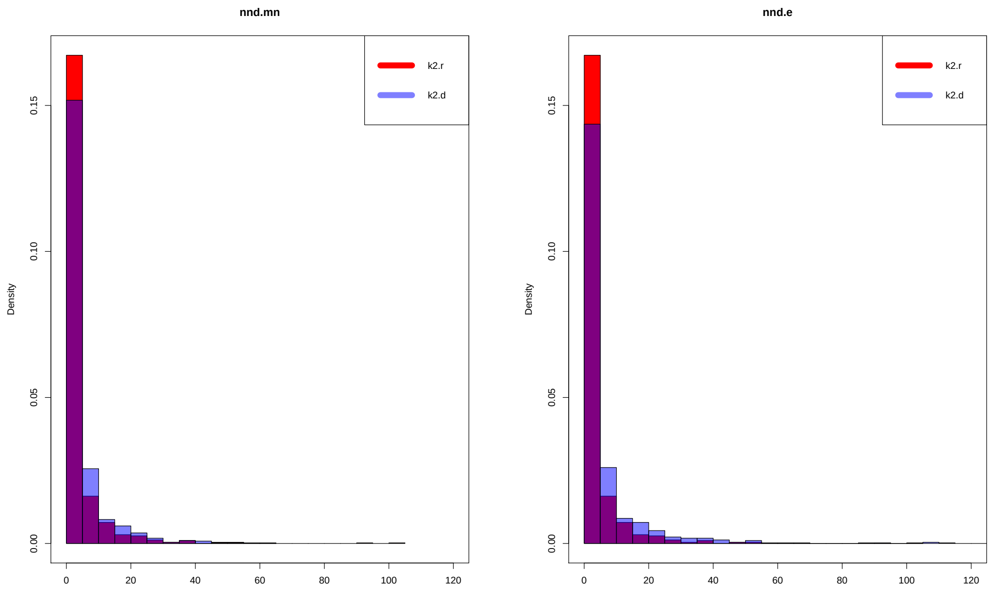
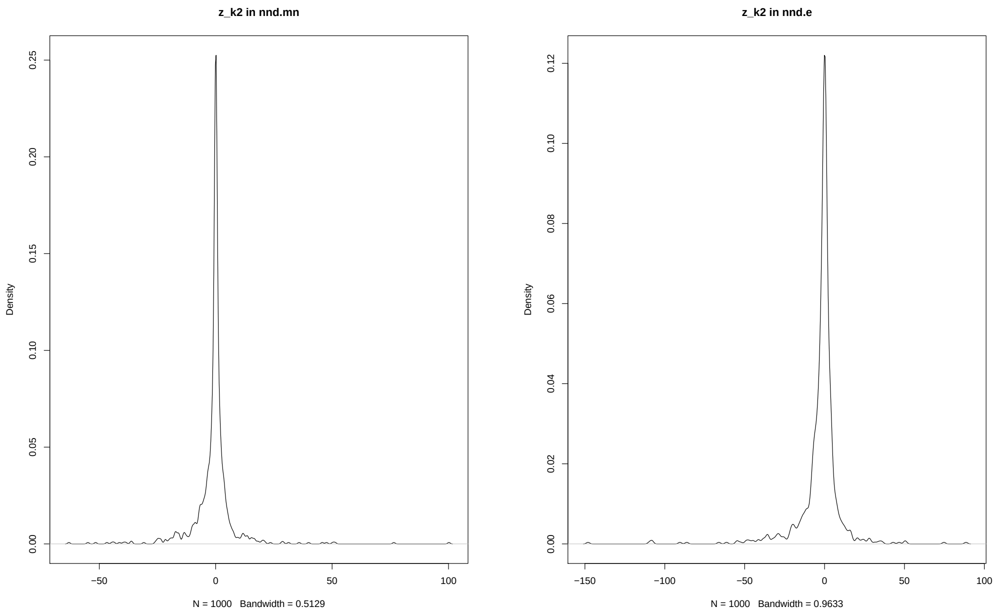


Figure 2.61 shows how the above-mentioned combinations perform well with respect to the variable K_2^R in controlling for the outliers and discretely associating the matching units pairs.

Figure 2.61: Simulation 4, distributions of z_{K2} in nnd imputation (without don. cl.)



For sake of brevity, distributions of K_1^R, K_1^D and K_2^R, K_2^D in the synthetic datasets generated by combinations nndc.mn, nndc.ms and nndc.e, and the respective distributions of the differences z_{K1}, z_{K2}, are omitted (both the imputations with and without donation classes), because they generate results which are highly similar to the combinations with the unconstrained SM imputation technique (i.e. the Nearest Neighbour Distance Hot Deck one).

Figure 2.62 shows imputation (with donation classes) results for variable K_1^R using the combinations rnd.mn, rnd.ms, and rnd.e.The former two generate a good synthetic dataset in terms of the overlap between the variables K_1^R and K_1^D with an overall tendency to properly estimate K_1^R. The combination rnd.e instead, presents an overestimate of K_1^R in the classes 30-40 and 40-50 (in which K_1^R is almost doubled) and a slightly tendency to underestimate K_1^R in the classes 10-20, 20-30, 50-60 and 60-70.

Figure 2.62: Simulation 4, distributions of K_1^R, K_1^D in rnd imputation (with don. cl.)

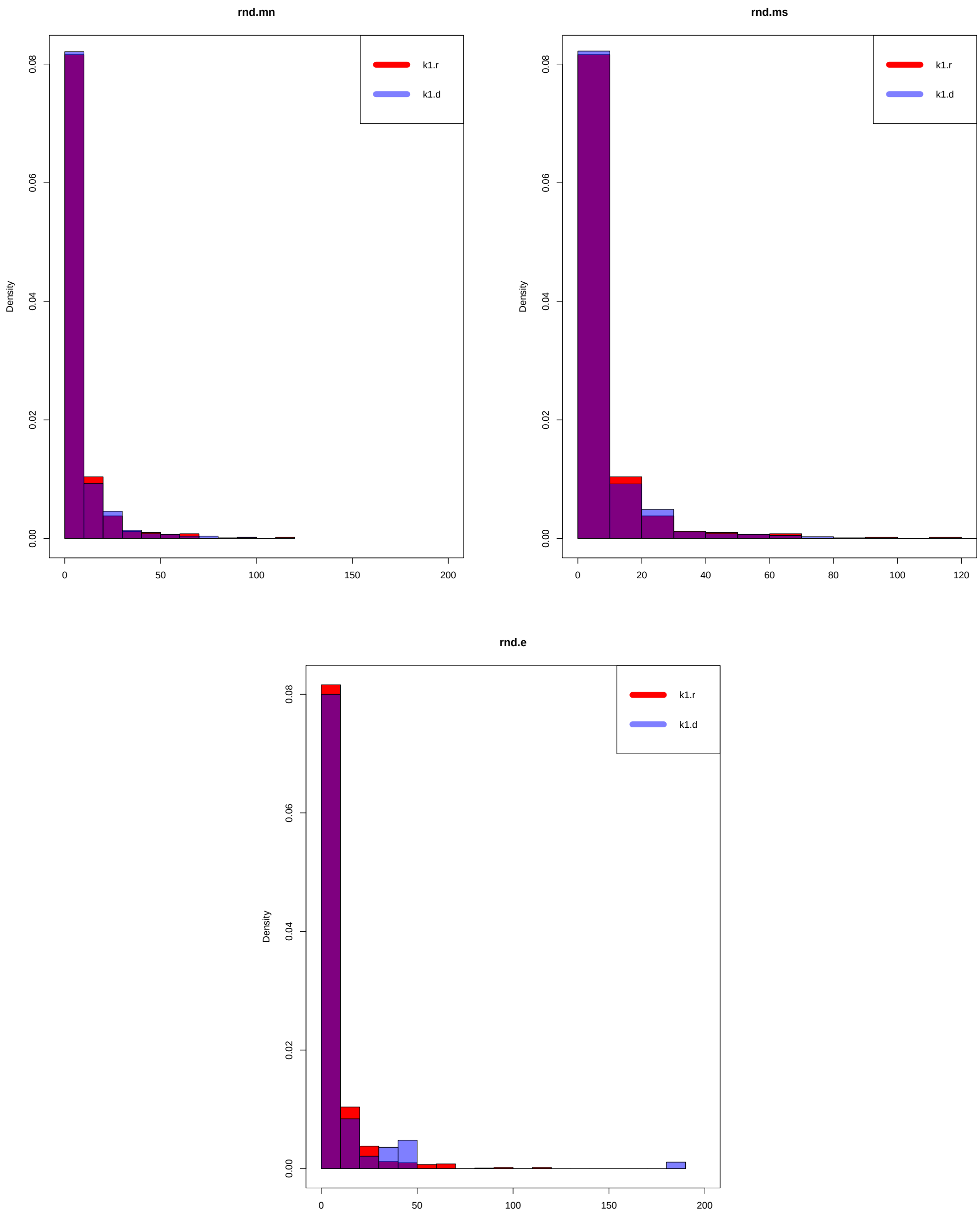


Figure 2.63 shows distribution of the differences z for the above-mentioned combinations with respect to the variable K_1^R; as we can notice, these combinations perform (rnd.e tends not to be as much good as the rnd.mn and rnd.ms), a good control of the outliers and also guarantee a good association of the matching units pairs.

Figure 2.63: Simulation 4, distributions of z_{K1} in rnd imputation (with don. cl.)

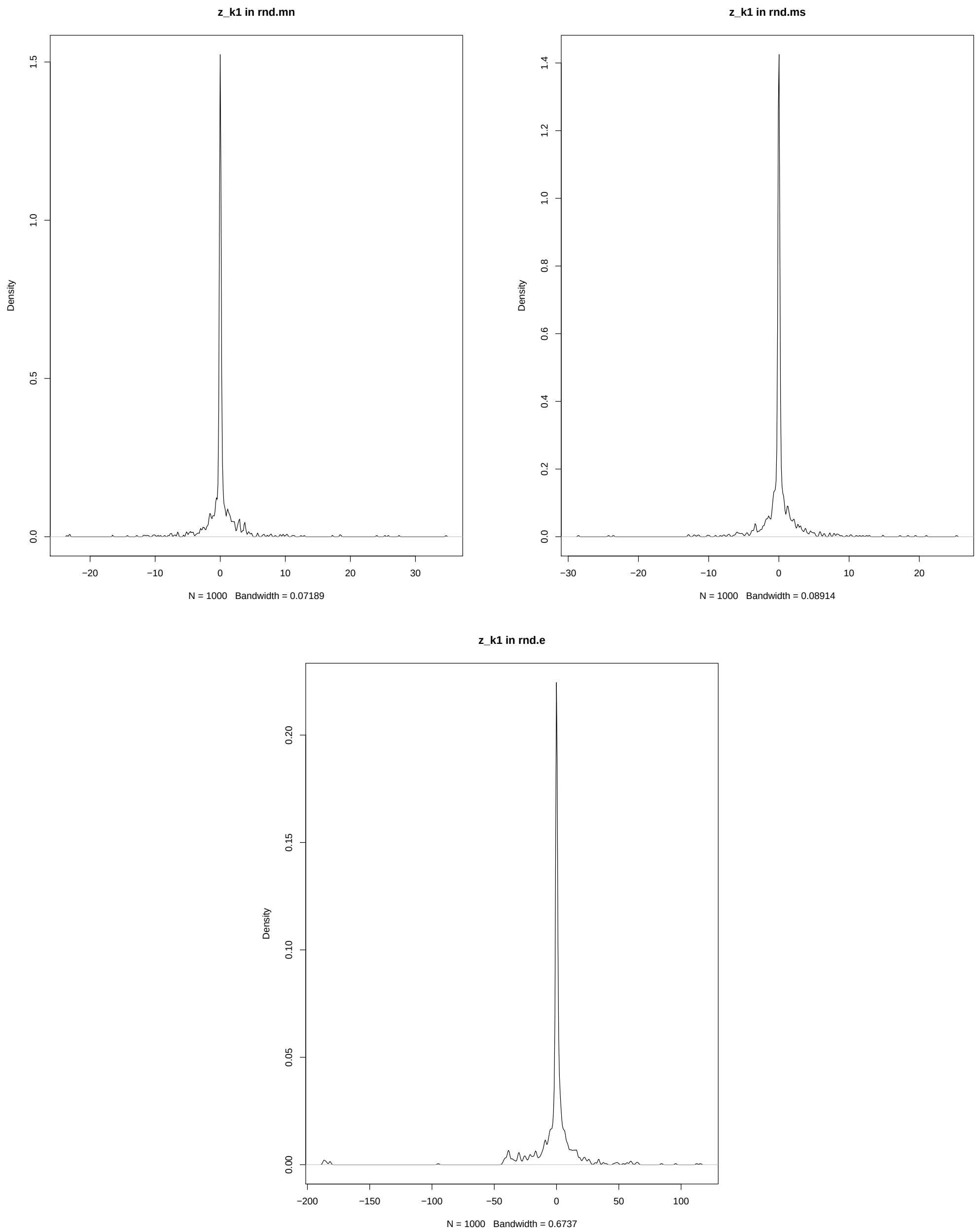
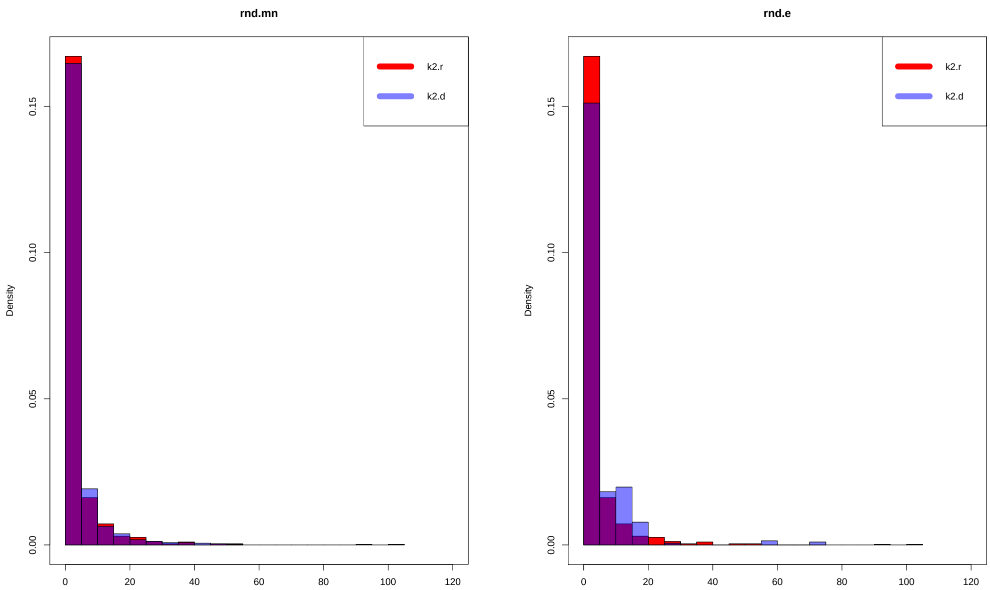


Figure 2.64 shows imputation (with donation classes) results for the combinations rnd.mn (and rnd.ms which is really similar, then omitted), and

rnd.e with respect to the variable K_2^R. We can notice an overall good overlap between K_2^R and K_2^D for the combinations rnd.mn (and rnd.ms), while rnd.e tends to overestimate the variable K_2^R (in the class 10-15 K_2^R is more than doubled whereas in the class 15-20 it is doubled).

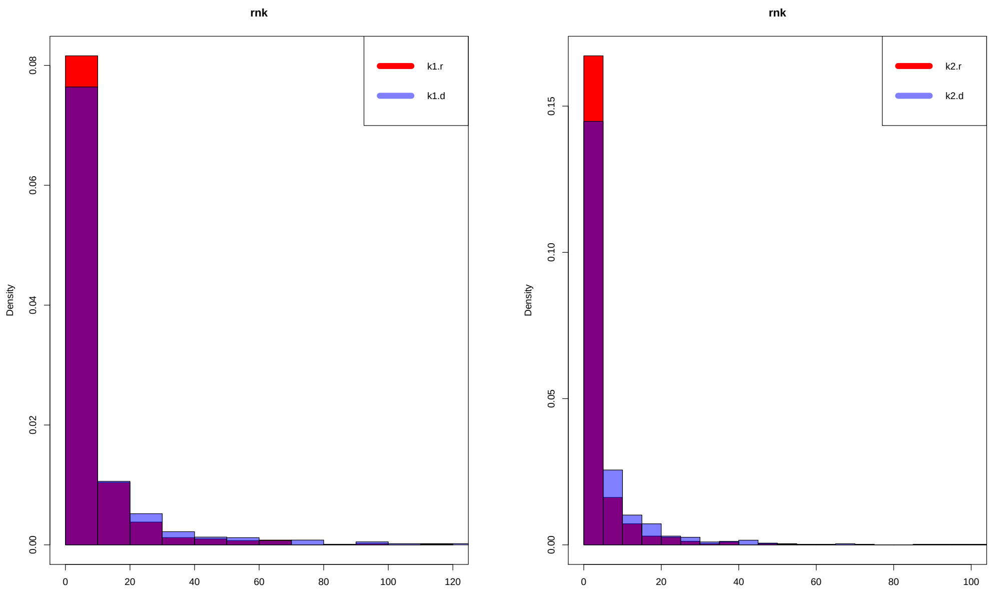
Figure 2.64: Simulation 4, distributions of K_2^R, K_2^D in rnd imputation (with don. cl.)



For sake of brevity, we decide to omit the distributions of K_1^R, K_1^D and K_2^R, K_2^D for the above mentioned combinations applied without the donation classes and also to omit the related distributions of the differences z_{K1} and z_{K2}. This is due to the fact that, generally, results from imputation without donation classes building related to these combinations are similar to the showed ones, just slightly worse in terms of the outliers control and for an overall tendency of overestimation.

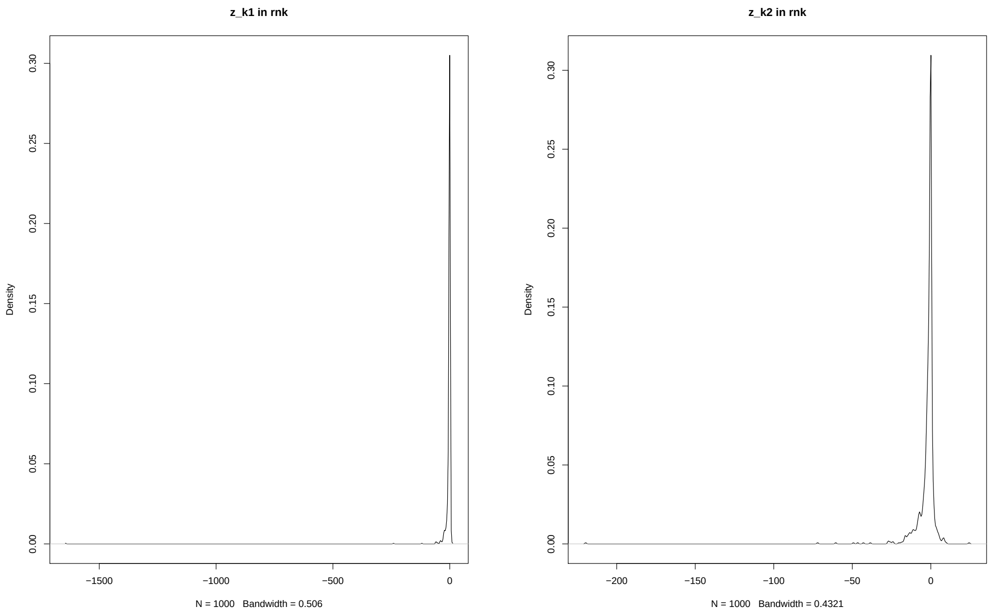
Figure 2.65 shows imputation (with donation classes) results for the variables K_1^R, K_2^R, obtained applying the rnk technique which, with the exception of the variable K_2^R (slightly overestimated), performs an overall good imputation.

Figure 2.65: Simulation 4, distributions of K_1^R, K_1^D and K_2^R, K_2^D in rnk imputation (with don. cl.)



Nevertheless, as figure 2.66 shows, rnk technique does not allow at all to control for the outliers, neither for the variable K_1^R nor for the variable K_2^R.

Figure 2.66: Simulation 4, distributions of z_{K1}, z_{K2} in rnk imputation (with don. cl.)



2.3.5 Summing up the imputation goodness validation

In order to validate the imputation results, i.e. to choose the best synthetic dataset generated by imputation using the different combinations, we evaluate, beyond the pre-post distributions of the originally present (in the recipient R), and the imputed (from the donor D) variables, the distributions of the differences z and their MSE values.

Simulations are made in order to test our expectations on the different combinations performances, taken into account the different characteristics of the recipient and the donor datasets. Previous to the simulation running our expectations were the following ones, i.e.:

1. being equal the dimensionality ratio between the recipient (R) and the

donor (D) datasets, their variability characteristics are crucial; specifically, the situation in which the variance of the matching variable(s) in the recipient dataset R is lower than the variance of the matching variable(s) in the donor dataset D, is always preferable;

2. in the unlucky case in which the variance of the matching variable(s) in the recipient dataset (R) is higher than the variance of the matching variable(s) in the donor dataset (D), the condition of a wider dimensionality ratio is always preferable;

3. being different the dimensionality ratio between the recipient and the donor datasets, the key assumption "the bigger, the best" present in the literature should hold;

4. the donation classes building helps to refine the imputation goodness.

The above-mentioned expectations are based on the assumptions that the wider is the difference in dimensionality between the recipient and the donor datasets, the greater is the choice among variables values to be used for associating two units and constitute a proper matching pair. Moreover, in order to properly create good matching units pairs, it is better to have a greater variability for the matching variable(s) in the donor dataset than in the recipient one. Finally, donation classes building, when possible, is strongly recommended because it benefits both the imputation goodness (more punctual units association), and the computational time for the generation of the synthetic dataset. All these expectations, with a remarkable exception (successively discussed), are confirmed.

