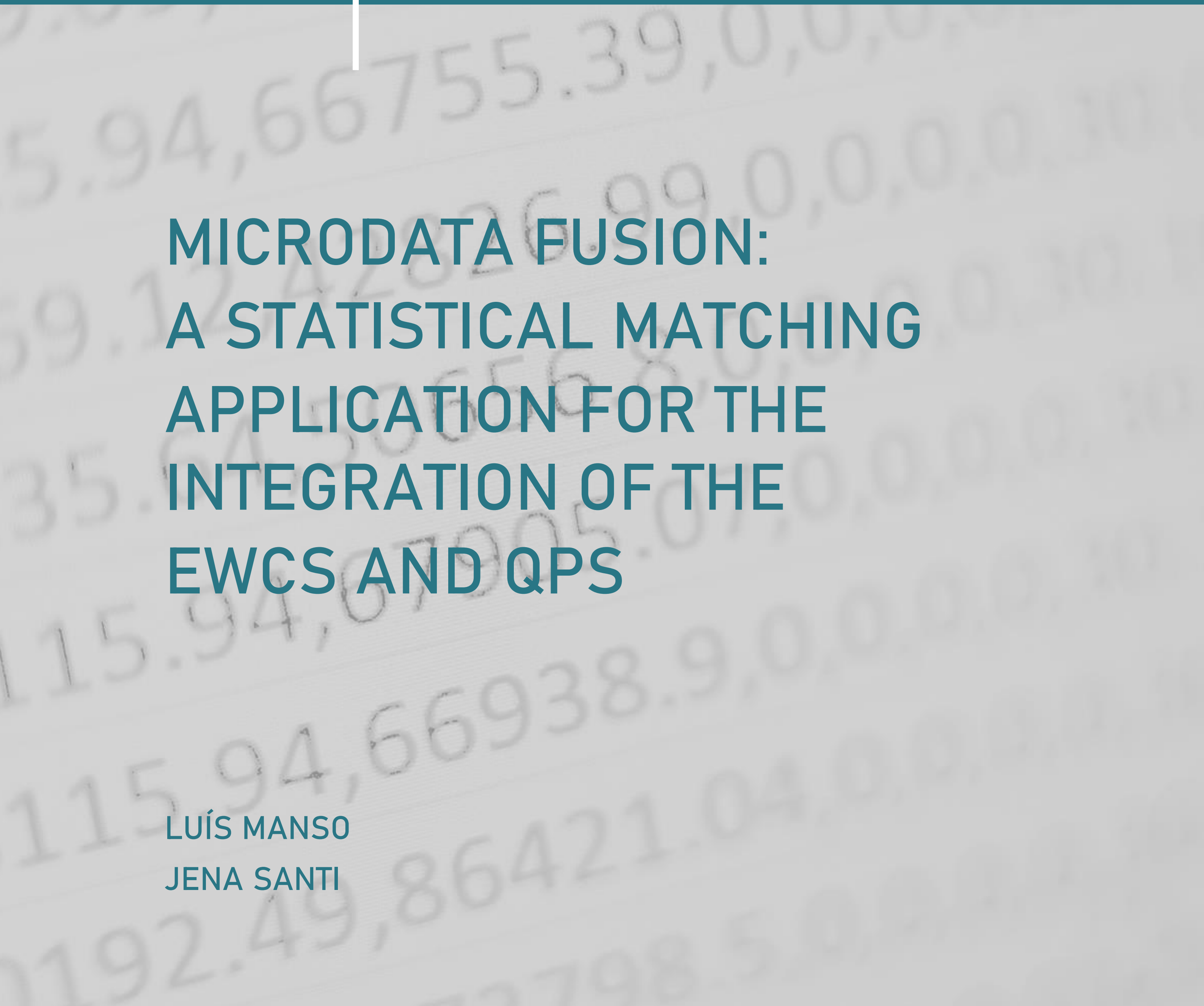
COLABOR
Laboratório Colaborativo para o Trabalho, Emprego e Proteção Social

4
JUN 2021

ESTUDOS COLABOR
WORKING PAPER



MICRODATA FUSION: A STATISTICAL MATCHING APPLICATION FOR THE INTEGRATION OF THE EWCS AND QPS

LUÍS MANSO
JENA SANTI

FICHA TÉCNICA

Título:

Microdata fusion: a statistical matching application for the integration of the EWCS and QPS.

Autores:

Luís Manso
Jena Santi

DOI:

10.5281/zenodo.4792281

Data de publicação:

Junho 2021

Cite este documento como:

Manso, L., Santi, J. (2021). *Microdata fusion: a statistical matching application for the integration of the EWCS and QPS*. Estudos CoLABOR, N.º4. CoLABOR.

CoLABOR
Laboratório Colaborativo
Para o Trabalho, Emprego
e Proteção Social
Rua das Taipas, n.º 1
1250-264 Lisboa

Índice de tabelas

Índice de figuras

Resumo

Microdados referem-se a bases de dados em que a microunidade constitui o elemento central de análise – indivíduos, famílias ou empresas. Estes dados são tradicionalmente recolhidos através de inquéritos, censos ou dados administrativos e permitem que utilizadores/investigadores analisem uma gama ampla de tópicos e relações entre subpopulações. As características dos dados são geralmente determinadas pelo objetivo orientador da recolha dos mesmos. Como tal, habitualmente, não cobrem todas as dimensões de análise em profundidade, o que cria a necessidade de recolha de informação através da realização de novos e dispendiosos inquéritos ou outros métodos de recolha de dados.

Em resposta a este problema, têm surgido vários métodos que procuram utilizar a informação existente dispersa por várias bases de dados. Este tipo de métodos procuram a integração de várias bases de dados, através de um conjunto de variáveis comum entre elas. Este documento apresenta uma análise dos métodos habitualmente utilizados para a integração de informação dispersa em microdados, com particular destaque na identificação da viabilidade de integração do Inquérito aos Quadros de Pessoal (QPS) e do Inquérito Europeu às Condições de Trabalho (EWCS). As técnicas aqui consideradas enquadram-se em três categorias distintas: (1) paramétricas; (2) não paramétricas; e (3) misto.

Os resultados desta análise sugerem que o EWCS e os QPS podem ser integrados com sucesso através de métodos de correspondência estatística. Como esperado, existe um custo de integração associado a este procedimento, que se reflete nas distribuições probabilísticas da nova base de dados sintética. De forma a integrar com sucesso as duas fontes de informação, é necessário proceder a um extenso procedimento de harmonização, que exige a agregação de algumas das variáveis continuas, que se traduz numa perda implícita da especificidade da informação contida nas bases de dados. Por último, derivado dos requisitos computacionais associados, não foi possível otimizar o processo de correspondência. A otimização ideal deveria ser obtida através de um algoritmo que resolve o problema de atribuição. No entanto, foi utilizada uma abordagem heurística para a otimização do nosso problema que minimiza as distâncias entre indivíduos nas duas bases de dados através de uma iteração sequencial.

Palavras-chave: Agregação de microdados; correspondência estatística; European Working Conditions Survey (EWCS); Inquérito Quadros de Pessoal (QPS)

Este estudo foi desenvolvido no âmbito do projeto "Trabalho, Teletrabalho, e Distanciamento Social em Situação de Pandemia", financiado pela Fundação para a Ciência e Tecnologia – call Research 4 COVID-19, projeto n.º 804.

Este texto foi submetido a avaliação científica interna e/ou externa. As análises, opiniões e conclusões nele expressas são da exclusiva responsabilidade do(s) autor(es).

Abstract

Microdata refers to data that has micro-units as the center unit of analysis, such as individuals, households or firms, commonly collected through surveys, census or administrative data. This type of data allows users/researchers to analyse a wide range of topics and to capture the intrinsic relationships between sub-populations. The characteristics and utility of the dataset is usually determined by the guiding objective of data collection. As such, datasets usually do not cover all dimensions in-depth, which creates the need for the new and costly surveys and other data collection methods.

More recently, data integration methods have been introduced as cost-effective way of obtaining a wider dataset that contains more dimensions. Essentially, these processes consist of the integration of distinct datasets based on a set of common variables. This document presents an overview of the problems and methods commonly used to integrate micro-data from different sources with a particular focus on identifying the feasibility of integrating the Quadros de Pessoal Survey (QPS) and the European Working Conditions Survey (EWCS). The techniques considered here fall into three distinct categories: (1) parametric; (2) non-parametric; and (3) mixed.

Our results suggest that the EWCS and the QPS can be successfully matched using statistical matching procedures. As expected, there is a cost of integration that is reflected in the probability distributions of the new synthetic dataset. In addition, to successfully integrate both datasets, there is a need for an extensive harmonization procedure, which may require the aggregation of continuous variables into categorical. Finally, we were unable to optimize our matching procedure due to the computational requirements for the application of an algorithm that can solve an assignment problem. Rather, we use a heuristic approach to the optimization of our problem. There is a clear trade-off between optimization and the computational requirements to carry out this procedure.

However, there is a need for extensive harmonization procedures identifying the matches between individuals in both datasets.

Keywords: Microdata fusion; statistical matching; European Working Conditions Survey (EWCS); Quadros Pessoal Survey (QPS)

This work was developed under the project "Work, Remote-Work and Social Distancing in a Pandemic" financed by the Foundation for Science and Technology - call Research 4 COVID-19, project n.º 804.

This text has been submitted to internal and/or external scientific evaluation. The analyses, opinions and instruments expressed therein are the sole responsibility of the author(s).

1. Introduction

Microdata is a term usually referred to describe a survey, census or administrative data sample that has a micro-unit of analysis. These micro-units often refer to individuals or households, where each observation is equivalent to a single individual/household and that may contain a weight to account for a representative sample of a specific population. Often, these data samples provide specific characteristics about individuals or households that are mostly determined by the guiding purpose of data collection. For instance, the EU-SILC, as the name suggests, provides information on individual characteristics regarding income and living conditions of the respondents. It contains variables on earnings, social security benefits, family conditions, etc. Similarly, to the EU-SILC there are several other micro-data sources that can be used to characterize individuals and populations on different topics – administrative social security data, the European Working Conditions Survey (EWCS), Quadros de Pessoal Survey (QPS), among others.

There are several benefits for Statistical Agencies (national or otherwise) to make micro-datasets available for the public in general. Often, after conducting a survey or collecting microdata for any purpose, Statistical Agencies provide an aggregate summary of the data in the form of tables or by means of other data visualization techniques to provide users with the highlights of the data. However, these agencies are often not equipped, or funded, to perform in depth analysis of data or to identify the range of research questions that can be derived from the dataset. Consequently, by making micro-datasets publicly available, Statistical Agencies play a key role in fostering research on a wide range of topics, often associated with the purpose of the survey/data collection. On the other hand, from a researcher perspective, microdata allows users to analyse and evaluate fine and intrinsic relationships, including interaction between different phenomena. For instances, the EU-SILC is often used to analyse the distributional effects of Social Security benefits, which are nearly impossible to capture using aggregate data.

The statistical infrastructure of social surveys that provide the basis for the finer analysis that aim to disentangle the intrinsic nature of relationships in contemporary societies is organized around specific surveys that cover many relevant aspects of the user’s necessities: income, consumption, labour-market, health, education, etc. However, no single survey can cover all these dimensions in-depth. For instance, consider the case of a researcher who wishes to engage in the estimation of poverty in Portugal. The classical approach to this problem would be to estimate poverty indicators based on the EU-SILC dataset. This survey contains a range of income indicators that allow the researcher to calculate equivalized household income and, consequently, the at-risk-of-poverty threshold. However, the EU-SILC is characterized by a small sample size, which in turn originates in large variances when considered at the regional level. On the other hand, the Census data consists of a much larger dataset, however, it does not contain the necessary income variables to allow for the necessary computations to estimate poverty indicators – the census data usually does not inquire the respondents regarding income.

This issue, leads to the main question of this working paper: is it feasible to integrate both of these micro-datasets into a single dataset containing the stronger characteristics of each one, and if so, what are the statistical procedures that would guarantee a high-quality match between individuals in each dataset? For this purpose, this working paper will aim to answer the following questions:

1. What are the existing methods used to fuse micro-datasets?
2. What are the specific criteria the datasets must contain to ensure a successful integration that allows researchers to derive conclusions and make inferences based on the new data?
    a. Does the EWCS and the QPS data meet these criteria?
    b. How do these surveys compare in terms of their common variables, sampling and population?
3. What is the ideal method that can be used to integrate the EWCS and the QPS based on the specific characteristics of the surveys?

2. Micro-fusion

Microdata fusion in the context used in this paper refers to the process of integrating data from different sources for statistical purposes. As the name suggests, it is concerned with the integration of microdata, i.e. data composed of micro-units, resulting in a dataset that is also composed of micro-units. This process encompasses a combination of theory, methods, and tools for creating a synergy of the information acquired in both datasets. The resulting dataset should, in theory, provide a wider range of analysis and accuracy in terms of prediction than it would have been possible if any of the sources were to be used individually. Although the dataset we aim to create is not necessarily being used for predictions, we are, in fact, attempting to widen the range of our analysis.

In the literature reviewed two different categories of micro-fusion were identified: record linkage and statistical matching. These two categories are used to answer specific problems that arise when performing data fusion. For instance, an important element that needs to be carefully considered when performing data fusion is the unit-composition of the datasets. Are the datasets we are attempting to fuse composed by the same units? If so, this would mean that we are primarily dealing with a case of integration between registers, such as administrative data for example. On the other hand, if the unit composition is different, it would most likely mean that our data belongs to one or more sample surveys. This initial distinction is crucial since it will provide guidance for the methods and tools that are necessary for the successful integration of both datasets.

3. Record linkage or object matching

Record linkage is essentially a combination of methodologies and tools used to match records that are believed to belong to the same unit or entity. Herzog *et al.* (2007) define record linkage as the bringing together of information from two records that are believed to relate to the same entity – such as the same family, individual or household. The challenge is specifically to bring together the records from the same individual entity. This type of linkage is called *exact matching* (Herzog *et al.*, 2007). For instance, the process of integrating the administrative social security dataset with the administrative tax revenue dataset is an example of a record linkage process. Since individual identifying numbers are available, it would be a simple task to integrate both datasets into a single one containing all the variables present in both datasets. However, this task becomes more challenging when there are no identification numbers that can be used to determine the exact match between units in the datasets. For this purpose, other variables are used as identifying variables, such as names, addresses or date of birth. Note that this process does not necessarily rely on a single identifying variable and can use a combination of variables to strengthen the linking process and to reduce possible errors that may arise from matching addresses or date of birth. Essentially, there are two specific types of record linkage processes: deterministic and probabilistic. The remainder of this section provides a brief review of both types.

3.1 Deterministic

Deterministic record linkage compares an identifier, or a group of identifiers, across databases to establish a link between units. In this method, a link is only established if all identifiers match between the two datasets. For example, linking two datasets that contain a citizen identification number is a simple process. When the numbers match, we can be relatively certain that it belongs to the same individual and therefore are able to establish a link between the two datasets. However, not all datasets contain such an obvious identification variable.

In these cases, it is necessary to use other identification variables, such as a person's name, postal-code, date of birth, etc. Herzog *et al.* (2007) call these variables weakened characteristics. The authors provide a very illustrative example of how this process worked in a specific project. In sum, researchers established a deterministic matching scheme that involved the first four characters of the name variable and the first five digits of the zip-code, generating a matching string of nine characters. This is a great example of different approaches that may be employed when considering record linkage. It is important to note that these matching schemes should be developed to meet the specific needs of the analysis that is to be performed, always with the aim of maximizing the number of correct matches. However, when developing a weakened match, this process will undoubtedly increase the number of false matches.

3.2 Probabilistic

Probabilistic record matching uses a slightly different approach however, the starting point is similar to deterministic record linkage. Initially, the purpose is to use a group of common identifiers to link two distinct files. For the sake of simplicity, we will use a simple example of probabilistic record matching, where a researcher wishes to bring files *A* and *B* together with the purpose of studying the relationship between gender and education – see Table 1. For this purpose, the researcher wishes to use first name and last name as key variables.

Table 1. Files A and B for matching

File A		
fname_A	lname_A	sex
Joana	Alcantara	f
João	António	m
André	Simões	m
Jorge	Jesus	m

File B		
fname_B	lname_B	educ
Joana	Alcantara	Ph.D
João	António	Ph.D
Andre	Simoes	MA
Jorge	Jeus	MA

The first step in performing probabilistic record matching is to merge the two files to compare key variable matches. Table 2, shows an example of a merge, also called join, of files *A* and *B* by first and last name. The column *agr* shows the agreement pattern between the two files. The first digit of this column indicates whether the first name agrees between both files (coded 1) or disagrees (coded 0). The second digit relates to the agreement on the last name field. This process is more intrinsic than the deterministic matching, since it allows for partial links, where the first name agrees but the last name does not for example. The researcher is now in a comparatively informed position. If the choice is to only accept full matches, this would produce the same results as a deterministic matching procedure. However, the researcher is now able to accept a lower threshold for accepting a link between two observations.

Table 2. Joined files A and B for comparative analysis of agreement

fname_A	lname_A	fname_B	lname_B	educ	sex	Agr
Joana	Alcantara	Joana	Alcantara	Ph.D	f	(1,1)
Joana	Alcantara	João	António	Ph.D	f	(0,0)
Joana	Alcantara	Andre	Simoes	MA	f	(0,0)
Joana	Alcantara	Jorge	Jeus	MA	f	(0,0)
João	António	Joana	Alcantara	Ph.D	m	(0,0)
João	António	João	António	Ph.D	m	(1,1)
João	António	Andre	Simoes	MA	m	(0,0)
João	António	Jorge	Jeus	MA	m	(0,0)
André	Simões	Joana	Alcantara	Ph.D	m	(0,0)
André	Simões	João	António	Ph.D	m	(0,0)
André	Simões	Andre	Simoes	MA	m	(0,0)
André	Simões	Jorge	Jeus	MA	m	(0,0)
Jorge	Jesus	Joana	Alcantara	Ph.D	m	(0,0)
Jorge	Jesus	João	António	Ph.D	m	(0,0)
Jorge	Jesus	Andre	Simoes	MA	m	(0,0)
Jorge	Jesus	Jorge	Jeus	MA	m	(1,0)

The simple dichotomy presented in this example does not fully reflect similarities between cases. For example, in the case of André Simões and Andre Simoes, this join assumes that there is no match, even though it is clearly the same individual and it was a record issue. In this case, probabilistic record matching can calculate how much any two fields disagree, partially agree, or totally disagree. Assuming 1 indicates complete agreement and 0 complete disagreement, we can determine partial agreement as a number between 0 and 1 that reflects how similar two data entries are. For instance, we can evaluate similarity between each one of the characters. This could be calculated as 1 minus the proportion of disagreement calculated as the number of characters in disagreement divided by the total number of characters in each name. In this case, the proportion of similarity between "André" and "Andre" would be: 1 – 1/5 = 0,8. And for "Simões" and "Simoes" we would have: 1 – 1/6 = 0,833. Consequently, the agreement for this case would be (0,8; 0,8333), situating this case in a partial agreement category. This is just a simple example of how this process could be operationalised. There are other, more complex matching procedures that can be used to evaluate the agreement between key variables. However, this example clearly illustrates the ability of the researchers to relax their assumptions and performing matches unable to achieve using a deterministic procedure. Since we can quantify the level of agreement between the two key variables in dataset *A* and *B*, this means we are also able to calculate probability scores for the matches.

Although useful in many cases, record linkage does not meet the necessary requirements for the problem exposed. As previously stated, the purpose of this working paper is to evaluate methods and tools that may allow for the fusion of the EWCS and QPS datasets. In this regard, the procedure of record linkage (deterministic or probabilistic) is not adequate, since these procedures are only possible to implement under the condition that both sets of units in the datasets are, at least, partially overlapping, which is not the case. For this reason, we will not expand on this procedure.

4. Statistical matching

Statistical matching is another method for the integration of data from different sources. In sum, statistical matching aims to integrate two (or more) datasets under the following conditions:

- The different datasets contain information on a group of common variables.
- Variables are not jointly observed.
- The units observed in the datasets are different (disjoint sets of units).

In this case, integration refers to the possibility of having joint information on the not jointly observed variables of the different sources. For this purpose, there are two distinct ways to achieve the desired outcome:

- **Micro approach:** The purpose of the micro approach is to build a synthetic file that contains all the variables of interest independent of the fact that these were observed and collected from different sources.
- **Macro approach:** The purpose of the macro approach is to use the source files to directly estimate the joint distribution function of the variables of interest which have not been observed in common.

Further, the approach to statistical matching can be: (1) parametric, where the relationship between variables is explicitly explained by means of a statistical model; (2) non-parametric, when no model is assumed; or (3) mixed, when there is a mix between parametric and non-parametric procedure. provides an overview of the methods for statistical matching regarding its approach and purpose.

Table 3. Review of statistical matching approaches

Objective of SM	Parametric	Non-Parametric	Mixed
Macro	• Methods for the estimation of model parameters in the presence of missing values	• Estimation of the empirical cumulative distribution. • Kernel density estimators	
Micro	• Conditional mean matching. • Stochastic regression imputation. • …	• Hot deck imputation procedures.	• Combination of the predictive mean matching with hot deck imputation. • …

Source: adapted from D´Orazio (2015).

Since the macro approach presented above is best suited to generate contingency tables of variables not jointly observed or to calculate correlation coefficients, this approach is not suitable for the purpose of our analysis. As such, the focus will be solely on the micro approach. However, it is important to note that the two approaches are not necessarily distinct, in the sense that the micro approach is always a by-product of an estimation of the joint distribution of all the variables of interest.

Before going any further, it is important to provide an explanation on the synthetic dataset generated in the micro-approach. Although the datasets that are being integrated were directly observed, the resulting dataset was not. As such, we must consider the resulting dataset as being composed by synthetic data, since all the variables were not directly observed together. It is important to consider that statistical matching is not a necessity. Rather, when a set of variables is not commonly observed in a dataset, researchers have the option to conduct a new survey that may incorporate the questions to capture the desired information. In this regard, statistical matching provides an answer to issues that are commonly associated with the process of gathering microdata through surveys:

1. It takes time to plan and conduct a survey.
2. There are specific costs that are associated with conducting a survey.
3. The need for new data may require several questions/variables, which can compromise the quality of the responses and, consequently, of the data itself.
4. Additional surveys increase the burden on the respondent, affecting data quality and the rate of non-response.

In this regard, considering the driving purpose of this working-paper, we identify the potential of statistical matching as a primary candidate for the integration of the EWCS with the QPS. The remainder of this working paper will review different approaches withing the framework of statistical matching for the integration of micro-datasets (parametric, non-parametric and mixed), identifying their strengths, weaknesses and above all, their suitability for the integration desired. For the purpose of identifying the most adequate method, several approaches will be tested and their results compared. The following section provides a definition of the statistical/mathematical framework for the statistical matching problem.

4.1 Definition of the statistical/mathematical framework for the statistical matching problem

For illustrative purposes it is often useful to define the statistical matching problem as a simple integration of two hypothetical and independent survey datasets A and B composed by random groups of variables (X, Y, Z) with a respective number of variables for each group denoted by *P*, *Q* and *R* respectively such that $X = (X_1, \dots, X_P)'$, $Y = (Y_1, \dots, Y_Q)'$ and $Z = (Z_1, \dots, Z_R)$' represents a vector consisting of these variables. Further, assume that each of these datasets is composed by a total number of observations denoted by n_A and n_B and where the units in A have Z missing and the units in B have Y missing. We can write the observed values of the units in sample A and B as:

$$(x_a^A, y_a^A) = \left(x_{a1}^A, \dots, x_{ap}^A \ \ , \ \ y_{a1}^A, \dots, y_{aq}^A\right), a = 1, \dots, n_a$$

$$(x_b^B, z_b^B) = \left(x_{b1}^B, \dots, x_{bp}^B \ \ , \ \ z_{b1}^B, \dots, z_{bq}^B\right), b = 1, \dots, n_b$$

When the objective is to obtain information on the joint distribution of (X,Y,Z), denoted as *A*∪*B*, from observe samples *A* and *B*, we are dealing with a statistical matching problem that's described in Table 4.

Table 4. Definition of the statistical matching problem when considering samples, A and B, to obtain AUB

Sample	Y_1	…	Y_q	…	Y_Q	X_1	…	X_p	…	X_P	Z_1	…	Z_r	…	Z_R
A	Y_{11}^A Y_{a1}^A $Y_{n_A1}^A$	… … …	Y_{1q}^A Y_{aq}^A $Y_{n_Aq}^A$	… … …	Y_{1Q}^A Y_{aQ}^A $Y_{n_AQ}^A$	X_{11}^A X_{a1}^A $X_{n_A1}^A$	… … …	X_{1p}^A X_{ap}^A $X_{n_Ap}^A$	… … …	X_{1P}^A X_{aP}^A $X_{n_AP}^A$					
B						X_{11}^B X_{b1}^B $X_{n_B1}^B$	… … …	X_{1p}^B X_{bp}^B $X_{n_Bp}^B$	… … …	X_{1P}^B X_{bP}^B $X_{n_BP}^B$	Z_{11}^B Z_{b1}^B $Z_{n_B1}^B$	… … …	Z_{1r}^B Z_{br}^B $Z_{n_Br}^B$	… … …	Z_{1R}^B Z_{bR}^B $Z_{n_BR}^B$
$\boldsymbol{A \cup B}$	Y_{11}^A Y_{a1}^A $Y_{n_A1}^A$	… … …	Y_{1q}^A Y_{aq}^A $Y_{n_Aq}^A$	… … …	Y_{1Q}^A Y_{aQ}^A $Y_{n_AQ}^A$	$X_{11}^{A\cup B}$ $X_{a\cup b1}^{A\cup B}$ $X_{n_{A\cup B}1}^{A\cup B}$	… … …	$X_{1p}^{A\cup B}$ $X_{a\cup bp}^{A\cup B}$ $X_{n_{A\cup B}p}^{A\cup B}$	… … …	$X_{1P}^{A\cup B}$ $X_{a\cup bP}^{A\cup B}$ $X_{n_{A\cup B}P}^{A\cup B}$	Z_{11}^B Z_{b1}^B $Z_{n_B1}^B$	… … …	Z_{1r}^B Z_{br}^B $Z_{n_Br}^B$	… … …	Z_{1R}^B Z_{bR}^B $Z_{n_BR}^B$

Source: adapted from D'Orazio *et al.* (2006), Statistical Matching: theory and practice.

Table 4 shows a practical example of the statistical matching of samples *A* and *B*, which clearly illustrates the problem of data integration of (X, Y, Z) in the absence of joint information for this groups of variables.

4.2 Parametric matching

Parametric statistical matching attempts to solve a multiple imputation problem using a parametric model, i.e. a model in which all its information is represented within its parameters. The underlying assumption of a parametric model is that the parameters represent all the information necessary for the prediction of unknown values.

Using the example above, considering the integration of *B* in *A*, with a resulting dataset composed of *AUB*, a parametric statistical matching approach consists of the development of a parametric model for this purpose. There is a plethora of models that could be used for this purpose. A simpler imputation procedure could be based on a regression model. Regression imputation can be classified into two different types:

- **Deterministic regression imputation** replaces missing values with the exact prediction of the regression model and does not consider variation around the slope, i.e. it does not take uncertainty into account.

- **Stochastic regression imputation** takes uncertainty into account by adding extra variance to predicted values from a regression model.

Essentially, deterministic regression imputation uses predicted values from a regression model as imputations for missing cases. The predicted value $\hat{z}_i$ is the best predictor for the *i-th* unobserved value z_i under the population model:

$$E(z_i) = a + \beta x_i, \qquad V(Z_i) = \sigma^2, \qquad Cov(z_i, z_j) = 0$$

There are two major shortcomings commonly associated with this type of imputation method. First, a simple regression model is, in most cases, hardly explanatory of real-world phenomena. Second, this approach to missing data imputation ignores predictive uncertainty. In this case, imputed values are often too precise and may lead to overfitting where the predictions are situated way too close to the regression slope. For this purpose, stochastic regression imputation is generally considered a better choice, since it considers the model error.

Stochastic regression imputation attempts to solve the shrinkage to the mean issue associated with the deterministic approach by adding small random disturbances to the predictions, which increase variability within the imputed values. Hu & Salvucci (2001) identify three commonly adopted methods that are used to draw small random disturbances:

1. Draw a random disturbance from a distribution with mean zero and variance derived from the observed data $N(0, \sigma^2)$.
2. Draw a random disturbance from residuals of the regression model.
3. Draw a random disturbance from residuals of respondents that have similar values on a pre-selected set of variables to protect against non-linearity and non-additivity in regression models.

By adding a small random disturbance to the deterministic imputation equation, we are left with:

$$E(z_i) = a + \beta x_i + \epsilon_i$$

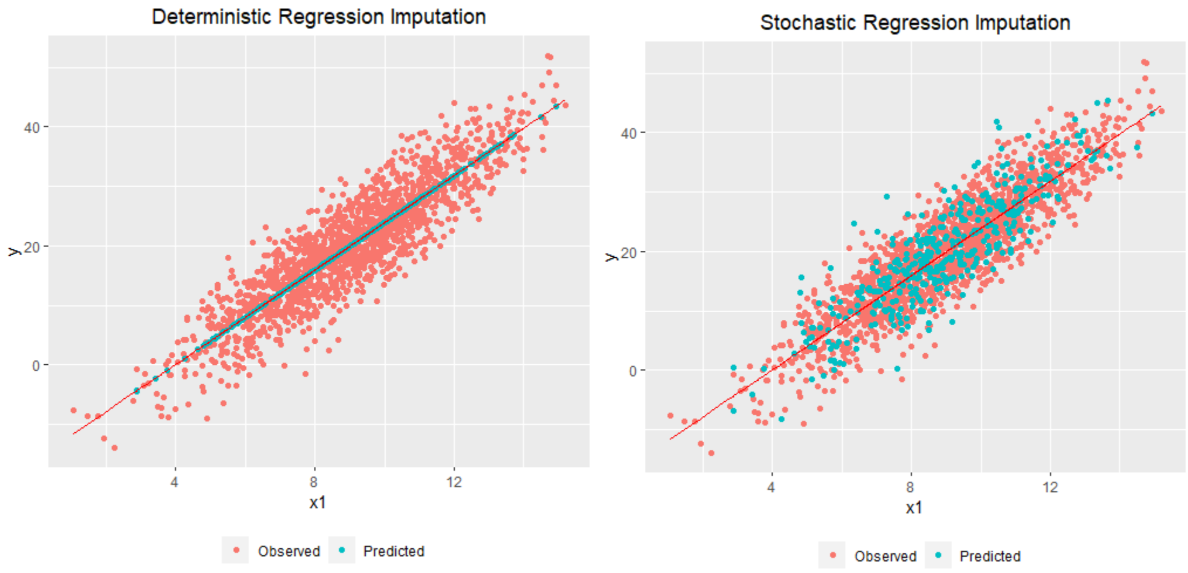
Where ϵ represents the identically and independently distributed (iid) vector of disturbance terms or errors. Although this method constitutes an improvement on deterministic regression imputation, it still has considerable drawbacks that need to be considered. First, stochastic regression imputation can often lead to implausible values since it fails to consider value restrictions (example: income should always be positive). Second, stochastic regression imputation performs poorly when dealing with heteroscedastic data, since it assumes the random error to have the same size for every part of the distribution, often resulting in error terms that are either too large or too small for the imputed values.

To illustrate the methods described, we have generated a simple example of an imputation procedure based on a randomly generated dataset. The dataset generated is composed of two vectors:

1. **y**: A normally distributed vector of 2000 observations with mean 20 and standard deviation 10.
2. **x1**: A normally distributed vector of 2000 observations with mean 5 added to a vector composed of 0,2 times the initial vector.

These variables were generated in this manner to ensure correlation that the vectors correlate. After generating the necessary vectors, we artificially generated 20% of randomized missing values in our y vector. We proceeded applying both deterministic and stochastic regression imputation to both datasets to illustrate the differences between them. For the imputation procedure, we used the R package Mice (van Buuren *et al.*, 2011). The results can be found in Figure 1.

Figure 1. Comparison between deterministic and stochastic regression imputation procedures



The example provided clearly illustrates the way imputation via deterministic regression can result in a biased result since the values are imputed very closely to the regression line that was used to determine it. On the other hand, stochastic procedures can account for the variability of data by taking an error measurement into account in the imputation of missing values.

4.3 Non-parametric matching

Non-parametric matching is predicated on the idea of establishing an imputation method that does not rely on the assumption of any parametric family of distributions for the variables of interest. In this regard, there is a particular set of non-parametric imputation procedures that constitute common practice for approach called *hot deck imputation procedures* (D'Orazio *et al.*, 2006). Essentially, hot deck procedures are a type of donor imputation procedure. The purpose of this type of imputation is to fill in missing values (recipient) with real live observed value (donor) (Ford, 1983). Letting $\tilde{Z}_i^A$ denote the score of the i^{th} unit on the target variable Z missing in file *A* and Z_{jd}^B

the score for the j^{th} unit of the target variable Z in file *B*. Let the index d denote the donor, we can write the generic formula for hot deck imputation as:

$$\tilde{Z}_i^A = Z_{jd}^B$$

The idea behind donor imputation is to find groups of observations that are similar in terms of a set of auxiliary variables. In this sense, the donor usually constitutes an observation that resembles the recipient in one or more common auxiliary variables (Eurostat, 2017). The types of hot deck procedures can, therefore, be distinguished with regard to the donor identification approach. Singh et. al. (1993) identify three distinct hot deck techniques in this regard:

i. Random hot deck;
ii. Rank hot deck;
iii. Distance hot deck.

There are several reasons associated with the growing popularity of hot deck statistical matching procedures. In his discussion paper on statistical matching, de Wall (2015) identifies four specific advantages of using hot deck techniques for the purpose of statistical matching. First, hot deck techniques yield realistic imputation values based on actually observed values. Second, imputed values will always be situated within the realm of possible values. Third, and perhaps one of the most important aspects, is the fact that it is not necessary to model the distribution of the missing data. Essentially, this third point eliminates the need for modelling assumptions that may not hold for a plethora of cases. Finally, hot deck procedures are relatively easy to understand and implement.

Given the growing importance of these techniques in the field of statistical matching, each of them will be reviewed in-depth to assess its suitability for the problem set out by this working paper. This working paper closely follows the description of hot deck procedures found in D'Orazio *et al.* (2006).

4.3.1 Random hot deck

Random hot deck techniques consist of choosing a donor observation at random from the donor file. In order to fine tune this procedure, the choice can sometimes be adjusted so that it is made within a subset of suitable donors regarding their auxiliary attributes. More specifically, units in both the recipient and donor files are usually grouped into homogenous subsets according to a set of common characteristics, which may include gender, geographical area, other demographic characteristics, etc.

Using our statistical framework defined in 0*.* we adapted the example present in D'Orazio *et al.* (2006) to illustrate how a random hot deck procedure might work in practical terms. For this purpose, let A be a dataset composed of 5 units, such that $n_A = 5$, with a set of three observed variables for each unit: gender, age and income. Let B denote a dataset composed of 10 observations, such that$n_b = 10$, with a set of three observed variables for each unit: gender, age and expenditure. Thus, we have a set of common variables $X = (X_1 = Gender, X_2 = Age)$ and two variables no jointly observed: $Y = Income$ and $Z = Expenditure$ (see Table 5).

Table 5. Example for random hot deck statistical matching problem of integration of B in A

Sample	Unit	Y	X_1	X_2	Z
A	a_1	22	F	27	
	a_2	19	M	35	
	a_3	47	M	41	
	a_4	41	F	61	
	a_5	17	F	52	
B	b_1		F	54	22
	b_2		M	21	17
	b_3		F	48	15
	b_4		F	33	14
	b_5		M	63	13
	b_6		F	29	15
	b_7		M	36	19
	b_8		M	55	24
	b_9		F	50	26
	b_{10}		F	27	18

Source: adapted from the example presented in D'Orazio *et al.* (2006).

The integration of *B* in *A* under a random hot deck procedure would mean that to each unit present in *A* would be assigned a donor unit chosen at random from B. Once this unit is assigned, the missing z value in *A* is imputed with the real value from Z in unit b. An example of a possible imputation using random hot deck procedures is found in Table 6.

Table 6. Example of a random hot deck statistical matching result for matching files A and B

Recipient	Donor	X_1^A	X_1^B	X_2^A	X_2^B	Y	Z
a_1	b_2	F	M	27	21	22	17
a_2	b_8	M	M	35	55	19	24
a_3	b_5	M	M	41	63	47	13
a_4	b_6	F	F	61	29	41	15
a_5	b_4	F	F	52	33	17	14

Source: adapted from the example presented in D'Orazio *et al.* (2006).

The results presented above represent a single combination of possible imputations. The total number of possible imputation combinations is given by $n_B^{n_A}$, which in the example presented above is 10^5 total possible combinations, and consequently, the same number of possible distributions. This poses a significant issue in the choice of combination to adopt. In order to deal with this issue, it is common to define specific homogenous groups that will limit the number of combinations substantially. For instance, if we use gender to define homogenous groups for males and females, the possible donor configuration is given by:

$$(n_M^B)^{n_M^A} + (n_F^B)^{n_F^A} = 4^2 + 6^3 = 232$$

As evidenced by the number the formula, this approach would reduce the number of possible donor combinations substantially. The inclusion of additional subgroups will continue to decrease the number of possible donor combinations. For instance, in the example presented above, this could be done by including age-groups as well as gender. Prediction via random hot deck within donation classes defined through a set of auxiliary variables is equivalent to estimating the conditional distribution of Z given X in *B* and drawing observations from it.

4.3.2 Rank hot deck

Rank hot deck procedure can be used in cases where there is one ordinal matching variable X. In this situation, rank hot deck exploits the order relationship between values of X (Singh *et al.*, 1993). Drawing on our previous example in 0., this would be done by ranking the files separately according to values to values of an X variable. When both files are of the same size, this would be a simple procedure. Consider file *A* as the recipient file and $n_b = kn_A$, with k integer, files would be matched by associating records that have the same rank. On the other hand, when files contain a different number of records, such as the case in our previous example, matching is performed by considering the cumulative distribution function of the distribution of X in the recipient file:

$$\hat{F}_X^A(x) = \frac{1}{n_A}\sum_{a=1}^{n_A} I(x_a \le x), \qquad x \in X$$

and in the donor file:

$$\hat{F}^B_X(x) = \frac{1}{n_B}\sum_{b=1}^{n_B} I(x_b \leq x), \qquad x \in X$$

Then, each observation in A is matched with the record in B that minimizes the difference between ranks, such that:

$$\left|\hat{F}^A_X(x^A_a) - \hat{F}^B_X(x^B_b)\right| = \min_{1\leq b\leq n_B}\left|\hat{F}^A_X(x^A_a) - \hat{F}^B_X(x^B_b)\right|$$

For instance, in our example presented above, if age is used as a matching variable, the units in sample *A* and *B* all ranked according to value for age. Table 7 shows how each file in our previous example is arranged when considering age as the matching variable.