Firstly, we see that a wider dimensionality ratio between the donor and the recipient datasets is determinant when the variance of the matching variables in the recipient dataset is higher than the variance of the matching variables in the donor one, as table 2.7 shows.

Table 2.7: MSE values of differences z (imputations 1, 2, 5, 6)

	don. cl.				no don. cl.			
	1 to 10		1 to 3		1 to 10		1 to 3	
	var(R) > var(D)				var(R) > var(D)			
	Imputation 1		Imputation 5		Imputation 2		Imputation 6	
	z_{K1}	z_{K2}	z_{K1}	z_{K2}	z_{K1}	z_{K2}	z_{K1}	z_{K2}
nnd.mn	101.536	9.617	102.534	10.017	176.171	83.896	182.890	90.273
nnd.ms	101.536	9.617	102.534	10.017	176.171	83.896	182.890	90.273
nnd.e	1,972.411	136.508	2,113.379	121.772	1,850.420	180.590	2,047.865	187.587
nndc.mn	101.527	9.608	102.679	10.293	175.903	83.628	183.459	90.858
nndc.ms	101.526	9.606	102.815	10.368	176.010	83.734	183.573	90.964
nndc.e	2,688.750	139.780	2,728.813	131.305	108.465	14.920	108.465	14.920
rnd.mn	1,000.011	15.570	1,186.610	19.674	1,253.199	85.351	1,192.059	73.047
rnd.ms	1,005.479	17.575	1,121.168	16.839	1,257.923	90.165	1,465.474	105.852
rnd.e	1,794.635	127.224	1,756.882	137.068	1,798.596	182.784	1,883.323	164.871
rnk	165.375	45.464	133.446	23.293	281.824	167.775	203.317	99.555

With the exception of combinations nnd.e and nndc.e for z_{K2} and rnd.e for both z_{K1} and z_{K2}, in the imputation with donation classes between R (recipient) and D (donor) characterised by the dimensionality ratios 1 to 10 and 1 to 3 and the var(R) > var(D), MSE values show how the bigger dimensionality ratio between the recipient dataset R and the donor dataset D is always preferable. We do not take into account the rnk technique imputation results since this technique systematically violate our expectations, often also representing the worst SM imputation technique for the control of the outliers. Furthermore, our expectations with respect to the dimensionality ratio conceived as a determinant factor for imputation goodness, find validity in simulation results even for the imputation without donation classes building.

As table 2.7 shows, with the exception of rnd.mn and rnd.e combinations, the case with the bigger dimensionality ratio is the preferable one.

Secondly, being the dimensionality ratio equal between R and D, to be determinant is the lower variance of the matching variables in the recipient dataset R with respect to the variance of the matching variables in the donor dataset D, as table 2.8 and table 2.9 show.

Table 2.8: MSE values of differences z (imputations 1, 2, 3, 4)

	don. cl.				no don. cl.			
	1 to 10				1 to 10			
	var(R) > var(D)		var(R) < var(D)		var(R) > var(D)		var(R) < var(D)	
	Imputation 1		Imputation 3		Imputation 2		Imputation 4	
	z_{K1}	z_{K2}	z_{K1}	z_{K2}	z_{K1}	z_{K2}	z_{K1}	z_{K2}
nnd.mn	101.536	9.617	9.532	9.528	176.171	83.896	77.918	77.904
nnd.ms	101.536	9.617	9.532	9.528	176.171	83.896	77.918	77.904
nnd.e	1,972.411	136.508	444.579	157.936	1,850.420	180.590	786.865	208.549
nndc.mn	101.527	9.608	9.466	9.465	175.903	83.628	84.813	84.770
nndc.ms	101.526	9.606	9.494	9.492	176.010	83.734	84.515	84.474
nndc.e	2,688.750	139.780	343.698	163.905	108.465	14.920	46.965	37.842
rnd.mn	1,000.011	15.570	8.273	7.295	1,253.199	85.351	78.321	81.351
rnd.ms	1,005.479	17.575	9.421	9.767	1,257.923	90.165	92.751	88.203
rnd.e	1,794.635	127.224	407.317	94.668	1,798.596	182.784	583.777	121.647
rnk	165.375	45.464	2,943.404	98.975	281.824	167.775	2,963.817	160.906

With the exception of combinations nnd.e and nndc.e for z_{K2}, in the imputation with donation classes between R (recipient) and D (donor) characterised by the dimensionality ratio 1 to 10, and the two different conditions of var(R) > var(D) and var(R) < var(D), MSE values show how, with the dimensionality ratio being equal between the recipient dataset R and the donor dataset D, the lower variance of the matching variables in the recipient dataset with respect to the variance of the matching variables in the donor one is always determinant, as table 2.8 shows. A less evident validity of this is found with respect to the imputation ran without donation classes. Indeed, in this case, not only the combinations nnd.e and nndc.e do violate our expectations but also the nndc.mn and nndc.ms.

Table 2.9: MSE values of differences z (imputations 5, 6, 7, 8)

	don. cl.				no don. cl.			
	1 to 3				1 to 3			
	var(R) > var(D)		var(R) < var(D)		var(R) > var(D)		var(R) < var(D)	
	Imputation 5		Imputation 7		Imputation 6		Imputation 8	
	z_{K1}	z_{K2}	z_{K1}	z_{K2}	z_{K1}	z_{K2}	z_{K1}	z_{K2}
nnd.mn	102.534	10.017	7.872	7.945	182.890	90.273	87.838	8.045
nnd.ms	102.534	10.017	7.872	7.945	182.890	90.273	87.838	8.045
nnd.e	2,113.379	121.772	477.174	158.138	2,047.865	87.587	666.437	205.484
nndc.mn	102.679	10.293	7.867	7.976	183.459	90.858	95.708	95.738
nndc.ms	102.815	10.368	7.913	8.022	183.573	90.964	77.219	77.183
nndc.e	2,728.813	131.305	420.386	169.801	108.465	14.920	46.965	37.842
rnd.mn	1,186.610	19.674	12.321	6.484	1,192.059	73.047	104.761	99.260
rnd.ms	1,121.168	16.839	9.950	16.915	1,465.474	105.852	85.926	87.745
rnd.e	1,756.882	137.068	573.707	106.418	1,883.323	164.871	334.443	76.499
rnk	133.446	23.293	2,834.001	86.592	203.317	99.555	2,953.937	43.025

Again, with the exception of combinations nnd.e and nndc.e for z_{K2}, in the imputation with donation classes between R (recipient) and D (donor) characterised by the dimensionality ratio 1 to 3, and the two different conditions of var(R) > var(D) and var(R) < var(D), MSE values show how, with the dimensionality ratio being equal between the recipient dataset R and the donor dataset D, the lower variance of the matching variables in the recipient dataset with respect to the variance of the matching variables in the donor one is always determinant, as table 2.9 shows. Even here, there is less evidence of this validity for the imputation without donation classes, not confirmed by combinations nnd.e, nndc.mn, nndc.e and rnd.mn.

Finally, and here it comes the only relevant violation of the previous expectations, we find evidence that a narrower dimensionality ratio between R (recipient) and D (donor), being the variance of the matching variables in R lower than the variance of the matching variables in D, can produce the best imputation results if the matching variables in the donor dataset have a proper variability, as the table 2.10 shows. In other words, oppositely to the common prescription of the SM imputation literature, the dimensionality bond between R and D (i.e. $n_R < n_D$), can be relaxed if the variance of the matching variable(s) in the recipient dataset R is lower than the variance of the matching variable(s) in the donor dataset D, and the variance of the matching variable(s) in the smaller of the two donor datasets is the wider one.

Table 2.10: MSE values of differences z (imputations 3, 4, 7, 8)

	don. cl.				no don. cl.			
	1 to 10		1 to 3		1 to 10		1 to 3	
	var(R) < var(D)				var(R) < var(D)			
	Imputation 3		Imputation 7		Imputation 4		Imputation 8	
	z_{K1}	z_{K2}	z_{K1}	z_{K2}	z_{K1}	z_{K2}	z_{K1}	z_{K2}
nnd.mn	9.532	9.528	7.872	7.945	77.918	77.904	87.838	88.045
nnd.ms	9.532	9.528	7.872	7.945	77.918	77.904	87.838	88.045
nnd.e	444.579	157.936	477.174	158.138	786.865	208.549	666.437	205.484
nndc.mn	9.466	9.465	7.867	7.976	84.813	84.770	95.708	95.738
nndc.ms	9.494	9.492	7.913	8.022	84.515	84.474	77.219	77.183
nndc.e	343.698	163.905	420.386	169.801	46.965	37.842	46.965	37.842
rnd.mn	8.273	7.295	12.321	16.484	78.321	81.351	104.761	99.260
rnd.ms	9.421	9.767	9.950	16.915	92.751	88.203	85.926	87.745
rnd.e	407.317	94.668	573.707	106.418	583.777	121.647	334.443	76.499
rnk	2,943.404	98.975	2,834.001	86.592	2,963.817	160.906	2,953.937	143.025

As table 2.10 shows, in the imputations with donation classes between R and D characterised by the dimensionality ratios 1 to 10 and 1 to 3, and the var(R) < var(D), MSE values show if the smaller dimensionality ratio between the recipient dataset R and the donor dataset D is preferable when the variance of the matching variables in the smaller donor dataset D is bigger than the variance of the matching variables in the other donor dataset D. Indeed, this is true with the exception of combinations nnd.e, nndc.e and rnd for both variables z_{K1} and z_{K2}. More evidence is nevertheless found with respect to the imputation without donation classes; indeed, as table 2.10 shows, we find validity with the exception of nnd.mn, nnd.ms, nndc.mn and rnd.mn combinations.

Taking into account the MSE values for the differences z, we find validity of our expectations; we do also find that the commonly prescribed imputation constraint related to choice of the recipient and the donor dataset (i.e. the bound of the dimensionality ratio between R and D), is not always true and can be relaxed. This can happen if we are in the case in which the donor with the smaller dimensionality ratio does have a higher variance for the matching variable(s) with respect to the recipient dataset.

Nevertheless, the analysis of the MSE values itself do not fit the purposes of the imputation goodness validation. Indeed, we take into account also the pre-post distributions of the imputed variables and the distributions of the differences z. Applying non-parametric micro SM imputation techniques, we have to focus also on descriptive statistics to validate the results goodness. Therefore, taking into account these others two tools, we have to consider that:

1. The combinations nnd.mn and nnd.ms applied both with and without the donation classes, generally perform a good imputation, presenting an optimal overlap between the "observed" variables and the "simulated" ones. These combinations perform also well with respect to the outliers control.

2. The nndc.mn and nndc.ms combinations, depending on the specific datasets characteristics, perform slightly similarly, showing sometimes

not significant (under)overestimation tendencies which are often more evident and sometimes statistically significant when donation classes are built.

3. The combination of both nnd and nndc techniques and the e distance function usually do not guarantee neither a proper estimation of the “observed”variable (often overestimating it), nor a good control of the outliers values. These performances always worsen without the imputation classes building.

4. The rnd.mn, rnd.ms and rnd.e applied both with and without the donation classes, usually perform well with respect of the overlap between the “observed”variables and the “simulated”ones (usually with the worst results obtained by the combination rnd.e). Nevertheless, generally they perform bad with respect of the outliers control, with a clearly significant lack of control usually manifested by the combination with the e distance function.

5. The rnk technique itself perform well always conditionally to the characteristics of the recipient and donor datasets at disposal. The overall tendency is to overestimate the “observed”variables, not guaranteeing at all the outliers control, neither with donation classes nor without them.

Taking into account the simultaneous consideration of the above-mentioned tools, form our simulation study, we find that the best synthetic datasets are to be selected among the ones generated by the combinations nnd.mn, nnd.ms, nndc.mn, nndc.ms, obviously considering the characteristics of the available datasets and the purposes of the structured SM imputation procedure.

2.4 Propensity Score Matching

The Propensity Score Matching (PSM) methodology is frequently used in the observational studies research context, in order to run causal effects analysis when randomized and experimental design analysis can not be planned. Indeed, PSM is useful to build for each treated unit i a counterfactual unit which has not been observed but which can be provided by control units similar to the treated ones in terms of observables characteristics that these two have in common.

Saying i, with $i = 1, \ldots, n$, the units which can (can not) receive a unique treatment (control) T, we should observe two different treatment outcomes for the outcome variable Y, observed for each i, such that $Y_i(0)$ is the outcome for the control units and $Y_i(1)$ is the outcome for the treated ones. PSM methodology is then structured upon three theoretical assumptions, i.e.:

- **Assumption 1.** Units do not interfere with each other so that treatment applied to one unit does not affect the outcome of another unit. This assumption, also called "Stable Unit Treatment Value Assumption" (SUTVA) (Rubin, 1977), states that there is only a single version of each treatment level for each unit, such that:

$$\underset{n_A\times Q}{\mathbf{Y}}{}^{A}(\underset{n_A\times S}{\mathbf{T}}{}^{A}) = \underset{n_A\times 1}{Y}(\underset{n_A\times 1}{T}) \ . \tag{2.3}$$

This exclusion restriction is not based upon the data themselves but on the previous knowledge about the research subject and does exclude the possibilities both that units interfere with each other and that there are multiple versions of the treatment T.

- **Assumption 2.** There is a set of common variables $\mathbf{X}$ such that, controlling for these common variables (covariates), both the potential outcomes are independent of the treatment status, conditional to the $\mathbf{X}$, such that:

$$[Y(0), Y(1)] \perp T|\mathbf{X} \ , \tag{2.4}$$

This assumptions is also called "Conditional Independence Assumption" (CIA) (Rubin, 1977) but it is also known as "unconfoundedness condition".

- **Assumption 3.** Under the theoretical framework of observational studies, the probability of a unit to be assigned to a treatment T, conditional to the set of observed covariates $\mathbf{X}$, is positive and lies between 0 and 1, i.e.:

$$0 < Pr(T = 1|\mathbf{X}) < 1 \tag{2.5}$$

This assumption is also called "common support condition" or "overlap condition" (Rosenbaum and Rubin, 1983). It basically means that, given the observed covariates, there is a positive probability for each unit of being both treated and control.

These three basic assumptions make it possible to think the assignment of units to the treatment as good as if it is random, selecting on observables characteristics related to each unit. Assumption of randomness with respect to the treatment assignment does require that all the most relevant variables to the probability of receiving the treatment may be observed and included in the list of the **X** covariates. This means that for each treated (control) unit we can find (i.e. construct) its unbiased counterfactual. Moreover, whereas the three assumptions hold, the probability of a unit to be assigned to treatment, it is equal to the probability of not receiving it, and this is true whenever there is sufficient overlap in the characteristics of treated and control units.

Saying τ the general effect of the treatment, we assume that both the treatment status and the control one are observed for each unit. Therefore, the causal effect of the treatment for each unit i results by easily solving the following equation:

$$\tau_i = Y_i(1) - Y_i(0) \ . \tag{2.6}$$

Being interested in knowing the average causal effect of the treatment in the population, we have to calculate:

$$\tau^{pop} = E\{Y(1) - Y(0)\} \ ; \tag{2.7}$$

being interested in knowing the average causal effect of the treatment in a sample, we have to calculate, instead:

$$\tau^{sam} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\} \ , \tag{2.8}$$

where the apexes *pop* and *sam* are referred, respectively, to the units in population and the units in the sample.

We can also be interested in knowing the average causal effect for the treated units both in the population and in the sample, such that, respectively, we have to calculate:

$$\tau^{pop,t} = E\{Y(1) - Y(0)|T = 1\} \ , \tag{2.9}$$

or

$$\tau^{pop,c} = E\{Y(1) - Y(0)|T = 0\} \ , \tag{2.10}$$

where the apexes t and c are referred to the treated and control units, respectively.

Finally, we can be interested in knowing the average causal effect for the control units in population and in the sample, such that, respectively, we have to calculate:

$$\tau^{samp,t} = \frac{1}{n_1}\sum_{i=1}^{n_1}\{Y_i(1) - Y_i(0)|T_i = 1\} \ , \tag{2.11}$$

$$\tau^{samp,c} = \frac{1}{n_0}\sum_{i=1}^{n_0}\{Y_i(1) - Y_i(0)|T_i = 0\} \ , \tag{2.12}$$

where $n_1 = \sum_i T_i$ and $n_0 = \sum_i (1 - T_i)$.

The problem of operating in the research context of the observational studies is that we do always observe for each unit i either $Y_i(0)$ or $Y_i(1)$. Constraints imposed by this peculiar research context do not allow researchers to plan an experimental design analysis and do not allow randomization. This is the main reason we usually have to resort to PSM methods.

If assumptions 2.3, 2.4, 2.5, respectively SUTVA, CIA and the overlap condition, hold, we can assume that the assignment mechanism of a unit to the treatment is strongly ignorable and, being assumptions 2.4 and 2.5 true, we can assume that for each unit i, being $\mathbf{X} = \{X_1, \ldots, X_l, \ldots, X_L\}$ the set of observed variables (covariates) for i, the two possible outcomes corresponding to treatment and control, i.e. $Y_i(0)$ and $Y_i(1)$, are independent form the assignment mechanism conditional to those observed covariates. Given the unit i, the two different outcomes which can not be both observed, are rather replaced by an observed outcome and a "missing" one, respectively defined by the apexes o and m. These two outcomes can be defined such that:

$$
\begin{aligned}
Y_i^o \equiv Y_i(T_i) &= T_i \cdot Y_i(1) + (1 - (T_i)) \cdot Y_i(0) \\
Y_i^m \equiv Y_i(1 - T_i) &= (1 - T_i) \cdot Y_i(1) + T_i \cdot Y_i(0) \ .
\end{aligned}
$$

The probability of unit i of being assigned to the treatment is tough:

$$
P_i(\mathbf{X}^A{}_i, Y_i(0), Y_i(1)) = \sum_{T_i=1} P(T; \mathbf{X}^A{}_i, Y_i(0), Y_i(1)) \ .
$$

Assuming that the set of functions $P_i(\cdot)$ can be written just in terms of a generic function $P(\cdot)$ which depends on the observed covariates $\mathbf{X}$ and the potential outcomes $Y(0)$, $Y(1)$ for all the units, we define the Propensity Score (PS) for the unit i as the average conditional probability of being assigned to a treatment T (Rubin, 1973). Following Rubin (1974), if assumption 2.4 holds, bias due to the observed covariates can be removed solely by conditioning on the PS. Then PS can be used in order to build, first, for each unit i the counterfactual outcome and, second, to estimate the treatment effect as the difference in outcomes for that unit.

Saying $\Im_w(i)$ the set of indices of the matched units (with opposite treatment status) for the unit i which result to be at least as close as the ones of the w-th match (or matched unit), with $w = 1, \ldots, W$, we define $No.\Im_w(i)$ as the number of elements in the set of indices $\Im_w(i)$. We then have that (Abadie et al., 2004):

$$\Im_w(i) = \{\{h = 1, \ldots, n\} \mid T_h = 1 - T_i, \|\mathbf{X}_h - \mathbf{X}_i\| \leq \delta_w(i)\},$$

where $\|\cdot\|$ is the norm of the differences among covariates values of the potential matching unit h with i and $\delta_w(i)$ is the distance among the covariates values of unit i from the w-th nearest matched unit with the opposite treatment.

We can estimate the potential outcome in the following way:

$$\hat{Y}_i(0) = \begin{cases} Y_i & , \quad \text{if } T_i = 0 \\ \frac{1}{No.\Im_w(i)} \sum_{h \in \Im_w(i)} Y_h(0) & , \quad \text{if } T_i = 1 \end{cases}$$

$$\hat{Y}_i(1) = \begin{cases} \frac{1}{No.\Im_w(i)} \sum_{h \in \Im_w(i)} Y_h(1) & , \quad \text{if } T_i = 0 \\ Y_i & , \quad \text{if } T_i = 1 \end{cases} .$$

Having built the potential outcome, we can estimate the treatment effect τ, i.e.:

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i(1) - \hat{Y}_i(0) \; = \; \frac{1}{n} \sum_{i=1}^{n} (2T_i - 1)\{1 + \; \Gamma_w(i)\} Y_i \; , \tag{2.13}$$

where $\Gamma_w(i)$ is the number of times the unit i is matched with the unit h, weighted for the number of matches h does have.

We then have the treatment effect for treated units and control ones, respectively, re-defined in the way that equations 2.11 and 2.12 are the following ones:

$$\tau^t = \frac{1}{n_1} \sum_{i=1}^{n_1} \{(Y_i - \hat{Y}_i(0)) | T_i = 1\} \; = \; \frac{1}{n_1} \sum_{i=1}^{n_1} \{T_i - (1 - T_i)\Gamma_w(i)\} Y_i \; , \tag{2.14}$$

and

$$\tau^c = \frac{1}{n_0}\sum_{i=1}^{n_0}\{(\hat{Y}_i(1) - Y_i)|T_i = 0\} \;=\; \frac{1}{n_0}\sum_{i=1}^{n_0}\{T_i\Gamma_w(i) - (1 - T_i)\}Y_i \;. \quad (2.15)$$

Chapter 3

Data Description

3.1 Introduction

Chapter 3 presents the complete description of the data sources and datasets used in our application both for the SM imputation and the PSM analysis. We take into account FADN 2009, SPA 2005 and CAP-IRE 2009 datasets, describing the data sources they belong to and the variables we choose to use.