Table 7. Files A and B with records ranked according to age

Sample	Unit	Y	X_1	X_2	Z	$\hat{F}^A_X(x^A_a)$	$\hat{F}^B_X(x^B_b)$
A	a_1	22	F	27		1/5	
	a_2	19	M	35		2/5	
	a_3	47	M	41		3/5	
	a_5	17	F	52		4/5	
	a_4	41	F	61		5/5	
B	b_2		M	21	17		1/10
	b_{10}		F	27	18		2/10
	b_6		F	29	15		3/10
	b_4		F	33	14		4/10
	b_7		M	36	19		5/10
	b_3		F	48	15		6/10
	b_9		F	50	26		7/10
	b_1		F	54	22		8/10
	b_8		M	55	24		9/10
	b_5		M	63	13		10/10

Source: adapted from the example presented in D´Orazio *et al.* (2006).

Using the matrix of $\left|\hat{F}_X^A(x_a^A) - \hat{F}_X^B(x_b^B)\right|$ found in Table 8, we are able to find the pairs of units that minimize the absolute difference between both cumulative distribution functions for files *A* and *B*. Having established a link between units that minimizes the difference between the cumulative distribution functions, it is possible to impute the missing values of Z in table *A* using the donor values of Z from the corresponding unit in *B*. The final matched file can be found in Table 9.

Table 8. Matrix of $\left|\hat{\mathrm{F}}_{\mathrm{X}}^{\mathrm{A}}(\mathrm{x}_{\mathrm{a}}^{\mathrm{A}}) - \hat{\mathrm{F}}_{\mathrm{X}}^{\mathrm{B}}(\mathrm{x}_{\mathrm{b}}^{\mathrm{B}})\right|$

Unit	b_2	b_{10}	b_6	b_4	b_7	b_3	b_9	b_1	b_8	b_5
a_1	0,1	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8
a_2	0,3	0,2	0,1	0	0,1	0,2	0,3	0,4	0,5	0,6
a_3	0,5	0,4	0,3	0,2	0,1	0	0,1	0,2	0,3	0,4
a_5	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0	0,1	0,2
a_4	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0

Table 9. Example of a rank hot deck statistical matching result for matching files A and B

Recipient	Donor	X_1^A	X_1^B	X_2^A	X_2^B	Y	Z
a_1	b_{10}	F	F	27	27	22	18
a_2	b_4	M	F	35	33	19	14
a_3	b_3	M	F	41	48	47	15
a_5	b_1	F	F	52	54	41	22
a_4	b_5	F	M	61	63	17	13

Source: adapted from the example presented in D'Orazio *et al.* (2006).

4.3.3 Distance hot deck

Distance hot deck is by far one of the most used procedures in early statistical matching procedures (D'Orazio *et al.*, 2006). In this approach to the data fusion, each of the units in the recipient file *A* is matched with a unit in the donor file *B* according to a distant measure computed based on the common auxiliary variables X. The simplest example of distance hot deck is to match files using a single continuous variable X. Essentially, the donor record would be selected so that it satisfises the following condition:

$$d_{ab^*} = \left|x_a^A - x_{b^*}^B\right| = \min_{1\leq \mathrm{b} \leq n_B} |x_a^A - x_b^B|$$

In the eventual case where two or more donors have the same distance from the recipient, one is chosen at random.

Using our example, considering the integration of *B* in *A* using continuous variable $X = Age$, we can develop a similar matrix to the one found in Table 8 that calculates the distance in terms of age from units in *A* to units in *B*. The distance matrix can be found in Table 10. Note that in the case of unit a_5 there are two equally distant units in *B*, which are b_1 and b_9. In this case, as previously explained, one of them is chosen at random. The final form of the matched file can be found in Table 11.

Table 10. Distance matrix of $\left|x_a^A - x_{b^*}^B\right|$

Unit	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}
a_1	27	6	21	6	36	2	9	28	23	**0**
a_2	19	14	13	2	28	6	**1**	20	15	8
a_3	13	20	7	8	22	12	**5**	14	9	14
a_4	7	40	13	28	**2**	32	25	6	11	34
a_5	**2**	31	4	19	11	23	16	3	**2**	25

Table 11. Example of a distance hot deck statistical matching result for matching files A and B

Recipient	Donor	X_1^A	X_1^B	X_2^A	X_2^B	Y	Z
a_1	b_{10}	F	F	27	27	22	18
a_2	b_7	M	M	35	36	19	19
a_3	b_7	M	M	41	36	47	19
a_4	b_5	F	M	61	63	41	13
a_5	b_1	F	F	52	54	17	22

Source: adapted from the example presented in D'Orazio *et al.* (2006).

Since the values in the donor file can be used more than one time, this type of distance hot deck approach is denominated *unconstrained* hot deck. On the other hand, we may choose to use each donor record only once. This would constitute a *constrained* distance hot deck procedure. This type of approach requires that the number of donors be greater than, or equal to, the number of recipients, i.e. $n_A \leq n_B$.. In the simplest case, where the number of donors is equal to the number of recipients, the donor patter should be such that:

$$\sum_{a=1}^{n_A}\sum_{b=1}^{n_B}(d_{ab}w_{ab}), \qquad w_{ab} \in \{0;1\}$$

where w_{ab} takes the value of 1 if the pair (a, b) is matched, and 0 if it is not. On the other hand, the linear programming problem becomes slightly more complex when there are more donors that recipients. In this case the set of constrains becomes:

$$\sum_{b=1}^{n_B} w_{ab} = 1, \qquad a = 1, \dots, n_A$$

$$\sum_{a=1}^{n_A} w_{ab} \leq 1, \qquad b = 1, \dots, n_B$$

The main advantage of constrained distance hot deck matching is that the imputed variable maintains its marginal distribution. In fact, when $n_A = n_B$, the marginal distribution is perfectly replicated for the imputed value, which in our case refers to variable Z.

4.4 Mixed methods matching

Up to this point we covered two of the three possible methods for micro statistical matching: parametric and non-parametric. The third method identified in D'Orazio (2015) is a mixed approach. Usually, this method consists of a two-step process that combines parametric and non-parametric procedures for statistical matching. More precisely, a mixed approach will initially adopt a parametric model and then obtain a complete synthetic dataset via non-parametric methods. The basic principle behind a mixed approach is to achieve a statistical matching procedure that exploits the benefits of both methods: (1) parametric models are more parsimonious and (2) non-parametric methods are more resilient against model misspecification. Using the methods described above in the parametric and non-parametric sections, we identify some approaches that could be used under mixed methods statistical matching for continuous variables. For categorical variables, a logistic regression model approach should be adopted.

Method 1: Deterministic regression on *A* with a nearest neighbour hot deck matching.

A simple approach to mixed methods would consist in estimating the regression parameters of Z_b on X_b and estimating $\tilde{Z}_a$ for every $n_a = 1, \dots, a$, such that:

$$\tilde{Z}_a = \alpha_Z + \beta_{ZX} X_a$$

This would constitute the parametric step for method 1. The second step, the non-parametric matching, would be achieved by finding the nearest neighbour between the estimated $\tilde{Z}_a$ and the observed Z_b through a hot deck procedure.

Method 2: Deterministic regression on *A* and *B* with a nearest neighbour hot deck matching.

Method two would be similar to method 1, however, the parametric procedure would estimate values for both $\tilde{Z}_a$ and $\tilde{Z}_b$, such that:

$$\tilde{Z}_a = \alpha_Z + \beta_{ZX} X_a$$

$$\tilde{Z}_b = \alpha_Z + \beta_{ZX} X_b$$

Consequently, the non-parametric procedure would find the nearest neighbour between the estimated $\tilde{Z}_a$ and $\tilde{Z}_b$through a hot deck procedure.

Method 3: Stochastic regression on A with a nearest neighbour hot deck matching.

This method would consist on estimating the regression parameters for a stochastic regression model of Z_b on X_b and estimating $\tilde{Z}_a$ for every $n_a = 1, \dots, a$, such that:

$$\tilde{Z}_a = \alpha_Z + \beta_{ZX} X_a + \epsilon_a$$

The second step, the non-parametric matching, would be achieved by finding the nearest neighbour between the estimated $\tilde{Z}_a$ and the observed Z_b through a hot deck procedure.

Method 4: Stochastic regression on *A* and *B* with a nearest neighbour hot deck matching.

Method two would be similar to method 3, however, the parametric procedure would estimate values for both $\tilde{Z}_a$ and $\tilde{Z}_b$, such that:

$$\tilde{Z}_a = \alpha_Z + \beta_{ZX} X_a + \epsilon_a$$

$$\tilde{Z}_b = \alpha_Z + \beta_{ZX} X_b + \epsilon_b$$

Consequently, the non-parametric procedure would find the nearest neighbour between the estimated $\tilde{Z}_a$ and $\tilde{Z}_b$through a hot deck procedure.

5. Comparing the EWCS and the QPS

5.1 Comparing datasets

5.1.1 The European Working Conditions Survey

The EWCS has been regularly carried out on a 5-year interval since its launch in 1990. The primary purpose of this survey is to provide an overview of working conditions in Europe to:

- Assess and quantify working conditions of both employees and the self-employed across Europe on a harmonized basis.
- Analyse the relationships between different aspects of working conditions.
- Identify groups at risk and issues of concern as well as progress.
- Monitor trends by providing homogeneous indicators of these issues.
- Contribute to European policy development on quality of work and employment issues.

The themes covered in the most recent waves of this survey include:

- Employment status.
- Working time duration and organization.
- Work organization.
- Learning and training.
- Physical and psychological risk factors.
- Work-life balance.
- Worker participation.
- Earnings and financial security.
- Work and health.

The 6th wave of this survey, the most recently available date for the EWCS, has 2015 as the reference year and it includes the following countries: EU28, Norway, Switzerland, Albania, the former Yugoslav Republic of Macedonia, Montenegro, Serbia and Turkey. The survey interviewed approximately 44 thousand workers spread out among the 35 countries.

For our analysis, we are particularly concerned with variables that can help us characterize occupations. Essentially, these variables contain information regarding the task content the way these occupations are organized. Thus, the EWCS contains valuable information that we wish to incorporate into our project. A list of variables can be found in Table 12.

Table 12. List of recoded variables from the EWCS (continues)

Dimension	Recoded variable	Description	EWCS Source
Sociodemographic	age	Continuous variable with values 15+.	Q2b
	gender	Binary categorical variable with values 1 (Male) and 2 (Female).	Q2a
	educ	Categorical variable with the following categories: 1. Early childhood education. 2. Primary education. 3. Lower secondary education. 4. Upper secondary education. 5. Post-secondary non-tertiary education. 6. Bachelor or equivalent. 7. Master or equivalent. 8. Doctorate or equivalent.	ISCED
	education_level	Categorical variable with the following categories: 1. Low education (educ 1-2). 2. Medium education (educ 3-5). 3. High education (educ > 5).	ISCED
	nuts2_loc	Location at the Nuts II region.	nuts2_loc
Employment	isco08_1	ISCO 08 Classification at 1 digit.	isco_08_1
	isco08_2	ISCO 08 Classification at 2 digits.	isco_08_2
	workstate	Categorical variable with the following categories: 1. Working. 2. Unemployed. 3. Disabled. 4. Currently on leave. 5. Retired. 6. Homemaker. 7. In education. 8. Other.	Q2c
	employee	Binary categorical variable with the following categories: 0. FALSE. 1. TRUE.	Q7

Table 12. List of recoded variables from the EWCS (conclusion)

Dimension		Recoded variable	Description	EWCS Source
		private	Binary categorical variable indicating the worker is currently employed in the private sector with the following categories: 0. FALSE. TRUE.	
		earnings_m_n	Continuous variable for NET earnings per month.	Q140_euro
		hours_w	Continuous variable for the number of hours worked per week	Q24
Employer		nace_rev2_1	Sector of economic activity – NACE Rev.2 1 digit.	nace_rev2_1
		nace_rev2_2	Sector of economic activity – NACE Rev.2 2 digits.	nace_rev2_2
		n_emp_est	Number establishment workers	Q16a
		N_emp_co	Number of company workers	Q16b
Tasks	**Physical**	pain	Categorical variable indicating the individual works in painful positions with the following categories: 1.	Q30a
		lift	Binary categorical variable indicating the individual works often physically lifts or moves people at work with the following categories: 0. FALSE. 1. TRUE.	Q30b
		load	Binary categorical variable indicating the individual carries or handles heavy loads at work with the following categories: 0. FALSE. 1. TRUE.	Q30c
	ICT	ict	Binary categorical variable indicating the individual often uses ICT's in their work	Q30i

See Annex I for recoding stata code.

5.1.2 The Quadros Pessoal Survey

The QPS is a compulsory survey of all firms conducted annually for the purpose of monitoring compliance with labour law provisions in Portugal. This dataset contains detailed information on every wage earner in Portuguese economy, as well as their employers. The data covers information regarding the following fields:

- Demographic information.
- Earnings.
- Hours worked.
- Profession.

One of the most important aspects of the QPS is the detailed portrait that it paints of the distribution of workers in Portugal across the country. The most important aspect of this dataset in the context of our analysis is the fact that this information is detailed at the NUTS III level and allows us to deepen our understanding of how occupations and, subsequently, tasks our distributed across the country. It is important to note that since this dataset is extended to every firm, it contains information on the whole universe of wage earners in Portugal. Additionally, the QPS contains a range of variables that pertain to the characteristics of the employer, which can be helpful in the context of our analysis.

For our analysis we have recoded a range of variables to achieve a characterization across two dimensions: (1) employee characterization and (2) employer characterization. Our recoded variables and their dependencies can be found in Table 13.

Table 13. List of recoded variables from the QPS (continues)

Dimension	Recoded variable	Description	QPS Source
Employee	age	Continuous variable ranging from 17-68 with open groups for <=17 and >=68 respectively.	idade_Cod
	gender	Binary categorical variable with values 1 (Male) and 2 (Female)	sexo
	educ1	Categorical variable with the following categories: 0. Technical/professional studies. 1. 1 Primary education not complete. 2. Primary education. 3. Secondary education. 4. Post-Secondary education. 5. Bachelor's degree. 6. Undergraduate degree. 7. Master's degree. 8. Doctorate degree. 9. Ignored (missing equivalent).	habil1
	education_level	Categorical variable with the following categories: 1. Low education (educ1 0-2). 2. Medium education (educ1 3-4). 3. High education (educ1 5-8). All ignored values were set to missing.	habil1

Table 13. List of recoded variables from the QPS (continues)

Dimension	Recoded variable	Description	QPS Source
	qualification_level	Categorical variable for the qualification level of the individual with the following categories: 1. Quadros superiores. 2. Quadros médios. 3. Encarregados, contramestres e mestres. 4. Profissionais altamente qualificados. 5. Profissionais qualificados. 6. Profissionais semi-qualificados. 7. Profissionais não qualificados. 8. Estagiários, praticantes e aprendizes. 9. Ignorado All ignored values were set to missing	nqual1
	cpp10_1	Portuguese Profession classification at 1 digit.	prof_1d
	cpp10_2	Portuguese Profession classification at 2 digits.	prof_2d
	cpp10_3	Portuguese Profession classification at 3 digits.	prof_3d
	cpp10_4	Portuguese Profession classification at 4 digits.	prof_4d
	earny_base	Base earnings for the month of October NOT including subsidies and supplements. Continuous variable.	rbase
	earny_paid	Paid earnings for the month of October including subsidies and supplements. Continuous variable.	rganho
	earny_reg_sub	Regular supplements and subsidies paid in October. These include lunch, shift and other subsidies. Continous variable.	prest_reg
	earny_ireg_sub	Irregular supplements and subsidies paid in October. These include all amounts that are not paid on a regular basis throughout the year (ex: vacation pay). Continuous variable.	prest_irreg
	hours_w	Number of hours regularly worked per week. Continuous variable.	hnormais
	hours_m	Number of hours effectively worked in the month of October. Continuous variable.	pnt

Table 13. List of recoded variables from the QPS (conclusion)

Dimension	Recoded variable	Description	QPS Source
Employer	nuts2_co	Company location at the Nuts II level.	nut2_emp
	nuts2_est	Establishment location at the Nuts II level.	nut2_est
	cae_co	Código de Actividade Económica of the company.	caem1l
	cae_est	Código de Actividade Económica of the establishment.	caest1l
	n_est	Number of establishments.	nest
	n_emp_co	Number of company workers (Independent + TCO).	pemp
	n_emp_est	Number of establishment workers (Independent + TCO).	pest
	n_tco_co	Number of company TCO employees.	tcoemp
	n_tco_est	Number of establishment TCO employees.	tcoest

See Annex II for recoding stata code.

5.2 Reconciling definitions

5.2.1 Population

5.2.1.1 EWCS

The target population for each country in the 6th wave of the EWCS were individuals aged 15 and over at the time of the survey living in private households and in employment. The Eurofound aimed to achieve a target sample size of 1000 respondents per country. For Portugal, a sample size of 1 037 respondents were achieved.

Since the survey was not extended to the entire population, a series of weighting steps was taken to ensure the results were representative at a variety of levels. The weighting steps were as follows:

Step 1: adjusting samples with design weights in a way that properly reflects probabilities of selection.

Step 2: adjusting for differences between the sample and population distributions on variables that are related to key outcomes (via post-stratification weighting).

Step 3: given the cross-national focus of the ECWS, the last step of the weighting adjustment consists of developing cross-national or population-size weights for each country covered.

Since a cross-national comparison is not the focus of the analysis, steps 1 and 2 are of particular importance. As such, we will be using a combination of weights defined in Step 1 and Step 2 - for more information on the way these weights are designed please refer to Eurofound (2015). The post-stratification weighting was done to ensure that the sample accurately reflects the socio-demographic structure of the target population across for weighting variables:

1. Age by gender using four age bands (15-24; 25-39; 40-59 and 60+).
2. NUTS II region. The LFS 2014 was also used in this case.
3. Industry, using NACE Sector as a proxy.
4. Occupation, using an 8-category approach based on ISCO at the 1-digit level.

All of the information on the variables used for the post-stratification weights was drawn from the Labour Force Survey (LFS) 2014. Therefore, the distributions in the EWCS at each of the levels presented above, should closely follow the same distributions of the LFS 2014. As suggested in the EWCS documentation, in order to achieve a total population size we will be using variable the individual weight *w4* which includes design and strata information combined into one weight value for each observation. In order to get frequency weights for each observation, we use the following formula:

$$w_{freq} = \frac{w_4 \times N_{in-work\ pop}}{N_{obs}}$$

Where $N_{in-work\ pop}$ refers to the total population working as an employee self-employed at the time the survey was conducted (3 710,6 thousand and 815,1 thousand respectively in Portugal – *INE*) and N_{obs} refers to the total number of observations in our survey – 1 037. Since the purpose of this exercise is to match the EWCS with the QPS, we have removed all observations for individuals that were not working as employees, i.e. those working as self-employed, since these are not considered in the latter.

In order to achieve a higher number of observations, a process of expansion was used to replicate each observation the same number of times as their frequency weight. Since we are unable to create a proportion of an observation, such as 0,8 for example, all weights were rounded to their closest integral value.