3.2 FADN 2009

the Farm Accountancy Data Network, also known as FADN, is an official administrative data source related to EU agricultural holdings (i.e. farms), which collects accountancy data from farm samples around EU Member States. It is considered to be, as reported on the FADN website, "the only source of microeconomic data of agricultural holdings that is harmonised among EU Member States" (Website, 2000). FADN is set up since 1965 by EU Council Regulation in order to allow the EU Commission to analyse CAP impacts and CAP changes which occur over time in the farms structure, the employment, business and income management. The Regulation establishes the annual organisation of the survey, which is carried out by EU Member States on their national farm populations. Farms are selected to be part of the survey according to sampling plans established by EU regional institutions, providing statistical representation of units on three different dimensions. Indeed, farm samples are stratified by:

- territory, using NUTS levels;
- agricultural holdings specialization, using TF14 classifications;
- agricultural holdings economic size, using economic classes defined by DG Agri and the ES6 Grouping.

Basically, stratification provides farm samples which are representative of EU farms population in terms of commercial weight, location, type of farming. As reported on FADN Website (2000):

> "the annual sample covers approximately 80,000 holdings; they represent a population of about 5,000,000 farms in the EU, which covers approximately 90% of the total Utilised Agricultural Area (UAA) and account for about 90% of the total agricultural production".

FADN data contain around 1,000 variables described in the so-called Farm Return questionnaire.

We use FADN data limited to the accounting year 2009, with respect to the Italian farm sample, made up of 10,743 units and 859 variables. We focus only on the Emilia-Romagna Region farm sample, made up of 1,054 units, taking into account variables which refer to farms physical and structural characteristics, location, type of crops and livestock head, labour force, but also farms income, costs, sales, purchases, assets, quotas, subsidies, farm-household characteristics and some variables connected with the uptake of CAP measures.

In table 3.1 we show the "useful" variables we choose to select within the FADN 2009 dataset in order to run the SM imputation with the CAP-IRE 2009 dataset and the consequent PSM causal effects analysis within the new generated dataset. Among the 859 variables at disposal, we focus on 76 variables, the ones we identify as key-variables with respect to the availability and relevance of the most similar present variables (or which could be constructed), in the CAP-IRE 2009 dataset. Table 3.1 reports variables codes in the original data source, their descriptions and a brief note about their quantitative/qualitative nature. Since FADN data source is English-based, none of these variables, codes and/or descriptions vary from the original ones. Symbols and abbreviations are explained in table footnotes.

Table 3.1: FADN 2009 chosen variables

Variable	Description	Notes
	Farm general information	
a18	organisational form	
a27	economic size	€
a32	organic farming	
a39	less favoured area	
a44	Structural Funds	
a45	environmental constraints	"NATURA 2000" areas

b48	UAA in owner occupation	ha
b50	UAA in share cropping	ha
cluaa	class of UAA	6 classes
nuts2	NUTS2 (IT Region)	
nuts3	NUTS3 (IT Province)	
se005	economic size	ESU
se030	UAA rented	ha
se410	Gross Farm Income	€
se631	SFP	€
sys02	extrapolation factors	
tf14	farm specialisation	13 classes
	Crops	
k120	common wheat	ha
k121	durum wheat	ha
k122	rye	ha
k123	barley	ha
k124	oats	ha
k125	summer cereals	ha
k126	maize	ha
k127	rice	ha
k129	dry pulses	ha
k128	others cereals	ha
k130	potatoes	ha
k131	sugar beat	ha
k133	hops	ha
k134	tobacco	ha
k135	industrial crops	ha
k136	fresh vegetables in open field	ha
k137	fresh vegetables in market garden	ha
k138	mushrooms	ha
k140	flowers open air	ha
k141	flowers protected	ha
k152	fruit	ha
k153	citrus orchards	ha
k154	olive groves	ha
k115	vines	ha
k156	permanent crops protected	ha
k157	nurseries	ha
se035	ha in cereals	
se050	ha in vineyards	
se055	ha in orchards	

se060	ha in olive groves	
a40	UAA under irrigation	ha
Livestock		
d23, . . . , d32	cattle	head
d40, d41	sheep	head
d38, d39	goats	head
d43, . . . , d46	pigs	head
d47, . . . , d49	poultry	head
Labour force		
se010	total labour input	AWU
se011	labour input	hours
se015	unpaid labour input	AWU
se016	unpaid labour input	hours
se020	paid labour input	AWU
se021	paid labour input	hours
se025	total UAA	ha
Farm activities		
se420	f-h income	€
Notes: *ha = hectare; UAA = Utilised Agricultural Area; € = amount of Euro; SFP = Single Farm Payment; ESU = European Size Unit; AWU = Annual Working Unit; f-h = farm-household.*		

3.3 CAP-IRE 2009

CAP-IRE 2009 is the core dataset we use in this work, since it is the one we choose as recipient dataset for the imputation of variables through SM imputation techniques, previously from FADN 2009 dataset and after from SPA 2005 one. The CAP-IRE 2009 survey has been produced within the CAP-IRE 2009 project (Website, 2008), financed by EU in FP7 in 2008-2010, coordinated by the Department of Agricultural Engineering and Economics (DEIAGRA) of the Alma Mater Studiorum University of Bologna, and partners from 9 EU countries. The main aim of the project was assessing the multiple impacts of CAP reform on Europe's rural economies, focusing on 11 case study areas. Considering solely the group of farms being beneficiaries of the Single Farm Payments (SFP), for each case study area farms were randomly selected following different sampling procedures chosen autonomously by each project partner. A total amount of 2,363 units are taken into account; collected variables describe the most important changing dynamics in farms structures and activities, farm-household characteristics, business, income, investments, innovations and labour force management, plus variables regarding future behaviour intentions about socio-environmental sustainability and several governance issues under different CAP scenarios.

We focus on the Emilia-Romagna Region farm sample, made up of 300 units and 239 variables. The sample was constructed by telephone interviews; farms were chosen by random selection from the regional list of SFP beneficiaries, stratified by:

- territory, following the altitude division among plain, hill and mountain;
- farms amount of the SFP;

In table 3.2 we show the "useful" variables we have at disposal in the CAP-IRE 2009 data source. We focus on 35 variables, the ones which could be used in order to run the SM imputation using both FADN 2009 and SPA 2005 data, but also which could be used in a second step for the PSM causal effects analysis. We report here variables codes in the original data source, their descriptions and a brief note about their quantitative/qualitative nature. Since the original dataset is Italian-based, we report here the variables with their translated codes and/or descriptions. Symbols and abbreviations are explained in table footnotes.

Table 3.2: CAP-IRE 2009 chosen variables

Variable	Description	Notes
Farm general information		
3.01	farm corporate organisation	
3.03	specialisation	16 modalities
3.07	PSR attendance for AES	
3.08	bio productions	
3.09a	TAA owned	ha
3.09b	TAA rented out	ha
3.09c	TAA rented in	ha
3.17a	SFP founds	€
3.17b	others founds	€
Crops		
no variables at disposal		
Livestock		
3.04a, ..., 3.04c	bovine	head
3.04h	ovine	head
3.04d, 3.04e	pigs	head
3.04g	adult poultry	head
Labour force		
2.03	highest education level in f-h	
2.04	agricultural education	
2.05	f-h members unemployed	No.
2.06a	f-h members part-time employed	No.
2.06b	f-h members full-time employed	No.
3.10a	full-time male employees extra f-h	No.
3.10b	part-time male employees extra f-h	No.
3.10c	full-time female employees extra f-h	No.
3.10d	part-time female employees extra f-h	No.
6.01	owner sex	
6.02	owner age	
6.04	owner education level	
Farm activities		
2.08	f-h income from agriculture	%
3.05	extra-agricultural activity	
3.06a	third party activities	
3.06b	food production/processing	
3.06c	products selling	
3.06d	services/leisure activities	
Notes: *ha = hectare; UAA = Utilised Agricultural Area; TAA = Total*		

Agricultural Area; No. = "number of"; € = amount of Euro; SFP = Single Farm Payment; ESU = European Size Unit; AWU = Annual Working Unit; f-h = farm-household.

3.4 SPA 2005

SPA 2005 data source (its full name is *"Indagine sulla Struttura e sulla Produzione delle Aziende Agricole"*), is a statistical survey produced by the Italian Institute of Statistics (ISTAT) on a regular basis (it should be done the third year after the General Census on the Italian Agriculture and, then, every two years). It is made on the basis of a representative sample of the Italian farms, drawn from the General Census on the Italian Agriculture, made every 10 years on the Italian agricultural holdings. SPA 2005 data are constructed using a questionnaire similar to the one used for the General Census analysis, slightly modified in order to take into account less variables than the Census ones. Data released are always properly manipulated, through a standardized procedure, in order to reduce the risks of privacy violations and made them available and accessible to universities and research institutions in the form of "elementary data for the research". SPA 2005 data are subjected to secrecy constraints; they are given to researchers following a precise bureaucratic scheme and under mandatory release constraints. SPA 2005 data contain 319 variables for a total amount of 47,780 units, representing, as the Methodological Note attached to data release package reports, "1,728,532 agricultural holdings in Italy". With respect to the other data sources we use, SPA 2005 dataset represents the most difficult data to manage, because of the few generic released information concerning both the sample construction and the pre-release variables aggregation and removal.

We focus on Emilia-Romagna Region farm sample, made up of 2,936 units and 319 variables. Emilia-Romagna Region farm sample is stratified by:

- territory, using NUTS levels;
- agricultural holdings size, considering the farm size in terms of UAA or LUs (Livestock Units);
- agricultural holdings economic size, using the Gross Farm Income.

In table 3.3 we show the "useful" variables we have at disposal in SPA 2005 data source. We focus on 74 variables, the ones which could be used in order to run the SM imputation with the CAP-IRE 2009 dataset and, then, the PSM causal effects analysis. We report here variables codes in the original data source, their descriptions and a brief note about their quantitative/qualitative nature. Since SPA 2005 data are Italian-based, we report here variables with translated codes and/or descriptions. Symbols and abbreviations are explained in table footnotes.

Table 3.3: SPA 2005 chosen variables

Variable	Description	Notes
Farm general information		
a03	environmental restrictions	
a06	OTE	52 modalities
a07	NUTS2 (IT Region)	
a09	extrapolation factors	
a11	UAA	ha
a12	Gross Standard Income	€
b0102	farm juridical personality	
h01	UAA	ha
cc01	UAA owned	ha
cc02	UAA rented	ha

cc05a	biological agriculture	
cc05f1	public funds for investments	
cc05f2	public funds for rural development	
	Crops	
d01	total cereals	ha
d07	rice	ha
d09	total dry pulses	ha
d10	potatoes	ha
d14	total garden open air	ha
d14a	garden open field	ha
d14b	industrial garden	ha
d15	garden protected	ha
g01	fruit	ha
g02	citrus orchards	ha
g03	olive groves	ha
g04	vine	ha
i03b	total irrigated area	ha
	Livestock	
j02, ..., j08	bovine	head
j09	sheep	head
j10	goats	head
j11, ..., j13	pigs	head
j13	pigs	head
j14, ..., j16	poultry	
	Labour force	
a13	AWU entrepreneur	
a14	AWU owner	
a15	AWU entrepreneur's spouse	
a16	AWU entrepreneur's family	
a17	AWU others full-time	
a18	AWU others part-time	
l011	entrepreneur sex	
l012	entrepreneur class of age	
l01a1	owner sex	
l01a2	owner class of age	
b03	owner education level	
l03c1t	entrepreneur's family AWU 0-25%	No.
l03c5t	entrepreneur's family AWU 100%	No.
l04a2t	extra family AWU 25-50%	No.
l04a5t	extra family AWU 100%	No.
l04b2t	extra family female AWU 25-50%	No.

l04b4t	extra family female AWU 75-100%	No.
Farm activities		
m01a	agritourism	
m01b	craftsmanship	
m01c	food production/processing	
m01d	wood processing	
m01e	aquaculture	
m01f	energy production	
m01g	third party activities	
Notes: *ha = hectare; UAA = Utilised Agricultural Area; TAA = Total Agricultural Area; No. = "number of"; € = amount of Euro; SFP = Single Farm Payment; OTE = Orientamento Tecnico-Economico; ESU = European Size Unit; AWU = Annual Working Unit; f-h = farm-household.*		

Tables 3.4, 3.5, 3.6 and 3.7 synthesize the correspondence among the "useful" variables (with respect to our research purposes), at disposal in the three datasets, referring to application possibilities of both the SM imputation and the PSM causal effects analysis methodologies. They refer to the four macro-areas of farms characteristics that we want to take into account: *i.* the farm general information, *ii.* the cultivated crops, *iii.* the labour force management and, *iv.* the farm activities (for sake of brevity, we omit the discussion of the macro-area referring to the livestock variables since we do not use it in our application). The four tables can give a brief but incisive idea of how tight the overlap among variables at disposal can be even when they are collected by institutions organised in a hierarchical structure and collaborating together in order to collect similar data. Moreover, they can also give the idea of the logic which guides the harmonization procedure described in chapter 4. All variables codes and descriptions in the figures are reported in the original language of the respective data source.

Table 3.4: Overlap among "farm general information" variables

SPA 2005		CAP-IRE 2009		FADN 2009	
code	description	code	description	code	description
a06	ote	3.03	specializzazione	tf14	farm specialisation
a07	NUTS2			nuts2	NUTS2
				nuts3	NUTS3
a09	fattore di estrapolazione			sys02	farms represented
				id	unique farm ID
				sys12	farms represented (cluster)
				sys13	sample farms (cluster)
				a27	economic size
				se005	economic size
				a39	less favoured area
				a41	altitude zone
				a44	structural funds
				a45	environmental constraints
a12	Reddito Lordo Standard			se410	Gross Farm Income
				se425	Farm Net Value Added/AWU
a11	SAU			cluaa	classes of UAA
b0102	personalità giuridica	3.01	forma societaria azienda	a18	farm organisational form
		3.02.a	proprietari out famiglia capo azienda		
		3.02.b	proprietari non parenti capo azienda		

Table 3.5: Overlap among "crops" variables

SPA 2005		CAP-IRE 2009		FADN 2009	
code	description	code	description	code	description
cc01	SAU prorietà	3.09.a	SAT proprietà	b48	UAA in occupation
cc02	SAU affitto	3.09.b 3.09.c	SAT affitto out SAT affitto in	se030	UAA rented
				se025	TAA
cc05a	SAU in bio	3.08	produzioni bio	a32	organic farming
				b50	UAA in share cropping
cc05f1	aiuti investimenti produttivi	3.17.a	finanziamenti 2008 SFP	se631	SFP
cc05f2	aiuti misure sviluppo rurale	3.17.b	finanziamenti 2008 altro		
d01	frumento				
				k120	common wheat
				k121	durum wheat
				k122	rye
				k123	barley
				k124	oats
				k125	summer cereals
				k126	maize
				k128	oth. cereals
				se035	area in ha - cereals
d07	riso			k127	rice
d09	proteaginose			k129	dry pulses
				k130	potatoes
				k137	fresh vegetables market gardens
				k136	fresh vegetables open field
				k140	flowers open field
				k131	sugar beat
				k133	hops
				k134	tobacco
				k135	industrial crops
				k138	mushrooms
				k141	flowers protected
				k156	permanent crops protected
				k157	nurseries
d14	ortive piena aria				
d14a	ortive campo pieno			k136	fresh vegetables open field
d14b	ortive industriali				
d15	ortive protette				
e	orti familiari			se055	area in ha - orchards
g01	frutteti			k152	fruit
g02	agrumi			k153	citrus orchards
g03	uliveti			k154	olive groves
				se060	area in ha - olive groves
g04	vigneti			k155	vines
				se050	area in ha - vineyards
i03b	superficie irrigata totale			a40	UAA under irrigation

Table 3.6: Overlap among "labour force" variables

SPA 2005		CAP-IRE 2009		FADN 2009	
code	description	code	description	code	description
l011	sesso conduttore				
l012	età conduttore				
a13	lavoro conduttore				
l013	% ore lavorate conduttore				
l01a1	sesso capo azienda	6.01	sesso capo azienda		
l01a2	età capo azienda	6.02	età capo azienda		
b03	titolo di studio capo azienda	6.04	titolo di studio capo azienda		
		2.04	istruzione agricola in famiglia		
a14	lavoro capo azienda				
l01a3	% ore lavorate capo azienda				
		2.03	titolo di studio più alto in famiglia		
l021	sesso coniuge				
l022	età coniuge				
a15	lavoro coniuge				
l023	% ore lavorate coniuge				
				se010	total labour input AWU
				se011	labour input hours
				se015	unpaid labour input AWU
				se016	unpaid labour input hours
				se020	paid labour input AWU
				se021	paid labour input hours
a16	lavoro familiari conduttore	2.06.a	familiari full-time		
		2.06.b	familiari part-time		
a17	lavoro altri continuato	3.10.a	dipendenti M full-time		
		3.10.c	dipendenti F full-time		
a18	lavoro altri saltuario	3.10.b	dipendenti M part-time		
		3.10.d	dipendenti F part-time		
l07	attività extra agricola conduttore o capo azienda	2.08	% reddito lordo famiglia da attività agricola		
l08	attività extra agricola coniuge				
l10	giorni di lavoro dipendenti				

Table 3.7: Overlap among "farm activities" variables

SPA 2005		CAP-IRE 2009		FADN 2009	
code	description	code	description	code	description
m01a	attività extra agriturismo	3.06.d	attività extra servizi/ricreative		
m01b	attività extra artigianato				
m01c	attività extra lavorazione alimenti	3.06.b	attività extra lavorazioni alimentari		
m01d	attività extra artigianato				
m01e	attività extra acquacoltura				
m01f	attività extra energie rinnovabili				
m01g	attività extra contoterzismo	3.06	attività extra contoterzismo		
m01h	attività extra altro				

Chapter 4

SM imputation application

4.1 Introduction

Chapter 4 describes the different applications of the combinations of non-parametric micro SM imputation techniques with not default distance functions. The applications are divided into the three following macro-steps, i.e.:

1. the datasets harmonization;
2. the imputation building and running;
3. the synthetic dataset analysis (i.e. the imputation goodness validation and the results discussion).

Each one of these steps is repeated for the four imputation applications we run and we define in the following way:

- Imp 1: FADN 2009 1 (donor) and CAP-IRE 2009 (recipient);

- Imp 2: FADN 2009 2 (donor) and CAP-IRE 2009 (recipient);
- Imp 3: FADN 2009 3 (donor) and CAP-IRE 2009 (recipient);
- Imp 4: SPA 2005 (donor) and the best synthetic dataset previously generated (recipient).

We build three different FADN 2009 donor datasets (see paragraph 4.2.1 for further details); we then run the SM imputation with CAP-IRE 2009 as the recipient dataset. Consequently we run an imputation between the SPA 2005 (donor) dataset and the best synthetic dataset chosen among the ones created by the previous SM imputation, generating the new final synthetic dataset named NEW CAP-IRE 2009. Each SM imputation application for the different donor datasets is structured upon a standardized procedure based on a descriptive analysis of each dataset at disposal (in order to analyse similarities and differences in the datasets structures, possible paths for SM imputation running, etc.). Secondly, we proceed to the datasets harmonization (considering the object of impacts evaluation, the time span, the observed units characteristics, the covariates influencing the treatment, etc.). Thirdly, we set the imputation itself choosing the matching variables, eventually building donation classes, choosing the variables to be imputed, deciding which combination of SM imputation technique and distance function has to be applied. Then, we run the SM imputation. Finally, we check the imputation results, validating the imputation goodness with respect to the synthetic datasets generated (see paragraph 4.3.2 for further details); among these synthetic datasets we choose the best one in order to use it for the imputation from the SPA 2005 (donor) dataset. After this last SM imputation application, we create the new generated dataset named NEW CAP-IRE 2009, the one we use for the PSM application.

4.2 Data harmonization

Data harmonization results to be, inevitably, a highly data-driven procedure, not so easily manageable through a standardized process, not even if all the data sources to be harmonized have the same reference framework and are produced by the same statistical agencies and/or for the same analysis purposes. Since the present work uses for the application part, two official administrative data sources managed by two different statistical agencies, built with different designs and through different reference frameworks, and a project survey which follows its own design and its own analysis purposes, the harmonization procedure difficulty certainly increases. Indeed, data harmonization among FADN 2009, CAP-IRE 2009 and SPA 2005, requires the managing of several practical problems, such as: *i.* the linguistic differences among the three data sources and the consequent differences in variables codes, modalities and descriptions, *ii.* the different expressed modalities that even the similar variables have (for example, farm owner age is expressed in years in CAP-IRE 2009 but in age classes in SPA 2005), *iii.* the need of proxy variables in order to cover for variables which are not exactly the same or similar at least, *iv.* the problem of treating the missing values and the outliers.

We present, anyway, a recursive harmonization procedure which can be applied to farm data sources even for further developments concerning others datasets and others farm samples. Data harmonization represents a key preliminary step for both the SM imputation and the consequent PSM analysis. Indeed, to properly work, both these procedures require homogeneous and complete datasets; the issue of missing values is, for example, a thorny one to face at. Moreover, computationally speaking, the solely presence of key useful variables between the donor and recipient datasets, represents an important benefit for the running of the SM imputation. In the present work, data harmonization is then pursued with two fundamental goals: first, it is a necessary preliminary step for the SM imputation and the PSM analysis; second, due to the absence of an official reference framework and/or a common archive on Italian farm data (neither in ISTAT nor in Eurostat -FADN data do not constitute a complete and fully reliable farm data source but for the accounting information-), we present an embryonic recursive procedure for farm data harmonization.

The first step of the data harmonization procedure, common to all the three data sources, consists in a mere translation of the variables codes, modalities and descriptions. We translate from Italian to English, coherently with the FADN 2009 data source framework. This first step completed, we progressively harmonize the three different data sources, as described in the following paragraphs.

4.2.1 FADN 2009 harmonization

FADN 2009 data for the Emilia-Romagna Region, originally concern a sample of 1,054 units and 859 variables, which we reduce through the harmonization procedure, to 937 units and 407 variables. Previous to data harmonization, we carry a complete descriptive analysis in order to decide the main variables dropping, concerning variables related to the questionnaire description (such as the variable indicating the year of the survey), the redundant variables (such as the variable indicating the country of the observed farm -when we do have the *nuts0* variable indicating the NUTS0 stratus-), and the unusable ones, such as the variable referring to the sampling clusters, unusable due to the inaccessibility of the complete and detailed FADN data methodological note and/or beyond of its use in a wider FADN data panel.