5.2.1.2 QPS

The QPS is a compulsory survey made to all registered private employers and public organizations with employment contracts under private law in Portugal. The information is filled in by the companies themselves in the declaration of IRS and is compulsory in nature. As previously mentioned, the survey contains information regarding all employees that work for the company. Contrary to the EWCS, the QPS does not contain information regarding independent workers. Additionally, since it is a compulsory survey, it contains information on the whole universe of workers stratified at different levels:

1. Age.
2. Gender.
3. NUTS II.
4. Industry, using CAE as proxy.
5. Profession, using CPP as proxy.
6. Education level.
7. Qualification level.
8. Earnings.
9. Hours worked.

5.2.1.3 Harmonization of population

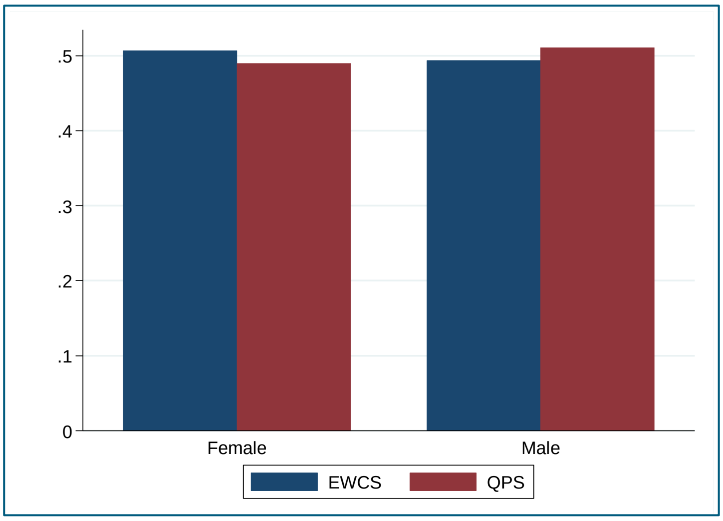
Since the population definitions are not necessarily a match, some steps were taken to harmonize both datasets with respect to the population. All of the procedures were done to the EWCS survey in order to preserve the more detailed information present in the QPS. The steps taken were as follows:

- Removal of all observations referring to self-employed workers.
- Removal of all observations referring to public-sector workers.
- Removal of all observations referring to workers situated in the Azores and Madeira regions.
- Removal of all observations referring to workers without a contract of unlimited duration (permanent contract).

With these four steps we achieve a similar population definition for both the EWCS and QPS datasets, which will improve the likelihood of a successful fusion.

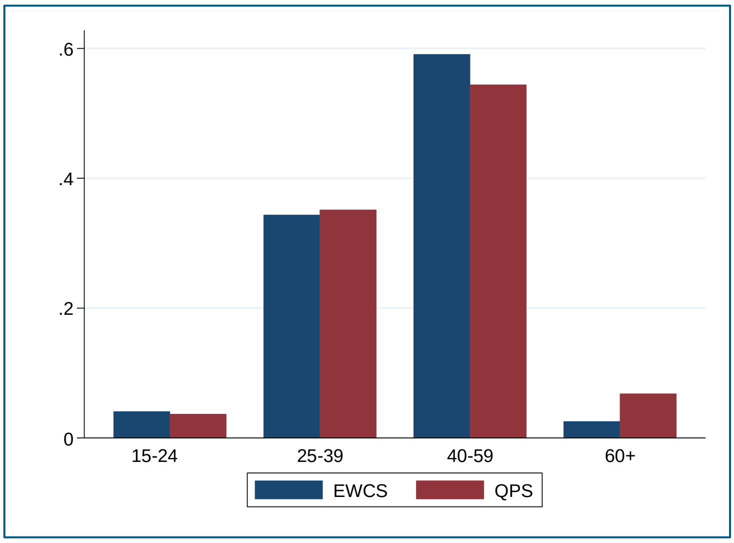
The population achieved after the procedures described above is 1 709 thousand workers in the EWCS and 1 798 thousand workers in the QPS. Since we have a considerable difference in the total number of observations (N) in each dataset, we use categorical distributions to compare the common variables in the EWCS and QPS. This allows us to compare the probability distributions for the occurrence of each category k for a set of carefully selected common variables.

Figure 2. Gender distribution: EWCS vs QPS



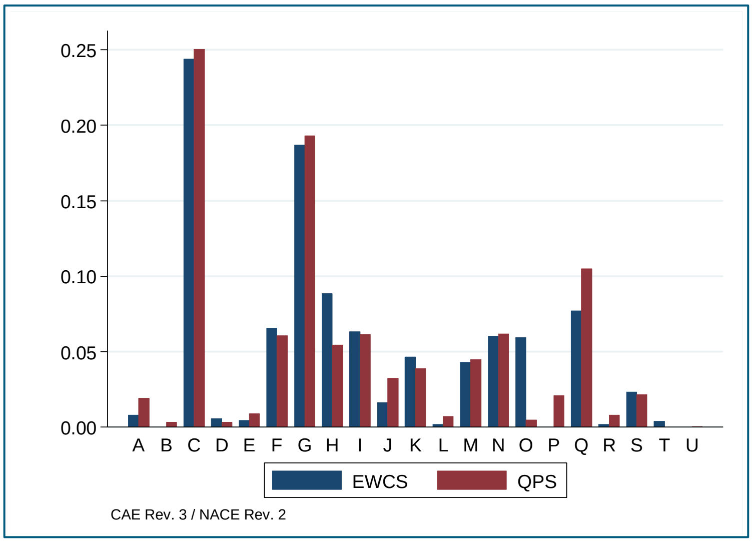
Regarding gender distribution, Figure 2 shows that there is a slight difference in the probability distribution between the QPS and the EWCS, where the QPS has a 0,49 probability of drawing a Female and a 0,51 probability of drawing a Male and the EWCS has 0,51 and 0,49 for Males and Females respectively.

Figure 3. Age groups distribution: EWCS vs QPS



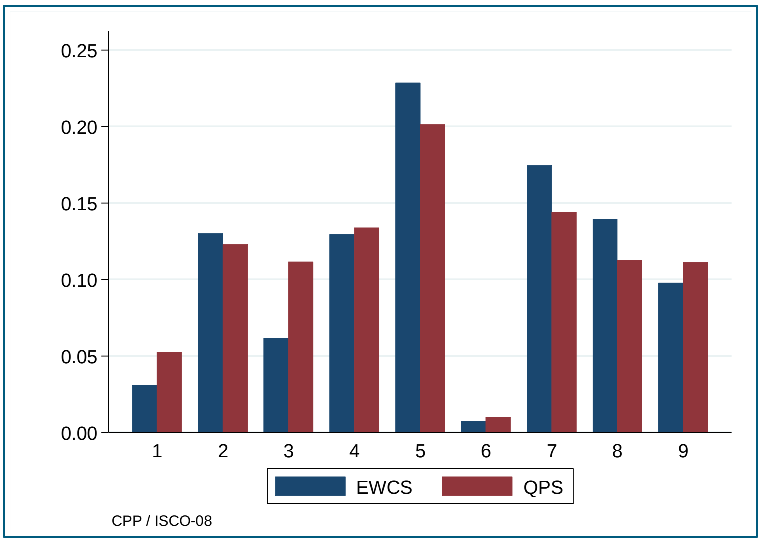
The second variable chosen is age-groups, where we consider 4 age bands (15-24; 25-39; 40-59 and 60+) similarly to what is done in the construction of the weights in the EWCS. As shown in Figure 3, there is a marginal difference between datasets, with a more accentuated difference in the older age groups (40-59 and 60+).

Figure 4. Economic activity distribution: EWCS (NACE Rev. 2) vs QPS (CAE Rev. 3)



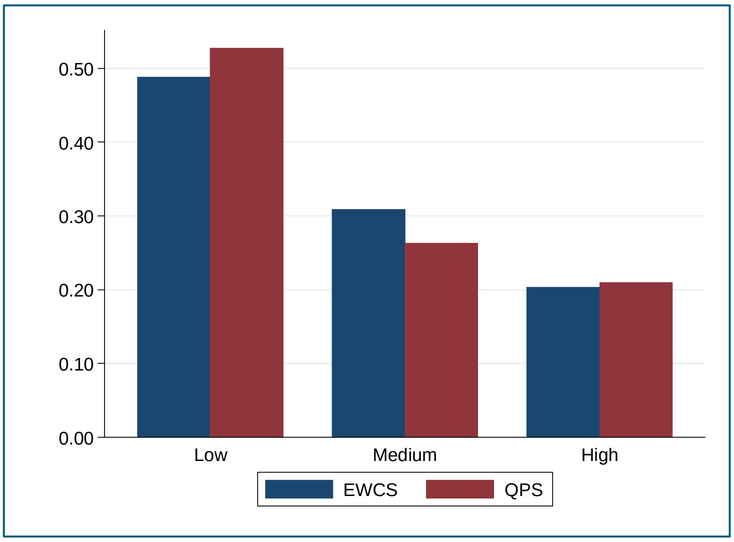
The third variable chosen was economic activity (Figure 4), which is represented by the NACE Rev 2 in the EWCS and by the CAE Rev. 3 in the QPS. The correspondence between these two classifications of the type of economic activity is direct and can be found in more detail in Appendix I. After careful analysis, we have concluded that there are some differences that need to be taken into account, especially noticeable in the H (Transportation and storage), O (Public administration and defense) and Q (Human health and social work activities) sectors. Further analysis is required to ensure compatibility between datasets at this level.

Figure 5. Occupation distribution: EWCS (ISCO-08) vs QPS (CPP)



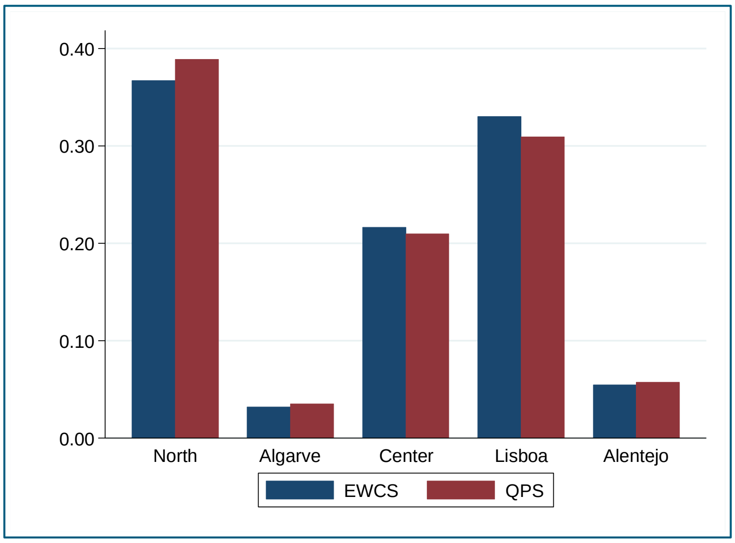
The fourth variable considered is the occupation type, which is represented by the ISCO-08 in the EWCS and by the CPP in the QPS. The correspondence between the ISCO-08 and the CPP is direct and can be found in Appendix II. As shown in Figure 5, there are some notable differences in the probability distributions across occupational groups that should be taken into consideration, more specifically, particular attention should be paid to category 3 (Technicians and associate professionals).

Figure 6. Education level distribution: EWCS vs QPS



The fifth variable taken into account is education level recoded into Low, Medium and High categories (see Table 12 and Table 13 for information on the recoding process). As shown in Figure 6, there are some noticeable differences in the Low and Medium education distributions, where the EWCS overestimates the presence of employees with a permanent contract and medium education, while underestimating those with Low education.

Figure 7. NUTS II distribution: EWCS vs QPS



Finally, regarding geographical distribution, represented in both datasets by the NUTS II variable, the probability distributions are very close, almost identical.

Although evaluating the differences between probabilistic distributions is an easy way to assess whether there are discrepancies across different categories between both datasets, aside from calculating absolute differences, it becomes hard to quantify the degree of similarity between variables. For this reason, we provide a range of measures that evaluate the similarity between these variables across datasets. The measures considered in this exercise are the dissimilarity index, overlap index, Bhattacharyya coefficient, Hellinger's distance and Pearson's Chi-Square.

Dissimilarity index*:* The dissimilarity index is a commonly used indicator that quantifies the degree of segregation between two populations and it is often used in demographic and population studies to quantify racial segregation in metropolitan areas (Lee, Minton, & Pryce, 2015). The dissimilarity index is between two categorical variables *A* and *B* with j categories is given by:

$$D = \frac{1}{2}\sum_{j=1}^{n}\left|P_{Aj} - P_{Bj}\right|$$

Where P_{Aj} and P_{Bj} represent the relative frequencies for category *j* in datasets *A* and *B* respectively. The dissimilarity index ranges between 0 and 1, with zero meaning minimum dissimilarity.

Overlap index: The overlap index is, essentially, the opposite of the dissimilarity index in the sense that it quantifies the degree to which two populations overlap. The overlap index between two categorical variables *A* and *B* with *j* categories is given by:

$$O = \sum_{j=1}^{n} \min\,(P_{Aj}, P_{Bj})$$

Where P_{Aj} and P_{Bj} represent the relative frequencies for category *j* in datasets *A* and *B* respectively. It is noteworthy that the overlap is the opposite of the dissimilarity index and can also be calculated as:

$$O = 1 - D$$

Similarly to the dissimilarity index, the overlap index ranges between 0 and 1, however, the value 1 represents maximum overlap.

Bhattacharyya coefficient: The Bhattacharyya coefficient is a measurement of the degree of similarity between two probabilistic distributions. The Bhattacharyya coefficient of two categorical variables *A* and *B* with *j* categories with a discrete and continuous probability distribution is given by:

$$B = \sum_{j}^{n} \sqrt{P_{Aj}\,P_{Bj}}$$

Essentially, the Bhattacharyya coefficient represents the overlap between the probabilistic distributions between two categorical variables and ranges between 0 and 1, where 1 indicates identical distributions.

Hellinger's distance: The Hellinger´s distance is closely related to the Bhattacharyya coefficient, since it uses probabilistic distributions to quantify distances between distributions through a contingency table. The Hellinger's distance for two categorical variables *A* and *B* with a discrete and continuous probability distribution is given by:

$$HD = \sqrt{\frac{1}{2}\sum_{j=1}^{K}\left(\sqrt{P_{Aj}} - \sqrt{P_{Bj}}\right)^2}$$

The measures identified up to this point will also be used to validate, to a certain extent, the statistical matching procedure by identifying how close the imputed distributions are from the original distributions. Table 14 presents the results for each of the measures presented, as well as the absolute difference in percentage points for the comparison between distributions of key matching variables between the EWCS and the QPS.

Table 14. Measures of similarity for common variables between the EWCS and the QPS (continues)

Variable	Absolute difference (p.p.)	Similarity Measures			
		Dissimilarity Index	Overlap	Bhattacharyya coef.	Hellinger dist.
Gender	.	0,02	0,98	1.00	0,01
Male	0,02	.	.	.	.
Female	0,02	.	.	.	.
Age group	.	0,05	0,95	0,99	0,08
15-24	0,00	.	.	.	.
25-39	0,01	.	.	.	.
40-59	0,05	.	.	.	.
60+	0,04	.	.	.	.
Sector of economic activity	.	0,10	0,90	0,98	0,14
A	0,01	.	.	.	.
B	0,00	.	.	.	.
C	0,01	.	.	.	.
D	0,00	.	.	.	.
E	0,00	.	.	.	.
F	0,01	.	.	.	.
G	0,01	.	.	.	.
H	0,03	.	.	.	.
I	0,00	.	.	.	.
J	0,02	.	.	.	.
K	0,01	.	.	.	.
L	0,01	.	.	.	.
M	0,00	.	.	.	.
N	0,00	.	.	.	.
O	0,05	.	.	.	.
P	0,02	.	.	.	.
Q	0,03	.	.	.	.

Table 14. Measures of similarity for common variables between the EWCS and the QPS (conclusion)

Variable	Absolute difference (p.p.)	Similarity Measures			
		Dissimilarity Index	Overlap	Bhattacharyya coef.	Hellinger dist.
R	0,01	.	.	.	.
S	0,00	.	.	.	.
T	0,00	.	.	.	.
U	0,00	.	.	.	.
Occupation	.	0,09	0,91	0,99	0,09
1	0,02	.	.	.	.
2	0,01	.	.	.	.
3	0,05	.	.	.	.
4	0,00	.	.	.	.
5	0,03	.	.	.	.
6	0,00	.	.	.	.
7	0,03	.	.	.	.
8	0,03	.	.	.	.
9	0,01	.	.	.	.
Education level	.	0,04	0,96	0,99	0,03
Low	0,04	.	.	.	.
Medium	0,05	.	.	.	.
High	0,01	.	.	.	.
NUTS II	.	0,03	0,97	1.00	0,02
North	0,02	.	.	.	.
Algarve	0,00	.	.	.	.
Center	0,01	.	.	.	.
Lisbon	0,02	.	.	.	.
Alentejo	0,00	.	.	.	.

As shown, the gender distributions are particularly close between both datasets. In fact, they are so similar that the Bhattacharyya coefficient gives a value of 1 when rounded to the closest integer by two decimals. The values for the dissimilarity index and the Hellinger's distance are also considerably small, which clearly demonstrates that both distributions are very close.

Regarding age-groups, there is a greater discrepancy between the distributions in each dataset. Although the Bhattacharyya coefficient remains particularly high, the Hellinger's distance and the dissimilarity index point to a less similar distribution. It is important to note that, in this case, the Bhattacharyya coefficient seems to be less sensitive to differences in the distribution of categorical variables than other measures. This can be somewhat connected to the fact that the Bhattacharyya coefficient uses probabilistic distributions to compute differences rather than frequencies or relative frequencies. Still, the values of each of the measures remains within the generally acceptable range.

The differences become more noteworthy when we analyse sector of economic activity. In fact, the differences are well displayed in all results except for the Bhattacharyya coefficient, which seems to maintain its lack of sensitivity to distributional differences, even when these are more evident. In this case, the dissimilarity index has risen to 0,1, while the Hellinger's distance is now at 0,14. These do not demonstrate an agreement between distributions for the same variable across datasets. This is probably due to the fact that there are a multitude of categories in this variable – 21 to be precise – which in practice should make differences more relevant as each category will be composed of a lower number of individuals. Therefore, in proportional terms, a difference of 1 individual will produce a higher impact than if the variable was simply composed by two categories for example.

Regarding occupation, there is an improvement in relation to the previous variable. Although the results are not as good as in gender and age-groups. Like in previous cases, the Bhattacharyya coefficient remains abnormally high. The dissimilarity index and the Hellinger's distance are both at 0,09. While this value is not ideal, it is not so high that would make us consider discarding this dimension from our matching variables.

The differences between probabilistic distribution of education level minimal. Essentially, the results show that the EWCS slightly overestimates the number of employees with a permanent contract that have a medium education level, at the detriment of those that have low education. The results in terms of distances are favourable and we conclude that the populations are very similar in terms of education-level distribution.

Finally, in terms of NUTS II distribution, the results are as good as those found for gender. The dissimilarity index is 0,03 and the Hellinger's distance is 0,02. Similarly to the gender comparison, the Bhattacharyya coefficient is 1, which would mean that the distributions are practically identical.

Through this analysis we conclude that while there are some notable differences, both datasets belong to the same, or a relatively similar, population and argue that accounting for the harmonization procedures enacted thus far, we are now in a position where the datasets can be integrated by means of statistical matching procedures. However, it is important to note that our results should be limited especially limited when considering the integration of variables across sector of economic activity.

5.3 Harmonization of common variables

After careful analysis of both datasets, we have concluded that there is a need for further harmonization between variables, more specifically in the way the categories are defined. The comparison of categories between variables in both datasets, as well as the harmonization procedures that were applied can be found in Table 15.