In order to properly run SM imputation between FADN 2009 and CAP-IRE 2009, considering that we can not build donation classes, we transform the *tf14* variable indicating farms specialisation in a quantitative one, a strategic operational choice. For FADN 2009 *tf14* variable, we decide to maintain the original modalities, regrouping or dropping some of them but keeping this variable and its modalities as framework reference for the other farms specialisation variables present both in CAP-IRE 2009 and SPA 2005. This choice is motivated by the fact that TF14 categories, defined by DG Agri, are or at least should be the reference categories indicating farms specialisation in the European agripolicy research context. Therefore, we rename each *tf14* modality keeping their core descriptions, deciding to drop units with modalities "37: specialist olives" and "70: mixed livestock", of the *tf14* variable. We drop farms with specialisation in olives because of the overlap issues with the recipient dataset CAP-IRE 2009 in which such specialized farms are missing (and this would weaken and/or obstruct the SM imputation application through non-parametric micro techniques). For similar reasons (no such modality does exist for the *specialisation* variable observed in CAP-IRE 2009), we decide the latter drop. We decide then to aggregate modalities "48: specialist sheep & goats", "49: specialist cattle" referred to farms specialised in bovine but not farms exclusively dairy, "50: specialist granivores", in the new created modality "50: livestock (no dairy)". Also, we aggregate modalities "38: various permanent crops combined", "60: mixed crops" and "80: mixed crops & livestock", in the modality "80: mixed crops & livestock". Table 4.1 shows the re-coding procedure of the *tf14* variable modalities.

Table 4.1: Re-coding scheme for the variable *tf14*

existing *tf14*		re-coded *tf14*	
farm specialisation		farm specialisation	
15	COP rice COP & rice	15	cereals
16	root crops cereals & root crops field vegetables field crops	16	seminative (others)
20	horticulture	20	horticulture
35	wine	35	wine
36	fruit	36	fruit
37	olives	DROPPED	
38	various permanent combined cereals fruit vine	80	mixed crops & livestock
45	milk	45	dairy
48	sheep & goats heep sheep & cattle goats various grazing livestock	50	livestock (no dairy)
49	cattle cattle rearing dairying, rearing & fattening		
50	granivores pigs poultry various granivores combined		
60	mixed crops market gardens & permanent field crops & market gardens field crops & permanent mixed crops mainly field field crops & vine	80	mixed crops & livestock
70	mixed livestock	DROPPED	
80	mixed crops & livestock	80	mixed crops & livestock

The other key variables we use both for the SM imputation and PSM analysis, beyond the *tf14* one, are all the "*k*" variables indicating hectares of cultivated crops and quantities of productions measured at different times. We drop around 390 "*k*" variables showing all missing entries, almost all referred to the produced quantities observed at different times. Then, we maintain only the "*k*" variables indicating the hectares of the specific crops cultivated by the farm, aggregating them by logic and taking into account the overall framework of variables at disposal among the three datasets. Table 4.2 summarizes the aggregation procedure of the "*k*" variables values.

Table 4.2: "*k*" variables values summed

original variable	description	new variable
k120	common wheat	cereals
k121	durum wheat	
k122	rye	
k123	barley	
k124	oats	
k125	summer cereals	
k126	maize	
k128	cereals (others)	
k130	potatoes	gardens
k136	fresh vegetables open field	
k137	fresh vegetables market gardens	
k140	flowers open air	
k138	mushrooms	under_glass
k141	flowers protected	
k156	permanent protected	
k140	nurseries	
k131	sugar beat	industrial_crops
k133	hops	
k134	tobacco	
k135	industrial crops	

With respect to the other "*k*" variables which are not aggregated, we merely rename them in the following way:

- *k127aa* → *rice*;

- *k129aa* → *dry_pulses*;
- *k152aa* → *fruit*;
- *k153aa* → *citrus_orchards*;
- *k154aa* → *olive_groves*;
- *k155aa* → *vine*.

Given these new aggregated variables indicating the hectares of the cultivated crops, we decide to generate the variable *uaa_tot* (representing the total UAA -Utilised Agricultural Area-), as the sum of the values of all the above-mentioned variables, in order to use it for the SM imputation between FADN 2009 (donor) and CAP-IRE 2009 (recipient).

A similar procedure is then run in order to re-code and aggregate (or rename) those variables indicating the livestock units. Nevertheless, since these variables are discarded from both the SM imputation and PSM analysis applications, for sake of brevity, we do omit them.

All the "*l*" and "*m*" variables are then dropped, due to the fact that we are not able to use them with respect to the other two datasets which lack of the variables indicating, respectively, quotas and rights but also crops subsidies and direct payments.

Final steps of the data harmonization concerning FADN 2009 focus on the building of three different FADN 2009 donor datasets, which are completely similar but for the way the respective *taa* variables are constructed. Indeed, the donor dataset FADN 2009 1 has the *taa* variable generated as the sum of the hectares of the cultivated crops expressed by the "*k*" variables, proportional to the TAA (Total Agricultural Area) of the Emilia-Romagna Region farms in the year 2009, as reported by the Regional Statistical Office (Website, 2004). Therefore we use the above-mentioned variable *uaa_tot*, adjusted by the ratio of UAA and TAA for the Emilia-Romagna Region farms in 2009. The donor dataset FADN 2009 2, instead, has the *taa* variable generated merely renaming the originally existing *se025* variable; in this case, we decide to use the original indication of the total UAA of the farm as if it was its TAA. Finally, FADN 2009 3 dataset has the *taa* variable generated as the sum of the hectares of the cultivated crops expressed by the "*k*" variables. Then, we use again the above-mentioned variable *uaa_tot*, without any adjustment. Data harmonization for FADN 2009 data ends with the check of the eventually empty cells; since they can not be processed by the SM imputation, we try to prevent the impossibility of SM imputation running by dropping units which have all the values of the renamed and/or aggregated crops variables, equal to 0. Also units with *taa* variable values equal to 0 are deleted.

Summing up, harmonization for FADN 2009 data concerns a previous descriptive analysis of the dataset, crucial in order to know its structure and the variables at disposal. We drop then the main useless variables, such as, for example, variables which are not useful for research purposes and/or variables which can not be used due to practical constraints and/or which characterise the observed units for the presence of several missing values. We carefully harmonize the *tf14* variable modalities (one of the most important matching variables), even dropping units with specific *tf14* variable modalities. We harmonize the "*k*" variables (ours imputation ones), we then build the different variables indicating the TAA of farms and which both characterise each donor dataset and is used as second matching variable.

4.2.2 CAP-IRE 2009 harmonization

CAP-IRE 2009 data are originally constituted by 300 units and 239 variables which we reduce, after the harmonization procedure, to 289 units and 77 variables. We explore the dataset through a complete descriptive analysis, useful in order to decide the main variables dropping. In CAP-IRE 2009, we initially drop several variables related to the questionnaire description, such as the ones indicating the date and the time of the survey, the interviewer name, the duration, etc. Moreover, we drop all the variables indicating the future behaviour intentions about socio-environmental sustainability and several governance issues under different CAP scenarios. Around 160 variables are then dropped being unusable for our research purposes.

In order to properly run the SM imputation between CAP-IRE 2009 and FADN 2009, considering that we can not build donation classes, we transform the *tf14* variable indicating farms specialisation in a quantitative one, following the same strategy cited in the previous paragraph. With respect to the FADN 2009 *tf14* variable, the harmonization procedure for the *specialisation* variable in CAP-IRE 2009 is deeper. We obviously modify the original modalities coherently with the ones expressed by the *tf14* variable in the FADN 2009 dataset, but aggregating more modalities than the ones aggregated for the *tf14* one. As we do for FADN 2009, in CAP-IRE 2009 we drop units with modalities “6: uliveti” and “77: non classificabile” of the *specialisation* variable. We decide then to aggregate modalities "9: bovini da ingrasso", "10: bovini da latte & ingrasso", "11: ovini & altri da pascolo", "12: avicoli", in the new created modality "50: livestock (no dairy)". Also, we aggregate modalities "7: colture permanenti miste", "13: colture miste", "16: colture & animali da pascolo", "17: colture miste & allevamento" in the renamed modality "80: mixed crops & livestock". Table 4.3 shows the re-coding procedure of the *specialisation* variable modalities (they are expressed as they are in the original Italian dataset).

Table 4.3: Re-coding scheme for the variable *specialisation*

existing *specialisation*		re-coded *tf14*	
farm specialisation		farm specialisation	
1	cereali oleaginose proteiche	15	cereals
2	altri seminativi	16	seminative (others)
3	orticole	20	horticulture
4	vigneti	35	wine
5	frutta & agrumi	36	fruit
6	uliveti	DROPPED	
7	permanenti miste	80	mixed crops & livestock
8	bovini da latte	45	dairy
9	bovini da ingrasso	50	livestock (no diary)
10	bovini da latte & ingrasso		
11	ovini & altri da pascolo		
12	avicoli		
13	colture miste	80	mixed crops & livestock
16	colture & animali da pascolo		
17	colture miste & allevamento		
77	non classificabile	DROPPED	

Being the crops variables the ones we choose to impute from the FADN

2009 (donor) dataset to the CAP-IRE 2009 (recipient) one (due to the lack, in this latter one, of the variables indicating the hectares of the cultivated crops), the last part of the data harmonization procedure for CAP-IRE 2009 ends with the fixing procedure of the missing values of the variables *land_owned*, *land_rent_out* and *land_rent_in*, which are all replaced, if present, with the value 0 indicating 0 hectares of TAA. Then, we create the variable *taa* as the sum of the values of the variables *land_owned* and *land_rent_in* subtracted by the values of the variable *land_rent_out*. Finally, units with the *taa* variable values equal to 0 are deleted.

A procedure similar to the one followed for the "*k*" variables in FADN 2009, is followed in order to re-code and aggregate or rename the variables indicating livestock units in CAP-IRE 2009 (these ones are, indeed, collected). Nevertheless, since these variables are discarded from both the SM imputation and PSM analysis applications, we decide to omit them.

Summing up, harmonization for CAP-IRE 2009 data concerns a previous descriptive analysis of the dataset, crucial in order to know its structure and the variables at disposal. We drop then the main useless variables which are not useful for the research purposes. We carefully harmonize the *tf14* variable modalities (one of ours most important matching variable), even dropping units with specific *tf14* variable modalities. Finally, we harmonize the different variables indicating the TAA of farms which we decide to use as our second matching variable.

4.2.3 SPA 2005 harmonization

SPA 2005 data are originally constituted by 2,936 units and 319 variables which we reduce, through the harmonization procedure, to 2,912 units and 260 variables. Previous to the data harmonization we carry out a complete descriptive analysis in order to decide the main variables dropping, concerning the variables related to the pre-release anonymisation procedures done by ISTAT, such as all the empty completely useless "*filler*" variables. Also another group of empty variables created by ISTAT during the pre-release procedures, filled with ":" missing symbols, are dropped. Most part of these variables, in the pre-released data source are key-informative variables which have to be sacrificed in order to respect the imposed privacy constraints and dropped and/or aggregated on a upper level of detail (here, more than 50 key-informative variables).

Since we can not build donation classes for the SM imputation from the SPA 2005 (donor) dataset to th CAP-IRE 2009 (recipient) one, we transform the *ote* variable indicating the farms specialisation in a qualitative variable. Even for this dataset, as it is for the CAP-IRE 2009 one, the harmonization procedure is deeper than that one we follow for the FADN 2009 dataset. First of all, we modify the original modalities coherently with the ones expressed by the *tf14* variable in FADN 2009, dropping units with modalities "33: olivicoltura", "711: poliallevamento per latte", "712: poliallevamento non latte", "721: granivori & bovini per latte", "722: granivori & erbivori non bovini", "723: granivori & misto", "511: suini" and "9: non classificabili" of the *ote* variable.

We aggregate, coherently with the previous aggregation of the variables modalities indicating the specialisation of farms both in FADN 2009 and CAP-IRE 2009, the modalities “421”, “422”, “431”, “432”, “441”, “444”, “502”, “503” in the new created modality “50 : livestock (no dairy)”. We also aggregate the modalities “34”, “601”, “602”, “603”, “604”, “605”, “606”, “811”, “812”, “813”, “814”, “821”, “822”, “823” in the modality “80 : mixed crops & livestock”. Table 4.4 shows the re-coding procedure for the *ote* variable modalities (they are expressed as they are in the original Italian dataset).

Table 4.4: Re-coding scheme for the variable *ote*

existing *ote*		re-coded *tf14*	
farm specialisation		farm specialisation	
131	COP	15	cereals
132	risicole		
133	COP & risicole		
141	sarchiate	16	seminative (others)
142	cereali & sarchiate		
143	orti pieno campo		
144	seminativi vari		
201	orti industriali	20	horticulture
202	floricoltura & ornamentali		
203	ortofloricole & risicole		
311	vini di qualità	35	wine
312	vini non di qualità & sarchiate		
313	vini combinati		
314	vini varie denominazioni		
321	frutta	36	fruit
33	olivicoltura	DROPPED	
34	varie permanenti combinate	80	mixed crops & livestock
411	latte	45	dairy
412	bovine da latte		
421	bovine	50	livestock (no dairy)
422	bovine da ingrasso		
431	latte & bovine per carne		
432	bovine per carne & latte		
441	ovini		
444	erbivori vari		
501	suini	DROPPED	
502	pollame	50	livestock (no dairy)
503	granivori combinati		
601	ortofloricoltura & permanenti	80	mixed crops & livestock
602	seminativi & ortofloricoltura		
603	seminativi & vigneti		
604	seminativi & permanenti		
605	policoltura & seminativi		
606	policoltura & ortofloricoltura		
811	miste seminativi & bovini		
812	miste bovini & seminativi		
813	miste & erbivori		
814	miste erbivori		
821	miste seminativi & granivori		
822	miste permanenti & erbivori		
823	miste & misti		
9	non classificabili	DROPPED	

In order to run the SM imputation between SPA 2005 and the best synthetic dataset generated from the above-mentioned SM imputation applications (i.e. the best one generated from the imputations between the FADN 2009 (donor) datasets and the CAP-IRE 2009 (recipient) one), we have to harmonize also the SPA 2005 dataset with respect to this consequent imputation.

We create the variable *legal_status* indicating the farms organisational form, distinguishing between the corporation (including family) farms and the sole proprietorship farms. We decide to reduce the seven modalities existing originally for the variable *legal_status* to only two because of the overlap issues which would be present during the SM imputation with the best synthetic dataset chosen. This decision does not imply a loss of information since the most part of the farms in the sample do have one of these two modalities. Then, we recode the *sex* variable from the existing modalities in order to create a dummy, create the variable *edu_agri* indicating the presence of the agricultural education for the farm owner (distinguished among "none"-"basic"-"practical" modalities). We aggregate and re-code the variable *age* modalities creating another variable *age_cl* which is expressed in age classes. Finally, we create a variable *crops* indicating the mere number of different crops cultivated by the farm.

For sake of brevity, we omit the discussion of the other variables involved in this further harmonization procedure, being the above-mentioned ones the most relevant with respect to both the SM imputation between SPA 2005 and the best synthetic dataset chosen and the PSM analysis application.

4.2.4 The chosen best synthetic dataset harmonization

In order to properly run the SM imputation between the best synthetic dataset among the ones generated from the imputations between the FADN 2009 donor datasets and the CAP-IRE 2009 recipient one, and the SPA 2005 dataset harmonized as previously described, we have to harmonize also this chosen dataset. Firstly, we create the "treatment" variable t indicating if the farm uptake AES ($t = 1$) or not ($t = 0$); secondly, we re-code the variable *legal_status* indicating the farm organisational form (harmonized with the correspondent variable in the SPA 2005 dataset). Then, the same process previously described for the variable *edu_agri* in the SPA 2005 dataset, is followed for the correspondent variable in the best synthetic dataset chosen. Fourthly, we create a dummy variable for the variable *sex*. With respect to the variable *age* we create a new variable *age_05* referred to the year 2005, plus a new variable *age_cl_05*, properly adjusted for the age 4 years before. Finally, we create a variable *crops* equal to the above-mentioned one in the SPA 2005 donor dataset.

For sake of brevity, we omit the discussion of the other variables involved in this further harmonization procedure, being the above-mentioned ones the most relevant with respect to both the SM imputation between SPA 2005 and the best synthetic dataset chosen and the PSM analysis application.

4.3 SM application and results

We conduct four different SM imputations (with the different combinations of the "hot deck" techniques and distance functions), among the four datasets we have at disposal, i.e.:

1. Imp 1: FADN 2009 1 (donor) and CAP-IRE 2009 (recipient);
2. Imp 2: FADN 2009 2 (donor) and CAP-IRE 2009 (recipient);
3. Imp 3: FADN 2009 3 (donor) and CAP-IRE 2009 (recipient);
4. Imp 4: SPA 2005 (donor) and the best synthetic dataset previously generated (recipient).

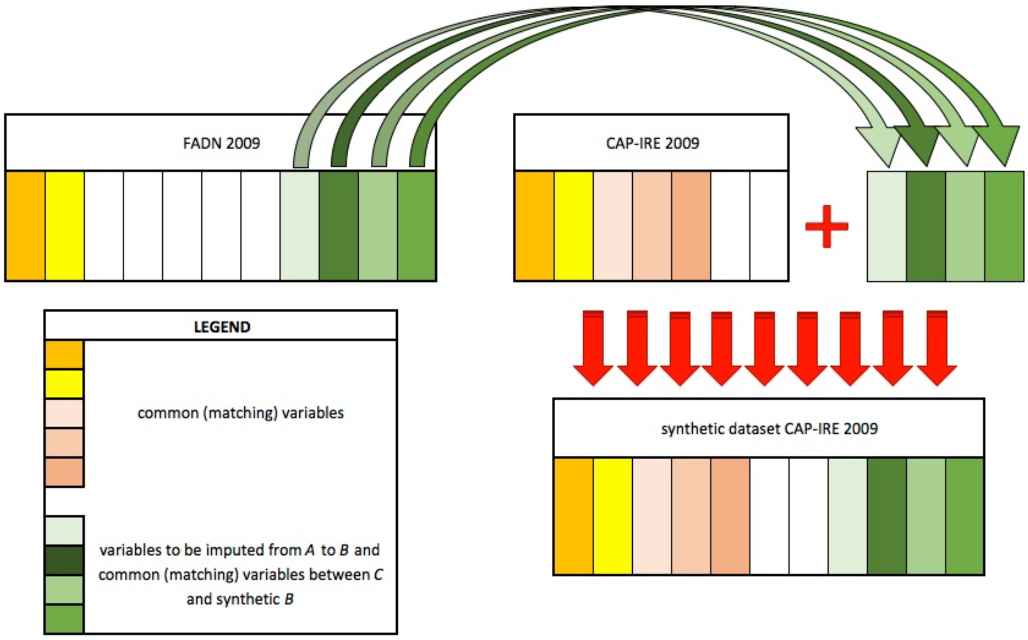
We divide the standard SM imputation process into three main steps:

1. the datasets harmonization;
2. the imputation building and running;
3. the synthetic dataset analysis (i.e. the imputation goodness validation and the results discussion).

The developing of the three main steps are similar for the SM imputation between the FADN 2009 (donor) datasets and the CAP-IRE 2009 (recipient) one. The SM imputation between the SPA 2005 (donor) dataset and the CAP-IRE 2009 (recipient) one, instead, differs slightly from these ones.

Figure 4.1 shows a schematic representation of Imp 1, Imp 2 and Imp 3 procedures, concerning the FADN 2009 1, FADN 2009 2 and FADN 2009 3 donor datasets and the CAP-IRE 2009 recipient one. Similarly to figure **??**, in figure 4.1 we have a common set of variables among which we choose the ones to be used as matching variables (i.e. the orange and the yellow ones). We decide then to impute the variables in green shades; this is done in order to create a synthetic dataset which is complete and homogeneous with respect to both the FADN 2009 and the CAP-IRE 2009 data.

Figure 4.1: SM imputation between FADN 2009 and CAP-IRE 2009



4.3.1 Imp 1, 2 and 3: building and running

Being FADN 2009 1, 2 and 3 the donor datasets and CAP-IRE 2009 the recipient one, we analyse the variables the two datasets do share. Since we do not have a sufficient number of shared variables between the two datasets, we choose the matching variables and the ones to be imputed into a shrink range of available possibilities. We also try to build donation classes, useful to better control the imputation process conditioning on them, but without successful results. Moreover, since units with modality "15" of the variable *tf14* (i.e. the farms which are specialised in cereals production), in the FADN 2009 dataset are lesser than in the CAP-IRE 2009 one, we decide to treat even the variable *tf14* as if it was a quantitative one. This has, obviously, consequences on the imputation goodness (that we are anyway able to check and control), but is detriment to the initial running of the imputation itself.

The imputation process consists in the setting of the matching variables, the choice of the proper combination of technique-distance function, the generation of the synthetic dataset and the extraction of donors and recipients ids, distances and, when it is the case, the number of donors available at the minimum distance. The only two shared variables between the donor and the recipient datasets are forcedly selected as matching variables: *tf14* and *taa*. The donation classes can not be neither defined nor built in order to try to better control the imputation process. Matching on the *tf14* and the taa variables, we recursively use the FADN 2009 1, 2 and 3 datasets as the donor ones and the CAP-IRE 2009 as the recipient, choosing the following variables to be imputed: *cereals*, *rice*, *dry_pulses*, *gardens*, *industrial_crops*, *under_glass*, *fruit*, *citrus_orchards*, *olive_groves*, *vine*, *se005*, *a40*, *se010*, *se011*, *se015*, *se016*, *se020* and *se021*.

In order to generate the synthetic dataset obtained as the aggregation of the imputed variables and the original ones previously present, we use combinations of the SM imputation techniques (the Nearest Neighbour Distance Hot Deck (nnd), the Constrained Nearest Neighbour Hot Deck (nndc), the Random Hot Deck (rnd), the Rank Hot Deck (rnk)) with the distance functions (Manhattan (mn), Mahalanobis (ms), Exact (e)). We stress that we adopt a particular approach for the rnk technique since this technique basically ranks the units (the donor and the recipient ones), in order to find and associate proper units pairs. Considering that one of the matching variable we have selected is taken into account as if it was a quantitative one (i.e. the variable *tf14*), we decide to try not to use the entire donor dataset to run the SM imputation but, instead, to divide both the donor and the recipient datasets into sub-datasets in which we keep recursively only those units with the same *tf14* modality. This way, we consider only the farms which have the same specialisation and, consequently, by ranking the units, the rnk technique takes into account each time just a specific modality of the variable *tf14* in the donors FADN 2009 1, 2 and 3 in correspondence of the same ones in the recipient CAP-IRE 2009. This is done to prove the performance of the rnk technique which, otherwise, without the bounds imposed by the choice of the matching variables as previously defined, systematically violates the correspondences.

For each synthetic dataset, we always extract donors and recipients ids, distances between donors and recipients and, if generated (it depends, indeed, on the kind of technique combined), the number of donors available at the minimum distance.

4.3.2 Imp 1, 2 and 3: imputation goodness validation

Imputation goodness validation is based on the robust strategy built with the simulation study. Nevertheless, for the discussion of the real data application results, we also use the “checking table”, which is the overall output obtained by the imputations ran (donors-recipients ids, distances, donors available at the minimum distance, etc.). For sake of brevity we do not discuss its use in details but we do attach it to the appendix. For each combination of SM imputation technique and distance function, this table reports all the donor and recipient matching units pairs ids, the distances between the matching donors and recipients associated, the eventually present number of available donors at the minimum distance. We use this tool as support of the imputation goodness validation strategy proposed, recursively observing the right correspondence between donors and recipients in terms of the values of the chosen matching variables, the existence both of the lowest distance between matching units pairs and the fewest number of donors at the minimum distance.

As we showed with the simulation study, the imputation goodness validation is based on a strategy which takes into account, first of all, the pre-post distributions of the matching variables. Due to the presence, in the recipient dataset (CAP-IRE 2009) of the variable *taa*, which indicates the Total Agricultural Area of the farm, and being the imputed variables the ones indicating the UAA of the cultivated crops, we sum the values of these latter ones creating a new "control" variable named *taa_imp*, adjusted by a 10% of its value. We then verify whether the distribution of the TAA in the original dataset (i.e. before the imputation), is as much closer as possible to the distribution of the TAA after the imputation. Secondly, we do analyse the correspondence between the modalities of the variable *tf14* previous and after the SM imputation application. Therefore, we calculate the differences "z", defined as the differences between the values of the TAA imputed from the FADN 2009 1, 2 and 3 (donor) datasets and the TAA originally present in the donor one (CAP-IRE 2009). We look at the distributions of these differences (with the expectation that they are as much closer to 0 as possible in order to have a good imputation fit), and, also, at their MSE values.

Table 4.5 shows the share of the variable *tf14* modalities which properly correspond with respect to the donor and the recipient matching units pairs. The best synthetic dataset generated is clearly the one that has the highest share.

Table 4.5: Share of the proper correspondence of the *tf14* modalities between donor and recipient units

Combination	Imp 1	Imp 2	Imp 3
nnd.mn	87.543%	89.619%	89.965%
nnd.ms	93.426%	97.578%	94.464%
nnd.e	99.308%	97.924%	99.308%
nndc.mn	65.744%	64.360%	66.090%
nndc.ms	78.547%	78.201%	78.547%
nndc.e	68.166%	67.820%	68.166%
rnd.mn	48.443%	54.325%	50.519%
rnd.ms	58.478%	60.901%	50.173%
rnd.e	98.616%	96.886%	94.810%
rnk	11.765%	12.803%	12.111%
rnk (sub-datasets)	100%	100%	100%

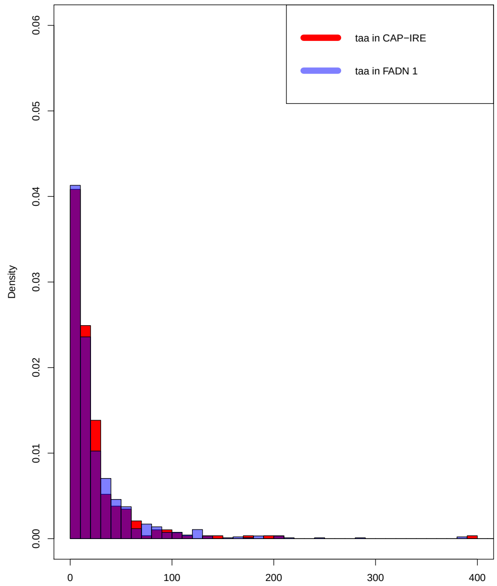
As the table shows, with the exception of the rnk technique applied to the sub-datasets expressly created in order to avoid the impossibility of building donation classes, and which report the 100% share of correspondence, the other techniques combined with the different distance functions, perform differently in the three imputations. As expected, the rnk technique applied neither building the donation classes nor creating the sub-datasets, performs the worst. It associates no more than the 12% of the units with the same *tf14* (i.e. the same specialised farms). Even the combinations of both the Manhattan and Mahalanobis distance functions within the matching algorithm of the rnd technique do not perform an overall good imputation, whereas the Exact distance function performs far better. Due to the small dimensionality of both the donor and, especially, the recipient datasets, the combinations based on the nndc technique, which constrains the Nearest Neighbour Distance Hot Deck (nnd) excluding each time the associated units, does not perform an optimal imputation (never reaching the minimum 85% of share). The best imputation results in terms of correspondence of the variable *tf14* modalities, are then obtained using the nnd technique differently combined with the three distance functions (with best results given by the application of the Exact distance function).

Taking into account the distributions of the TAA before and after the SM imputation, we look for the best overlap among the variables related to the TAA (i.e. *taa* and *taa_imp*).

Figure 4.2 shows the distributions of the variable *taa* in the CAP-IRE 2009 (recipient) and FADN 1 (donor) datasets. As we can see, the two datasets have a similar mean for the variable *taa* but the donor one has a double variance. The correspondence of this matching variable is almost good, with a significant lower presence of *taa* in the class 20-30 and also slightly under-correspondences in other classes.

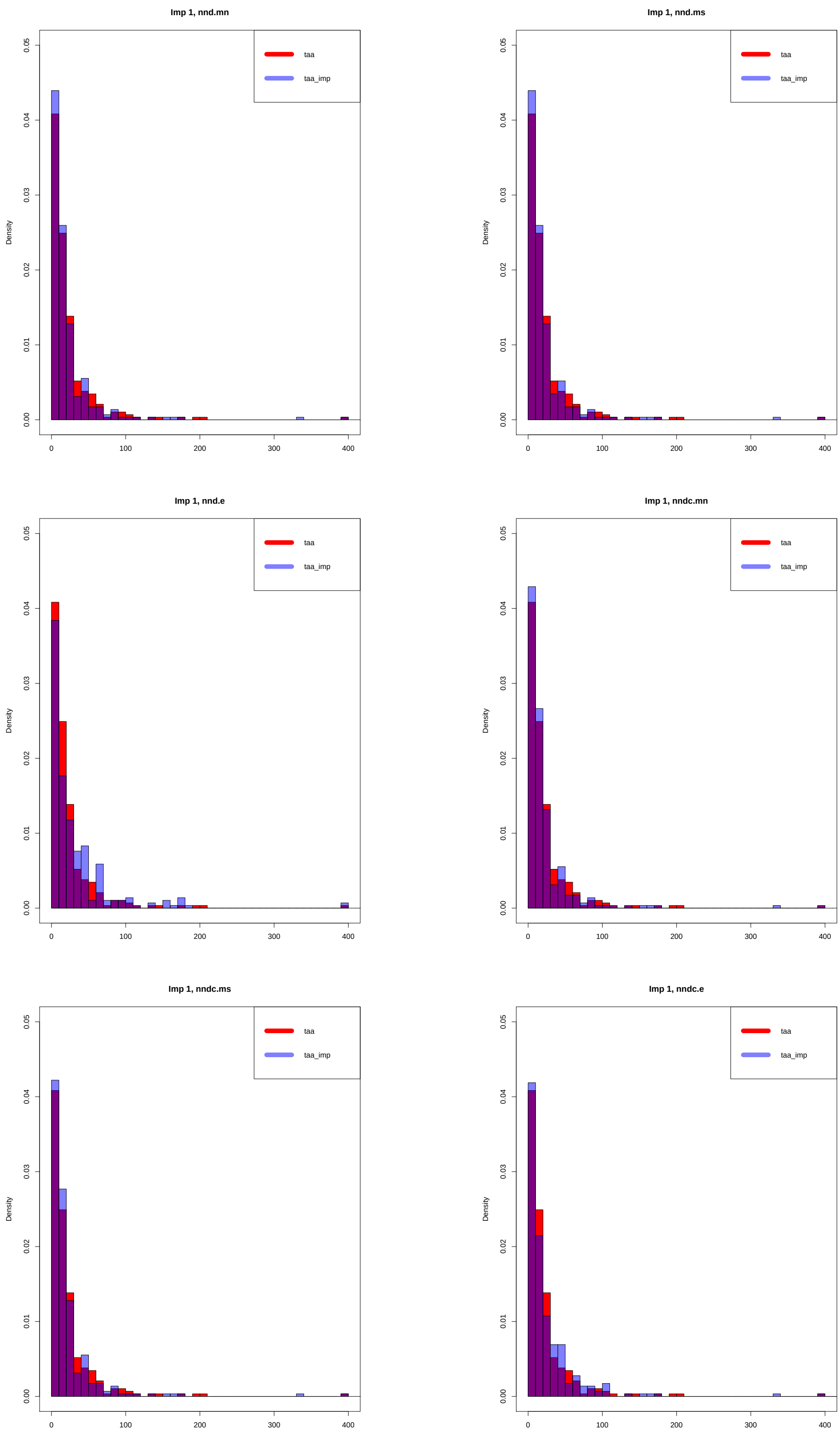
Figure 4.2: Imp 1, variable *taa* in CAP-IRE and FADN 1

taa		
	CAP-IRE	FADN 1
mean	25.972	27.994
var	2043.819	4845.096
min	1	0.038
max	470	1670.384



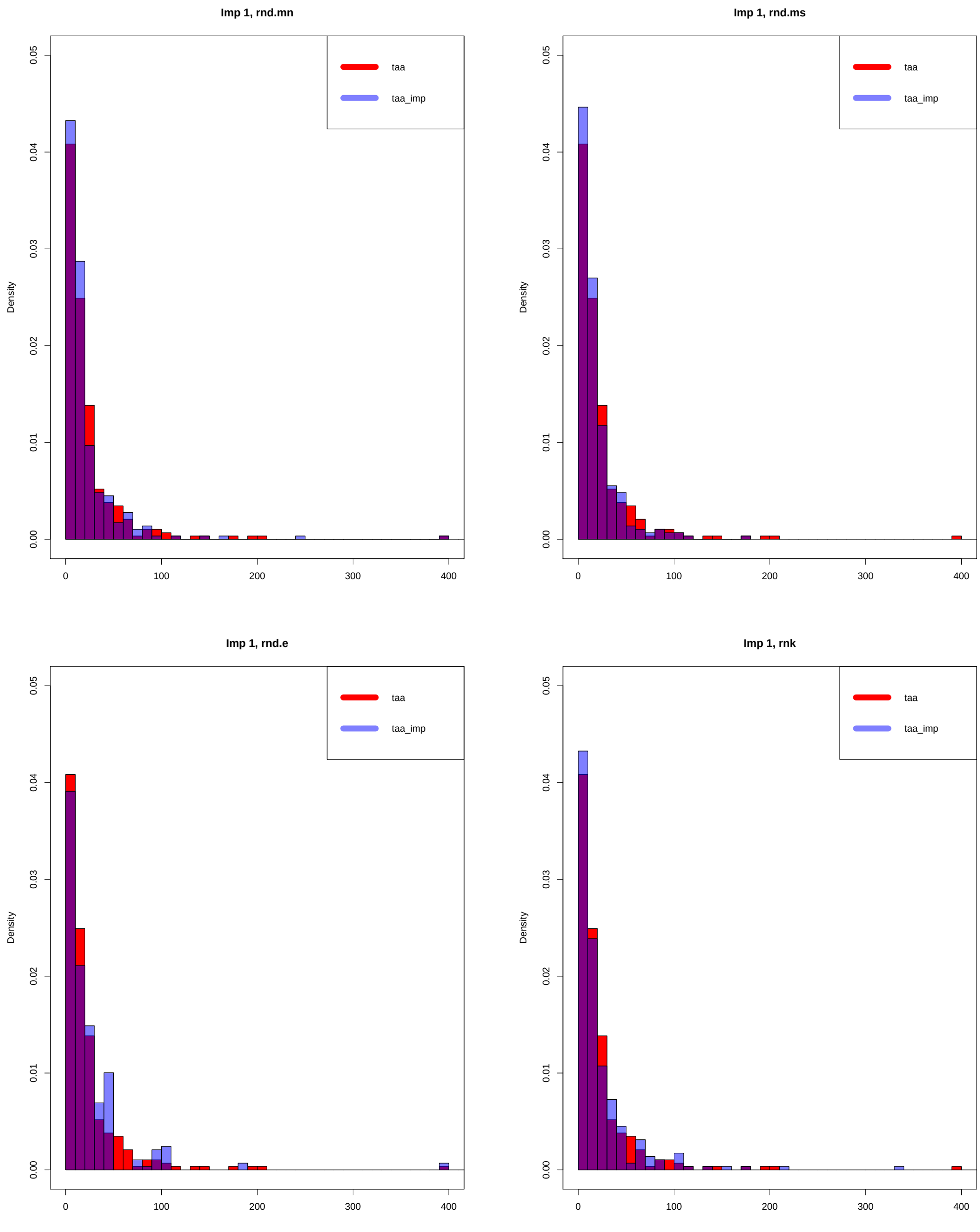
Taking into account the synthetic datasets generated using FADN 2009 1 as the donor dataset, in figure 4.3 we can see the imputation results from the different combinations of the techniques nnd and nndc with respect to the pre-post imputation distributions of the Total Agricultural Area (TAA). As we can see, nnd.mn and nnd.ms perform really similar in Imp 1 generating synthetic datasets in which the variable *taa* is overestimated in the class 0-10 whereas it is slightly underestimated in the classes 20-30, 30-40 and 50-60. The same results but more pronounced with respect to the same classes of values are obtained for the imputations with nndc.mn and nndc.ms. Worse results are obtained by the combination nnd.e (which shows both significant underestimates -for example in the class 10-20- and overestimates -for example in the classes 30-40, 40-50 and 60-70), whereas the nndc.e produces the same but less pronounced results.

Figure 4.3: Imp 1, *taa* and *taa_imp* in nnd and nndc



With respect to the rnd combinations and the rnk technique application, figure 4.4 shows that rnd.mn and rnd.ms produce mediocre synthetic datasets with the latter combination mitigating the (under)overestimation tendencies of the previous one. The combination rnd.e instead, shows a significant overestimate of the variable *taa* in the class 40-50 end two even more significant underestimates in the classes 50-60 and 60-70. The rnk technique does perform a mediocre imputation with the variable *taa* significantly underestimated at least in two classes.

Figure 4.4: Imp 1, *taa* and *taa_imp* in rnd and rnk

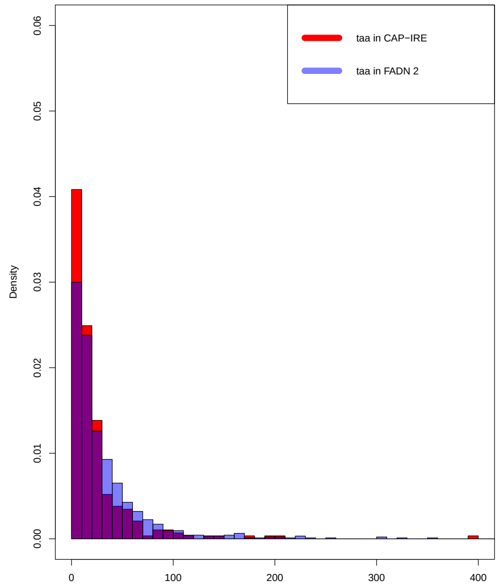


We omit the discussion of the imputation results obtained applying the rnk technique to the several sub-datasets since they are slightly worse than the ones obtained by the rnk application without the sub-datasets.

Figure 4.5 shows the distributions of the variable *taa* in the CAP-IRE 2009 (recipient) and FADN 2 (donor) datasets. The two datasets have different means for the variable *taa* and the donor one has a double variance. The correspondence of this matching variable is almost good, over-corresponded in the donor dataset with the significant exception of the lower presence of the variable *taa* in the class 0-10.

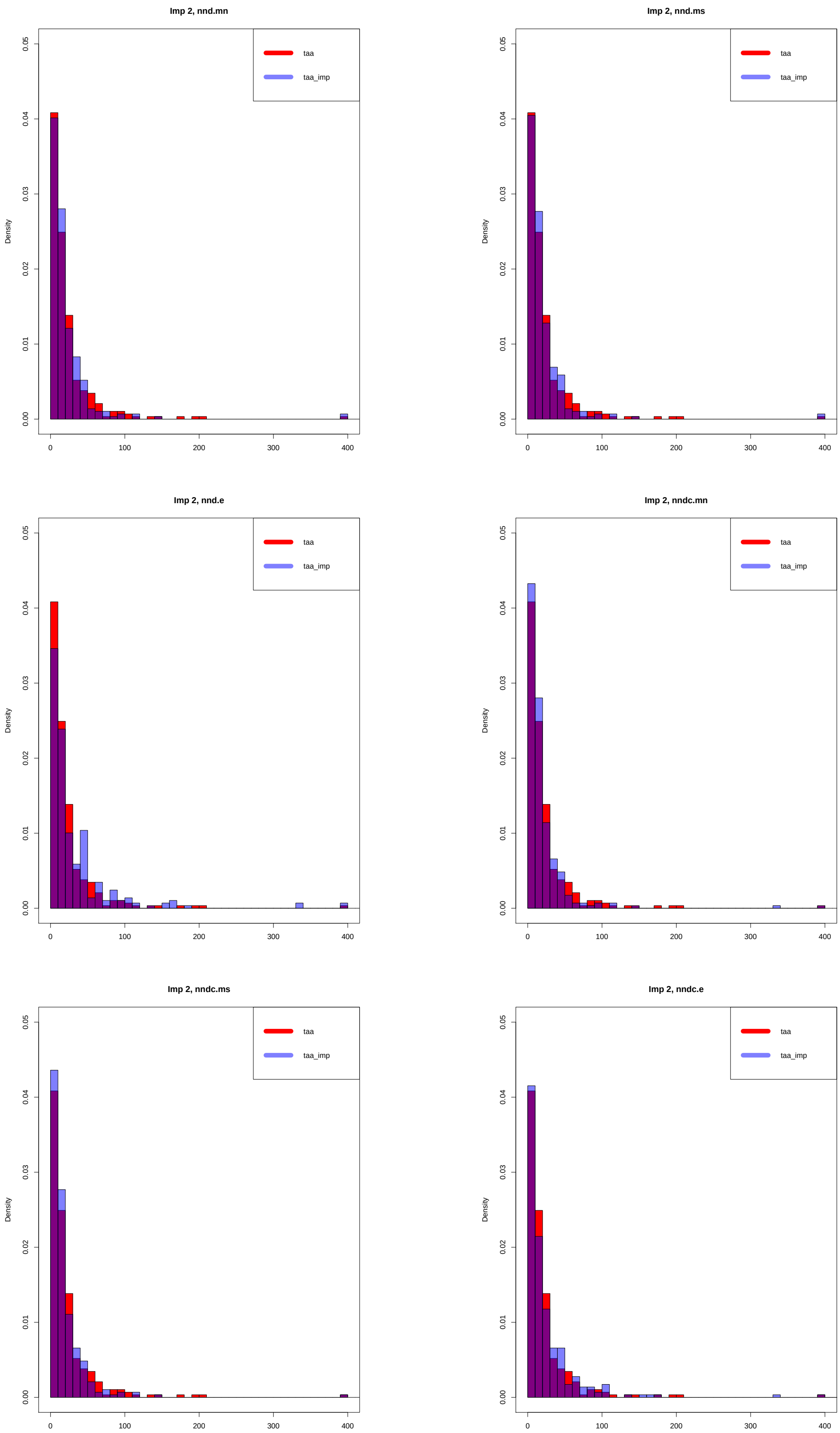
Figure 4.5: Imp 2, variable *taa* in CAP-IRE and FADN 2

taa		
	CAP-IRE	FADN 2
mean	25.972	33.874
var	2043.819	4847.691
min	1	0.445
max	470	1659.242