Table 15. Harmonization of categories for common variables between the EWCS and the QPS (continues)

Dimension	QPS	EWCS	Harmonization process
Age	Continuous variable ranging from 17-68 with open groups for <=17 and >=68 respectively.	Continuous variable with values 15+.	Recoded EWCS values for age to create open groups such that: - If age<=17 all values were replace by 17. - If age >=68 all values were replaced by 68.
Education (ISCED)	Categorical variable with the following categories: 0. Technical/professional studies. 1. 1 Primary education not complete. 2. Primary education. 3. Secondary education. 4. Post-Secondary education. 5. Bachelor's degree. 6. Undergraduate degree. 7. Master's degree. 8. Doctorate degree.	Categorical variable with the following categories: 1. Early childhood education. 2. Primary education. 3. Lower secondary education. 4. Upper secondary education. 5. Post-secondary non-tertiary education. 6. Bachelor or equivalent. 7. Master or equivalent. 8. Doctorate or equivalent.	Both variables contain a similar classification. As such no harmonization process was necessary, since both variables are recoded into a three-category variable (low, medium, high) that serves as proxy for education. For the purpose of harmonization, we assume the following correspondence: 0-NA; 1-1; 2-2; 3-(3,4); 4-5; (5,6)-6; 7-7; 8-8.
Education level	Categorical variable with the following categories: 1. Low education (educ1 0-2). 2. Medium education (educ1 3-4). 3. High education (educ1 5-8).	Categorical variable with the following categories: 1. Low education (educ 1-2). 2. Medium education (educ 3-5). 3. High education (educ > 5).	The education_level variable reflects the correspondence assumed in the educ1 variable. We assume the following: - **Low** education refers to those who have completed up to primary education; - **Medium** education refers to those that have completed up to post-secondary non-tertiary education.

Table 15. Harmonization of categories for common variables between the EWCS and the QPS (conclusion)

Dimension	QPS	EWCS	Harmonization process
			High education refers to those that have completed an undergraduate degree or above.
Occupation	CPP10	ISCO-08	These variables have a direct match.
Earnings	Paid GROSS earnings for the month of October including subsidies and supplements. Continuous variable.	Continuous variable for NET earnings per month.	Since the EWCS only contains NET earnings, and the variable has a high degree of non-response, earnings was not considered as a possible matching variable. A factor that contributed to this decision is the lack of information necessary to calculate NET from GROSS and vice-versa.
Sector of economic activity	CAE Rev. 3	NACE Rev. 2	These variables have a direct match.

5.4 Definition of the statistical matching problem of the integration of the EWCS and the QPS

The choice of matching variables is a crucial step in ensuring the successful match between two or more datasets. Some authors argue that the choice of matching variables is, in fact, the most important step in the process of statistical matching in order to ensure the validity of results, surpassing even the matching technique (Leulescu & Agafitei, 2013). In practice, all shared common variables may be used in the matching process, however, this may have a detrimental effect on the match since it may undermine the predictive power of the model employed. This is particularly relevant when using parametric models. Rather, it is necessary to carefully select the variables that are connected at the same time with *Y* and *Z*.

At this point it becomes important to define our matching problem statistically. Let *A* represent a subset of the QPS composed of a group of variables X_A and Y_A, such that X_A is composed by a group of common variables between the QPS and the EWCS and Y_A is composed by a single continuous variable that captures gross earnings:

Table 16. Variables $\mathrm{X_A}$

Variables X_A	Description
Gender	Gender of respondents: 1. Male 2. Female
Age	Continuous variable for age.
Age squared	Continuous variable of the square of age
Age group	Age group of respondents: 1. 15-24 2. 25-39 3. 40-59 4. 60+
Education level	Level of education by categories: 1. Low 2. Medium 3. High
CAE Rev. 3	Sector of economic activity (see Appendix I for specific categories)
CPP 10	Occupation (see Appendix II for specific categories)

Table 17 Variables $\mathrm{Y_A}$

Variables Y_A	Description
Earnings	Paid GROSS earnings for the month of October including subsidies and supplements. Continuous variable.

Let B represent a subset of the EWCS composed of a group of variables X_B and Z_B, such that X_B is composed by a group of common variables between the QPS and the EWCS and Z_B is composed by four categorical variables that belong to the EWCS and attempt to capture the degree to which occupations involve the following risks/activities:

Table 18. Variables $\mathrm{X_B}$

Variables X_B	Description
Gender	Gender of respondents: 1. Male 2. Female
Age group	Age group of respondents: 1. 15-24 2. 25-39 3. 40-59 4. 60+
Education level	Level of education by categories: 1. Low 2. Medium 3. High
NACE Rev. 2	Sector of economic activity (see Appendix I for specific categories)
ISCO-08	Occupation (see Appendix II for specific categories)

Table 19. Variables $\mathrm{Z_B}$

Variables Z_B	Description
Q30a	Tiring or painful positions
Q30b	Lifting or moving people
Q30c	Carrying or moving heavy loads
Q30i	Use of Information and Communication Technologies (ICTs)

Each of the categorical variables in $\mathrm{Z_B}$ has the same categories, which are as follows: (1) “All the time”; (2) “Almost all the time”; (3) “Around 3/4 of the time”; (4) “Around 1/2 of the time”; (5) “Around 1/4 of the time”; (6) “Almost never”; and (7) “Never”. Having defined the variable groups (*X, Y, Z*), we are able to write our statistical matching problem as follows:

Table 20. Illustration of the statistical matching problem

	$earr$	…	sex	age	age^2	age_g	$educ$	eco	occ	…	$pain$	$lift$	$load$	ict
A	$earn$	…	sex_1^A	age_1^A	age_1^{2A}	$\mathrm{age\_g}_1^{2A}$	educ_1^A	eco_1^A	occ_1^A					
	$earn$	…	sex_a^A	age_a^A	age_a^{2A}	$\mathrm{age\_g}_a^{2A}$	$educ_a^A$	eco_a^A	occ_a^A					
	$earn$	…	$sex_{n_A}^A$	$\mathrm{age}_{n_A}^A$	$\mathrm{age}_{n_A}^{2A}$	$\mathrm{age\_g}_{n_A}^{2A}$	$\mathrm{educ}_{n_A}^A$	$eco_{n_A}^A$	$\mathrm{occ}_{n_A}^A$					
B			sex_1^B	age_1^B	age_1^{2B}	$\mathrm{age\_g}_1^{2B}$	educ_1^{2B}	eco_1^B	occ_1^B	…	$pain_1^B$	$lift_1^B$	$load_1^B$	ict_1^B
			sex_b^B	age_b^B	age_b^{2B}	$\mathrm{age\_g}_b^{2B}$	educ_b^{2B}	eco_b^B	occ_b^B	…	pain_b^B	$lift_b^B$	$load_b^B$	ict_b^B
			$sex_{n_B}^B$	$\mathrm{age}_{n_B}^B$	$\mathrm{age}_{n_B}^{2B}$	$\mathrm{age\_g}_{n_B}^{2B}$	$\mathrm{educ}_{n_B}^{2B}$	$\mathrm{eco}_{n_B}^B$	$\mathrm{occ}_{n_B}^B$	…	$\mathrm{pain}_{n_B}^B$	$lift_{n_B}^B$	$load_{n_B}^B$	$ict_{n_B}^B$
$\boldsymbol{A} \cup \boldsymbol{B}$	$earn$	…	$sex_1^{A\cup}$	$\mathrm{age}_1^{A\cup B}$	$\mathrm{age}_1^{2A\cup B}$	$\mathrm{age\_g}_1^{2A\cup B}$	$\mathrm{educ}_1^{2A\cup B}$	$\mathrm{eco}_1^{A\cup B}$	$\mathrm{occ}_1^{A\cup B}$	…	$pain_1^B$	$lift_1^B$	$load_1^B$	ict_1^B
	$earn$	…	$sex_b^{A\cup}$	$\mathrm{age}_b^{A\cup B}$	$\mathrm{age}_b^{2A\cup B}$	$\mathrm{age\_g}_b^{2A\cup B}$	$\mathrm{educ}_b^{2A\cup B}$	$\mathrm{eco}_b^{A\cup B}$	$\mathrm{occ}_b^{A\cup B}$	…	pain_b^B	$lift_b^B$	$load_b^B$	ict_b^B
	$earn$	…	$sex_{n_B}^{A\cup}$	$\mathrm{age}_{n_B}^{A\cup B}$	$\mathrm{age}_{n_B}^{2A\cup B}$	$\mathrm{age\_g}_{n_B}^{2A\cup B}$	$\mathrm{educ}_{n_B}^{2A\cup B}$	$\mathrm{eco}_{n_B}^{A\cup B}$	$\mathrm{occ}_{n_B}^{A\cup B}$	…	$\mathrm{pain}_{n_B}^B$	$lift_{n_B}^B$	$load_{n_B}^B$	$ict_{n_B}^B$

It is important to note that this is an ideal representation of our problem, i.e. a representation where all the main common variables are used in order to fuse both datasets. However, this may not be the case. Optimally, the common variables should contain all of the information necessary to explain the association shared between *Z* and *Y*. From this perspective, the inclusion of all common variables from *A* and *B* that possess some degree of explanatory power seems to be a reasonable decision. However, it is important to note that each additional variable significantly complicates the computational procedure and can have a negative impact on the quality of the results. Therefore, parsimony is recommended in the selection of matching variables. This explains the initial screening of matching variables done up to this point. The variables were chosen specifically due to their importance in the EWCS, in terms of the representativeness of the survey.

Following the methodology for the identification of matching variables found in D'Orazio (2014) we perform two types of tests according to the type of response variable. When the response variable is continuous, we look at its correlation with the predictors using Spearman's rank correlation coefficient, which also allows us to identify nonlinear relationships. On the other hand, when the response variable is categorical and all the predictors are also categorical, D'Orazio (2014) alerts to the need for using Chi-Squared based association measures.

5.4.1 Predictors of earnings in the QPS

Following the methodology described above, we obtained the Spearman's rank correlation coefficient between earnings in the QPS and each of the predictors previously described. The results are shown in Table 21.

Table 21. Spearman rank correlation for response variable earny

Predictors	Rsquared	F-Statistic	DF	DF2	p	Rsquared-Adjusted	N
age	0,001	1 082,73	1	1 792 322	0	0,001	1 792 324
age2	0,001	1 082,73	1	1 792 322	0	0,001	1 792 324
educ	0,199	445 336,8	1	1 792 322	0	0,199	1 792 324
age_g	0,002	3 024,66	1	1 792 322	0	0,002	1 792 324
gender	0,041	76 177,14	1	1 792 322	0	0,041	1 792 324
eco	0,002	4 008,99	1	1 792 322	0	0,002	1 792 324
occ	0,228	530 337,8	1	1 792 322	0	0,228	1 792 324

A quick analysis of the adjusted RSquared values indicates that the best predictors of earnings in our data are, in fact, education level (*educ*) and occupation (*occ*).

5.4.2 Predictors of categorical variables in the EWCS

Following the methodology described above we decided to use the Cramer's V as a measure of association between two categorical variables. This measure is based on the Pearson's Chi-Squared statistic and provides a value 0 and 1, where 1 means that the predictor perfectly explains the response variable. The results of the analysis can be found in Table 22.

Table 22. Cramer's V analysis of categorical variables and predictors in the EWCS

	age_g	sex	eco	occ	educ
pain	0,1463	0,222049	**0,26941**	0,211311	0,2463337
lift	0,222225	0,240306	**0,345798**	0,20918	0,1272191
load	0,205047	**0,263403**	**0,26048**	**0,280245**	**0,2723909**
ict	0,178201	0,124963	**0,308234**	**0,353423**	**0,4381348**

A quick analysis of the results show that all variables are significant in the prediction of the categorical variables in the EWCS. Although it is important to note that the variable age-groups (*age_g*) does not behave as expected and may not be as suitable as the remaining identified variables. Still, we will be including this variable in the matching process and test whether it affects the procedure in a positive or negative way. From our analysis, we identify the following as matching variables for the integration of the EWCS and the QPS data:

Table 23. Selected matching variables

Variable	Code
Age group	age_g
Gender	sex
Sector of economic activity	eco
Occupation	occ
Education level	educ

6. Application of parametric, non-parametric and mixed methods for the statistical matching of the EWCS and the QPS

As mentioned throughout this working paper, the main purpose of this analysis is to figure out the feasibility of integrating the EWCS and the QPS and to identify a suitable procedure for it among. So far, we have concluded that the with harmonization procedures and some data cleaning, it is possible to integrate these datasets and identified a set of common variables that can be used as matching variables in this process.

Additionally, we have identified three groups of statistical procedures for matching these datasets: parametric, non-parametric and mixed models – see Section 4 for a description of the methods identified. This section will analyse the results of the statistical matching of these datasets under the umbrella of each of these procedures.

6.1 Parametric techniques

6.1.1 Logistic regression

The application of parametric techniques described in section 4, focusses mainly in ways of modelling continuous variables. As previously mentioned, our response variables are constituted of ict, pain, load and lift, which are categorical variables composed of 7 different categories. While linear regression models may be suited to predict the value of a numeric variable based on its relationship to one or more independent variables, it is not well suited for every type of problem. More specifically, linear regression is not well suited to solve problems where the response variable is categorical. In these cases, it is common to use a logistic regression model. While linear regression models seek to predict a numerical variable, logistic regression seeks to predict the probability of a categorical response variable.

Rather than modelling the response variable directly, logistic regression will model the probability of a particular response value, or category. Applying this to our problem we can model the probability of each outcome of our response variable. For the sake of simplicity in our tests, we have recoded all of our response variables into binary variables comprised by the following categories: 1 “Often” and 0 “Not often/Never”. Illustrating how this approach works, if we were to model use of information and communication technologies, represented by variable ict, using education level as a predictor, the model would be represented as:

$$\Pr(ict = 1 \mid edu)$$

Generalizing the equation in terms of X and Y would give us:

$$\Pr(Y = 1 \mid X)$$

Essentially, what this means is that we are predicting the probability of Y given X. Since we are modelling probabilities, we would expect the value to range between 0 and 1, such that a prediction of 0,8, for example, would be interpreted as an 80 percent likelihood of an event of occurring – in our case that event would be that the individual would use ICTs at work often. For this purpose, we could use a straight-line function, such as the one used in linear regression, to calculate these probabilities. This function would be defined as follows:

$$\Pr(Y = 1 \mid X) = \beta_0 + \beta_1 X$$

From a practical perspective, this model would provide us some values. However, the fitted line of a linear approach to a binary response variable would not be suited, as it would comprise several limitations.

Essentially, under this approach, there is the possibility of obtaining both negative probabilities and probabilities that exceed 1. Although these values could be transformed to accommodate our problem – negative values transformed to zero and values exceeding one transformed to one – this is simply not a good fit for our binary classification problem.

To overcome this challenge, it is necessary to use a non-linear function for the regression line. One of the functions that allows us to achieve this is a logistic function:

$$p(X) = \Pr(Y = 1 \mid X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

This function provides an output that will always be situated between zero and 1. Overall, the logistic function produces an S shaped curve, also known as the *Sigmoid Curve*, which does a better job at describing our data than a straight line.

One more important aspect of logistic regression that needs clarification before proceeding with our analysis refers to the way the model is interpreted regarding its coefficients. For instance, in linear regression, the coefficient β_1 captures the average expected increase in Y from a unit change in X. On the other hand, in logistic regression, the same coefficient captures the corresponding change in the log-odds of $\Pr(Y = 1 \mid X)$ as a result of a unit change in X.

6.1.1.1 Modelling approach

To model a logistic regression for our response variables we employed a stepwise procedure, where groups of dummy variables are introduced in a stepwise fashion. This process allows for the identification of variables that improve and reduce the suitability of our model. Although all categorical variables were transformed into dummies for the purpose of the logistic regression, they are introduced as groups of variables that are composed by all the categories that make the original categorical variable. For example, gender was transformed into two binary (0 “No” 1 “Yes”) variables: $isMale$ and $isFemale$. Our response variables were also recoded into binary variables. The relationship between the previous variable and the binary response variable can be found in Table 24.

Table 24. Recodification of response variable

Original variable	Binary response variable
1 – All the time	1 – Often
2 – Almost all the time	
3 – Around ¾ of the time	
4 – Around ½ of the time	
5 – Around ¼ of the time	0 – Not often/Never
6 – Almost never	
7 – Never	

For the creation of our models our initial dataset was split into two smaller datasets: training (75% of the data) and test (25% of the data). As the name of the datasets suggests, the training dataset is used to train our model, while the test dataset is used to test the accuracy and predictive power of our model on a dataset that was not included in the modelling process. At this point, we can check whether the distribution between the categories in our binary response variable is balanced.

Table 25. Distribution between classes of ICT usage: original

Dataset	1 – Often	0 – Not often/Never
EWCS	37,81	62,19
train_ewcs	40,99	59,01
test_ewcs	36,75	63,25

As clearly shown in Table 25, there is a clear imbalance between classes in our response variable. Although that may not be of much importance in our original and test datasets, it is an important factor in our training dataset since it will create a bias in our model towards the dominating class. In order to address this issue, we use a function that generates new artificial cases for the minority class based on a k-nearest neighbors approach (see SMOTE R function for more details). After we apply the procedure, we are left with a more balanced training dataset, as evidenced in Table 26.

Table 26. Distribution between classes of ICT usage: adjusted

Dataset	1 - Often	0 - Not often/Never
EWCS	37,81	62,19
train_ewcs	50,00	50,00
test_ewcs	36,75	63,25

6.1.1.2 Model selection

Based on the modelling approach, six models were developed. Each of the models is composed by a different group of predictors, that are introduced in a stepwise fashion from one model to the next. The models considered are as follows:

- Model 1: education
- Model 2: education + gender
- Model 3: education + gender + agegroups
- Model 4: education + gender + agegroups + nuts2
- Model 5: education + gender + agegroups + nuts2 + occupation
- Model 6: education + gender + agegroups + nuts2 + eco_sector
- Model 7: education + gender + agegroups + nuts2 + occupation + eco_sector

For model evaluation three criterion were analysed: *p-values*, AIC and predictive accuracy. In this respect, *p-values* are the same as the *p-values* we find in linear regression models. A *p-value* is the probability that an observed value could have occurred simply by chance. The lower the *p-value* the greater the statistical significance of the predictor. The AIC is the Akaike Information Criterion. It is a quantification of how well the model does in explaining the variability in our data. This measurement is often used when comparing tow models built from the same data. Usually, the model with the lower AIC is preferred. Finally, the predictive accuracy of our model can be calculated by predicting our test dataset. By making an actual prediction, we can quantify the results of our model in terms of what percentage of our test data it was able to successfully predict.

Table 27. Logistic regression model selection for ICT (continues)

		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
P Values	Intercept	0,00***	0,00***	0,00***	0,00***	0,9879	0,98704	0,9891
	eduLow	0,00***	0,00***	0,00***	0,00***	0,00***	0,00***	0,00***
	eduMed	0,00***	0,00***	0,00***	0,00***	0,00***	0,00***	0,00***
	age25-39		0,565	0,522398	0,437324	0,5390	0,74357	0,6636
	age40-59		0,155	0,128490	0,076364$^{.}$	0,9414	0,05797$^{.}$	0,4000
	age60+		0,934	0,961925	0,589621	0,2383	0,28312	0,4441
	genderMale			0,00***	0,00***	0,0436*	0,00156**	0,3834
	nuts2Alg				0,722665	0,4211	0,31983	0,3962
	nuts2Cen				0,746464	0,5702	0,17986	0,2154
	nuts2Lis				0,354214	0,9080	0,92914	0,3283
	nuts2Nor				0,700453	0,0949$^{.}$	0,10836$^{.}$	0,0899
	occ2					0,9862		0,9914
	occ3					0,9856		0,9909
	occ4					0,9849		0,9906
	occ5					0,9872		0,9925
	occ7					0,9887		0,9934
	occ8					0,9887		0,9929
	occ9					0,9887		0,9933
	ecoC						0,98628	0,9926
	ecoD						0,98132	0,9867
	ecoE						0,99963	0,9997
	ecoF						0,98545	0,9921
	ecoG						0,98467	0,9914
	ecoH						0,98551	0,9926
	ecoI						0,98664	0,9925
	ecoJ						0,98371	0,9920
	ecoK						0,98331	0,9919

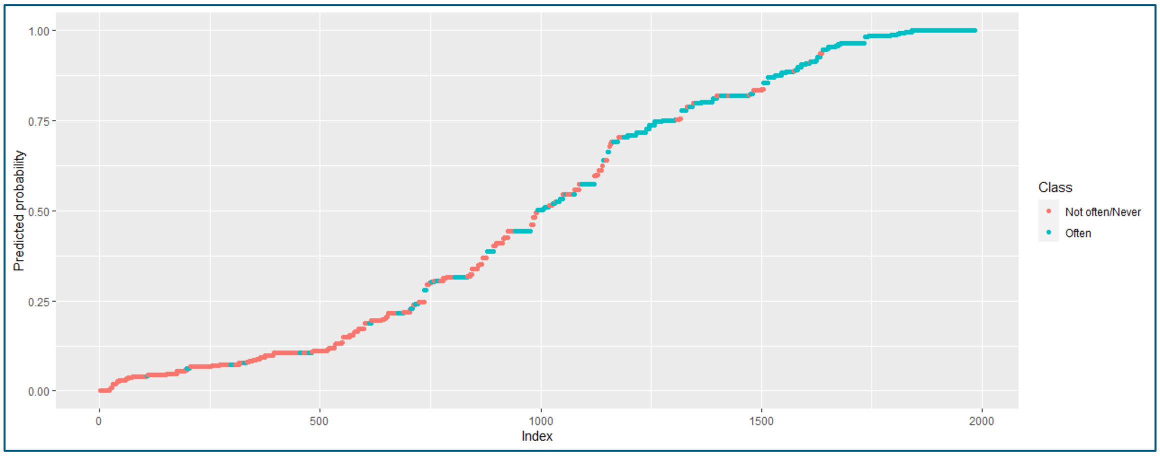
Table 27. Logistic regression model selection for ICT (conclusion)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
ecoM						0,98462	0,9841
ecoN						0,97282	0,9920
ecoQ						0,98509	0,9926
ecoR						0,98577	0,9932
ecoS						0,98541	0,9923
AIC	1790,8	1790,7	1778,9	1781,6	1272,2	1571,7	1159,6
Predictive accuracy	0,8059701	0,8059701	0,8152985	0,8320896	0,9067164	0,8488806	0,9011194

Sig. Values: 0.001***; 0,01**; 0.05*; 0.1.

Table 27 shows the result for each of the models. As shown, the model with the most predictive accuracy and the smallest AIC is Model 7, which includes all the predictors available. However, a quick analysis of the *p-values* for Model 7 shows there are variables with tremendously high values. Often, high *p-values* are related to multicollinearity between variables. Consequently, this model was tested for multicollinearity using the variance inflation factor, or VIF. The VIF is often used as a measure of multicollinearity for models with more than one predictor. As a rule of thumb, any predictor with a VIF value that exceeds 5 should be considered for removal. Both variables *occ* and *eco* have a VIF value of 12.29 and 10,38, respectively. Considering the removal of each variable, presented in Model 5 and Model 6, the choice was done for the model that presents the highest predictive accuracy, which is Model 5, with a predictive accuracy of 0,9067. Although the *p-values* for Model 5 remain abnormally high, both the AIC and the predictive accuracy surpass the results of every other model when multicollinearity is accounted for and resolved. As such, Model 5 was selected for the estimation of ICT usage in the QPS.

Figure 8. Indexed probabilities for ICT



For illustrative purposes we modelled the probabilities every individual in the test dataset and their actual values for the *ict* variable. Figure 8 shows the probabilities calculated for every individual indexed in an increasing order and their actual observed class. It appears that the model developed has done a good job predicting the usage of ICT's in the work context. Those that often use ICT's (turquoise) are predicted to have a high probability of using ICT's, while those that do not use ICT's often (orange) are predicted to have a low probability of using ICT's at work.

6.1.1.3 Parametric statistical matching results

To assess the validity of our results we have employed the same methodology used to compare the populations – see section 5.2.1.3. The results can be found in Table 28. The table provides absolute values, which are not comparable since the total size of each dataset is considerably different. However, to ensure comparability, probability distributions were included. These distributions are for the total cases of ICT and by control variables (age-groups, gender, occupation, sector of economic activity and NUTS II).

Table 28. Comparison between ICT usage imputed by logistic regression (QPS) and observed (EWCS) (continues)

	QPS				EWCS			
	Absolute		Probability		Absolute		Probability	
	Often	Not often	Often (prob)	Not often (prob)	Often	Not often	Often (prob)	Not often (prob)
total	667 578	1 060 039	0,39	0,61	814	1 328	0,38	0,62
agegroups								
15-24	15 651	48 989	0,01	0,03	24	55	0,01	0,03
25-39	262 581	347 881	0,15	0,20	368	386	0,17	0,18
40-59	363 110	575 115	0,21	0,33	409	848	0,19	0,40
60+	26 236	88 054	0,02	0,05	13	39	0,01	0,02
gender								
F	347 870	490 912	0,20	0,28	455	578	0,21	0,27
M	319 708	569 127	0,19	0,33	359	750	0,17	0,35
educ								
Low	118 044	793 494	0,07	0,46	86	961	0,04	0,45
Med	213 221	241 428	0,12	0,14	346	311	0,16	0,15
High	336 313	25 117	0,19	0,01	382	56	0,18	0,03

Table 28. Comparison between ICT usage imputed by logistic regression (QPS) and observed (EWCS) (continues)

	QPS				EWCS			
	Absolute		Probability		Absolute		Probability	
	Often	Not often	Often (prob)	Not often (prob)	Often	Not often	Often (prob)	Not often (prob)
occ								
Agriculture	200	17 613	0,00	0,01	0	17	0,00	0,01
Clerical	212 732	17 635	0,12	0,01	266	26	0,12	0,01
Craft	3 450	253 267	0,00	0,15	14	387	0,01	0,18
Elementary	5 074	187 547	0,00	0,11	7	188	0,00	0,09
Managers	90 193	0	0,05	0,00	66	0	0,03	0,00
Operator	1 216	196 540	0,00	0,11	14	306	0,01	0,14
Professionals	183 772	17 187	0,11	0,01	224	40	0,10	0,02
Service	19 002	327 864	0,01	0,19	119	331	0,06	0,15
Technicians	151 939	42 386	0,09	0,02	104	33	0,05	0,02
eco								
A	1 886	32 488	0,00	0,02	0	19	0,00	0,01
C	90 584	358 694	0,05	0,21	89	470	0,04	0,23
D	6 064	0	0,00	0,00	13	0	0,01	0,00
E	459	15 470	0,00	0,01	0	11	0,00	0,01
F	34 464	73 921	0,02	0,04	55	96	0,03	0,05
G	142 393	203 967	0,08	0,12	197	231	0,09	0,11
H	33 251	64 261	0,02	0,04	61	143	0,03	0,07
I	16 358	93 822	0,01	0,05	21	120	0,01	0,06
J	53 420	4 808	0,03	0,00	32	6	0,02	0,00
K	67 089	2 524	0,04	0,00	101	4	0,05	0,00
M	79 849	28	0,05	0,00	97	0	0,05	0,00
N	28 284	82 409	0,02	0,05	44	95	0,02	0,05
Q	82 417	105 808	0,05	0,06	73	102	0,05	0,00
R	14 061	136	0,01	0,00	4	0	0,02	0,05
S	16 999	21 703	0,01	0,01	23	31	0,04	0,05

Table 28. Comparison between ICT usage imputed by logistic regression (QPS) and observed (EWCS) (conclusion)

	QPS				EWCS			
	Absolute		Probability		Absolute		Probability	
	Often	Not often	Often (prob)	Not often (prob)	Often	Not often	Often (prob)	Not often (prob)
nuts2								
Alentejo	27 990	70 646	0,02	0,04	28	91	0,01	0,04
Algarve	17 259	43 454	0,01	0,03	27	43	0,01	0,02
Center	118 280	243 997	0,07	0,14	149	312	0,07	0,15
Lisbon	275 742	258 068	0,16	0,15	287	416	013	0,19
North	228 307	443 874	0,13	0,26	323	466	0,15	0,22

Looking at the total distribution, the results are very similar. Those classified as Often in the QPS have a probability of 0,32 compared with a probability of 0,38 in the EWCS. Regarding the control variables, the biggest discrepancies are found in the variables that presented the lowest ranks in the analysis of the degree of similarity between populations, i.e. the EWCS and the QPS populations, which were occupation and sector of economic activity. To ensure that our results are favourable, we have also analysed the common measures of similarity between populations.

Table 29. Measures of similarity between ICT + predictor between the imputed by logistic regression (QPS) and observed (EWCS) distributions

Variable	Dissimilarity index	Overlap	Bhattacharyya coeff,	Hellinger dist,
agegroups	0,08	0,92	0,99	0,09
gender	0,03	0,97	1,00	0,02
edu	0,06	0,94	1,00	0,06
occ	0,13	0,87	0,98	0,14
eco	0,09	0,91	0,99	0,10
nuts2	0,07	0,93	1,00	0,06

The results are shown in Table 29. A brief analysis shows that the highest degree of dissimilarity is found when we compare the populations by ICT and type of occupation between observed and imputed distributions, with a 0,13 dissimilarity index and a Hellinger's distance of 0,14. The best performing variable groups in terms of similarity are gender and education, with a dissimilarity of 0,3 and 0,6 respectively.

Finally, when considering ICT by occupation, sector of economic activity or NUTS II, the dissimilarity index is slightly higher, bit always under the 0,1 mark. Although not ideal, the results for these final three control variables still represent a high degree of similarity.

Overall, the results of this statistical matching are very positive. Although, it is important to note that while the populations are similar regarding ICT usage, we were force to transform it into a binary variable. This entails a loss of variable specificity that was associated with categorical variable with more than two dimensions. Although we were successful in the statistical matching procedure, we lost a high degree of information along the way due to the transformation of a variable with 7 categories into a binary YES/NO variable. To solve for this problem, we could consider the use of multinomial logistic regression. However, a quick look at our data revealed that we do not have enough observations to perform this estimate. By using a crosstabulation between all seven categories of the original ICT variable and each of the predictors, we find several empty cells, which would severely affect the validity of our model.

6.2 Non-parametric techniques

6.2.1 Random hot deck

6.2.1.1 Modelling approach

The first non-parametric technique considered is Random hot deck. This procedure is described in section 4.3.1. An attempt at modelling this procedure was carried out using the R package StatMatch. However, due to our large sample sizes and the coding structure of the StatMatch package, the computational requirements for this procedure were too extensive. Consequently, we wrote a user-defined function that was able to perform the match accordingly and was more efficient regarding its computational requirements. Essentially, this new function performs a random match between two observations in two different datasets according to a group of common variables.

Initially, all common variables and statistically significant variables were considered. However, it is important to note that in this procedure we are using the ICT variable with all its categories. A crosstabulation between the ICT variable and each of the common variables shows that there exist several empty cells. This makes it extremely difficult to match the cases. As such, we used only the common variables that did not display any empty cells, which are gender and NUTS II. This has some considerable consequences for the validity of our results. However, we decided to proceed with the match and analyse the results for comparability purposes.

In order to reduce the computational effort of the matching procedure, the QPS dataset is split into 17 equally sized datasets according to the observation number – the first 100 thousand are selected in the first dataset, the second 100 thousand are selected in the second dataset and so forth. After the split is completed, each individual in all the split datasets is matched with all individuals with the same value for

Gender and Nuts II in the EWCS dataset. Once matched, a random observation is selected as the donor for the ICT information.

Since the selection is performed at random according to NUTS II and Gender, it is expected that the first cases to extinguish a categorical class would be more accurate. For example, if all females for the Lisbon Metropolitan Area are ordered first in the dataset, it is most likely that they will extinguish all Lisbon cases, which consequently means that the match is done using females from other NUTS II regions. For this purpose, to improve the matching procedure, this random hot deck process is open, meaning that each donor can be selected more than once. While this will ensure integrity for the match itself, it will have a detrimental effect in the new distribution of ICT in the artificial dataset created.

6.2.1.2 Random hot deck statistical matching results

To assess the validity of our results we have employed the same methodology used to compare the populations – see section 5.2.1.3. The results can be found in Table 30. The table provides absolute values, which are not comparable since the total size of each dataset is considerably different. However, to ensure comparability, probability distributions were included. These distributions are for the total cases of ICT and by control variables (age-groups, gender, occupation, sector of economic activity and NUTS II).

Table 30. Comparison between ICT usage imputed via Random hot deck (QPS) and observed (EWCS) (continues)

ICT	QPS							EWCS						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
total	0,22	0,11	0,02	0,03	0,04	0,13	0,44	0,22	0,11	0,02	0,03	0,04	0,13	0,44
agegroups														
15-24	0,01	0,00	0,00	0,00	0,00	0,01	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,03
25-39	0,08	0,04	0,01	0,01	0,02	0,05	0,15	0,13	0,03	0,01	0,01	0,02	0,04	0,12
40-59	0,12	0,06	0,01	0,02	0,02	0,07	0,24	0,09	0,08	0,01	0,02	0,03	0,09	0,28
60+	0,01	0,01	0,00	0,00	0,00	0,01	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,01
gender														
F	0,14	0,05	0,01	0,02	0,02	0,04	0,21	0,13	0,05	0,01	0,02	0,02	0,04	0,21
M	0,08	0,06	0,01	0,01	0,03	0,09	0,23	0,09	0,06	0,01	0,02	0,03	0,09	0,23

Table 30. Comparison between ICT usage imputed via random hot deck (QPS) and observed (EWCS) (continues)

ICT	QPS							EWCS						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
edu														
Low	0,11	0,06	0,01	0,01	0,02	0,07	0,23	0,01	0,01	0,01	0,01	0,02	0,09	0,34
Mid	0,06	0,03	0,00	0,01	0,01	0,03	0,12	0,09	0,05	0,01	0,01	0,02	0,03	0,09
High	0,05	0,02	0,00	0,01	0,01	0,02	0,09	0,12	0,04	0,00	0,01	0,00	0,01	0,01
occ														
Agriculture	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,00	0,00	0,01	0,00	0,00	0,00
Clerical	0,03	0,01	0,00	0,00	0,01	0,02	0,06	0,07	0,03	0,00	0,01	0,01	0,01	0,00
Craft	0,03	0,02	0,00	0,00	0,01	0,02	0,06	0,03	0,01	0,00	0,00	0,00	0,00	0,01
Elementary	0,02	0,01	0,00	0,00	0,00	0,02	0,05	0,08	0,04	0,01	0,00	0,00	0,00	0,01
Managers	0,01	0,01	0,00	0,00	0,00	0,01	0,02	0,01	0,03	0,01	0,01	0,02	0,05	0,09
Operator	0,03	0,01	0,00	0,00	0,00	0,01	0,05	0,00	0,00	0,00	0,00	0,00	0,00	0,01
Professionals	0,03	0,01	0,00	0,00	0,01	0,01	0,05	0,00	0,00	0,00	0,00	0,01	0,04	0,13
Service	0,05	0,02	0,00	0,01	0,01	0,02	0,09	0,00	0,00	0,00	0,00	0,00	0,03	0,11
Technicians	0,02	0,01	0,00	0,00	0,01	0,02	0,05	0,00	0,00	0,00	0,00	0,00	0,01	0,08
eco														
A	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01
C	0,06	0,03	0,01	0,01	0,01	0,04	0,11	0,01	0,02	0,00	0,01	0,01	0,04	0,17
D	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00
E	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01
F	0,01	0,01	0,00	0,00	0,00	0,01	0,03	0,01	0,01	0,00	0,00	0,01	0,01	0,03
G	0,04	0,02	0,00	0,01	0,01	0,03	0,09	0,04	0,03	0,02	0,01	0,01	0,06	0,04
H	0,01	0,01	0,00	0,00	0,00	0,01	0,03	0,02	0,01	0,00	0,00	0,01	0,01	0,05
I	0,02	0,01	0,00	0,00	0,00	0,01	0,03	0,01	0,00	0,00	0,00	0,01	0,01	0,04
J	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,00	0,00	0,00	0,00	0,00
K	0,01	0,00	0,00	0,00	0,00	0,00	0,02	0,04	0,01	0,00	0,00	0,00	0,00	0,00
M	0,01	0,00	0,00	0,00	0,00	0,01	0,02	0,04	0,01	0,00	0,00	0,00	0,00	0,00
N	0,02	0,01	0,00	0,00	0,00	0,01	0,03	0,02	0,00	0,00	0,00	0,01	0,01	0,03
Q	0,03	0,01	0,00	0,00	0,00	0,01	0,05	0,02	0,01	0,00	0,01	0,00	0,00	0,04
R	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
S	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01

Table 30. Comparison between ICT usage imputed via random hot deck (QPS) and observed (EWCS) (conclusion)

ICT	QPS							EWCS						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
nuts2														
Alentejo	0,11	0,03	0,01	0,01	0,01	0,06	0,16	0,01	0,00	0,00	0,00	0,00	0,01	0,03
Algarve	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,01
Center	0,02	0,04	0,00	0,00	0,01	0,04	0,10	0,02	0,04	0,00	0,00	0,01	0,04	0,10
Lisbon	0,08	0,03	0,00	0,02	0,02	0,03	0,14	0,08	0,03	0,00	0,02	0,02	0,03	0,15
North	0,01	0,00	0,00	0,00	0,00	0,01	0,03	0,10	0,03	0,01	0,01	0,01	0,06	0,15

As expected, the results show a bigger discrepancy between the two populations. This is mainly due to two specific reasons. First, the random draw, or random sampling, of a donor does not consider all of the properties of the recipient. Rather, it subdivides the dataset by specific categories, namely gender and NUTS II for the purpose of computational efficiency instead of accurate matching. This clearly results in clear mismatches for all the remaining characterization variables. Second, since ICT is not recoded into a binary variable, such as the parametric matching for example, the matching procedure is more complex, which is likely to be expressed in the form of the discrepancies found in the table above. To assess the validity of our results, we also considered measures of similarity between populations.

Table 31. Measures of similarity between ICT + predictor between the imputed via random hot deck (QPS) and observed (EWCS) distributions

Variable	Dissimilarity index	Overlap	Bhattacharyya coeff,	Hellinger dist,
agegroups	0,14	0,86	0,98	0,15
gender	0,01	0,99	1,00	0,01
edu	0,30	0,70	0,91	0,30
occ	0,40	0,60	0,83	0,41
eco	0,28	0,72	0,90	0,32
nuts2	0,03	0,97	1,00	0,02

The results can be found in Table 31. As shown, the control variables gender and NUTS II that were used to partition the dataset have very low values for both the dissimilarity index and the Hellinger's distance. However, the remaining variables show considerable differences between population.