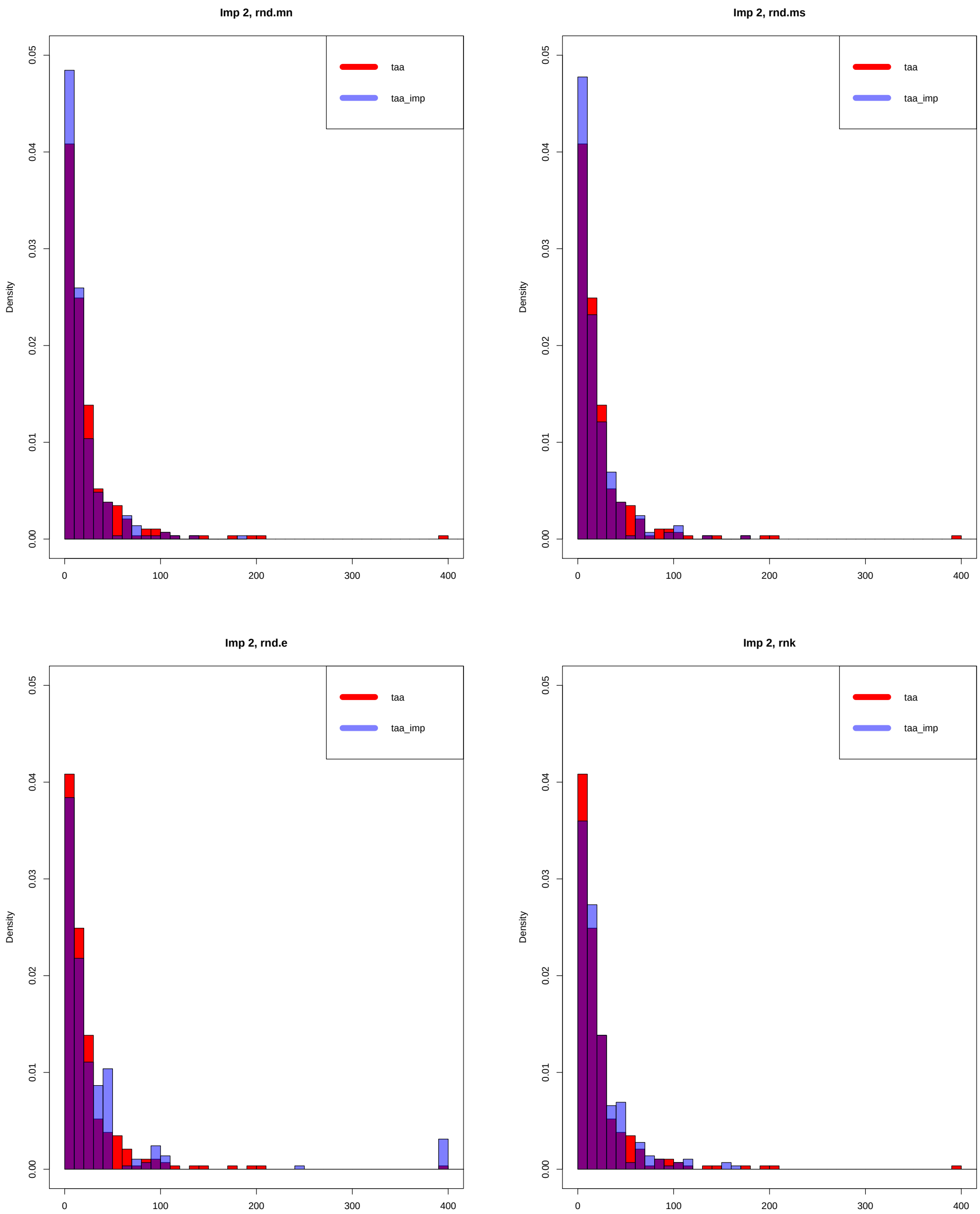
Taking into account the synthetic datasets generated using FADN 2009 2 as the donor dataset, in figure 4.6 we can see the imputation results from the different combinations of the techniques nnd and nndc with respect to the pre-post imputation distributions of the Total Agricultural Area (TAA). The combinations nnd.mn and nnd.ms perform really similar in Imp 2; the variable *taa* is overestimated in the class 50-60. Really similar results are obtained for the imputations with nndc.mn and nndc.ms. The combination nnd.e instead, generates a synthetic dataset in which the variable *taa* is significantly overestimated in the class 40-50 (more than doubled), but also underestimated in the classes 0-10 and 20-30. The same results but far more diminished result form the combination nndc.e.

Figure 4.6: Imp 2, *taa* and *taa_imp* in nnd and nndc



The rnd.mn and rnd.ms combinations, as figure 4.7 shows, generate two synthetic datasets in which there is a significantly overestimate of the variable *taa* in the class 0-10 but also (even if only slightly significant, for the first class, in the latter combination), in the classes 20-30 and 50-60.

Figure 4.7: Imp 2, *taa* and *taa_imp* in rnd and rnk



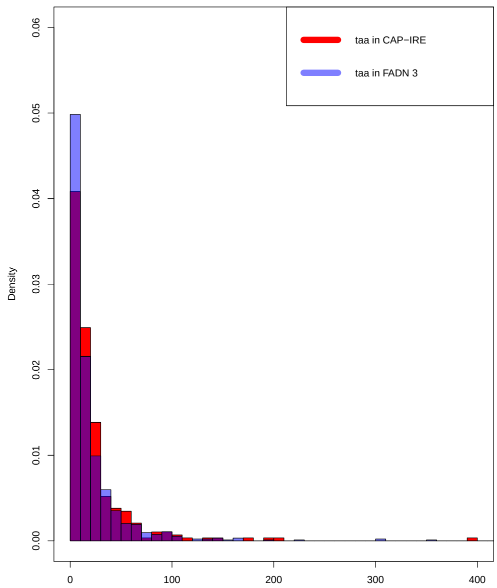
Again, we omit the discussion of the imputation results obtained applying the rnk technique to the several sub-datasets.

Figure 4.8 shows the distributions of the variable *taa* in the CAP-IRE

2009 (recipient) and FADN 3 (donor) datasets. The two datasets present characteristics more similar to the ones in the Imp 1; the means for the variable *taa* are closer and the donor dataset has a far lower variance (always higher than the variance of the matching variable in the recipient dataset, anyway). The correspondence of this matching variable is almost good, with a pronounced over-correspondence in the donor dataset for the class 0-10 (but a higher frequency of the variable *taa* in the recipient dataset for the class 20-30).

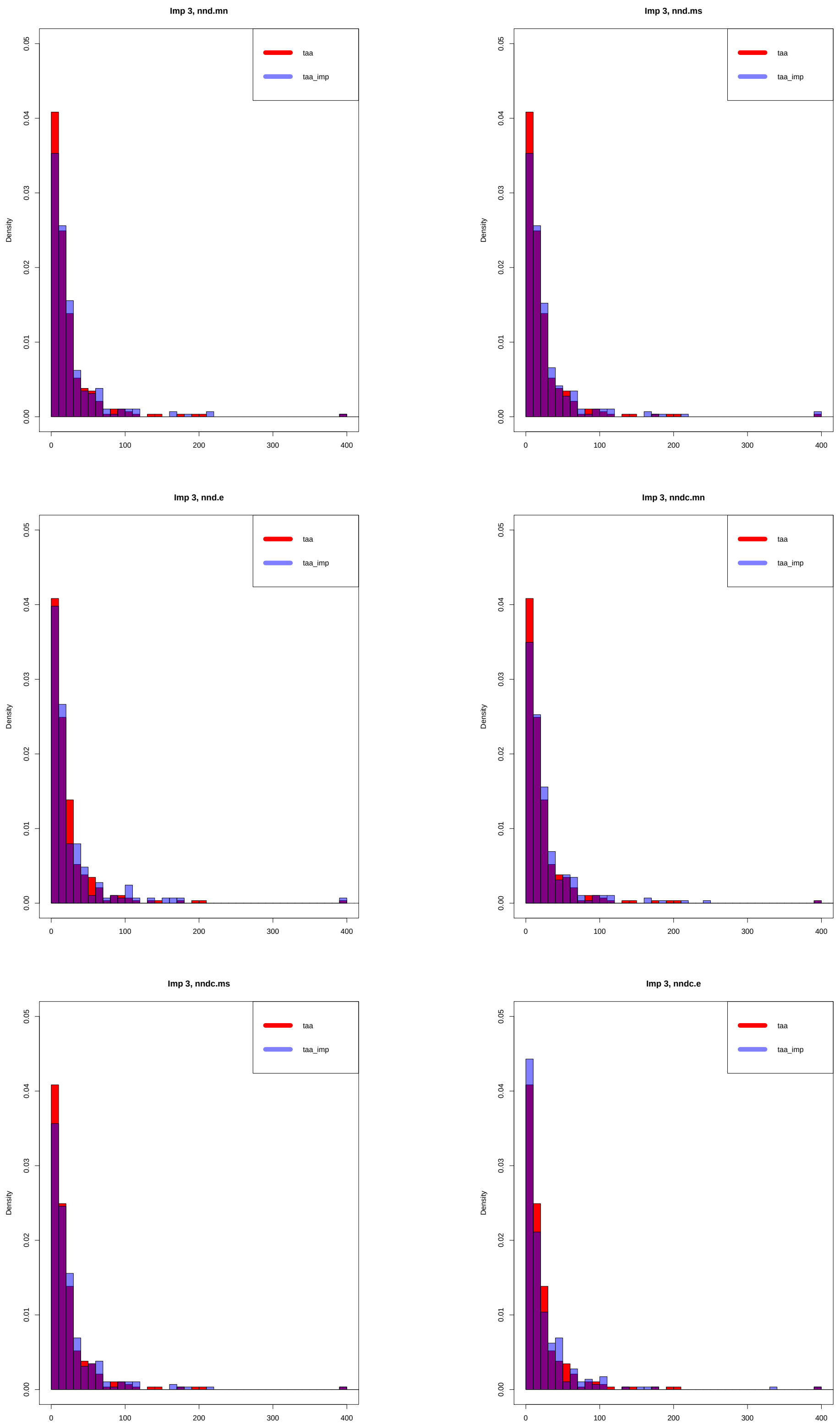
Figure 4.8: Imp 3, variable *taa* in CAP-IRE and FADN 3

taa		
	CAP-IRE	FADN 3
mean	25.972	21.891
var	2043.819	2962.892
min	1	0.5
max	470	1306.246



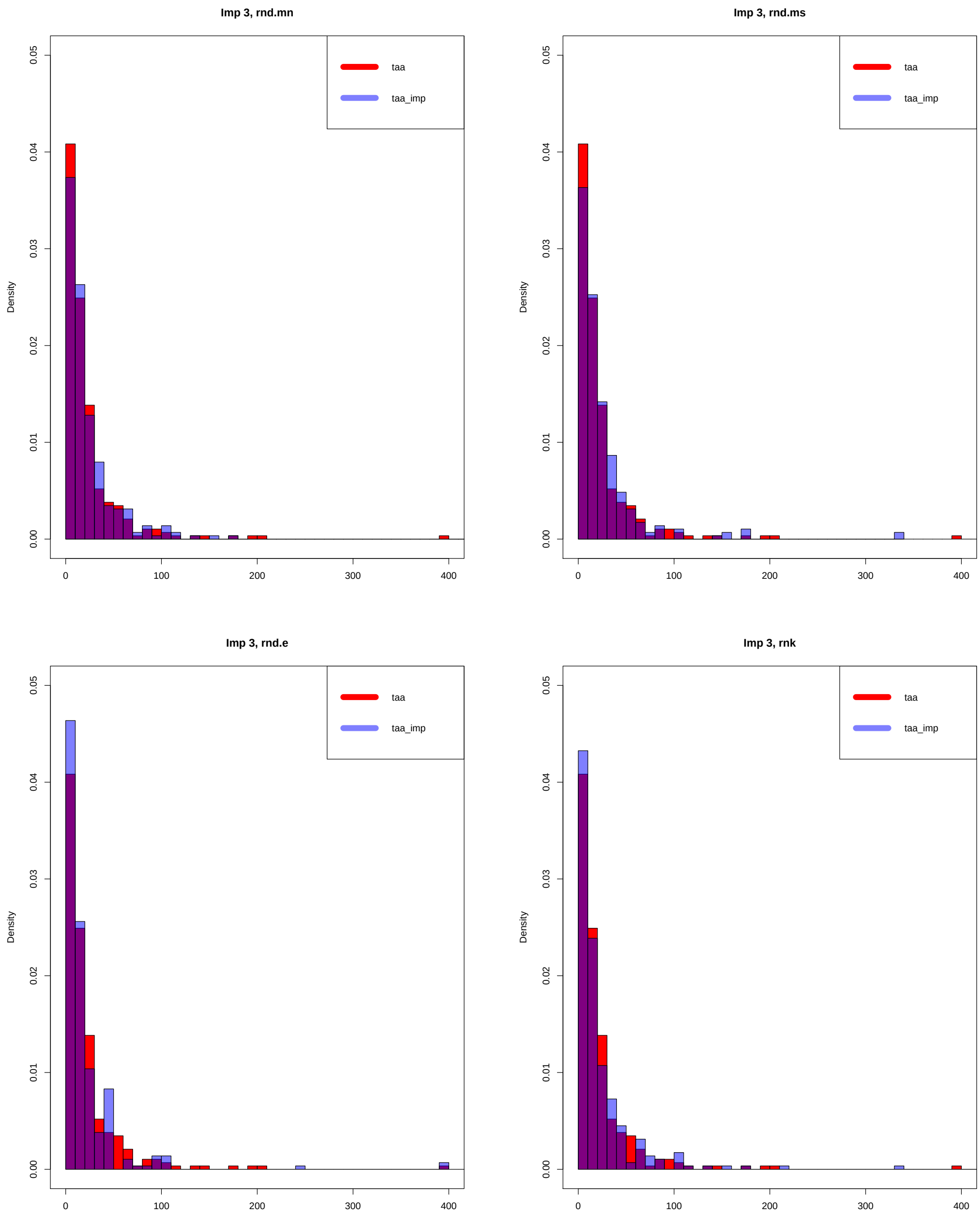
Taking into account the synthetic datasets generated using FADN 2009 3 as the donor dataset, in figure 4.9 we can see the imputation results from the different combinations of the techniques nnd and nndc which perform really good. With the exception of an underestimate in the class 0-10, there is an overall good estimate of the variable *taa*, with the better results showed by nnd.ms. Really similar results, with more pronounced overestimated values of the variable *taa*, are obtained with the combinations nndc.mn and nndc.ms. Always mediocre, the combination nnd.e generates a synthetic dataset in which the variable *taa* is significantly underestimated in the class 20-30 with slightly overestimates for others values. Similar results are showed by the combination nndc.e.

Figure 4.9: Imp 3, *taa* and *taa_imp* in nnd and nndc



Even the rnd.mn and rnd.ms combinations, as figure 4.10 shows, generate good synthetic datasets (there are an underestimate in the class 0-10 and an overestimate in the class 30-40). Both the rnd.e and the rnk perform bad estimates (for example in the classes 20-30 and 50-60 the variable *taa* is significantly underestimated or, in the class 40-50 it is underestimated).

Figure 4.10: Imp 3, *taa* and *taa_imp* in rnd and rnk



We omit the discussion of the imputation results obtained applying the rnk technique to the several sub-datasets.

The imputation goodness validation requires also to take into account the distributions of the differences z. Figures 4.11, 4.12 and 4.13 show them with respect to the three imputations; we omit to show the combination nnd.ms for the three imputations since they are really similar to the nnd.mn ones in Imp 1, Imp 2 and Imp 3 whereas the combination of the nndc technique for them is really similar to the nnd ones. As we can see, the combinations nnd.e, nndc.e and even rnd.e apparently allow for a good control of the outliers values (even if they do not associate really good matching units pairs). The combinations nnd.mn (and the really similar nnd.ms omitted), perform good associations even if they do not properly control for the outliers as it does, instead, for example, the combination rnd.ms.