6.2.2 Distance hot deck

6.2.2.1 Modelling approach

The second non-parametric technique considered is distance hot deck. This procedure is described in section 4.3.3. To prepare our data for the matching, we must first normalize it. Once again, we attempted to use the StatMatch package, however the computation requirements were too high to run in a normal computer. Consequently, we had to create our own function for this procedure. For this purpose, every categorical variable was transformed into binary variables. For instance, the variable edu was transformed into three dummy variables ($edu1$, $edu2$ and $edu3$) with values 0 or 1 according to the education level of the individual. This process was repeated for every variable while taking accounting for the number of categories that compose each variable. Essentially, this transformation process allows us to consider these variables as numerical and therefore can calculate distances between the EWCS and the QPS.

The distances between observations were calculated in absolute terms by subtracting the value of every variable in the QPS with the value of the same variable in the EWCS. For instance, assuming an observation in the QPS has $edu1 = 0$ and another observation in the EWCS has $edu1 = 1$, the absolute distance between these variables is quantified as 1. After these distances are computed, we add up all the variable distances to obtain the total distance between observations. The donor selection is done by selecting the observation that has the smallest absolute distance from the recipient. If more than one observation is selected as a donor, we apply a random draw to select the final donor.

Contrary to the random hot deck procedure described in 6.2.1.1, we have opted to keep this procedure closed. This means that each observation can only be selected as a donor one time. Since our donor file is smaller than our recipient file, once all donors have been selected in the EWCS at least one time, they are set as able for selection a second time to account for the differences in population size.

6.2.2.2 Distance hot deck statistical matching results

To assess the validity of our results we have employed the same methodology used to compare the populations – see section 5.2.1.3. The results can be found in Table 32. Due to the differences in population size, the table does not show absolute differences. Rather, to ensure comparability, probability distributions were included. These distributions are for the total cases of ICT by control variables (age-groups, gender, occupation, sector of economic activity and NUTS II).

Table 32. Comparison between ICT usage imputed via distance hot deck (QPS) and observed (EWCS) (continues)

ICT	QPS							EWCS						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
total	0,22	0,11	0,02	0,03	0,05	0,13	0,44	0,22	0,11	0,02	0,03	0,04	0,13	0,44
agegroups														
15-24	0,01	0,00	0,00	0,00	0,00	0,00	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,03
25-39	0,12	0,03	0,01	0,01	0,01	0,04	0,13	0,13	0,03	0,01	0,01	0,02	0,04	0,12
40-59	0,09	0,07	0,01	0,02	0,03	0,07	0,25	0,09	0,08	0,01	0,02	0,03	0,09	0,28
60+	0,01	0,00	0,00	0,00	0,00	0,01	0,04	0,00	0,00	0,00	0,00	0,00	0,00	0,01
gender														
F	0,13	0,05	0,01	0,02	0,02	0,05	0,21	0,13	0,05	0,01	0,02	0,02	0,04	0,21
M	0,09	0,05	0,01	0,02	0,03	0,08	0,23	0,09	0,06	0,01	0,02	0,03	0,09	0,23
edu														
Low	0,02	0,02	0,01	0,01	0,02	0,09	0,35	0,01	0,01	0,01	0,01	0,02	0,09	0,34
Mid	0,08	0,04	0,01	0,01	0,02	0,03	0,08	0,09	0,05	0,01	0,01	0,02	0,03	0,09
High	0,12	0,04	0,00	0,01	0,00	0,01	0,02	0,12	0,04	0,00	0,01	0,00	0,01	0,01
occ														
Agriculture	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,00	0,00	0,01	0,00	0,00	0,00
Clerical	0,06	0,03	0,00	0,00	0,01	0,01	0,02	0,07	0,03	0,00	0,01	0,01	0,01	0,00
Craft	0,00	0,00	0,00	0,00	0,01	0,03	0,11	0,03	0,01	0,00	0,00	0,00	0,00	0,01
Elementary	0,01	0,00	0,00	0,00	0,00	0,01	0,08	0,08	0,04	0,01	0,00	0,00	0,00	0,01
Managers	0,02	0,01	0,00	0,01	0,00	0,01	0,01	0,01	0,03	0,01	0,01	0,02	0,05	0,09
Operator	0,00	0,00	0,00	0,00	0,00	0,02	0,08	0,00	0,00	0,00	0,00	0,00	0,00	0,01
Professionals	0,07	0,02	0,00	0,01	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,04	0,13
Service	0,02	0,02	0,01	0,01	0,02	0,04	0,09	0,00	0,00	0,00	0,00	0,00	0,03	0,11
Technicians	0,04	0,02	0,00	0,00	0,01	0,01	0,03	0,00	0,00	0,00	0,00	0,00	0,01	0,08

Table 32. Comparison between ICT usage imputed via distance hot deck (QPS) and observed (EWCS) (conclusion)

ICT	QPS							EWCS						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
eco														
A	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01
C	0,02	0,02	0,00	0,01	0,01	0,04	0,17	0,01	0,02	0,00	0,01	0,01	0,04	0,17
D	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00
E	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01
F	0,01	0,01	0,00	0,00	0,00	0,01	0,04	0,01	0,01	0,00	0,00	0,01	0,01	0,03
G	0,05	0,02	0,01	0,01	0,01	0,05	0,05	0,04	0,03	0,02	0,01	0,01	0,06	0,04
H	0,01	0,01	0,00	0,00	0,00	0,01	0,03	0,02	0,01	0,00	0,00	0,01	0,01	0,05
I	0,01	0,01	0,00	0,00	0,01	0,01	0,04	0,01	0,00	0,00	0,00	0,01	0,01	0,04
J	0,02	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,00
K	0,03	0,01	0,00	0,00	0,00	0,00	0,00	0,04	0,01	0,00	0,00	0,00	0,00	0,00
M	0,03	0,01	0,00	0,00	0,00	0,00	0,00	0,04	0,01	0,00	0,00	0,00	0,00	0,00
N	0,01	0,00	0,00	0,00	0,01	0,01	0,03	0,02	0,00	0,00	0,00	0,01	0,01	0,03
Q	0,03	0,01	0,00	0,01	0,00	0,01	0,05	0,02	0,01	0,00	0,01	0,00	0,00	0,04
R	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
S	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01
nuts2														
Alentejo	0,01	0,01	0,00	0,00	0,00	0,01	0,03	0,01	0,00	0,00	0,00	0,00	0,01	0,03
Algarve	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,01
Center	0,04	0,02	0,00	0,01	0,01	0,03	0,10	0,02	0,04	0,00	0,00	0,01	0,04	0,10
Lisbon	0,10	0,04	0,01	0,01	0,02	0,03	0,10	0,08	0,03	0,00	0,02	0,02	0,03	0,15
North	0,07	0,03	0,01	0,01	0,02	0,05	0,20	0,10	0,03	0,01	0,01	0,01	0,06	0,15

The results show a great improvement when compared to the random approach for non-parametric matching. The distribution between categories of ICT is identical to the distribution in the donor dataset.

Some discrepancies can be found between specific variables, which are relatively high when compared with the parametric approach. However, it is important to note that this match is done on a categorical variable composed of seven distinct categories rather than a binary variable. To ensure the validity of our results, we also considered measures of similarity between populations.

Table 33. Measures of similarity between ICT + predictor between the imputed via distance hot deck (QPS) and observed (EWCS) distributions

Variable	Dissimilarity index	Overlap	Bhattacharyya coeff,	Hellinger dist,
agegroups	0,07	0,93	0,99	0,10
gender	0,01	0,99	1,00	0,01
edu	0,05	0,95	0,99	0,07
occ	0,14	0,86	0,96	0,19
eco	0,12	0,88	0,97	0,18
nuts2	0,13	0,87	0,99	0,12

The results can be found in Table 33. As shown, distribution of ICTs by *gender*, *education level* and *age-groups* are very similar. On the other hand, *occ*, *eco* and *nuts2* are found to have a higher dissimilarity index and Hellinger's distance measures.

6.2.3 Distance hot deck with clustering analysis

6.2.3.1. Modelling approach

Extending on the previous technique, we have introduced clustering analysis to improve our matching. Ideally, this optimization would be done via an assignment solver, such as the Hungarian Algorithm, also called the Kuhn-Munkres algorithm , for instance, however, our datasets are very large, which invalidates our possibility to approach this problem correctly due to the lack of computational resources . As such, we adopt clustering approach to statistical matching by creating clusters of individuals in both datasets that are most similar and ensuring that the matches occur between these individuals.

The clustering technique adopted here consists of iterating a set of data by its degree of similarity, which depends on the definition of the problem and the algorithm used. Essentially, it structures the data to constitute partitions of objects or, in this case, individuals, that verify high intra-group homogeneity and high heterogeneity between groups, therefore aggregating common observations in the same group. This process aims to increase the correspondence between individuals in both datasets, while at the same time contributing to the minimization of the sum of distances calculated by the matching procedure.

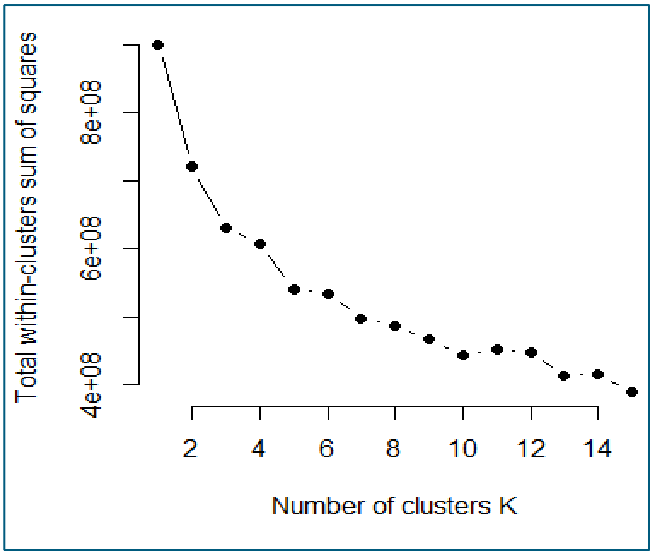
Table 34. Description of clustering methods

Method	Description
K-means	The k-means algorithm is considered one of the most popular, reliable and effective algorithms and is a type of unsupervised learning, which is used when it exists unlabeled data (data has not been tagged with labels identifying characteristics, properties or classifications). Is a partitional algorithm and it minimizes the clustering error. K-means subdivides data into clusters based on nearest means values. For determining the optimal division of these data is necessary that the distance between observations must be minimized. However, K-Means is suitable for numerical variables because it is calculated using the Euclidian distance that is only suitable for numerical data.
K-modes	In as extension of k-means clustering algorithm and is used to clustering categorical data. K-modes algorithm uses (1) a simple matching dissimilarity measure to deal with categorical objects, (2) modes instead of means for clusters, (3) and uses a frequency-based method to update modes in the clustering process to minimize the clustering cost function. With thes e extensions the k-modes algorithm enables the clustering of categorical data in a fashion similar to k-means and for that reason, it preserves the efficiency of the k-means algorithm.
K-Prototype	Is proposed by Huang (1998) and combines the ideas of k-means and k-modes algorithms. The k prototype algorithm, by defining a combined dissimilarity measure, allows the clustering of objects described by mixed numeric and categorical attributes.

Clustering methods are differentiated according to the type of variables that are being analysed. Table 34 describes the most commonly used clustering methods according to the variable types. Since our datasets group two types of variables, numeric and categorical, we opted by applying a K-Prototype clustering algorithm. This type of clustering approach is based on partitioning and its algorithm is an improved form of the K-Means and K-Mode clustering algorithm that allows for the handling of mixed data.

To prepare for the clustering procedure, we first merge both datasets into a single dataset. This allows us to cluster individuals from different datasets into the same cluster. Using the elbow graph approach we identify five as the optimal number of clusters. To prepare for the clustering procedure, we first merge both datasets into a single dataset. This allows us to cluster individuals from different datasets into the same cluster. Using the elbow graph approach we identify five as the optimal number of clusters.

Figure 9. Elbow chart (k- prototype)



Once the clusters are identified, we create five distinct datasets aggregating all individuals according to cluster for each of our initial datasets. We proceed by identifying the matches by iterating through all individuals in each of the dataset and finding the corresponding matching individual that minimizes the distances calculated – similar to the approach used in distance hot deck.

6.2.3.2 Distance hot deck on clusters statistical matching results

To assess the validity of our results we have employed the same methodology used to compare the populations – see section 5.2.1.3. The results can be found in Table 30. Similarly to previous sections, Table 34 presents the probability distributions of individuals across different socioeconomic indicators.

A quick analysis of our results shows that, contrary to what would be expected, our clustering approach does not outperform our simple distance hot deck approach. Although there is no significant deterioration of the match between the datasets, they are slightly worse. We attribute this detrimental effect in our match to the uneven number of individuals from each dataset on the clusters. For instance, when a cluster has more QPS individuals than EWCS individuals, this means that the some of the EWCS individuals will be forcefully matched more than one time, creating an overrepresentation effect that is reflected in the results of comparing both datasets. The opposite happens when a cluster is composed by more EWCS individuals than QPS individuals. In this case, after all QPS individuals are matched, there will be leftover individuals that will not be attributed a matching ID for the QPS, creating an underrepresentation effect of these individuals in our results. This problem would be solved with the implementation of the Hungarian Solver approach. Unfortunately, due to the computational requirements for this procedure, we are unable to employ this method. Consequently, we are disregarding this method as viable solution to improve our statistical matching procedure.

Table 35. Comparison between ICT usage imputed via distance hot deck on clusters (QPS) and observed (EWCS) (continues)

ICT	QPS							EWCS						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
total	0,23	0,10	0,02	0,03	0,04	0,13	0,45	0,22	0,11	0,02	0,03	0,04	0,13	0,44
agegroups														
15-24	0,01	0,00	0,00	0,00	0,00	0,00	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,03
25-39	0,10	0,03	0,01	0,01	0,02	0,04	0,14	0,13	0,03	0,01	0,01	0,02	0,04	0,12
40-59	0,11	0,06	0,01	0,02	0,02	0,07	0,26	0,09	0,08	0,01	0,02	0,03	0,09	0,28
60+	0,01	0,01	0,00	0,00	0,00	0,01	0,04	0,00	0,00	0,00	0,00	0,00	0,00	0,01
gender														
F	0,12	0,05	0,01	0,02	0,02	0,05	0,21	0,13	0,05	0,01	0,02	0,02	0,04	0,21
M	0,11	0,05	0,01	0,01	0,03	0,08	0,23	0,09	0,06	0,01	0,02	0,03	0,09	0,23
edu														
Low	0,05	0,03	0,01	0,02	0,02	0,08	0,30	0,01	0,01	0,01	0,01	0,02	0,09	0,34
Mid	0,07	0,04	0,01	0,01	0,01	0,03	0,09	0,09	0,05	0,01	0,01	0,02	0,03	0,09
High	0,10	0,03	0,00	0,01	0,01	0,01	0,05	0,12	0,04	0,00	0,01	0,00	0,01	0,01
occ														
Agriculture	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,00	0,00	0,01	0,00	0,00	0,00
Clerical	0,05	0,02	0,00	0,00	0,01	0,01	0,04	0,07	0,03	0,00	0,01	0,01	0,01	0,00
Craft	0,01	0,01	0,00	0,00	0,00	0,02	0,10	0,03	0,01	0,00	0,00	0,00	0,00	0,01
Elementary	0,01	0,01	0,00	0,00	0,00	0,01	0,07	0,08	0,04	0,01	0,00	0,00	0,00	0,01
Managers	0,02	0,01	0,00	0,00	0,00	0,01	0,02	0,01	0,03	0,01	0,01	0,02	0,05	0,09
Operator	0,01	0,01	0,00	0,00	0,00	0,02	0,07	0,00	0,00	0,00	0,00	0,00	0,00	0,01
Professionals	0,06	0,02	0,00	0,00	0,01	0,01	0,03	0,00	0,00	0,00	0,00	0,01	0,04	0,13
Service	0,03	0,02	0,01	0,01	0,01	0,04	0,08	0,00	0,00	0,00	0,00	0,00	0,03	0,11
Technicians	0,04	0,02	0,00	0,00	0,01	0,01	0,04	0,00	0,00	0,00	0,00	0,00	0,01	0,08
eco														
A	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01
C	0,03	0,02	0,00	0,01	0,01	0,03	0,17	0,01	0,02	0,00	0,01	0,01	0,04	0,17
D	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00

Table 35. Comparison between ICT usage imputed via distance hot deck on clusters (QPS) and observed (EWCS) (conclusion)

ICT	QPS							EWCS						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
E	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01
F	0,01	0,01	0,00	0,00	0,00	0,01	0,04	0,01	0,01	0,00	0,00	0,01	0,01	0,03
G	0,05	0,02	0,01	0,01	0,01	0,05	0,05	0,04	0,03	0,02	0,01	0,01	0,06	0,04
H	0,01	0,01	0,00	0,00	0,00	0,01	0,03	0,02	0,01	0,00	0,00	0,01	0,01	0,05
I	0,01	0,00	0,00	0,00	0,01	0,01	0,03	0,01	0,00	0,00	0,00	0,01	0,01	0,04
J	0,02	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,00
K	0,03	0,01	0,00	0,00	0,00	0,00	0,01	0,04	0,01	0,00	0,00	0,00	0,00	0,00
M	0,03	0,01	0,00	0,00	0,00	0,00	0,01	0,04	0,01	0,00	0,00	0,00	0,00	0,00
N	0,01	0,00	0,00	0,00	0,00	0,01	0,03	0,02	0,00	0,00	0,00	0,01	0,01	0,03
Q	0,03	0,01	0,00	0,01	0,00	0,01	0,05	0,02	0,01	0,00	0,01	0,00	0,00	0,04
R	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
S	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01
nuts2														
Alentejo	0,01	0,01	0,00	0,00	0,00	0,01	0,03	0,01	0,00	0,00	0,00	0,00	0,01	0,03
Algarve	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,01
Center	0,04	0,02	0,00	0,01	0,01	0,03	0,10	0,02	0,04	0,00	0,00	0,01	0,04	0,10
Lisbon	0,10	0,04	0,00	0,01	0,02	0,03	0,10	0,08	0,03	0,00	0,02	0,02	0,03	0,15
North	0,07	0,03	0,01	0,01	0,02	0,05	0,20	0,10	0,03	0,01	0,01	0,01	0,06	0,15

The same conclusion can be drawn when we analyse the measures of similarity between populations presented in Table 36. Overall, the statistics get worse when compared to the previous method.

Table 36. Measures of similarity between ICT + predictor between the imputed via distance hot deck on cluster (QPS) and observed (EWCS) distributions

Variable	Dissimilarity index	Overlap	Bhattacharyya coeff,	Hellinger dist,
agegroups	0,10	0,90	0,99	0,12
gender	0,04	0,96	1,00	0,03
edu	0,13	0,87	0,97	0,17
occ	0,24	0,76	0,92	0,29
eco	0,14	0,86	0,96	0,20
nuts2	0,13	0,87	0,99	0,12

6.3 Mixed Methods

6.3.1 Modelling approach

Under mixed methods, we used the same modelling approach presented in 6.2.2. The difference between the two processes lies in the introduction of a parametric model. Initially, we use the same methodology developed for the parametric logistic regression to estimate a binary variable for ICT usage in the QPS survey. This variable is then included in the calculation of the distances between observations in the EWCS and the QPS. This process adds an extra distinctive factor between observations that is based on the introduction of a parametric model estimation, which according to the predictive ability of the parametric model can add some welcome complexity to the selection procedure ensuring more accurate matches.