Figure 4.11: Imp 1, distributions of z

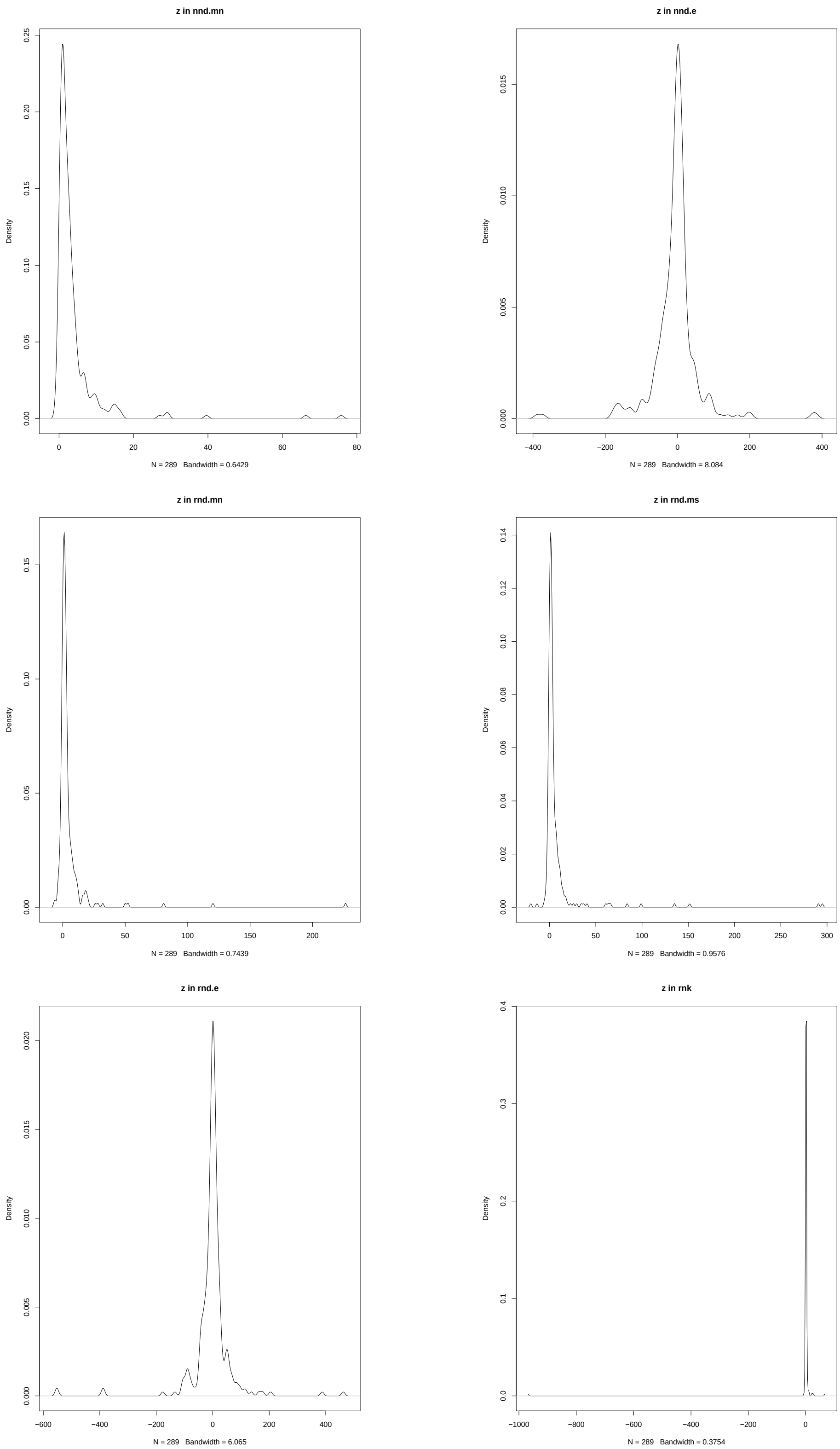


Figure 4.12: Imp 2, distributions of z

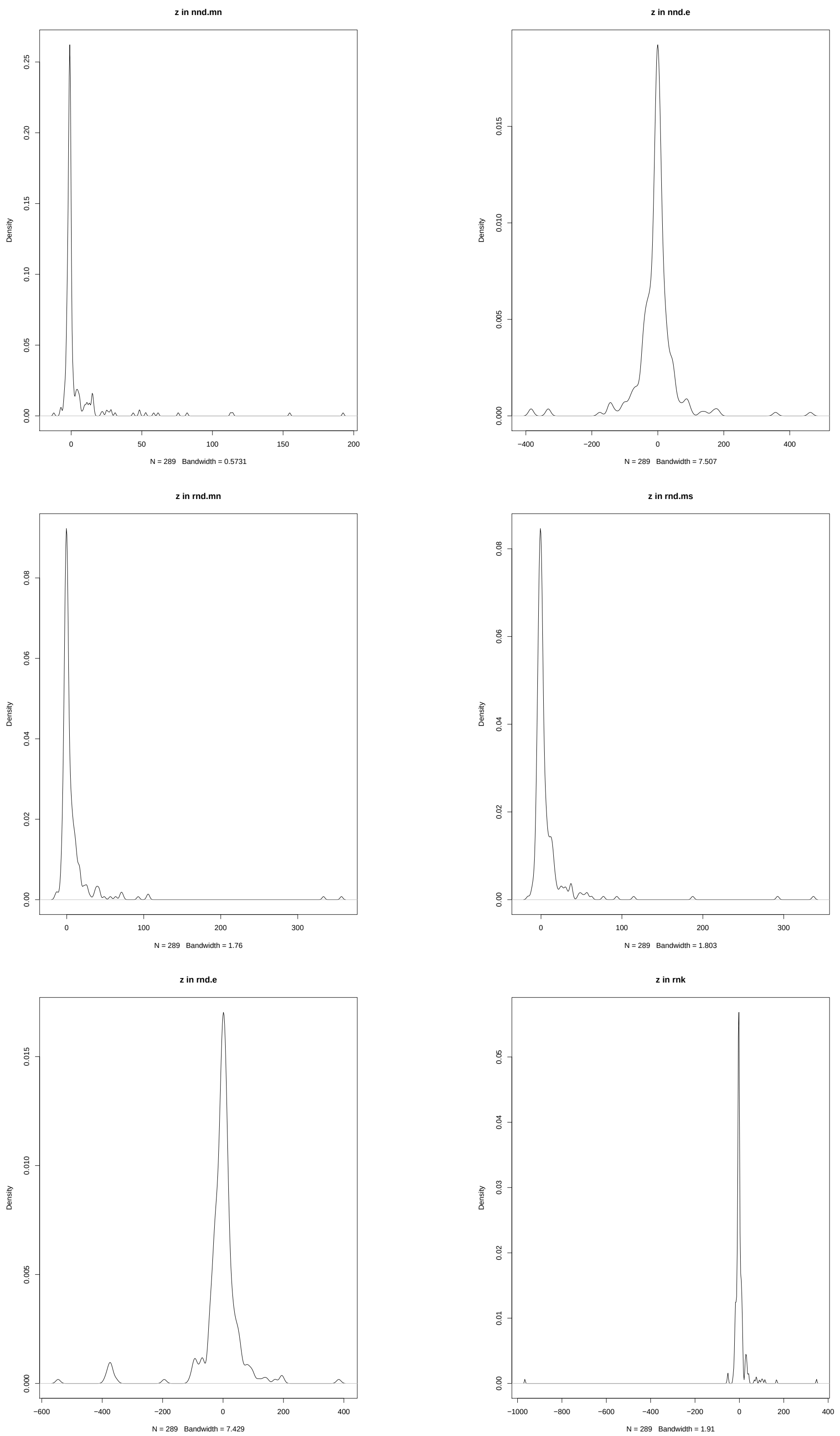
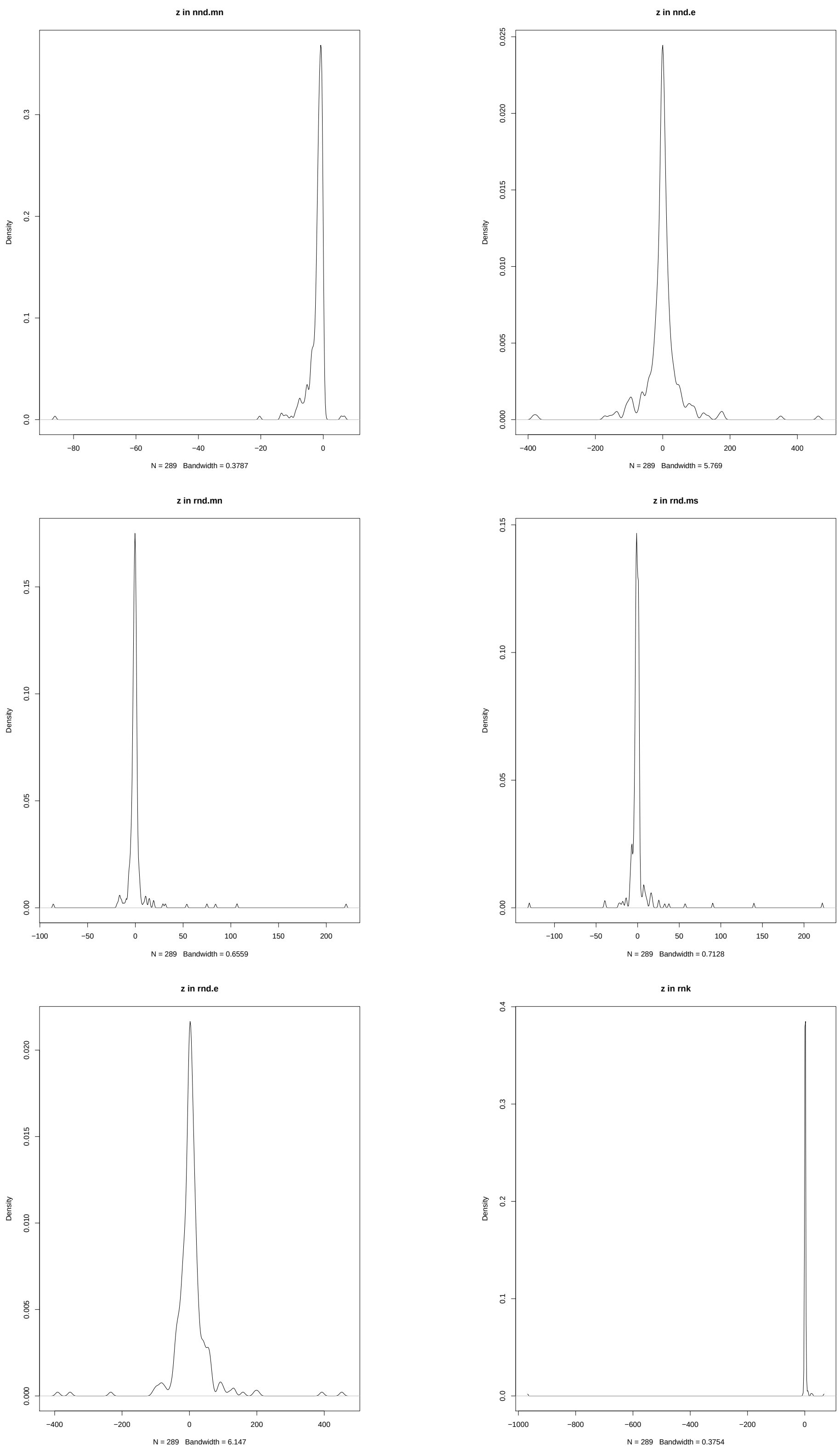


Figure 4.13: Imp 3, distributions of z



Finally, in order to complete the imputation goodness validation (and consequently choose the best synthetic dataset generated through imputation), we look at the MSE values referred to the differences z, as table 4.9 shows.

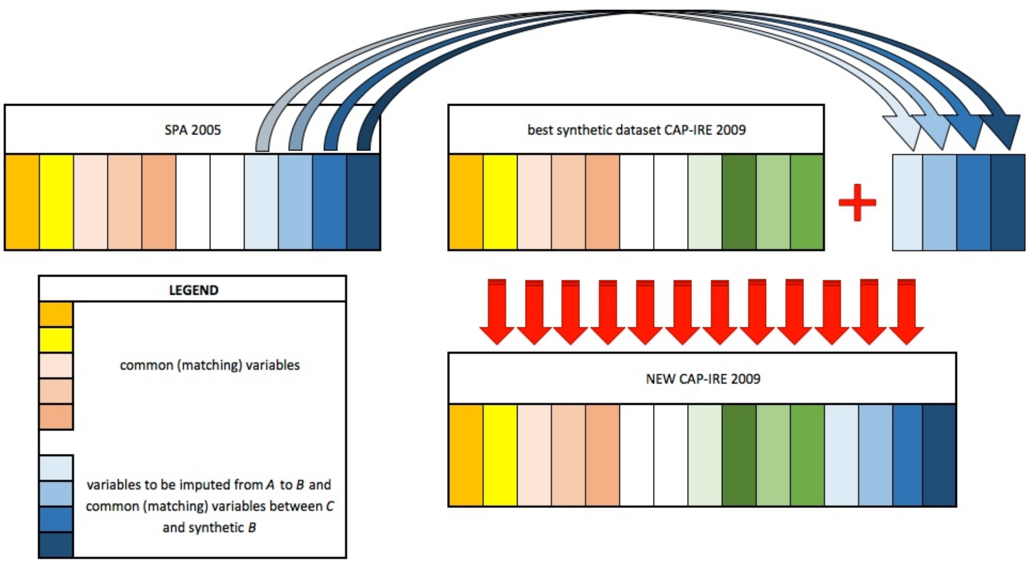
Table 4.9: MSE values for differences z in Imp 1, Imp 2, Imp 3

Combination	Imp 1	Imp 2	Imp 3
nnd.mn	67.775	438.064	36.451
nnd.ms	67.941	446.445	32.905
nnd.e	4397.971	7927.983	4114.501
nndc.mn	67.983	485.649	41.211
nndc.ms	68.218	1044.226	39.314
nndc.e	4765.217	4571.652	4561.145
rnd.mn	302.058	1113.074	314.189
rnd.ms	883.241	1059.131	383.124
rnd.e	6129.157	7116.171	3793.971
rnk	3261.124	4114.427	3261.124
rnk (sub-datasets)	10178.441	5042.652	10178.441

The best synthetic dataset, considering the simultaneous validity of the above-mentioned tools for analysing the imputation results, and looking at the MSE values, we decide to chose the synthetic dataset generated in Imp 3 using the combination nnd.ms (even the combination nnd.mn does perform well), i.e. the combination of the Nearest Neighbour Distance Hot Deck and the Mahalanobis distance function. This one is selected to be the best synthetic dataset and used for the Imp 4 which generates the final NEW CAP-IRE 2009. This choice is motivated by the fact that it presents a high correspondence between the variable *tf14* modalities between the donor and recipient units (almost 90%), an optimal pre-post distributions overlap of the *taa* and the *taa_imp* variables and, finally, a good MSE value for the differences *z*.

Figure 4.14 shows a schematic representation of the Imp 4 concerning the SPA 2005 (donor) dataset and the new generated one. Similarly to Figure 4.1, in Figure 4.14 we have a common set of variables among which we select as matching variables the orange, the yellow and the ones in pink shades. We can see that the synthetic dataset resulting from Imp 4 presents also the variables previously imputed from FADN 2009 to CAP-IRE 2009 (those in the green shades). Basically, then, what we do is to impute others variables (those in the blue shades), from SPA 2005 to the synthetic dataset selected in order to definitely build the NEW CAP-IRE 2009.

Figure 4.14: SM imputation between SPA 2005 and NEW CAP-IRE 2009



For sake of brevity, Imp 4 is not discussed in details but directly used for the PSM analysis application.

Chapter 5

PSM analysis

5.1 Introduction

Chapter 5 shows the application of PSM methods to the dataset NEW CAP-IRE 2009 generated by integration through non-parametric micro SM imputation techniques (combined with different not default distance functions). The main goal of this application, taking into account the fact that the CAP-IRE 2009 dataset was not expressly designed and produced for policy impacts evaluation purposes, is to show how, under the observational studies research context, it is fruitful to preserve observed data from different available data sources and integrate them for causal effects analysis purposes.

5.2 PSM application

Table 5.1 shows the treated and control groups present in the NEW CAP-IRE 2009 dataset, defined by taking into account as "treatment"variable (t

is equal to 0 if the unit is a control, 1 if the unit is treated), the farms AES uptake. As we stressed previously in this work, the choice of the treatment variable is due to the fact that AES uptake is the only detriment variable present in the new generated dataset, that can be used as plausible treatment for our PSM application purposes.

Table 5.1: Treatment and control groups in NEW CAP-IRE 2009

t	Frequency	Percent
0	178	62.46
1	107	37.54
Total	285	100.00

We stress that the farm sample is really small with respect to the most well-known applications of the PSM methodology for impacts evaluation and/or causal effects analysis present in the literature (usually, the total sample taken into account is not lower than 1,000-1,200 units). Nevertheless, the sample is representative of the Emilia-Romagna Region farms. In order to calculate the Propensity Score (PS) for the consequent PSM analysis, following both the literature prescriptions and the previously theoretical and empirical findings discussed in paragraph 5.3, we verify which of the observed covariates are significant for the treatment uptake (and simultaneously not affecting it since they are information on pre-treatment units status). For sake of brevity, we decide to show and discuss the best PS estimation obtained; we stress that, contrarily to our expectations, with respect to the significant variables which can determine farms bent to uptake the treatment (i.e. to uptake AES), the observed covariates concerning farm owner's characteristics are not significant at all and consequently discarded from the PS estimation.

These are the variables *sex*, *age*, *age*^{2}, but also covariates concerning the farms characteristics related both to the year 2009 and 2005 and potentially considered to be relevant for the AES uptake, such as the variables *tf14* indicating farms specialisation, *crops* indicating the number of crops cultivated by the farm, the variables indicating the amount of UAA dedicated to the single crops such as *cereals*, *rice*, *fruit* etc., *organic_production_05* indicating whether the farm had or not UAA in biological agriculture in the year 2005, *irrigated_uaa_05* indicating how much UAA the farm had under irrigation in the year 2005, *sfp_05* indicating the Single Farm Payment (SFP) status related to the year 2004.

Results from the best estimated PS are showed in table 5.2.

Table 5.2: Covariates for the Propensity Score estimation

t	Coef.	Std. Err.	z	P> $\|z\|$	[95% Conf.	Interval]
edu_agri	-0.4762151	0.3404621	-1.40	0.162	-1.143509	0.1910783
edu_owner	0.1151748	0.1174926	0.98	0.327	-0.1151065	0.3454562
legal_status	0.7781763	0.3035089	2.56	0.010	0.1833098	1.3730431
organic_prod	0.9203007	0.4247685	2.17	0.030	0.0877698	1.7528321
sfp_08	-1.0158061	0.3753656	-2.71	0.007	-1.7515091	-0.2801028
sfp_ha	0.0023965	0.0010404	2.30	0.021	0.0003572	0.0044357
size_esu	0.0002945	0.0002771	1.06	0.288	-0.0002486	0.0008377
irrigated_uaa	0.0141936	0.0079123	1.79	0.073	-0.0013143	0.0297015
gfi	-0.0000163	7.84e-06	-2.08	0.038	-0.0000316	-9.07e-07
ffi	0.0000153	7.95e-06	1.92	0.054	-2.90e-07	0.0000309
awu_total_input	0.5139779	0.1836111	2.80	0.005	0.1541088	0.8738471

We can notice that the significant covariates for the PS estimation, i.e. the most relevant observable units characteristics determining the units treatment uptake, result to be the farm legal status (0 for the corporation -including family-, 1 for the sole proprietorship), the presence of biological agriculture, the SFP status in the year 2008, the amount of SFP per hectare (expressed in Euro), the Gross Farm Income (GFI -expressed in Euro-) and the farm work total input (expressed in Annual Working Unit -AWU-). Slightly non significant instead, are the farm amount of UAA irrigated and the Family Farm Income (FFI -expressed in Euro), i.e. the amount of income produced by the agricultural activity by the farm family. Not significant in this PS estimation but proven to be determinant for the treatment uptake elsewhere in our application, are the variables *edu_agri* and *edu_owner*, indicating respectively the presence of an agricultural education for the farm owner and his educational level, and also the farm size expressed in European Size Unit (ESU).

We estimate the PS carefully checking both for the common support region ([0.15253599, 0.92872733]), and the satisfaction of the balancing property, building 6 blocks which ensure that the mean PS is not different for the treated and the control groups within each of them, as table 5.3 shows.

Table 5.3: Estimated Propensity Score blocks

	Percentiles	Smallest		
1%	0.1555662	0.1525361		
5%	0.1732742	0.1540191		
10%	0.1977189	0.1555662		
25%	0.240842	0.1555836		
50%	0.3459594	0.3810212		
		Largest		
75%	0.4875058	0.8949416	Obs.	279
90%	0.6514582	0.8990094	Std. Dev.	0.1751154
95%	0.7501219	0.9220944	Variance	0.0306654
99%	0.8990094	0.9287273	Pseudo R^2	0.1840

Table 5.4 shows the number of the treated and control units (being the balancing property satisfied) in each block, i.e.:

Table 5.4: Treated and control units in Propensity Score blocks

Inferior of block of PS	*t(0)*	*t(1)*	Total
0.152536	23	7	30
0.2	66	13	79
0.3	38	30	68
0.4	35	32	67
0.6	8	19	27
0.8	2	6	8
Total	172	107	279

For sake of brevity, we show only the most significant result obtained with the optimal Average Treatment Effect for the Treated (ATT) estimator, i.e. the radius estimator (with a caliper of 0.1). The impact variable we choose to show, among the ones we thought to be potential impact variables (discussed in paragraph 5.3), with respect to the treatment, is the total amount of land rent in by the farm (expressed in hectares). As table 5.5 shows, there is a negative (significant) effect of the AES uptake on the land rented in by the "treated"farms.

Table 5.5: Average Treatment Effect for the Treated (ATT)

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
land_rent_in	Unmatched	8.30841	7.18539	1.12302	2.78536	0.40
	ATT	8.35577	12.31989	-3.96412	2.93514	-1.35

Beyond the (significant) negative effect of the AES uptake on the amount of land rented in by farms, we stress that even other (significant) specifications of the ATT estimator (tinier calipers radius estimators, kernel estimators, nearest neighbour estimators without replacement, etc.), significantly prove the presence of a treatment effect for the treated units on the hectares of rented land in a circumstantial range ([-1.95326, -3.96412]).

After having estimated the ATT we properly check the satisfaction of the balancing property between the treated and the control groups, for each one of the covariate used for the PS estimation, as table 5.6 shows.

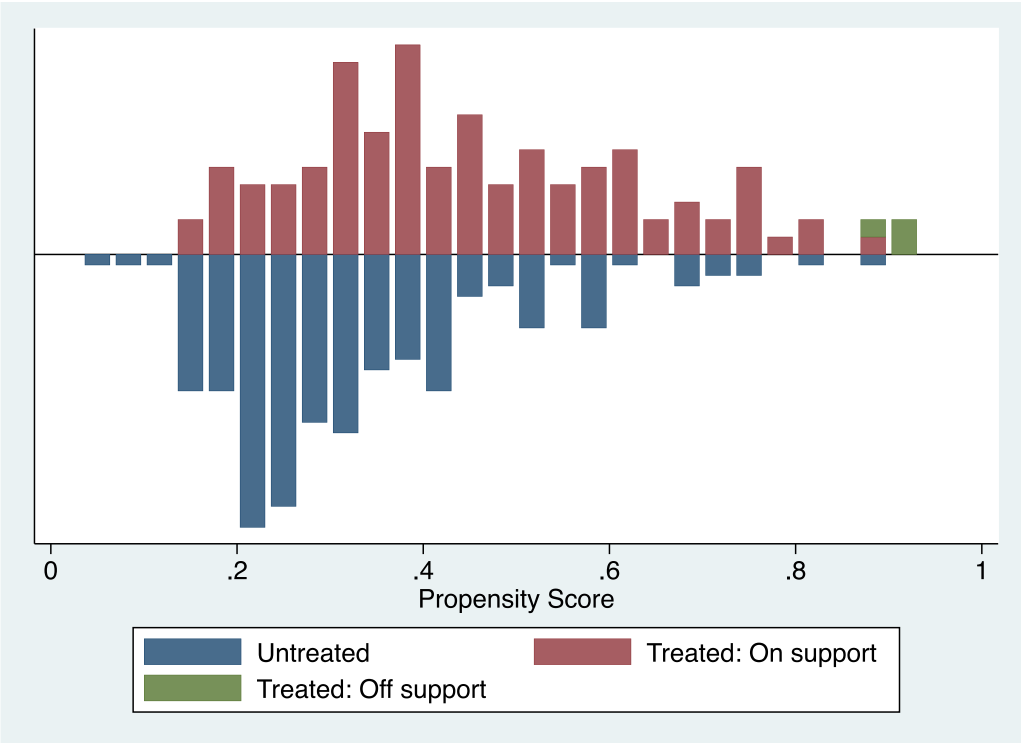
Table 5.6: Balancing property for (un)matched treated and control units

Variable	(Un)Matched	Mean		% bias	% reduct	t-test	
		Treated	Control		$\|bias\|$	t	P> $\|t\|$
edu_agri	U	0.24299	0.23034	3.0		0.24	0.808
	M	0.23077	0.23427	-0.8	72.3	-0.06	0.953
edu_owner	U	2.3271	2.0731	21.2		1.73	0.084
	M	2.28851	2.24172	3.9	81.6	0.28	0.782
legal_status	U	0.39252	0.19663	43.8		3.68	0.000
	M	0.37511	0.35588	4.3	90.2	0.28	0.776
organic_prod	U	0.16822	0.06742	31.5		2.71	0.007
	M	0.16346	0.15654	2.2	93.1	0.14	0.892
sfp_08	U	0.74766	0.81461	-16.2		-1.34	0.181
	M	0.74038	0.72927	2.7	83.4	0.18	0.857
sfp_euro	U	6801.9	3161.5	27.6		2.31	0.022
	M	6305.8	5508.1	6.0	78.1	0.33	0.745
sfp_ha	U	155.66	137.17	12.5		1.03	0.302
	M	155.55	157.04	-1.0	91.7	-0.07	0.948
size_esu	U	326.85	224.05	9.2		0.71	0.481
	M	307.06	288.43	1.7	81.9	0.11	0.916
irrigated_uaa	U	13.027	7.0613	23.5		2.02	0.044
	M	12.103	10.962	4.5	80.9	0.28	0.777
gfi	U	1.2e+05	79308	17.4		1.36	0.176
	M	1.2e+05	1.0e+05	5.3	69.6	0.35	0.730
ffi	U	87049	56665	14.6		1.13	0.260
	M	83267	73821	4.5	68.9	0.29	0.772
awu_total_input	U	2.1545	1.4337	34.4		2.90	0.004
	M	1.9571	1.7794	8.5	75.4	0.61	0.544

The balancing property checking puts in evidence which covariates are well balanced after matching (i.e. which covariates have a percentage bias after matching in absolute value lower than the 5%). As we can see in table 5.6, almost all the covariates are well balanced after matching, with the exception of *awu_total_input*, *sfp_euro* and, even if only slightly higher of the 5%, the variable *gfi*. In order to validate the good balance among the covariates in the two different groups for both the unmatched and the matched units, we look also at the variance ratio of treated and controls, which is supposed to lay in the range [0.68, 1.57]. We find that the variance ratio is significantly outside the range, with respect to the unmatched (treated and control) units, for the covariates *organic_prod* (2.23), *irrigated_uaa* (2.26) and *awu_total_input* (1.68). Moreover, taking into account both the distribution of the absolute bias and its mean reduction (before and after matching), we can consider the balancing property satisfied. Indeed, the mean bias for the unmatched sample is equal to 31.2 whereas the mean bias for the matched is equal to 3.8 ($< 5\%$); the absolute bias for the former is equal to 78.5 whereas the latter have a bias of 12.3 (significantly lower than the 25%). We can conclude that the PSM performs well and satisfactorily in balancing the covariates between the treated and control groups, so reducing the bias before and after matching.

Finally, we check the overlap between the treated and control units, represented in the table 5.1. Performing the PS estimation, we discard three treated units which are off the common support, obtaining a discrete overlap between matched treated and control units with the exception of the lowest values of the PS and for both the PS block 0.65 and 0.75. The significant lack of overlap for the upper blocks of the PS is due to the small sample dimensionality.

Figure 5.1: Propensity Score overlap between treated and control groups



5.3 PSM results discussion

Our PSM application is ran with the key purpose of demonstrating how consequently useful can be the integration of different (farm) data sources for the policy impacts evaluation analysis. Nevertheless, we also have to take into account agricultural economics prescriptions and theoretical findings in order to justify the PSM analysis and both the PS and the ATT estimations.

As we stressed in paragraph 5.2 and also previously in the present work, the choice of the treatment variable is due to the variables constraints imposed by data at disposal and by the fact that the CAP-IRE 2009 dataset was not thought and design for policy impacts evaluation purposes. Therefore, the only detriment variable that can be used as “treatment”is the farm uptake of AES, an information collected with respect to the full package of AES Measures (i.e. the farm AES uptake is conceived by the variable t as all the agri-environmental measures simultaneously), following the approach in (Arata and Sckokai, 2016).

The selection procedure of the most significant covariates for the PS estimation, among the variables at disposal in the NEW CAP-IRE 2009 dataset, follows the prescriptions of the literature. Among the most relevant variables, with respect to the participation in the treatment, there are:

- structural variables concerning the farms;
- the variables referred to farm owner's (and/or the farm household) characteristics;
- input variables;
- the variables referred to the farms geographical characteristics and/or location.

The most recent works like Kirchweger and Kantelhardt (2012), Chabé-Ferret and Subervie (2013), Ratinger et al. (2013) and Arata and Sckokai (2016), use structural variables such as the farm size (expressed in UAA and/or in ESU), the number of cultivated crops, the amount of UAA for each crops cultivated in the farm, the presence of biological productions, the Gross Farm Income, the Family Farm Income (or its proxies), the legal status or the farm organisational form. Kirchweger and Kantelhardt (2012), Chabé-Ferret and Subervie (2013) and Arata and Sckokai (2016) also use as covariates for the

PS estimation, the ones concerning the farms owner's characteristics, such as the age (adjusted), the educational level, the agricultural education (if present or not) and the characteristics of the farm household members. Finally, an "older" work as Pufahl and Weiss (2009), and Chabé-Ferret and Subervie (2013) and Arata and Sckokai (2016) use input variables such as equipments, buildings, expenditures, labour input, use of chemicals and geographical characteristics of the farms such as the altitude, variables for the plain-hill-mountain and regional/national location variables.

Following the above-mentioned works, we use the same covariates for the PS estimation, discarding the not significant ones through a stepwise strategy. Contrarily both to what prescribed from the previous findings and to our expectations, covariates related to the farm owner such as the sex, the age (adjusted) and his labour input, are evidently not significant in the PS estimation. Furthermore, we discard for the same reason of significance, the variables indicating the farm specialisation, the number of crops cultivated, the UAA of the single crops and the farm location (plain-hill-mountain).

Taking into account the outcome variables, with the exception of Chabé-Ferret and Subervie (2013) which has at disposal an undoubtedly fruitful amount of data, we follow Pufahl and Weiss (2009), Kirchweger and Kantelhardt (2012) and Arata and Sckokai (2016), analysing whether the treatment has impacts on farms structural changes, job and employment dynamics swing and farm activities diversification. Among all the potentially affected outcome variables (belonging to our expectations), we find a significant negative effect of the treatment (the AES uptake), on the amount of land rented in by the farms. Our results do not comply with Pufahl and Weiss (2009);

nevertheless, considering that:

> “the work takes into account the AE programmes under the period 2000-2005 whit respect to 32,000 farms in German LAND-Data and (...) the sample is not representative for Germany as large-scale and full-time farm entreprises are over represented”,

the positive effect of the AE programmes uptake on the farm land growth rates for the treated units, has to be contextualize. Indeed, as the authors themselves highlight, the higher farm land growth rates of participants in AE programmes can be due to the programme eligibility criteria which can have fostered the land growth rates. For example, farms specialised in cattle livestock, in order to participate in AE programmes (for which it is required to not exceed a certain threshold of cattle livestock density), tend to expand grassland maintaining the number of cattle per hectare stable. This land growth rates are then mainly achieved by renting additional land for the years of AE programmes programming (5 years). Furthermore, Pufahl and Weiss (2009) stresses that:

> “there is not a clear relationship between the individual treatment effect and the conditional probability of participation in AE programmes for changes in (...) rented land and (...) the magnitude of the treatment effect is heterogeneous between farms of different size and varies with programme duration”.

Taking into account Arata and Sckokai (2016), results of the farms AES uptake on the land rented in during the period 2003-2006, show that there is an average increase of the farm size (mainly due to the rented land) in the subsample of the treated group characterised by the share of AES on farm revenue larger than the 5% both in the UK and Italy. Nevertheless, both in Spain and Germany, with respect to the same subsamples, the effect is non-negative but not significant. Furthermore, as the above-mentioned work states:

> "in Italy, where the most widespread measues are organic farming and low-input agriculture, the increase of farm size is likely due to the attempt to offset the decrease in the output value per hectare (...) and it turns out that the increase in the average farm size may be explained by this factor, since in all the other cases the difference is not statistically significant".

Taking into account the heterogeneity issue for all the farm samples taken into account in the above-mentioned works, and the fact that, with respect to the Italian farm sample in FADN data, the information on the AES uptake is aggregated, i.e. there is not any information available either on which is the scheme applied by each farm or on the hectares dedicated to AES measures, but also considering the fact that our sample is more homogeneous and does take into account a Regional farm sample (i.e. a unique RDP with more or less homogeneous AES measures), validity of the negative effect of the treatment on the treated units, can be found beyond the statistical significance. Finally, we stress that a more robust analysis concerning the AES effects on land tenure and land allocation on homogeneous farm samples should be carried on having at disposal disaggregated data on AES uptake (which measures are implemented, where, how much hectares are committed, which are the other relevant covariates to take into account -such as the household characteristics, the farm specialisation, the farm productivity and the farm productivity factors, etc.).

Our PSM application, anyway, does not have the key purpose of evaluating policy impacts effects; rather, its goal is to demonstrate how fruitful can be the integration of the two different methodologies of the SM imputation and the PSM analysis when we have to deal with the observational studies research context. In that sense, the orthodox and robust PSM analysis carried out in the present work, shows how significant and profitable is the use, by preservation through non-parametric micro techniques, of different farm data sources.

Conclusions

This work analyses the methodological issues related to the non-parametric micro Statistical Matching (SM) imputation techniques theoretical framework and by their usefulness with respect to both the computational speeding and the preservation of the observed (real) data integrated from different data sources. Considering the different data issues discussed in chapter 1 (data availability, accessibility, collection costs, etc.) several ongoing researches could be fruitfully implemented and further developed resorting to different data sources integration methodologies.

In the most recent years, the non-parametric micro SM imputation "hot deck" techniques have found a large applicability. Nevertheless, in spite of the numerous practical applications, a proper improvement of the SM imputation theoretical framework has been lacking. "Hot deck" methods result to be largely unexplored both with respect to their theoretical formalization and the functioning of the matching algorithms with the application of different not commonly used distance functions.

Our main aim then, is to propose a coherent theoretical framework for potentially new combinations, within the matching algorithms of the above-mentioned SM imputation techniques, of not commonly used (not default) distance functions. We propose to combine the "hot deck" methods with the Manhattan, Mahalanobis and Exact distance functions. The research objective is to study and discuss the integration of different data sources using these combinations, taking into account both the different characteristics of the datasets at disposal and the different matching possibilities (for example, the dimensionality ratio between recipient and donor, the variance of the matching variable(s) in the recipient and the donor datasets, the donation classes building, etc.). The combinations of the Nearest Neighbour Distance Hot Deck, the Constrained Hot Deck and the Random Hot Deck with the above-mentioned distance functions and the Rank Hot Deck technique itself, validated in our simulation study using the proposed strategy, show evidence of the better performances of the nnd/nndc.mn and nnd/nndc.ms combinations with respect to their "estimation power".

Due to the absence in the SM literature, of a robust recursive strategy for the imputation goodness validation, we elaborate and propose, through the simulation study, such a procedure, which is structured upon three linked validation tools. This work explores new hypothesis on the SM imputation performances due to the different characterisation of recipient and donor datasets in four simulated scenarios (i.e. different dimensionality ratios between donors and recipients, different matching variable(s) variability and the possibility -or not- to run the imputations with donation classes). The simulation study is carried out in order to decide which tools are useful in order to build a procedure which properly and significantly validate the imputation goodness. The final proposal consists in the analysis of the pre-post distributions of the matching variable(s) chosen, the analysis of the differences "z" between the values of the variables observed for the recipient and the donor matching units pairs, the analysis of the MSE values of these differences. The simultaneous use of these tools within the imputation goodness validation strategy should then guide the choice of the best synthetic dataset generated through imputation.

The application of both the above-mentioned combinations of SM imputation techniques and distance functions concerns three different farm data sources (two official administrative ones and a project survey). Considering the specific practical problems related to the integration of different farm data sources, but also the need of a previous data harmonization, we present a reference framework for different farm data sources harmonization. Such a procedure, indeed, is essential for the application of both the SM imputation and PSM analysis since it allows the researchers to properly homogenize data at disposal and set the imputation running in the optimal way.

The new dataset generated through integration from two consequent SM imputations is consequently used to run the Propensity Score Matching (PSM) analysis. We stress that our purpose with respect to the PSM analysis application, is to demonstrate the usefulness of using a causal effects analysis method (specifically designed for the observational studies research context) after having integrated (i.e. preserved) different observed information. In our (clearly data-driven) PSM application, we choose as "treatment" variable the farms Agri-Environmental Schemes (AES) uptake. In spite of the agricultural economics literature prescriptions, we relax the orthodoxy of the economics hypothesis, being forced to use the solely variables at disposal. Moreover, we stress that neither the donor datasets nor the recipient one were originally designed and produced for policy impacts evaluation purposes. The treatment effects analysis concerns the following outcome variables: the farms structural changes, the swing of job and employment dynamics, the farm activities diversification. Through a robust and rigorous PSM application to the new generated dataset, we find a (negative) significant effect of the treatment on the farms land rented in.

This work has four macro-objectives: *i.* the study and discussion of new combinations of SM imputation techniques and distance functions, *ii.* the proposal of a recursive strategy for the imputation goodness validation when non-parametric techniques are used, *iii.* the proposal of a reference framework for different farm data sources harmonization and, *iv.* the consequential application of both the SM imputation and PSM analysis to a new generated dataset concerning farms. Pursuing these research objectives, four main points of strength emerge, i.e.:

1. The work implements the discussion and the theoretical formalization of the non-parametric micro SM imputation techniques, both exploring the possible new combinations of techniques and not default distance functions, and proposing a statistically effective and robust strategy for goodness imputation validation.

2. Through the simulation study, we define significant guidelines for evaluating the imputation performances, with respect to the different recipient and donor datasets characteristics (and the influence they potentially have on the imputation results). Moreover, with respect to the few consolidated prescriptions offered by the SM imputation literature,

we present a significant not-compliant finding related to the commonly accepted idea “the biggest, the best”.

3. The work approaches the specific integration case of different farm data sources, with respect of which there are only few relevant applications in the literature.

4. We robustly and significantly integrate two methodologies, the SM imputation one and the PSM analysis method, which are distinctly used but can jointly applied under the observational studies research context.

Taking into account the weaknesses of the present work, considering both our initial research objectives and the ongoing developments of the work, we have to stress that:

1. The work was originally thought in order to study and propose further developments and implementations of the currently debated SM imputation techniques. The idea was to do it also taking into account, in a suitable and statistically innovative way, the use of the sample weights within the non-parametric micro SM imputation techniques. Nevertheless, due to several issues and to the fact that we were re-directed toward others research perspectives, we gradually left this problem aside.

2. The theoretical formalization of the proposed combinations of the non-parametric micro SM imputation techniques and distance functions, despite constituting a first coherent effort, is still embryonic and can be further implemented and completed.

3. The application of the PSM, in spite of being robust and rigorous and even finding significant results, does not properly take into account the agricultural economics findings and prescriptions with respect to the AES literature, weakening the policy impacts evaluation purposes commonly addressed in the agricultural economics.

Considering both the points of strength and the weaknesses of the work, we consider that it could be further implemented toward different (but simultaneous) paths. Firstly, we could further develop the combinations of the "hot deck" techniques with the distance functions to properly consider the cases in which we want to use discrete matching variables and/or mixed discrete-continuous matching variables in our imputation. Moreover, we could try to re-consider the actual issue of sample weights and try to explore their usage within the non-parametric micro SM imputation techniques. Finally, with respect to the practical side of the SM imputation, the further implementation of these methods could expressly point to accounting for the time span dimension (i.e. we could try to construct a complete pre-post treatment cross-action data or a farm panel data, focusing, with respect to farm data, to specific RDP periods).

References

Abadie, A., D. Drukker, J. L. Herr, and G. W. Imbens

2004. Implementing matching estimators for average treatment effects in Stata. *STATA Journal*, 4(3):290–311.

Almus, M. and D. Czarnitzki

2003. The effects of public R&D subsidies on firms' innovation activities: the case of Eastern Germany. *Journal of Business and Economic Statistics*, 21(2):225–236.

Andridge, R. R. and R. J. A. Little

2010. A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64.

Arata, L. and P. Sckokai

2016. The impact of Agro-environmental Schemes on farm performance in five EU Member States: a DID-matching approach. *Land Economics*, 92(1):167–186.

Ashenfelter, O.

1978. Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, 60(1):47–57.

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.

Ballin, M., M. D'Orazio, M. Di Zio, M. Scanu, and N. Torelli

2009. Statistical matching of two surveys with a common subset. *ISTAT Report Published*.

Becker, S. O. and A. Ichino

2002. Estimation of average treatment effects based on propensity scores. *STATA Journal*, 2(4):358–377.

Bloschl, G.

2005. *Encyclopedia of Hydrological Sciences*, chapter 9 - Statistical Upscaling and Downscaling in Hydrology. John Wiley & Sons.

Chabé-Ferret, S. and J. Subervie

2013. How much green for the buck? Estimating additional and windfall effects of French agro-environmental schemes by DID-matching. *Journal of Environmental Economics and Management*, 65(1):12–27.

Chen, J. and J. Shao

2000. Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16(2):113–131.

Coli, A., F. Tartamella, G. Sacco, I. Faiella, M. Scanu, M. D'Orazio, M. Di Zio, I. Siciliani, S. Colombini, and A. Masi

2005. La costruzione di un archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'indagine Banca d'Italia sui bilanci delle famiglie italiane. Technical report, ISTAT.

Defrancesco, E., P. Gatto, F. Runge, and S. Trestini
2008. Factors affecting farmers' participation in Agri-environmental Measures: a Northern Ialian perspective. *Journal of Agricultural Economics*, 59(1):114–131.

Dehejia, R. H. and S. Wahba
1999. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

Dehejia, R. H. and S. Wahba
2002. Propensity score matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161.

Donatiello, G., M. D'Orazio, D. Frattarola, A. Rizzi, M. Scanu, and M. Spaziani
2016. Statistical matching of EU-SILC and HBS at ISTAT: where do we stand for the production of official statistics. Technical report, ISTAT.

D'Orazio, M.
2008. Reliability responses at the 14th Italian Population and Housing Census in 2001. Technical report, ISTAT.

D'Orazio, M.

2014. L'integrazione di fonti dati mediante statistical matching. Technical report, ISTAT.

D'Orazio, M.

2016. *StatMatch*. CRAN, 1.2.4 edition. Technical manual.

D'Orazio, M. and E. Catanese

2016. Evaluating revenues and economic growth for farms producing renewable energies: an investigation based on integration of FSS and EOAH 2013 survey data. Conference paper of the 7th International Conference on Agricultural Statistics, Rome - Italy.

D'Orazio, M., M. Di Zio, and M. Scanu

2006. *Statistical Matching: Theory and Practice*. John Wiley & Sons.

D'Orazio, M.

2015. Integration and imputation of survey data in R: the StatMatch package. *Romanian Statistical Review*, 63(2):57–68.

Fellegi, I. P. and A. B. Sunter

1969. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Gallagher, E. D.

1999. *COMPAH documentation*. University of Massachusetts at Boston. Technical manual.

Gower, J. C.

1971. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871.

Jaraitė, J. and A. Kažukauskas

2012. The effect of mandatory agro-environmental policy on farm fertiliser and pesticide expenditure. *Journal of Agricultural Economics*, 63(3):656–676.

Kirchweger, S. and J. Kantelhardt

2012. Improving farm competitiveness through farm-investment support: a propensity score matching approach. Conference paper of the 131st EAAE Seminar on “Innovation for Agricultural Competitiveness and Sustainability of Rural Areas”, Prague - Czech Republic.

Kleijn, D. and W. J. Sutherland

2003. How effective are European agri-environmental schemes in conserving and promoting biodiversity? *Journal of Applied Ecology*, 40(6):947–969.

Kopeinig, M., M. Caliendo, and S. Kopeinig

2005. Some practical guidance for the implementation of propensity score matching. *Open access publication of ZBW*, 1588(1):1–32.

Kum, H. and T. Masterson

2008. Statistical matching using propensity scores: theory and application to the Levy Institute measure of economic well-being. *Levy Economics Institute Papers*, 53(5):1–33.

Little, R. J. A. and D. B. Rubin

2002. *Statistical Analysis with Missing Data*. John Wiley & Sons.

Liu, X. and L. Lynch

2011. Do agricultural land preservation programs reduce farmland loss? Evidence from a propensity score matching estimator. *Land Economics*, 87(2):183–201.

Marconi, V., M. Raggi, and D. Viaggi

2015. Assessing the impact of RDP Agri-Environment Measures on the use of nitrogen-based mineral fertilizers through spatial econometrics: the case study of Emilia-Romagna (Italy). *Ecological Indicators*, 59:27–40.

Mardia, K. V. and P. E. Jupp

1979. Maximum likelihood estimators for the matrix Von-Mises-Fisher and Bingham distributions. *The Annals of Statistics*, 86(1):599–606.

Pufahl, A. and C. R. Weiss

2009. Evaluating the effects of farm programmes: results from propensity score matching. *European Review of Agricultural Economics*, 36(1):79–101.

Raggi, M., D. Viaggi, F. Bartolini, and A. Furlan

2015. The role of policy priorities and targeting in the spatial location of participation in Agri-Environmental Schemes in Emilia-Romagna (Italy). *Land Use Policy*, 47(1):78–89.

Rässler, S.

2012. *Statistical Matching: a Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Springer Science & Business Media.

Ratinger, T., T. Medonos, and M. Hruska

2013. An assessment of the differentiated effects of the investment support to agricultural modernisation: the case of the Czech Republic. *AGRIS on-line papers in Economics and Informatics*, 5(4):153–165.

Rosenbaum, P. R. and D. B. Rubin

1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and D. B. Rubin

1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rubin, D. B.

1973. Matching to remove bias in observational studies. *Biometrics*, 29(1):159–183.

Rubin, D. B.

1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–702.

Rubin, D. B.

1977. Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioural Statistics*, 2(1):1–26.

Rubin, D. B.

1980. Bias reduction using Mahalanobis-metric matching. *Biometrics*, 36(2):293–298.

Rubin, D. B.

2004. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31(2):161–170.

Rubin, D. B.

2005. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331.

Singh, A. C., H. Mantel, M. Kinack, and G. Rowe

1993. Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19(1):59–80.

Sisto, A.

2006. Propensity score matching: un'applicazione per la creazione di un database integrato ISTAT-Banca d'Italia. *POLIS on-line papers*, 63(2):1–26.

Tancredi, A. and B. Liseo

2011. A hierarchical Bayesian approach to record linkage and population size problem. *Annals of Applied Statistics*, 5(2):1553–1585.

Udagawa, C., I. Hodge, and M. Reader

2014. Farm level costs of Agri-Environment Measures: the impact of entry level stewardship on cereal farm incomes. *Journal of Agricultural Economics*, 65(1):212–233.

Uthes, S. and B. Matzdorf

2013. Studies on agro-environmental measures: a survey of the literature. *Environmental Management*, 51(1):251–266.

Vavreck, L. and D. Rivers

2008. The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties*, 18(4):355–366.

Viaggi, D., C. Signorotti, V. Marconi, and M. Raggi

2015. Do Ågri-Environmental Schemes contribute to high nature value farmland? A case study in Emilia-Romagna (Italy). *Ecological Indicators*, 59:62–69.

Website

2000. FADN official website. *http://ec.europa.eu/agriculture/rica/*. Accessed June 2016.

Website

2004. E-R Statistical Office official website. *http://statistica.regione.emilia-romagna.it*. Accessed May 2016.

Website

2008. CAP-IRE project website. *http://www.unibo.it/en/research/projects-and-initiatives/Unibo-Projects-under-7th-Framework-Programme/cooperation-1/socio-economic-sciences-and-humanities-ssh/cap-ire*. Accessed September 2016.

Website

2016. EC official website. *http://ec.europa.eu/agriculture/rurdev/index˙en.htm*. Accessed September 2016.

Winkler, W. E.

2005. *Overview of record linkage and current research directions.* U.S. Bureau of the Census. Technical manual.

Winship, C. and S. L. Morgan

1999. The estimation of causal effects from observational data. *Annual Review of Sociology*, 25(1):659–706.

List of Acronyms

AES - Agri-Environmental Schemes

CAP - Common Agricultural policy

e - Exact distance function

mn - Manhattan distance function

ms - Mahalanobis distance function

nnd - Nearest Neighbour Distance Hot Deck

nndc - Contrained Nearest Neighbour Hot Deck

PSM - Propensity Score Maching

RDP - Rural Development Policy

rnd - Random Hot Deck

rnk - Rank Hot Deck

SM - Statistical Matching (imputation)

Appendix

Due to the huge size of the Checking Table file, this part is available behind request to the author.