6.3.2 Mixed methods statistical matching results

To assess the validity of our results we have employed the same methodology used to compare the populations – see section 5.2.1.3. The results can be found in Table 37. Due to the differences in population size, the table does not show absolute differences. Rather, to ensure comparability, probability distributions were included. These distributions are for the total cases of ICT by control variables (age-groups, gender, occupation, sector of economic activity and NUTS II).

Table 37. Comparison between ICT usage imputed via mixed approach (QPS) and observed (EWCS) (continues)

ICT	QPS							EWCS						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
total	0,22	0,10	0,02	0,03	0,05	0,13	0,44	0,22	0,11	0,02	0,03	0,04	0,13	0,44
agegroups														
15-24	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,03
25-39	0,12	0,03	0,01	0,01	0,02	0,05	0,13	0,13	0,03	0,01	0,01	0,02	0,04	0,12
40-59	0,09	0,07	0,01	0,02	0,02	0,07	0,25	0,09	0,08	0,01	0,02	0,03	0,09	0,28
60+	0,01	0,00	0,00	0,00	0,00	0,01	0,04	0,00	0,00	0,00	0,00	0,00	0,00	0,01
gender														
F	0,13	0,05	0,01	0,02	0,02	0,05	0,21	0,13	0,05	0,01	0,02	0,02	0,04	0,21
M	0,09	0,05	0,01	0,02	0,03	0,08	0,23	0,09	0,06	0,01	0,02	0,03	0,09	0,23
edu														
Low	0,02	0,02	0,01	0,01	0,02	0,09	0,35	0,01	0,01	0,01	0,01	0,02	0,09	0,34
Mid	0,07	0,04	0,01	0,01	0,02	0,03	0,08	0,09	0,05	0,01	0,01	0,02	0,03	0,09
High	0,13	0,04	0,00	0,01	0,00	0,01	0,01	0,12	0,04	0,00	0,01	0,00	0,01	0,01

Table 37. Comparison between ICT usage imputed via mixed approach (QPS) and observed (EWCS) (continues)

ICT	QPS							EWCS						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
occ														
Agriculture	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,00	0,00	0,01	0,00	0,00	0,00
Clerical	0,06	0,04	0,01	0,00	0,00	0,00	0,02	0,07	0,03	0,00	0,01	0,01	0,01	0,00
Craft	0,00	0,00	0,00	0,00	0,01	0,03	0,11	0,03	0,01	0,00	0,00	0,00	0,00	0,01
Elementary	0,00	0,00	0,00	0,00	0,00	0,01	0,09	0,08	0,04	0,01	0,00	0,00	0,00	0,01
Managers	0,02	0,01	0,00	0,01	0,00	0,00	0,00	0,01	0,03	0,01	0,01	0,02	0,05	0,09
Operator	0,00	0,00	0,00	0,00	0,00	0,02	0,09	0,00	0,00	0,00	0,00	0,00	0,00	0,01
Service	0,01	0,01	0,01	0,01	0,02	0,05	0,10	0,00	0,00	0,00	0,00	0,00	0,03	0,11
Technicians	0,05	0,02	0,00	0,01	0,00	0,01	0,03	0,00	0,00	0,00	0,00	0,00	0,01	0,08
eco														
A	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,01
C	0,02	0,02	0,00	0,01	0,01	0,04	0,17	0,01	0,02	0,00	0,01	0,01	0,04	0,17
D	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00
E	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01
F	0,01	0,01	0,00	0,00	0,00	0,01	0,04	0,01	0,01	0,00	0,00	0,01	0,01	0,03
G	0,05	0,02	0,01	0,01	0,01	0,05	0,05	0,04	0,03	0,02	0,01	0,01	0,06	0,04
H	0,01	0,01	0,00	0,00	0,00	0,01	0,03	0,02	0,01	0,00	0,00	0,01	0,01	0,05
I	0,01	0,00	0,00	0,00	0,01	0,01	0,04	0,01	0,00	0,00	0,00	0,01	0,01	0,04
J	0,02	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,00
K	0,03	0,01	0,00	0,00	0,00	0,00	0,00	0,04	0,01	0,00	0,00	0,00	0,00	0,00
M	0,03	0,01	0,00	0,00	0,00	0,00	0,00	0,04	0,01	0,00	0,00	0,00	0,00	0,00
N	0,01	0,00	0,00	0,00	0,01	0,01	0,04	0,02	0,00	0,00	0,00	0,01	0,01	0,03
Q	0,03	0,01	0,00	0,01	0,00	0,01	0,05	0,02	0,01	0,00	0,01	0,00	0,00	0,04
R	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
S	0,01	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01

Table 37. Comparison between ICT usage imputed via mixed approach (QPS) and observed (EWCS) (conclusion)

ICT	QPS							EWCS						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
nuts2														
Alentejo	0,01	0,00	0,00	0,00	0,00	0,01	0,03	0,01	0,00	0,00	0,00	0,00	0,01	0,03
Algarve	0,01	0,00	0,00	0,00	0,00	0,01	0,02	0,01	0,00	0,00	0,00	0,00	0,00	0,01
Center	0,04	0,02	0,00	0,01	0,01	0,03	0,10	0,02	0,04	0,00	0,00	0,01	0,04	0,10
Lisbon	0,10	0,04	0,01	0,01	0,02	0,03	0,10	0,08	0,03	0,00	0,02	0,02	0,03	0,15
North	0,07	0,03	0,01	0,01	0,01	0,05	0,19	0,10	0,03	0,01	0,01	0,01	0,06	0,15

The results are very positive. The similarity between distributions is particularly striking regarding age-groups, education level and gender. On the other hand, variables with more categories, such as occupation or sector of economic show greater discrepancies. This is perhaps related to the numbers present in each cell, which are replaced with different categories in the matching procedure if they are all allocated in the EWCS and there are still individuals left to match in the QPS. To ensure the validity of our results, we also considered measures of similarity between populations.

Table 38. Measures of similarity between ICT + predictor between the imputed via mixed approach (QPS) and observed (EWCS) distributions

Variable	Dissimilarity index	Overlap	Bhattacharyya coeff.	Hellinger dist.
Agegroups	0,07	0,93	0,99	0,10
gender	0,02	0,98	1,00	0,02
edu	0,05	0,95	1,00	0,07
occ	0,12	0,88	0,98	0,16
eco	0,12	0,88	0,97	0,17
nuts2	0,13	0,87	0,99	0,11

The results can be found in Table 38. As shown, distribution of ICTs *by gender, education level* and *age-groups* are very similar. On the other hand, *occ, eco* and *nuts2* are found to have a higher dissimilarity index and Hellinger's distance measures.

7. Discussion

This working paper proposed a clear set of goals since its genesis. Among these, the central guiding line has been to identify the feasibility of merging the QPS and EWCS surveys. For this purpose, we conducted an exploratory methodological procedure that matched both surveys by a common set of variables using different statistical matching techniques. The techniques considered here fall under three distinct categories: parametric, non-parametric and mixed. The results presented in Section 6 show that almost all techniques were relatively successful in matching the distinct datasets, although with different levels of validity. To test these procedures, we statistically matched the variable ICT usage in work context. This variable is composed of 7 categories that determine the extent of ICT usage.

The first technique tested was parametric and consisted of a statistical match between the QPS and the EWCS using a logistic regression approach. An evaluation of the results shows that this method was successful in maintain a high degree of similarity between populations in the new synthetic dataset when compared with the EWCS using control variables. However, one major drawback of this approach lies in the need to reconfigure our categorical variable into a binary variable. Although successful in maintain an approximate distribution when compared with its original distribution, this method surely entails a considerable level of loss of information through the matching procedure. However, if this is not a problem, it should be seriously considered. When compared with other methods, it is very simple to implement and not very computationally intensive. The relationship between easy to implement, computational requirements and validity of results is very good, making this method one of the best go to approaches to match the QPS and EWCS. Second, we moved to the non-parametric methods. Here we tested too distinct approaches that fall under the umbrella of hot deck procedures. The first was random hot deck. This approach was not ideal, which is clearly evidenced by comparing the resulting synthetic dataset with the original distributions in the EWCS. In this case, perhaps due to the lack of guiding categories – only NUTS II and gender were used – the similarity between populations is extremely low, when considering control variables composed by many categories, which is the case of occupation and sector of economic activity.

The second non-parametric approach considered was distance hot deck. The results from this approach are considerably better than its counterpart – random hot deck. The computation of distances between observations ensures that the matches are considerably more accurate, which is clearly evidenced in the evaluation of similarity between the new synthetic population and the original EWCS. When compared with the parametric approach, the results are slightly worse. However, it is important to consider that in the hot deck approach, the original variable is not reconfigured into a binary variable, which means that all seven categories are present in the new synthetic dataset. Consequently, this means that there is no loss of information during the matching procedure. Additionally, using a categorical variable with 7 distinct categories rather than a binary variable is bound to add some error caused primarily by the reduced number of observations in each categorical cell. This is easily illustrated by creating crosstabulations between the binary variable and each of the control variables and repeating this process for the categorical variable. The total population in each individual cell will be considerably reduced in the latter, which affects the number of donors available for selection and can lead to the selection of alternative donors.

Finally, the mixed approach considered uses both parametric and non-parametric techniques to perform the matching between the QPS and the EWCS. In sum, it uses the logistic regression model to estimate the binary ICT variable in the QPS and uses this variable as a distinctive factor between individuals with similar distances. This approach has the best performance regarding similarity measures when the categorical variable is included. In comparison with the distance hot deck, distribution of ICT by the control variables composed by various categories, namely occupation and sector of economic activity, see a slight improve in this approach.

Our results suggest that the EWCS and the QPS can be successfully matched using statistical matching procedures. However, it is important to note that there was an extensive harmonization process involved in the application of these statistical matching models. One of the most important aspects of this process, that needs to be clearly stated in our results, is that we have opted to match only the populations that were similar in these datasets. In this regard, our results are only valid for the statistical matching of the EWCS and the QPS when considering employees with a permanent contract. This is an important limitation of the techniques discussed throughout this working paper.

Another aspect that needs to be considered is the fact that the non-common QPS variables and the imputed EWCS are never observed together. As such, without the assistance of a third dataset where these variables are included, we are forced to assume that they are conditionally independent. In simplistic terms, the conditional independence assumption (CIA henceforth) assumes that given the knowledge of matching variables *X*, knowledge of imputed variables *Y* provides no information on non-common variables *Z*. According to D'Orazio *et al.* (2006), this is a particularly strong assumption to make and rarely holds in practice. The CIA can be avoided using auxiliary information from a third independent dataset where these variables are observed together. However, the only non-common variable in the QPS that we wish to consider in further analysis is income. To test the independence between our non-common variables, earnings and ICT, we have conducted a Pearson's Chi Square test of independence in our synthetic dataset created by the mixed approach. The results show a *p-value* that is below 2,2e-16, therefore rejecting the null hypothesis and conclude that there is a relationship between these variables. Despite our ability to calculate the relationship between variables using Pearson's Chi Square test of independence, it would be preferred to have a third dataset where the variables are observed together.

Finally, it is important to note that the approach used here does not produce the optimal matching results. These results would only be achievable with the introduction of a solver for the assignment problem that minimizes the sum of distances between individuals in both datasets. Rather, we use a heuristic approach that allows us to minimize distances based on the iteration of individuals in the QPS dataset. This means that the sum of distances obtained through this process is directly linked to the order of the individuals in this dataset and may not represent the optimal matching. The lack of computational resources for this task is the main limitation in the use of a solver such as the Hungarian Algorithm for instance. More work needs to be done in order to be able to integrate this approach with a Hungarian Solver and optimize this problem.

Bibliography

D'Orazio, M. (2014). *Statistical matching and imputation of survey data with StatMatch*. https://www.researchgate.net/publication/263888033_Statistical_Matching_and_Imputation_of_Survey_Data_with_StatMatch

D'Orazio, M. (2015). Integration and imputation of survey data in R: the StatMatch package. *Romanian Statistical Review*, 63(2), 57–68. http://www.revistadestatistica.ro/wp-content/uploads/2015/04/RRS2_2015_A06.pdf

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). The conditional independence assumption. In Groves, R., Kalton, G., Rao, J., Schwarz, N. & Skinner, C. (Eds.), *Statistical Matching: Theory and practice*. John Wiley & Sons, Ltd., pp. 13–60. https://onlinelibrary.wiley.com/doi/10.1002/0470023554.ch2

de Wall, T. (2015). *Statistical matching: Experimental results and future research questions*. https://www.researchgate.net/publication/288291792_Statistical_matching_Experimental_results_and_future_research_questions

Eurofound (2015). *6th European Working Conditions Survey: Weighting report*. Publications Office of the European Union. https://www.eurofound.europa.eu/sites/default/files/ef_survey/field_ef_documents/6th_ewcs_2015_-_weighting_report.pdf

Eurostat (2008). *NACE Rev. 2 - Statistical classification of economic activities in the European Community*. https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF

Eurostat (2017). *Handbook on Methdology of Modern Business Statistics*. https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Ford, L. (1983). An overview of hot deck procedures. In Madow, G. , Olkin, I. & Rubin, D. (Eds.), *Incomplete data in sample surveys.* Academic Press Inc., 185–207.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). Record Linkage – Methodology. In *Data quality and record linkage techniques*, 81–92. https://doi.org/10.1007/0-387-69505-2_8

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery 2, 283–304. https://doi.org/10.1023/A:1009769707641*

Hu, M., & Salvucci, S. (2001). *A study of imputation algorithms*. Working Paper NCES No. 2001-17. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200117

ILO (2011). *International Standard Classification of Occupations: structure, group definitions and correspondence tables*. Publications of International Labour Office. https://www.ilo.org/public/english/bureau/stat/isco/docs/publication08.pdf

INE (2007). *Classificação Portuguesa das Actividades Económicas Rev. 3*. Instituto Nacional Estatística. https://www.ine.pt/ine_novidades/semin/cae/CAE_REV_3.pdf

INE (2011). *Classificação Portuguesa das Profissões 2010*. Instituto Nacional Estatística. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=107961853&PUBLICACOESmodo=2&xlang=pt

Lee, D., Minton, J., & Pryce, G. (2015). Bayesian inference for the dissimilarity index in the presence of spatial autocorrelation. *Spatial Statistics*, *11*, 81–95. https://doi.org/10.1016/j.spasta.2014.12.001

Leulescu, A., & Agafitei, M. (2013). *Statistical matching: a model based approach for data integration. Eurostat - Methodologies and Working papers.* https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/KS-RA-13-020?inheritRedirect=true

Singh, A. C., Mantel, H., Kinack, M., & Rowe, G. (1993). Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, *19(1)*, 59–79.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45* (3), 1–67. https://doi.org/10.18637/jss.v045.i03

Appendix I

This appendix provides a comparison between the NACE Rev.2 (Eurostat, 2008) and the CAE Rev.3 (INE, 2007) classifications. (continue)

Section	NACE Rev. 2	CAE Rev. 3
A	Agriculture, forestry and fishing	Agricultura, produção animal, caça, floresta e pesca
B	Mining and quarrying	Indústrias extractivas
C	Manufacturing	Indústrias transformadoras
D	Electricity, gas, steam and air conditioning supply	Electricidade, gás, vapor, água quente e fria e ar frio
E	Water supply; sewerage, waste management and remediation activities	Captação, tratamento e distribuição de água; saneamento gestão de resíduos e despoluição
F	Construction	Construção
G	Wholesale and retail trade; repair of motor vehicles and motorcycles	Comércio por grosso e a retalho; reparação de veículos automóveis e motociclos
H	Transportation and storage	Transportes e armazenagem
I	Accommodation and food service activities	Alojamento, restauração e similares
J	Information and communication	Actividades de informação e de comunicação
K	Financial and insurance activities	Actividades financeiras e de seguros
L	Real estate activities	Actividades Imobiliárias
M	Professional, scientific and technical activities	Actividades de consultoria, científicas, técnicas e similares
N	Administrative and support service activities	Actividades administrativas e dos serviços de apoio
O	Public administration and defence; compulsory social security	Administração Pública e Defesa; Segurança Social Obrigatória
P	Education	Educação

Section	NACE Rev. 2	CAE Rev. 3
Q	Human health and social work activities	Actividades de saúde humana e apoio social
R	Arts, entertainment and recreation	Actividades artísticas, de espectáculos, desportivas e recreativas
S	Other service activities	Outras actividades de serviços
T	Activities of households as employers	Actividades das famílias empregadoras de pessoal doméstico
U	Activities of extraterritorial organisations and bodies	Actividades dos organismos internacionais e outras instituições extra-territoriais

Appendix II

This appendix provides a comparison between the ISCO-08 (ILO, 2011) and the CPP (INE, 2011).

Group	ISCO-08	CPP
1	Managers	Representantes do poder legislativo e de órgãos executivos, dirigentes, directores e gestores executivos
2	Professionals	Especialistas das actividades intelectuais e científicas
3	Technicians and associate professionals	Técnicos e profissões de nível intermédio
4	Clerical support workers	Pessoal administrativo
5	Service and sales workers	Trabalhadores dos serviços pessoais, de protecção e segurança e vendedores
6	Skilled agricultural, forestry and fish	Agricultores e trabalhadores qualificados da agricultura, da pesca e da floresta
7	Craft and related trades workers	Trabalhadores qualificados da indústria, construção e artífices
8	Plant and machine operators, and assemblers	Operadores de instalações e máquinas e trabalhadores da montagem
9	Elementary occupations	Trabalhadores não qualificados

O CoLABOR – Laboratório Colaborativo para o Trabalho, Emprego e Proteção Social é uma instituição de investigação científica reconhecida pela Fundação para a Ciência e Tecnologia, que conta com uma equipa multidisciplinar de investigadores altamente qualificados.

O CoLABOR tem quatro objetivos centrais: apoiar a conceção e reformulação de políticas nas suas áreas temáticas; capacitar as instituições, incluindo a administração pública, as empresas e as instituições do terceiro setor; qualificar o emprego, mediante a formação de quadros e a criação de emprego científico; contribuir para debate público nas áreas do trabalho e da proteção social, através de formas de divulgação eficazes e inovadoras dos resultados da investigação que leva a cabo.

O CoLABOR concretiza estes objetivos através de uma agenda ambiciosa de aprofundamento do conhecimento científico em torno de três eixos temáticos centrais: o trabalho e emprego; a proteção social e os equipamentos e respostas sociais. Nesta agenda, destacam-se as seguintes prioridades: o estudo dos impactos das novas tecnologias sobre o trabalho e a proteção social; a reflexão sobre a adequação e sustentabilidade de diferentes modelos de proteção social; e a avaliação de equipamentos e respostas sociais.

Transversalmente a estas áreas temáticas, o CoLABOR desenvolve e mantém a DataLABOR, uma plataforma digital de sistematização, análise crítica, visualização de informação estatística e jurídica de âmbito internacional, nacional, regional e local nas áreas do trabalho, emprego e proteção social.

Para desenvolver a sua atividade, o CoLABOR conta com o apoio dos seus associados, onde se contam diversas instituições universitárias e de investigação, instituições do terceiro setor e empresas.

Associados

Centro de Estudos Sociais

Instituto de Direito Económico, Financeiro e Fiscal

Iscte - Instituto Universitário de Lisboa, CIES - Iscte

Confederação Nacional das Instituições de Solidariedade

Santa Casa da Misericórdia de Lisboa

Delta Cafés – Sociedade Gestora de Participações Sociais, SA

Mota-Engil, SGPS, S.A.

SONAE Corporate, S.A.

Cofinanciado por: