Statistical Matching

2014

Organismo Autónomo del

EUSKO JAURLARITZA
GOBIERNO VASCO

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA
www.eustat.eus

Edition:
EUSTAT
Euskal Estatistika Erakundea
Basque Statistics Institute

Publication:
EUSTAT
Euskal Estatistila Erakundea
Basque Statistics Institute
Donostia-San Sebastián 1
01010 Vitoria-Gasteiz

© **Administration of the Basque Country**

First Edition
I/2015

Printing and Bounding:
Printing and Reprography Service. Basque Government.

ISBN: 978-84-7749-483-6

Legal Deposit: VI-795/2014

STATISTICAL MATCHING

Ines Garmendia Navarro

ines.garmendia@gmail.com

Traducido por: Ofilingua, S.L.

EUSKAL ESTATISTIKA ERAKUNDEA
BASQUE STATISTICS INSTITUTE

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.eus
www.eustat.eus

Presentation

In the field of Official Statistics great many efforts have been made in studying data integration techniques in order to provide its users with the information of the highest quality possible. In this respect, Eustat organized its XXVI International Statistics Seminar under the theme "Statistical Matching: Methodological issues and practice with R-StatMatch”.

This publication seeks to present the work done in the context of a research internship in this field. The book is divided into four main sections: the first is devoted to the methodology and the second describes the techniques in the R environment. The third is dedicated to exposing the statistical matching of two independent surveys from Eustat, namely the Living Conditions Survey and the Population in Relation to Activity Survey. Finally, the fourth and final section describes the development of an own R package.

Vitoria-Gasteiz, December of 2014

Josu Iradi Arrieta

General Director of EUSTAT

Contents

Chapter

1

Introduction

The contents of this Technical Manual is the result of work carried out thanks to a grant for training and research into mathematical and statistical methodologies for the subject of *Statistical Matching,* which was awarded in 2012 by the Basque Statistics Institute - Euskal Estatistika Erakundea.

Statistical matching[1] is a methodology that allows you to create integrated statistics and combined indicators to provide information on independent surveys that refer to the same population of interest. The main advantage is that the information coming from distinct surveys can be used more efficiently, and that it can be kept in separate data files.

This methodology covers a wide variety of statistical techniques that are diverse and from different origins, such as the imputation of missing data, the quantification of uncertainty and the theory of complex sampling. These techniques are constantly developing, and many of them are available in the free R software environment, a platform that is becoming more and more prominent in the academic world, in industry and, gradually, in official statistics.

The manual is organised as follows: the first chapter contains this introduction, the second is about the methodology and presents the main techniques as well as numerous recommendations on how to implement them. The third chapter details the possibilities offered by the R environment to implement the techniques. The fourth chapter outlines a genuine case of statistical matching between two independent EUSTAT surveys, the Survey on Population with Relation to Activity and the Survey on Living Conditions. This serves to illustrate the main stages that a match must comprise, as well as the type of results that are obtained. The fifth chapter outlines the development of a suitable R package, and the sixth and final chapter provide conclusions. Finally, there is an annex that contains a series of tables with numerical results.

In conjunction with the creation of this manual, a suitable R package was developed to make it easier to implement the methodology presented here. The incentive is that currently, the functions used to tackle each phase of a match are in distinct packages. This means the user has to construct his/her code on the basis of distinct "philosophies" that depend on the package that is being dealt with at any given moment. The intention

[1] In English, the terminology is varied: the following terms can be used synonymously depending on the source and the context: *statistical matching, data fusion, file merging, survey linking and synthetic matching*

was to use this suitable package, `micromatch`, to overcome this difficulty, offering the user a unique interface where the statistical matching concepts remain unequivocally reflected, thus enabling packages that have already been tested and contrasted (e.g. `StatMatch` and `mice`) to be used more efficiently. The package was presented during the VI-grade R Users Conference in Santiago de Compostela (23 and 24 October 2014). It is also available on Eustat’s website.

KEY WORDS: Statistical matching, data fusion, missing data imputation, R

Chapter

2

Methodology

The foundations of statistical matching

Statistical matching comprises a series of techniques aimed at obtaining integrated statistics of indicators or variables gathered from diverse sources, generally sample surveys carried out on the same population.

In the most general case, these are based on independent sample surveys that refer to the same population (for instance, residents of the Basque Country in 2014), each of which measures a set of dimensions or indicators separately (lifestyles, employment status, income...). When creating the match, the surveys should share a series of variables or common measurements, usually basic socio-demographic variables such as age, gender or level of education.

When the data gathered from two independent[2] surveys is "linked", a situation such as the one described in Figure 1 arises. Block Z represents common variables (socio-demographic and others) gathered by both surveys. The other two blocks, X and Y, represent items from each questionnaire. If the information in common block Z is available for all registers (this is the basic premise of the match) in the two blocks specified (X and Y), only the registers for each survey will contain informed values. So, two large blocks of "missing" or unobserved values (the lighter segments in Figure 1) arise.

[2] For simplicity, this manual has only considered the match from two surveys, but the methodology can be applied to more than two surveys.

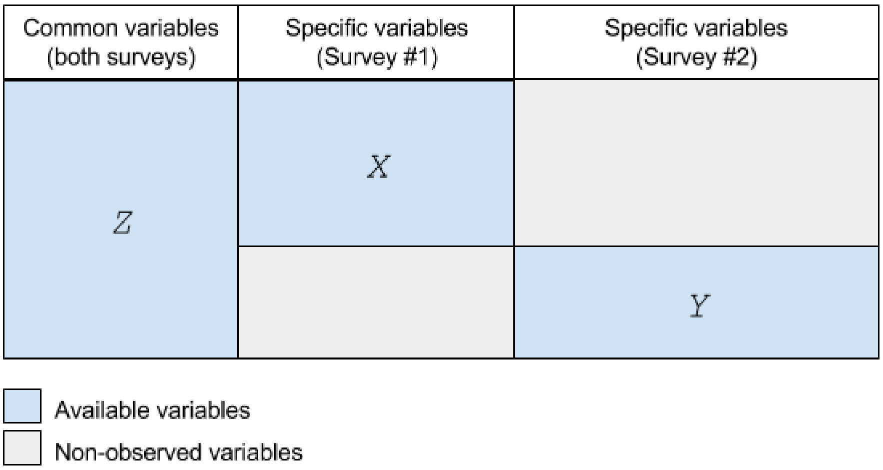

Common variables (both surveys)
Specific variables (Survey #1)
Specific variables (Survey #2)
Z
X
Y
Available variables
Non-observed variables

Figure 1 *The chart that emerges upon linking the registers of the two surveys that share a block Z of common variables. Blocks X and Y represent specific items or items that are not shared between the surveys. This process is called concatenation.*

Starting point:

- * There is a block Z of common variables, i.e. variables shared by both files.
- * There are two blocks, X and Y, of specific variables that are unobserved and obtained jointly: X only appears in file #1, and Y only appears in file #2.
- * The probability of a population unit appearing in both samples is close to zero, and we can ignore it.

Statistical matching sets out a situation that is considerably different from other techniques, such as **register fusion**, in which the objective is to identify identical units between files (for example, between a census and an administrative file). Statistical matching is distinctive because it is based on independent sample surveys. As a result, the situation is to a certain degree the opposite: at the outset, we *know* that the units are distinct, but we are *looking* to find "similar units" with the goal of matching *not the units*, but the *variables* from the surveys.

In short, statistical matching seeks to match specific variables from independent sample surveys that refer to the same population of interest, using information shared between them as a "bridge". Next, we will see the main approaches to this problem as well as the solutions that have been adopted over time.

Approaches and methods

When performing statistical matching, one of the following approaches is generally adopted (M. D'Orazio, M. Di zio & M. Scanu, 2008):

- The **macro approach** seeks to create straightforward estimates from specific variables, for instance a correlation coefficient between variable X and another variable Y, or a joint marginal distribution.

- The **micro approach** aims to create a synthetic file containing complete information with all these specific variables from the files, and for all the registers. This synthetic file is subsequently used to perform joint analyses that refer to variables initially found in separate files.

Today, there are multiple statistical methods for matching surveys. In effect, since the origins of this methodology in the 1960s, multiple solutions have been developed in the field of market research (in Europe) and in the field of official statistics (in the USA and Canada from the 1970s). (For any reader who might be interested, the third chapter of the book by S Rässler (2002) offers a brief history of statistical matching up to 2002).

Below we will examine the main methods within micro and macro approaches. Additionally, given its relevance to official statistics, we will examine the treatment of surveys with complex survey designs.

Micro methods

The aim of the micro approach is to generate a synthetic file with complete information for all the specific variables deemed of interest, and for all registers. A. Leulescu and M. Agafitei (2013) highlight four large groups of micro methods, which we will go on to describe next.

Hot-deck methods

Throughout the history of statistical matching, the hot-deck methods family has been by far the most utilised. It involves a combination of non-parametric measures, in other words, measures that presuppose no initial statistical distribution for the variables,

The procedure is as follows: for each register in one of the files (designated *recipient file*), one or various registers are sought in the other file (designated *donor file*) that are the most similar in terms of common variables (age, gender, level of education...) The values that refer to the donor register that has been found are *imputed* in the recipient register (see Figure 2).

The main characteristic of the hot-deck procedure is that the values imputed are always real values, in other words, they correspond to values that have really been observed and gathered in the donor file.

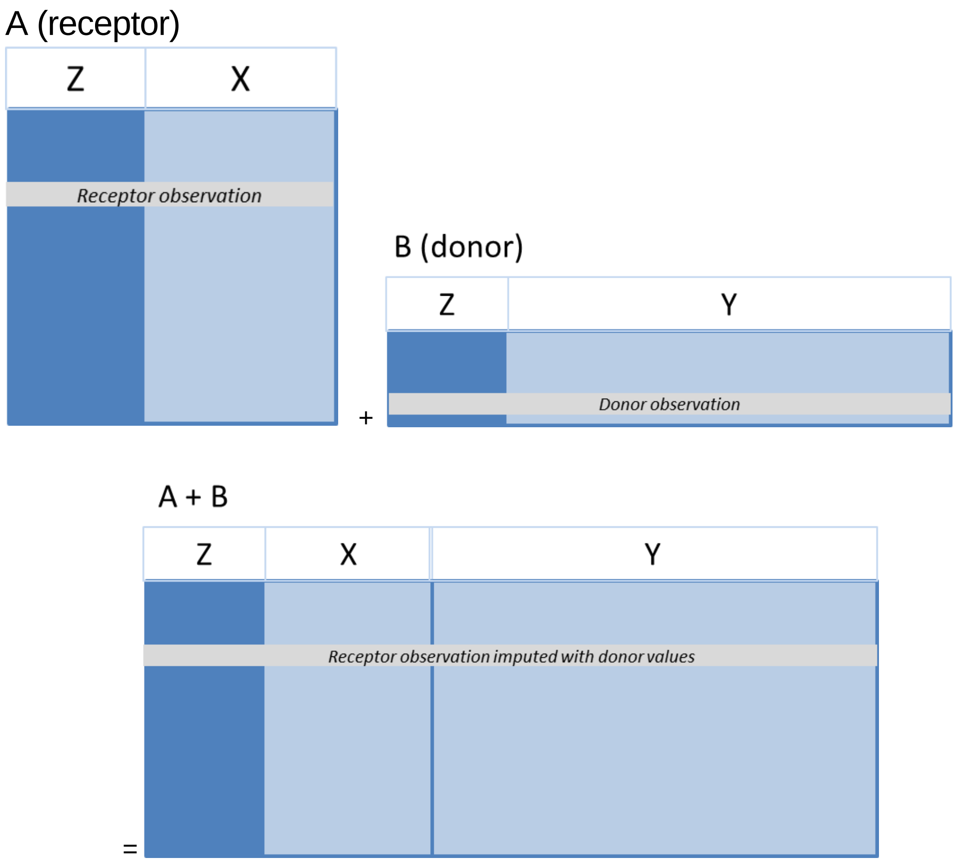


Figure 2 *The simplified premise behind hot-deck imputation For each recipient register, the most similar donor is sought, and the corresponding value is "imputed" into it.*

Often, and depending on the data, it turns out not to be possible to find an identical donor for all variables for each recipient. In this case, it is customary to define a *distance* based on the common variables so as to be able to look for pairs of similar registers. The distance will have a distinct mathematical formulation according to the nature of the variables (categorical or numerical) and depends on other considerations; see M. D'Orazio et al. (2008), Annex C

As Table 1 indicates, the hot-deck imputation algorithm admits diverse variables according to the way they are implemented: definition of the strata, restrictions so as to not repeat the use of donor registers etc. By conveniently customising these variables, a good hot-deck algorithm will correctly reproduce, in the imputed recipient file, the distributions observed in the donor file.

HOT-DECK METHODS

Classification

- Micro method.
- Non-parametric.
- Frequentist.

Algorithm

- For each recipient register in A, a donor is sought in B - one that is the most similar in terms of common Z variables. The Y values of the donor register are imputed into the recipient register.
    - When there is no donor register in B that is identical to the recipient, a distance is introduced to find one that is the most similar. There are many distance functions: the Manhattan distance, the Euclidean distance, the Gower distance (for the mixed case of categorical and numerical variables)...

Variants

- Define strata: a separate hot-deck is used for determined levels of common variables. For example, it is customary to perform a hot-deck for each gender, or for each age and gender intersection: the donor registers can be found within each stratum. By doing this, incoherent results are avoided and distance calculation becomes considerably simpler.

- Restrict the use of donors: In principle, the B registers can be used more than once as donors, which introduces the risk of altering the original distributions. To avoid this, *restricted hot-deck* methods have been developed that introduce the restriction of not using donors more than once.

- Instead of using one sole donor for each recipient, it is possible to use the more similar "*K*" registers. The imputed value in these cases is a combination of the aforementioned K values. Another variant is to take the registers that are at a determined distance, "*d*".

Calculation

- R > StatMatch > NND.hotdeck(), RANDwNND.hotdeck(), rankNND.hotdeck()

References

[1] M. D’Orazio et al. (2008). Annex C: Selection of distances (p 34-45)
[2] A. Leulescu et al. (2013).

Table 1. *Hot-deck methods family*

When it comes to matching independent surveys via hot-deck, as has been repeatedly demonstrated in literature, it is important to bear in mind that a situation referred to as *conditioned independence hypothesis* will be implicitly assumed.

By this hypothesis, any existing relation (albeit unobserved) between specific variables X and Y is gathered by partial (observed) relations between the variables (Z,X) and (Z,Y) As long as the search for donor-recipient pairs is exact (i.e. if for each recipient register, the algorithm is able to find an exact donor in all the common variables), according to the said hypothesis, then the hot-deck method will correctly reproduce a "real" relation (unobserved) between the specific variables (X and Y) In this ideal situation, the information supplied by the Z variables is sufficient to reproduce the relation between X and Y.

Nevertheless, this hypothesis is restrictive and is not always fulfilled. Let us take an example: survey "A" measures individual income, whilst another survey, "B", registers employment status ("unemployed", "employed", "inactive"). The age variable is used to match these surveys via hot-deck, whereby file A (recipient) is filled with the "employment status" variable from B (donor). Then, in the A file that is imputed, the income distribution that is conditioned by a specific employment status (e.g. inactive individuals) solely reflects that distribution's dependence on age. In other words: the file that is imputed in this form (i.e. using only the age variable) does not reflect all the dependence of income on inactive employment status, and the fact it is greater than the extent to which income it is determined by age.

In practice, we must try to move closer to this hypothesis, using all available information to do so. As we will see throughout this technical manual, the greatest challenge posed by statistical matching is looking for the best strategy for reaching the ideal situation of conditional independence.

Methods based on regression

In contrast to hot-deck methods, methods based on linear regression are purely parametric, i.e. they assume a specific statistical model for the variables. In this case, the hypothesis of conditional independence is understood as the fact that the function of joint distribution is the product of marginal distribution functions, which is:

$$f(x,y,z) = f(x|z) \times f(y|z) \times f(z)$$

(In other words: it is assumed that data in the partial files are sufficient to construct a file complete with all the variables).

Regression imputation uses the regression model to obtain predicted values for "missing" observations. In this way, a real (observed) value is not imputed, but rather an estimate based on common information is imputed. But this simple procedure presents disadvantages: sensitivity in the presence of an inadequate, specified model, regression to the average and the risk of underestimating variance (resulting from taking values situated on the line of regression).

A solution is stochastic regression imputation, which consists of introducing a random, residual value to reflect variance more adequately. This type of variation, as we will soon see, serves as a basis for developing mixed, more sophisticated methods.

Mixed methods

Mixed methods arise when combining the advantages of the two previous approaches: the non-parametric hot-deck methods, which are robust and specify no explicit model a priori; and the parametric methods based on regression, which are slower given they do not critically depend on the variables chosen to calculate distances.

Within this family, we must highlight the imputation method known as *predictive mean matching*, introduced by Rubin (1986). In this procedure, the "missing" values in the recipient file are imputed based on values predicted by a regression. More specifically: first of all, a regression from X to Y in donor file B is calculated. With this equation, a predicted mean value, $\hat{X}$, is calculated in recipient file A. Then, a hot-deck method based on distance $d(X,\hat{X})$ is used to look for donor-recipient pairs. Lastly, the values observed are imputed. In summary, *predictive mean matching* consists of a hot-deck imputation, but is based on mean values obtained by regression models.

Another mixed method worth highlighting is the *propensity score* (S. Rässler (2002), A. Leulescu et al. (2013). Both files, donor and recipient, "extend" with an additional variable of value 1 for all the registers of file A (the donor) and value 0 for all the registers of file B (recipient). By joining all the registers in a single file, a logit or probit model can be estimated by taking the additional added value as a dependent variable, and the common variables between files are taken as independent variables. Score propensity is defined as the estimated conditional probability of a unit belonging to one of the files. Finally, the match is performed by selecting the most similar donors according to propensity score values.

Methods based on multiple imputation

Multiple imputation was introduced by Rubin in the 1970s in the field of missing values, and has been often used in the context of statistical matching. The idea is to extract plausible m-values greater than 1 for each missing value (unobserved) instead of just one value, thereby reflecting the uncertainty over the said value The m values extracted are used to perform a pooling to produce a single value, as well as an estimate of uncertainty (or *intra-imputation variance*) regarding that value (A. Leulescu et al. 2013).

In the context of statistical matching, multiple imputation has generally been used to obtain files of complete data. But multiple imputation can be used in more complex contexts. This is the case for *sequential regression multiple imputation*, in which independent models are used to impute each variable through a series of iterations.

Although these advanced methods can be computationally costly, they potentially offer a great deal of flexibility. For instance, it would be possible to impute many variables at once. Today, it is possible to use these variable with relative ease thanks to the fact they are used in free R software packages, such as `mice`.. (For more information, consult the third section: Software).

Macro methods

The aim of macro methods is to obtain a direct estimate of any parameter of interest related to specific variables X and Y. To illustrate this type of procedure, we are going put ourselves in the hypothetical situation where we have two separate files referring to the same population.

- A: containing a simple, random sample with n_A observations of the variables Z and X.
- B: containing a simple, random sample with n_B observations of the variables Z and Y.

In the most simplified case, we are going to suppose that the trivariant distribution (unobserved) is a normal distribution with parameters:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_Z \\ \mu_X \\ \mu_Y \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_Z^2 & \sigma_{ZX} & \sigma_{ZY} \\ \sigma_{XZ} & \sigma_X^2 & \sigma_{XY} \\ \sigma_{YZ} & \sigma_{YX} & \sigma_Y^2 \end{pmatrix}$$

To characterise the joint distribution, we can use file A to estimate $\sigma_{ZX}$and file B to estimate σ_{ZY}. To estimate σ_Z^2, we can use the registers from A or the ones from B, or perhaps even better, the $n_A + n_B$ registers from the concatenated file, $A \cup B$.

If the objective is to obtain an estimate for σ_{XY} - and in the absence of an additional file, C, with observations of the joint distribution - it is necessary to introduce an additional hypothesis, such as the hypothesis of conditional independence. Using this hypothesis, the partial covariances would be enough to calculate σ_{XY}:

$$\sigma_{XY} = \frac{\sigma_{ZX}\,\sigma_{ZY}}{\sigma_X^2}$$

In the `StatMatch` package, the macro methods have been implemented in the functions `mixed.mtc()` and `comb.samples()`. The latter allows us to take into account the sample design for complex samples (see the following section).

Specific methods for complex sample designs

Frequently, files A and B that are to be matched correspond to complex samples, i.e. they come from surveys that do not perform simple, random sampling from units of the population.

There are various ways to introduce the sample design into procedures of statistical matching. Here, we highlight the Renssen procedure, implemented in the function `comb.samples()` from the R `StatMatch` package. This procedure involves a series of consecutive calibrations of the weights connected to the registers for each of the files, and that reflect the sample design. (In the terminology of complex sampling, the calibration is basically the recalculation of the weights in such a way that values are obtained that resemble as far as possible the theoretical design values, whilst also fulfilling a set of conditions. For instance, reproducing known reference totals on the population).

The Renssen procedure is generally used with categorical variables and with a macro objective (i.e. to estimate contingency table $X \times Y$of unobserved variables jointly). During the whole procedure, which comprises two phases, files A, B and auxiliary file C (where applicable) are kept separate.

In the first phase, the weights in A and in B (w_A and $w_B$respectively) are recalculated in order to obtain the totals for common variables X (well known or estimated using the same files, A and B) In the second phase, two cases are considered:

○ If there is an auxiliary file C with complete information, the weights in this file, w_C, are calibrated to align with the totals for A and B (after their respective calibrations in step 1), and after this, an estimate of the following is calculated: $X \times Y$

○ If there is no C file, the hypothesis of conditional independence is used to obtain an estimate.

For any reader who might be interested, the reference *"Old and new approaches in statistical matching when samples are drawn with complex survey designs"* (D'Orazio et al. , 2010) compares the Renssen calibration procedure with other similar methods.

Phases of statistical matching

Regardless of the method chosen, statistical matching involves following a series of phases that are closely connected to the development of a sample survey. It is important to keep in mind that the matching method is solely one of these phases, and it is often not the most important (A. Leulescu et al. 2013).

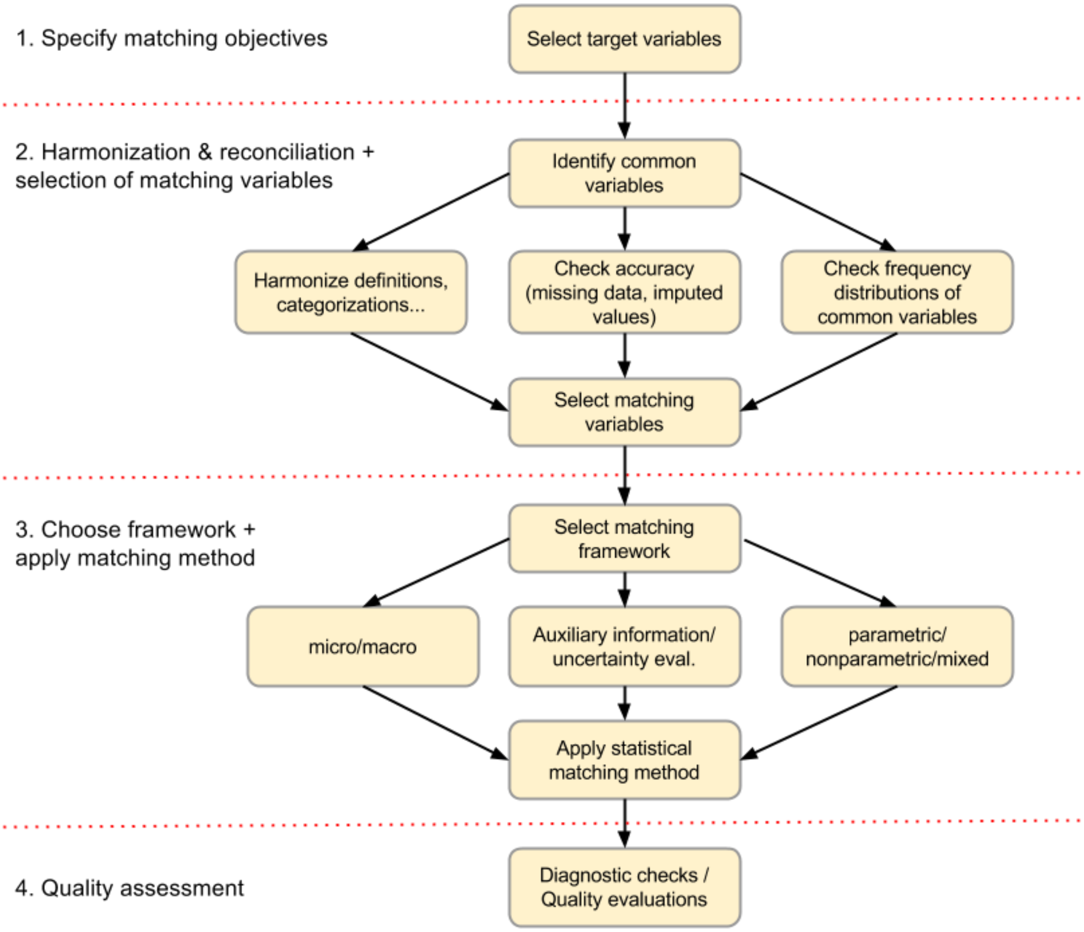

1. Specify matching objectives
Select target variables
2. Harmonization & reconciliation + selection of matching variables
Identify common variables
Harmonize definitions, categorizations...
Check accuracy (missing data, imputed values)
Check frequency distributions of common variables
Select matching variables
3. Choose framework + apply matching method
Select matching framework
micro/macro
Auxiliary information/ uncertainty eval.
parametric/ nonparametric/mixed
Apply statistical matching method
4. Quality assessment
Diagnostic checks / Quality evaluations

Figure 3. *Phases of statistical matching "Report WP2 ESS-net, Statistical Methodology Project on Integration of Surveys and Administrative Data.", page 34. Adapted for this manual.*

In the **first phase**, it is necessary to set an objective for the match: (i) It must be established whether the match will be micro or macro; and (ii) the specific variables to be matched must be fixed beforehand. These decisions are crucial since they will determine the subsequent phases.

The **second phases** involves tackling two main tasks Firstly, the coherence between the data and the samples must be studied. This entails performing an in-depth analysis of the rate of harmonisation and reconciliation between the sources - including a study of the following aspects as a minimum (M D'Orazio et al. 2008):

- Concordance in the definition of the units and in the reference period.
- Concordance of the common variables or measurements and their classifications (in the case of categorical variables).
- Total and partial nonresponse: treatment of the missing data.
- Error calculation (bias and precision of the samples)
- The (possible) processing of the original variables in the synthetic indicators.

There will usually be discrepancies in one or several of these points, primarily in the point that refers to the gathering of the data (measured variables with distinct categories). It is also to be hoped that treatment after the data is gathered (weight calibrations, calculation of synthetic indicators) is distinct for each sample. Taking all these aspects into account, by identifying an initial list of equivalent variables, or by using a comparable definition between the two sources, a comparative study of the empirical distributions observed in the two files will be performed, and those variables that cannot be harmonised will definitely be discarded.

After this, evaluate the extent to which the potential common variables selected provide relevant information for predicting target specific variables. Let's take a specific example: if we wanted to connect the self-perception of health (measured in the first survey) with the level of income (measured in the second), we would have to analyse what potential, common variables (age, gender, level of education) were closely linked (i.e. are predictive) to the aforementioned variables at the same time.

Ideally, a list of common variables should be chosen that contains a *high concordance rate* between the samples, such as the ones that *are predictive* for the specific variables of our objective. Additionally, it is recommended not to introduce redundant variables in the selection (for example, categorised age and uncategorised age).

Occasionally, in order to make use of all available information, it will be convenient to generate variables derived from the original ones.

The success of the match will largely depend on the quality of the selection of common variables. Hence, this phase ends up being especially critical. There are various statistical tools that may be of use for the best possible selection of variables for matching: see list in Table 2.

TOOLS FOR SELECTING VARIABLES

- Z: common variable
- X: specific variable, file A
- Y: specific variable, file B

1. Evaluating the concordance rate

   a) Categorical Z variables

   o Similarity index
   o Superposition index
   o Bhattacharyya coefficient
   o Hellinger distance
   o Graphical analysis: bar charts, pie charts...

   Calculation: R > StatMatch > comp.prop()

   b) Numerical Z variables

   o Descriptive statistics: minimum, maximum, medium, standard deviation, variation coefficient, percentiles
   o Graphic analysis: qqplots, histograms, density functions considered superposed

   Calculation: Multiple R packages, for example: Hmisc > describe()

2. Evaluation of the predictive value (relevance)

   a) constant or categorical ordinal X or Y, constant or categorical ordinal Z

   o Spearman correlation coefficient corrected

   b) constant or categorical ordinal X or Y, categorical nominal Z

   o Coefficient of determination corresponding to Eta-squared (related to the Kruskal-Wallis test)

   c) Categorical nominal X or Y, categorical nominal or ordinal Z

   o Association measurements based on the statistic Chi-squared, such as Cramer's V

   o Measurements based on the reduction of variance (reduction provided by variance) or entropy reduction ()

   Calculation: R > Hmisc > spearman2() or also StatMatch > pw.assoc()

3. Redundancy evaluation

   a. Redundancy analysis to discard predictors that provide similar information

b. Exploratory methods based on the clustering of variables.

Calculation: R > Hmisc > redun(), varclus()

4. Multivariant methods

a. Generic statistical methods: regression analysis for constant variables, classification and regression trees (C&RT) Use with care.

Calculation: R > randomForest

b. Specific methods for statistical matching: select variables that reduce uncertainty the most in estimating the parameter of joint distribution.

When the variables are categorical, the *Frèchet bounds* can be used.

Calculation: R > StatMatch > Fbwidths.by.x()

References

A. Leulescu et al. (2013)
A. Agresti (2014).

Table 2. *Tools for selecting variables*

In the **third phase**, an appropriate matching method must be chosen to obtain a fused synthetic file (if micro), or an estimate of any statistical parameter of interest (if macro). (These methods have been examined in the previous section, “Approaches and methods”) The choice of the method depends on the kind of information available: for example, if an auxiliary file, C, is available, which contains full information (about a survey carried out years ago on a similar population), then this information can be used to improve the result of the match.

The **fourth and final phase** involves validating the results so as to ensure that the fused file is applicable. Given its importance, this phase is discussed in detail in the next section.

Validity of the results

When it comes to evaluating the validity of a statistical match (as to the degree of applicability[3]; the last phase of Figure 2), we must take into account all phases of the

[3] Here, we are talking about *validity*, not *efficiency*: no criterion that resembles the criterion of the tiniest average quadratic error is being used in the way it is used in other statistical fields; instead, the objective here is to assess different levels of reproduction and preservation of the original distributions and associations.

process, especially: the quality and coherence of the original sources, the assumptions made regarding the conditional distributions (i.e. assume the hypothesis of conditional of conditional independence), and the matching method itself.

S. Rässler (2002) established four levels to evaluate the validity of a match systematically. Before that, and in order to make it easier to outline these concepts, we are going to assume that we have obtained a fused file (via the micro approach). These considerations apply analogously to the macro approach.

Validity levels

Level 1: Preserving individual values

After the match, the real values (unobserved) of the Y variables are precisely reproduced in the recipient file: $y_i = \hat{y}_i$ For $i = 1,2,\dots,n_A$, where y_i are the real (unobserved) values, and where $\hat{y}_i$ are the imputed values, with n_A being the size of the recipient file A. The aim is to calculate the number of times that the imputed value matches the "real" value so as to calculate a "hit rate".

Level 2: Preserving joint distributions

After matching, the real (unobserved), joint distribution of the three combinations for the variables X, Y and Z is reflected correctly in the imputed file. So: $f(X,Y,Z) = \hat{f}(X,Y,Z)$ where f denotes the observed joint distribution and $\hat{f}$ denotes the distribution obtained in the imputed file.

Level 3: Preserving the structure of correlations

After the match, the fused file preserves the structure of correlations and higher-order moments. So: $cov(X,Y,Z) = \widehat{cov}(X,Y,Z)$where cov denotes the matrix of variances-covariances observed, and $\widehat{cov}$ denotes the matrix calculated in the fused file.

Level 4: Preserving marginal distributions

After matching, the marginal joint distributions that are observed in the donor file can be reproduced correctly in the imputed file. In particular, the following occurs: $f(Y) = \hat{f}(Y)$ and $f(Y|Z) = \hat{f}(Y|Z)$, where f denotes the marginal distributions observed and that are real, with $\hat{f}$ denoting the imputed distributions.

Discussion

Given that there will generally be no "real" Y values in the imputed file (indeed, it would make no sense to establish a match), the **first validity level** is, in general, not analysed. In fact, this level is only worth analysing under a simulation analysis: a file A is divided in two artificially in such a way that in one of them, a series of "Y variables" is removed , whilst the Z variables are kept. Then, the variables that have been removed "are recovered" by the match, and a "hit rate" is calculated.

In a real case (i.e. not simulated), it will generally be desirable to obtain a valid synthetic file to perform combined statistical analyses, and in this sense, the second, third and fourth levels are more relevant.

The **second level** (preserving the joint distribution) and the **third level** (preserving the correlation structure or, more generally, the higher-order moments) guarantee the validity of the synthetic file as to its ability to reflect adequate statistics concerning the real relation that variables in the population have. However, these levels are also not directly comparable and for the same reason: by definition, distribution form $f(X,Y,Z)$ and the correlation structure are unknown.

In fact, only the **fourth level** can be directly contrasted in practice. Carrying out this level ensures that the marginal distributions observed in the donor file are reproduced correctly in the recipient file. If robust procedures and quality data is always used, it is relatively simple to reach this level; indeed, it is the minimum requirement for any matching practice (A. Leulescu et al. , 2013).

As we have already shown, merely creating a synthetic file that reaches the fourth validity level does automatically mean that this file will accurately reflect relations between the unobserved files jointly. Hence, an additional effort is required that will depend on whether auxiliary information is available. If this is the case, the additional information will have to be integrated in the match to increase the validity of the synthetic file. If this is not the case, it will be advisable to perform an uncertainty analysis.

Bearing in mind the relevance of auxiliary information when it comes to improving the results of a match, we will now concentrate on a specific section.

Integration of auxiliary information

We have shown that the use of auxiliary information can considerably improve the results of a match. For instance, in M. D'Orazio et al. (2008), monthly deciles for net income are used to improve certain, more detailed estimates of income and expenses.

Auxiliary information can come from varied sources:

a. A file of additional data, C, with observations of (X, Y, Z), possibly from previous years or from other independent sources

b. Parametric auxiliary information in the form of an independent external estimate

c. A priori information on the phenomenon being studied. (The typical case is that of logical restrictions placed on the values that the variables can take.)

In StatMatch, there are various functions that allow you to introduce auxiliary information. For instance, the function comb.samples () is designed to calculate contingency tables for categorical variables observed in separate files. The information of an auxiliary file, C, with joint observations, can be introduced via the parameter svy.C to improve estimates.

Anther function of the same package is mixed.mtc (), which admits parameter rho.yz[4] so as to provide an (external) estimate a priori of the correlation between variables that have not been jointly observed.

Uncertainty analysis

If there is no auxiliary information, it is advisable to use methods to estimate the uncertainty of a match.

In the context that concerns us, the word *uncertainty* refers to vagueness due to the fact there is a possible rank of values compatible with the data observed, for those relations between variables that have not been jointly observed (which is also known as a *lack of identification*).

As we have previously seen, the closer the relation between specific and common variables in each of the files, the lesser degree of uncertainty there will be in the match. There are several alternatives for assessing the uncertainty in a match:

- In the case of categorical variables, it is possible to calculate the *Frèchet bounds*, which provide lower and higher levels for the cells in contingency tables. The interval between the levels contains all the values that are compatible with the data observed. In StatMatch, this calculation is performed in the function Frechet.bounds.cat()

- Multiple imputation is the natural context for assessing uncertainty. Effectively, using this tool, especially in a Bayesian[5] context, it is possible to analyse the sensitivity of the results towards different initial hypotheses regarding the conditional relations at Z. This route can be explored with the help of the mice package (Buuren, S. and Groothuis-Oudshoorn, K. 2011).

[4] The notation of this manual does not match the StatMatch notation: Z denotes the common variable, X, and Y denotes the specific variables.

[5] The book by S. Rässler (2002) deals extensively with this topic.

Chapter

3

Software

Many of the methods presented in this manual have been implemented over many packages in the R statistical computation environment. Some of these packages are aimed at the world of official statistics (the case with `StatMatch`) or at the world of survey data analysis (the case with `survey`), others (such as `Hmisc`) are generic and offer multiple functions for analysing data.

Table 3 contains a non-exhaustive list of available R packages with various functions for matching independent surveys.

R PACKAGES FOR MATCHING INDEPENDENT SURVEYS

- `StatMatch`**:** by Marcello d'Orazio (ISTAT), issued partly as the result of two projects on the integration of data performed within the European Statistical System, see ESSnet[6]. This package is specifically aimed at matching and imputing data from independent surveys. It provides functions covering various phases of statistical matching, primarily:

    - non-parametric methods of hot-deck imputation,
    - mixed methods based on predictive mean matching,
    - methods for dealing with complex samples,
    - methods for exploring uncertainty in the context of a match.

- `survey`: by Thomas Lumley. This contains a wide variety of tools for analysing data from complex samples: descriptive statistics, tests, generalised linear models, Cox models, factorial analyses and primary components etc.

- `Hmisc`: *Harrell Miscellaneous* by Frank E Harrell Jr, with contributions from Charles Dupont. This contains functions covering various aspects of data analysis: advanced graphs, tabular creation, *clustering* of variables, manipulation of character variables, recoding of variables.

- `mice`: *Multiple Imputations via Chained Equations*, by Stef van Buuren. This implements multiple imputation based on Fully Conditional Specification (FCS), implemented by the MICE algorithm. The idea is that each variable is assigned its own imputation model. The package provides models for constant variables (*predictive mean matching*, normal), dichotomous variables (logistic regression), categorical variables in no order (*multinomial logistic regression*) and ordinal categorical variables

[6] The *Data Integration* projects (12/2009-12/2011) and ISAD: *Integration of Survey and Administrative Data* (12/2006-06/2008), both fronted by ISTAT.

(proportional *odds*). The package also offers diagnostic graphs for inspecting the results of the imputations.

- Amelia: *Amelia II: A Program for Missing Data*, by James Honaker, Gary King and Matthew Blackwell. This contains functions to perform multiple imputation of surveys, and is performed on an algorithm based on the *bootstrap* technique, which was created by the same authors. It is more advanced than other similar solutions and it allows you to deal with several variables at once. It contains a GUI or a Graphic User Interface that can be employed by users who are not dealing with R.

- BaBooN: *Bayesian Bootstrap Predictive Mean Matching - Multiple and single imputation for discrete data*, by Florian Meinfelder. This contains two versions of the algorithm *Bayesian Bootstrap Predictive Mean Matching* for multiple imputation of missing data. It is advisable to use the second variant for situations such as statistical matching (or data fusion), or situations in general in which the different variables show the same pattern of missing data.

References

StatMatch
Marcello D'Orazio (2013). StatMatch: Statistical Matching (aka data fusion). http://CRAN.R-project.org/package=StatMatch

D'Orazio, M. (2013). Statistical Matching and Imputation of Survey Data with StatMatch: StatMatch drawing.

survey
Thomas Lumley (2012) survey: analysis of complex survey samples. http://CRAN.R-project.org/package=survey

mice
Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL http://www.jstatsoft.org/v45/i03/
www.multiple-imputation.com

Hmisc
Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2014). Hmisc: Harrell Miscellaneous. R package version 3.14-3. http://CRAN.R-project.org/package=Hmisc

Amelia
James Honaker, Gary King, Matthew Blackwell (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), 1-47. URL http://www.jstatsoft.org/v45/i07/.

BaBooN
Florian Meinfelder (2011). BaBooN: Bayesian Bootstrap Predictive Mean Matching – Multiple and single imputation for discrete data. http://CRAN.R-project.org/package=BaBooN

Table 3. *R packages for matching independent surveys.*

In chapter 5: opment][Ref] Development of a suitable R package, a proper package, micromatch, is described. It is based on previous packages and has been in development throughout this very project, and it is also available on Eustat's website.

Capítulo
4

Practical application

Statistical matching of EUSTAT sample surveys: the Living Conditions Survey and the Population Survey In Relation to Activity

The goal of this study is to test techniques presented in the chapter on methodology (page 6) in the real case of two of EUSTAT's independent sample surveys: The Living Conditions Survey (2009) and the Survey on Population in Relation to Activity (4th Quarter of 2009).

Using the most recent available data from these surveys, we will illustrate step-by-step the main phases of the match, from the definition of the aims, then onto the analysis of the coherence between sources and the selection of matching variables, finishing with a brief validation and presentation of the results. All the calculations shown have been carried out with free R software.

Description of the surveys

The **Survey on Population in Relation to Activity** (hereinafter referred to as PRA) is a constant, quarterly panel that EUSTAT has been running since 1985. The goal of which is to become familiar with the characteristics and the dynamics of the labour market in the Basque Country. The PRA carries out probability sampling on a panel of dwellings, and is carried out quarterly (see in Bibliography: methodology file). The current sample amounts to approximately 5,000 dwellings (which affects a total of about 13,500 individuals), and with a rotation of eight times a quarter.

The survey has two main aims:

● Obtaining constant statistical information about the volume and characteristics of the main groups, which can in turn be used to classify the population of the Basque Country according to participation in distinct economic activities, as well as their situation changes.

● Obtaining statistical information about the main demographic and social characteristics of this population, as well as the degree of participation in activities that are not economically productive.

One of the principal results of the PRA is a population classification that is more detailed in relation to that population's activity than the basic distinction between people who are Employed, Inactive and Unemployed (see Table 1).

PRA SYSTEM OF CLASSIFICATION

→ **Population engaged in activity:** people who are carrying out activities to produce goods and services, people who are involved in performing domestic chores, people who are studying at university and people who are doing military service. These people can be divided into two groups:

> ═ **Population engaged in work activity:** people who are working. The notion of work applies to all activity carried out with remuneration or benefit, that is, all remunerated work in the context of an employer-employee relationship or all independent work. It can equally be applied to an unpaid job in the family (family support).
>
> ═ **Population engaged in non-work activity:** This includes those people who do not have a job and who are not at the service of an employer, and who carry out household tasks, are studying or doing military service. They are classified according to whether or not they are seeking a work activity and by their subjective availability to incorporate themselves into it. There are two distinct subcategories:
>
> > ≡ Population engaged in non-work activities: people who are not seeking employment and dedicate themselves exclusively to carrying out household tasks, studying or doing military service.
> >
> > ≡ The rest: people engaged in non-work activity who are looking for work

→ **Population not engaged in activity:** this is divided into two subsections:.

> ≡ **Population strictly unemployed:** people looking for work who are available to occupy a job post immediately.
>
> • **Population strictly retired and others:** People who, due to their age or physical situation, are not engaged in any activity (pensioners, those unfit for work, etc.).

Table 1. *Division of the population in relation to the type of activity as recorded in the PRA survey. The data corresponding to people who have participated in the questionnaire is extended to the population as a whole.*

The **Survey on Living Conditions** (hereinafter, ECV) is sample survey that EUSTAT has carried out every five years since 1989, the goal of which being to provide up-to-date information about family living conditions, individual living conditions and the living conditions the whole of the Basque Country find themselves in.

The ECV uses two types of questionnaire (an individual one and a familial one) and is based on a random, stratified sample split into two stages (see in Bibliography: methodology file). In the first stage, the dwellings in a stratum (geographical area) are selected, on which questions will be answered in the family questionnaire. In the second stage, one person in the dwelling is randomly selected, who will answer the individual questionnaire. The initial sample size is 7,500 dwellings.

The ECV pursues three specific objectives:

1. Becoming familiar with conditions relating to health, education, free time and social relations between individuals

2. Describing the state of the physical and social environment in the area or zone where people live

3. Analysing family relations and family economic resources, as well as their household amenities.

ECV-PRA match

In the following sections, there is a description of each phase involved in matching the EUSTAT ECV and PRA surveys, following the diagram in Figure 3 (section **¡Error! No se encuentra el origen de la referencia.**).

Available data and reference population

At the time this technical manual was drawn up, the most recent available, published data from the ECV corresponds to the final quarter of 2009. Hence, in order to perform the match, PRA data is being accordingly used that corresponds to that period.

In both surveys, people aged 16 or above have been chosen. So, our reference population is people aged 16 and above who were residents of the Basque Country during the last quarter of 2009. The available sample consists of 12,658 observations in the PRA and 5242 in the ECV. Once people under 16 are omitted, the total of observations decreases to 10,865 and 4,749 respectively.

Phase 1: Set an objective for the match

The objective of this study is to assess the possibility of providing integrated statistics that combine aspects related to living conditions and lifestyle to the labour market. To do this, the basis will be information provided independently by the PRA and ECV surveys.

To be precise, the goal is to obtain a synthetic file that combines variables from both surveys. To this end, a hot-deck imputation will be performed in which the ECV survey acts as the recipient survey, with the PRA as the donor survey. The primary variable of the PRA - the division we showed in Table 1 - is imputed into the ECV survey file.

So, the primary PRA variable, specifically the division of the population according to activity type (Table 1) is added to the ECV survey (with all the items that record different aspects such as level of education, health status, social relations, environment and economic situation...). This synthetic file will provide the opportunity to analyse distinct living conditions according to the division of the labour market. This is something that can not be performed directly given the fact the variables correspond to independent surveys (files).

To simplify the objective of this study, 6 specific variables from the ECV have been selected that cover the main dimensions recorded by this survey, see Table 2. (In the Annex: Index A includes a table that provides the origin and treatment of these variables in relation to the microdata files.)

SPECIFIC ECV VARIABLES (Sample)

- Health problems: {1-Health problem present; 2-No health problem}
- Languages spoken: {1-Spanish only; 2-Spanish and others; 3-Spanish and Basque; 4-Spanish, Basque and others}
- Free time: {1-Less than 2 hours; 2-2-4 hours; 3-More than 4 hours}
- Objective economic situation: {1-Poor; 2-Normal; 3-Good}
- Vehicle ownership {1-None; 2-One; 3-Two or more}
- Household amenities[a] {1-Limited amenities; 2-Sufficient amenities}

[a] Aggregate levels

Table 2. *Sample of specific variables from the ECV, alongside their categories. See Annex: File A.*

Phase 2: Select matching variables

In this phase, the objective is to select an optimum subset of variables (designated common or matching variables) between all those variables (original or derived) that are shared between the ECV and PRA surveys.

Phase 2-1: Meta-analysis of the questionnaires

In order to identify all the variables that (potentially) contain the same information, a meta-analysis of the questionnaires is performed. Considering that the objective is to try to impute the PRA division within the ECV survey, as well as the usual socio-demographic variables such as age, gender or family size, we are specifically looking for variables that give information on work activity. (These variables are known as proxy variables or "common-specific" variables). This is a viable option in this case seeing as there is a "Work conditions" section in the ECV survey, from which some indicators can be extracted.

As a result, the following common variables have been identified between ECV and PRA during the 4th quarter in 2009:

Social-demographic variables

- Age
- Gender
- Family size

Variables related to level of education

- People undertaking accredited studies (hereinafter, "Student Y/N")
- People undertaking distance studies
- Illiterate people (those who cannot read or write)

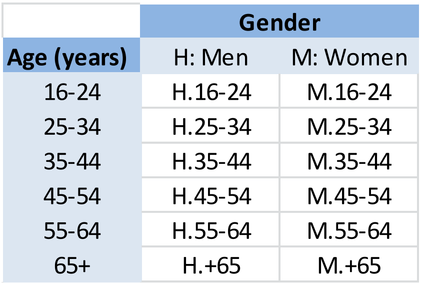
Variables associated with the relation to activity

- Identification of unemployed/employed/inactive individuals
- Hours worked by employed individuals
- People looking for work (hereinafter "Looking for work Y/N")
- Devotion to household chores

In order for the reader to trace the origins of this information and its subsequent treatment, the codes of the variables in the microdata files, as well as the levels of aggregation, have been included in the Annex: File B.

Phase 2-2: Coherence study

After the meta-analysis, a coherence study based on the marginal distributions observed was performed for the variables identified in both questionnaires. To begin with, two variables were discarded: "Illiterate people" and "People undertaking distance studies" due to the probability of occurrence being too low. The coherence of marginal distributions has been analysed both in a global sense (without taking into account other variables) and by groups defined by twelve clear-cut strata or the intersections of the age and gender variables, see Table 3.



	Gender	
Age (years)	H: Men	M: Women
16-24	H.16-24	M.16-24
25-34	H.25-34	M.25-34
35-44	H.35-44	M.35-44
45-54	H.45-54	M.45-54
55-64	H.55-64	M.55-64
65+	H.+65	M.+65

Table 3. *Coding for the Age and Gender strata used in this study.*

Distinguishing the strata is essential because, for most variables, the implications are distinct within each age and gender group. For instance, for the strata M +65 (male) and F +65 (female), the variable "Student Y/N" provides no information, and so should not be included in the match; conversely, in the strata M. 16-24 and F. 16-24 , this variable is indispensable.

In accordance with the recommendations of A. Leulescu et al (2013), in order to compare the distributions observed, a series of empirical measurements have been utilised. In this study, the Hellinger distance has been used, which takes values between 0 (equal distributions) and 1 (maximum possible dissimilarity). The results - in general and by sex and age group - are shown in the Annex: File C.

Phase 2-3: Predictive value study

To complete the selection of variables, the predictive value of the variables with regards to the specific variables has been studied. The idea is to select those variables that provide valuable information for performing the match.

Just like the coherence study, the predictive capacity of the common variables has been analysed in a global sense (including all the observations) as much as it has been analysed within twelve strata defined by the variables of age and gender. As in the case for ECV-PRA, all the variables are categorical, and association measurements based on the Chi-squared statistic have been used, such as Cramer's V. Again, the results are shown in the Annex: File C.

Phase 3: Apply a matching method

Lastly, a hot-deck method has been used for each stratum as per the selected variables. We will now see an illustration of the procedure for one of the strata: men aged between 25 and 34 (inclusive).

Example: Hot-deck method within the stratum W. 25-34

For this segment, the **common variables selected** are "Looking for work Y/N" and "Devotion to household chores". The variable "Family Size", though concordant, provides no information in this segment seeing as nearly all the families only have one member (FS = 1). There is not sufficient concordance in the variables "Employed Individuals" and "Inactive Individuals" and the variable "Unemployed Individuals" is not relevant to job seeking.

The recipient file contains 359 registers, the donor file 746. For each recipient file (i.e. any man aged between 25 and 34 who has responded to the ECV survey), a donor register (i.e. any man aged between 25 and 34 who has responded to the PRA survey) is sought that is the most similar in terms of variables selected. The code in `R > StatMatch` can be seen in Table 4:

EXAMPLE: HOT-DECK IMPUTATION IN `R > StatMatch`

# Step 1 - Look for donor-recipient pairs

```
out.nnd <- NND.hotdeck(data.rec = rec, data.don = don,
        dist.fun = “Gower”, match.vars = c(“BUSQ”,”DOM”),
        constrained = TRUE)
```

# Step 2 - Impute the recipient file (i.e. `ecv` filtered with the selected stratum)

```
fecv <- create.fused(data.rec  =  rec, data.don = don,
        mtc.ids = out.nnd$mtc.ids, z.vars = “PRA”)
```

where:

- `rec`: file with registers filtered from the ECV (men aged between 25 and 34).
- `don`: file with registers filtered from the PARA (idem).
- `match.vars`: list of common variables selected (in the example: BUSQ, DOM).
- `z.vars`: list of specific variables, in this case it is the only one, "PRA"
- `NND.hotdeck()`: Search for similar donors
- `dist.fun = “Gower”`: the Gower distance is used. (See package documentation).
- `constrained = TRUE`: indicates that the algorithm is restricted i.e. each donor register is used only once
- `create.fused()`: This function generates the ecv file that is enlarged with the imputed values
- `mtc.ids:` This contains the donor-recipient correspondence for generating the fused file
- z.vars: Imputed variables (in this case, just one: the "PRA" division)

Result

`fecv`: Initial `rec` file enlarged with the PRA variable.

Table 4. *Example of hot-deck imputation for the ECV-PRA match, stratum M. 25-34: (Men aged between 25 and 34).*

Phase 4: Assess the quality of the results

Lastly, the results are assessed, which involves comparing the marginal distributions observed with the imputed distributions (fourth validity level). The global results can be viewed in Figures 4-1 (global results) and 4-1 (results by Age and Gender).

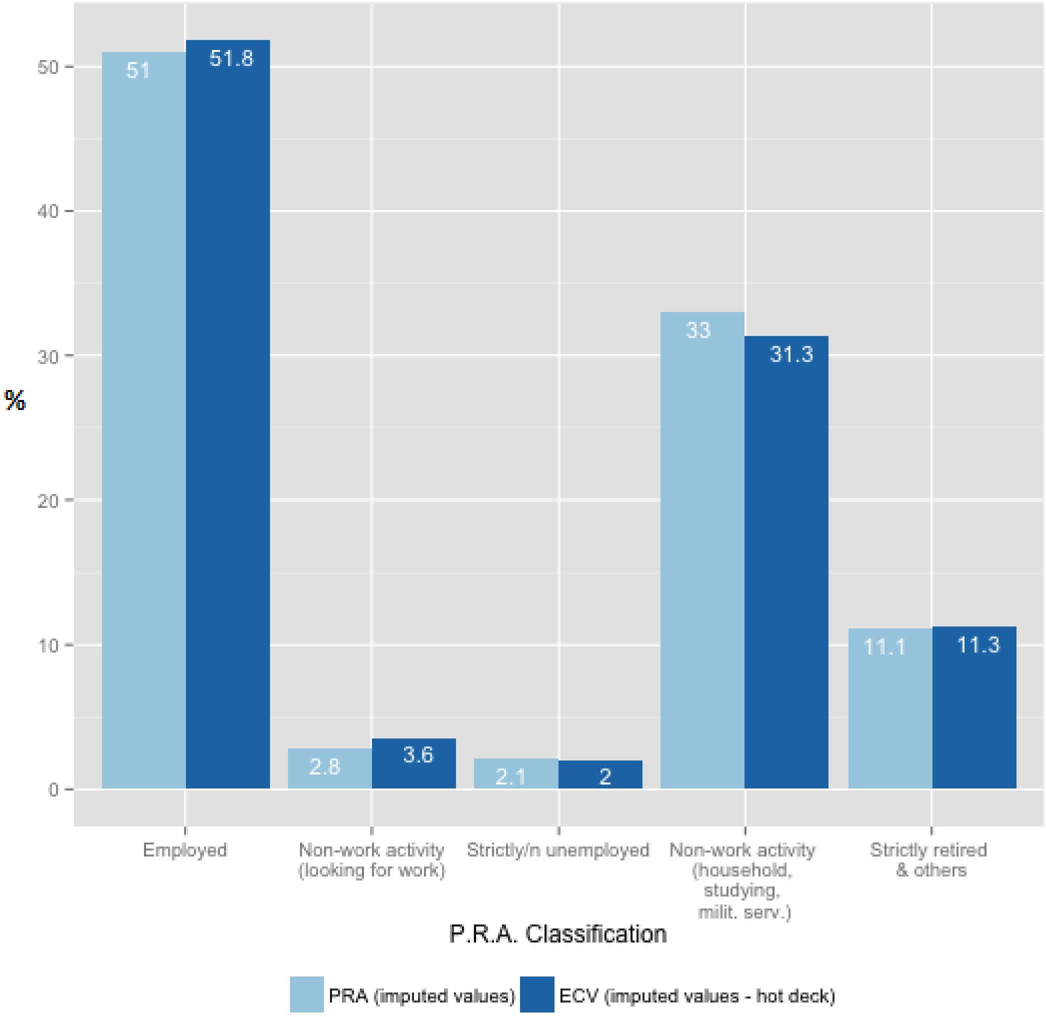


Figure 4-1. *Results after the ECV-PRA match. Real, observed marginal distribution (PRA survey, donor) compared with the imputed distribution (ECV survey, recipient).*

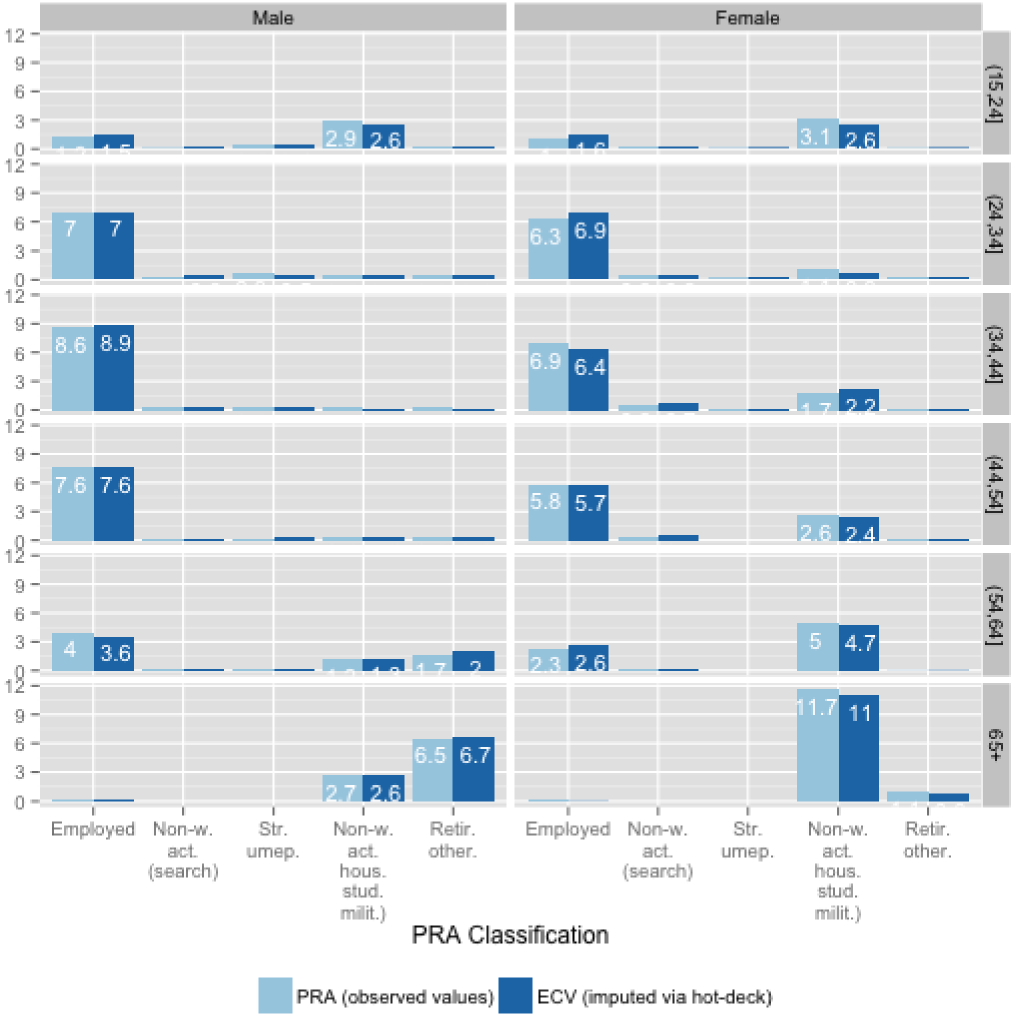


Figure 4-2. *Results after the ECV-PRA match by Gender and Age Groups. Real, observed marginal distribution (PRA survey, donor) compared with the imputed distribution (ECV survey, recipient).*

Results

For the objectives set and for the sample of specific variables selected, the result is a series of contingency tables that we will show next. The results are presented for both the total population and for determined age of gender strata that we have selected for illustrative purposes.

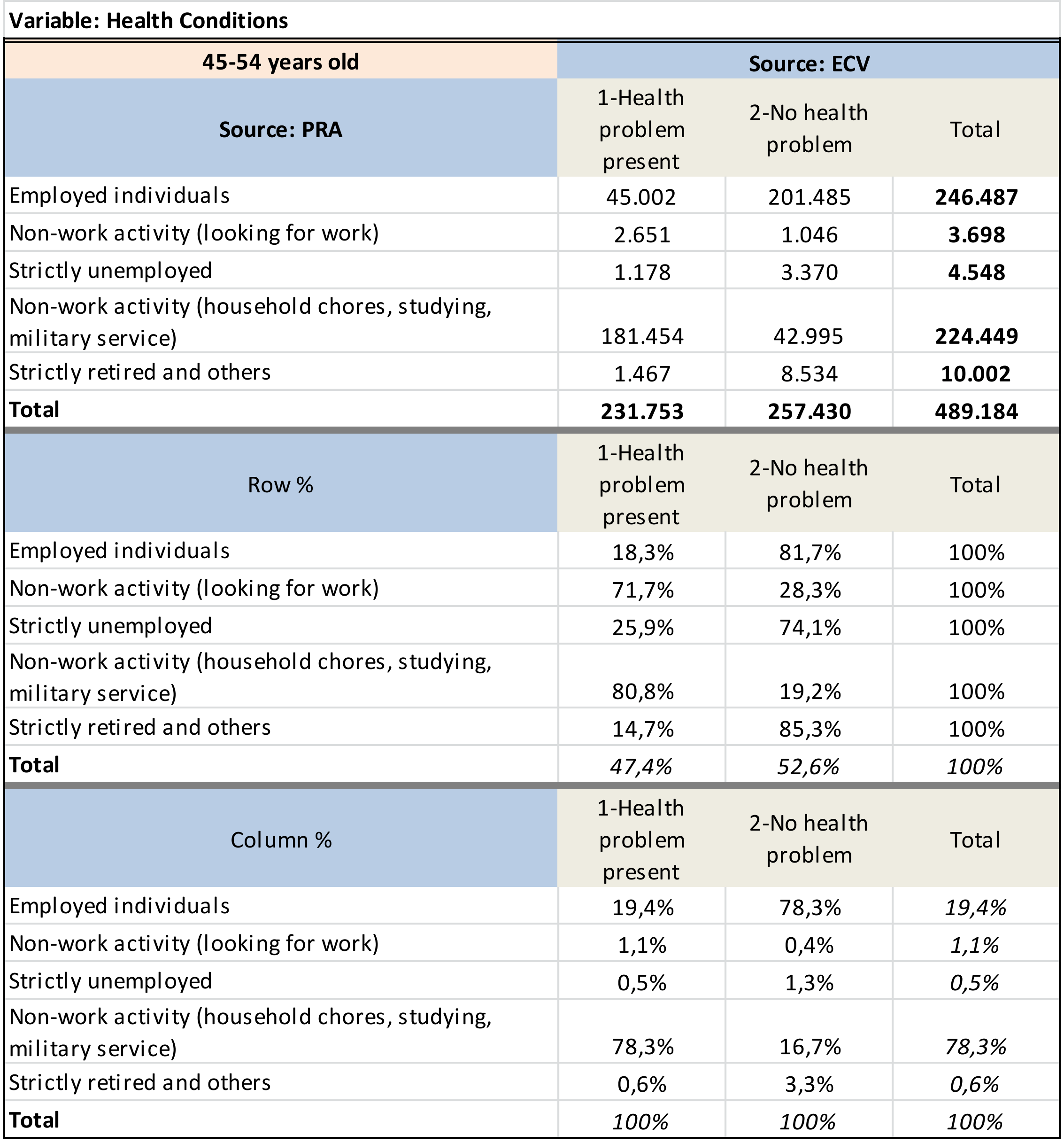
The interest of these tables lies in the fact they allow us to explore the living conditions (provided by the ECV) according to the division of the labour market (provided by the PRA). These variables were initially located in separate files, and, without additional information, it is only possible to "intersect them" using a statistical matching technique. The strategy followed here has involved imputing the main variable from the PRA (the donor) into the ECV survey (the recipient) by means of hot-deck imputation according to age and gender strata.

Contingency tables for exploring items from the Survey on Living Conditions (origin: ECV) according to the Relation to Activity (origin: PRA).

Variable: Health Conditions

Source: PRA	**Source: ECV**		
	1-Health problem present	2-No health problem	Total
Employed individuals	154.204	806.109	**960.312**
Non-work activity (looking for work)	14.633	51.196	**65.830**
Strictly unemployed	7.553	30.079	**37.633**
Non-work activity (household chores, studying, military service)	181.454	399.219	**580.673**
Strictly retired and others	86.556	121.987	**208.543**
Total	**444.400**	**1.408.590**	**1.852.991**
Row %	1-Health problem present	2-No health problem	Total
Employed individuals	16,1%	83,9%	100%
Non-work activity (looking for work)	22,2%	77,8%	100%
Strictly unemployed	20,1%	79,9%	100%
Non-work activity (household chores, studying, military service)	31,2%	68,8%	100%
Strictly retired and others	41,5%	58,5%	100%
Total	*24,0%*	*76,0%*	*100%*
Column %	1-Health problem present	2-No health problem	Total
Employed individuals	34,7%	57,2%	*51,8%*
Non-work activity (looking for work)	3,3%	3,6%	*3,6%*
Strictly unemployed	1,7%	2,1%	*2,0%*
Non-work activity (household chores, studying, military service)	40,8%	28,3%	*31,3%*
Strictly retired and others	19,5%	8,7%	*11,3%*
Total	*100%*	*100%*	*100%*

Result 1-1. *Health conditions (Source: ECV) vs PRA Division (Source: PRA). Totals for the population as a whole, and row and column percentages.*



Variable: Health Conditions

45-54 years old	**Source: ECV**		
Source: PRA	1-Health problem present	2-No health problem	Total
Employed individuals	45.002	201.485	**246.487**
Non-work activity (looking for work)	2.651	1.046	**3.698**
Strictly unemployed	1.178	3.370	**4.548**
Non-work activity (household chores, studying, military service)	181.454	42.995	**224.449**
Strictly retired and others	1.467	8.534	**10.002**
Total	**231.753**	**257.430**	**489.184**
Row %	1-Health problem present	2-No health problem	Total
Employed individuals	18,3%	81,7%	100%
Non-work activity (looking for work)	71,7%	28,3%	100%
Strictly unemployed	25,9%	74,1%	100%
Non-work activity (household chores, studying, military service)	80,8%	19,2%	100%
Strictly retired and others	14,7%	85,3%	100%
Total	*47,4%*	*52,6%*	*100%*
Column %	1-Health problem present	2-No health problem	Total
Employed individuals	19,4%	78,3%	*19,4%*
Non-work activity (looking for work)	1,1%	0,4%	*1,1%*
Strictly unemployed	0,5%	1,3%	*0,5%*
Non-work activity (household chores, studying, military service)	78,3%	16,7%	*78,3%*
Strictly retired and others	0,6%	3,3%	*0,6%*
Total	*100%*	*100%*	*100%*

Result 1-2. *Health conditions (Source: ECV) vs PRA Division (Source: PRA). Stratum selected: Individuals aged between 45 and 54 (inclusive). Totals and percentages of rows and columns*

Variable: Languages known

Source: PRA	**Source: ECV**				
	1- Spanish only	2- Spanish and others	3- Spanish and Basque	4- Spanish, Basque and others	Total
Employed individuals	310.088	181.823	165.623	302.778	**960.312**
Non-work activity (looking for work)	19.586	17.525	10.651	18.067	**65.830**
Strictly unemployed	10.867	4.874	4.762	17.129	**37.633**
Non-work activity (household chores, studying, military service)	265.572	55.413	112.938	146.750	**580.673**
Strictly retired and others	113.300	27.900	46.692	20.651	**208.543**
Total	**719.413**	**287.535**	**340.666**	**505.375**	**1.852.992**
Row %	1- Spanish only	2- Spanish and others	3- Spanish and Basque	4- Spanish, Basque and others	Total
Employed individuals	32,3%	18,9%	17,2%	31,5%	100%
Non-work activity (looking for work)	29,8%	26,6%	16,2%	27,4%	100%
Strictly unemployed	28,9%	13,0%	12,7%	45,5%	100%
Non-work activity (household chores, studying, military service)	45,7%	9,5%	19,4%	25,3%	100%
Strictly retired and others	54,3%	13,4%	22,4%	9,9%	100%
Total	*38,8%*	*15,5%*	*18,4%*	*27,3%*	*100%*
Column %	1- Spanish only	2- Spanish and others	3- Spanish and Basque	4- Spanish, Basque and others	Total
Employed individuals	43,1%	63,2%	48,6%	59,9%	*51,8%*
Non-work activity (looking for work)	2,7%	6,1%	3,1%	3,6%	*3,6%*
Strictly unemployed	1,5%	1,7%	1,4%	3,4%	*2,0%*
Non-work activity (household chores, studying, military service)	36,9%	19,3%	33,2%	29,0%	*31,3%*
Strictly retired and others	15,7%	9,7%	13,7%	4,1%	*11,3%*
Total	*100%*	*100%*	*100%*	*100%*	*100%*

Result 2-1. *Knowledge of Languages (Source: ECV) vs PRA Division (Source: PRA) . Totals for the population as a whole, and row and column percentages.*

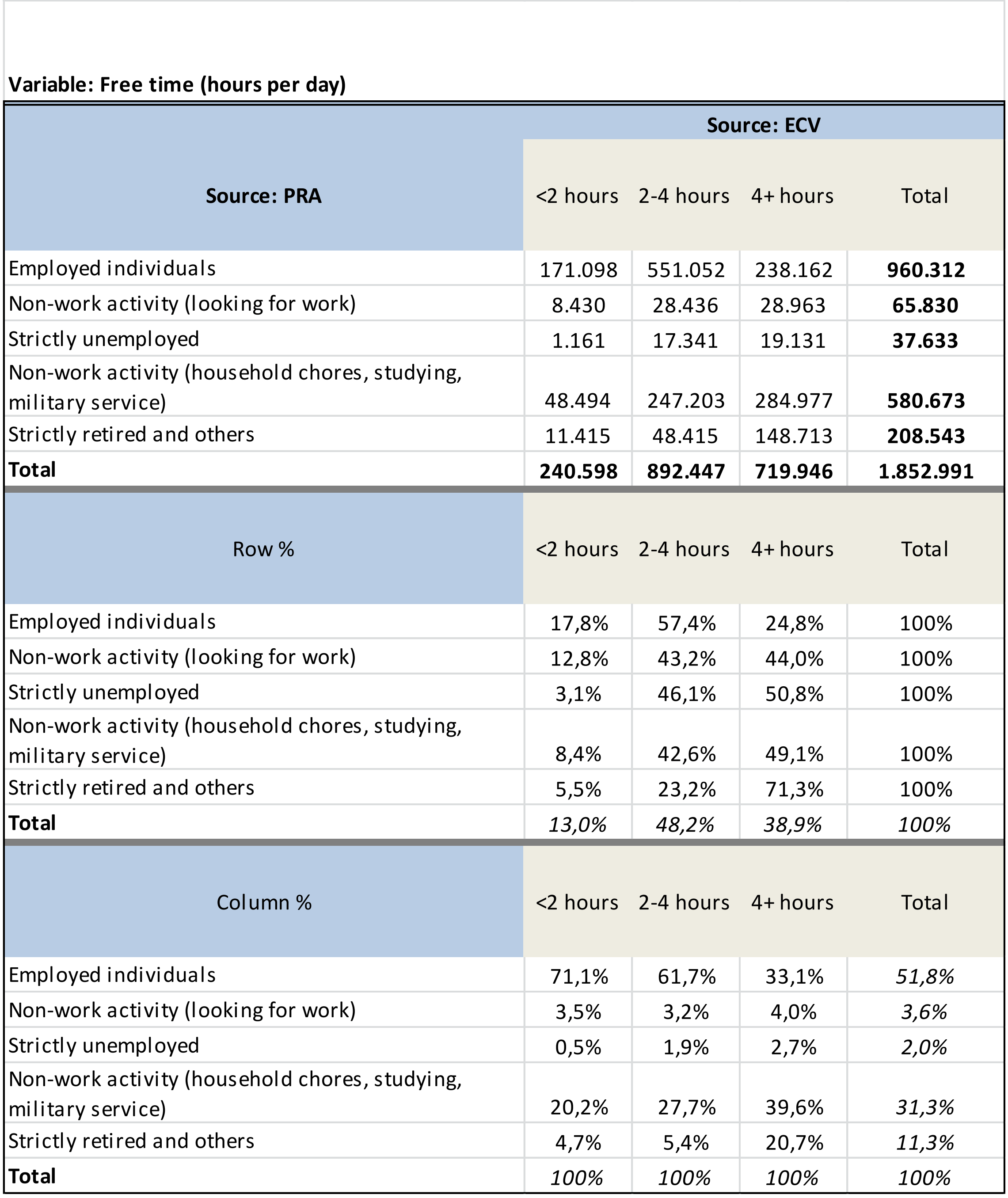
Variable: Languages known

35-44 years old	**Source: ECV**				
Source: PRA	1- Spanish only	2- Spanish and others	3- Spanish and Basque	4- Spanish, Basque and others	Total
Employed individuals	76.100	59.203	50.488	96.088	**281.879**
Non-work activity (looking for work)	7.081	4.679	4.096	3.464	**19.319**
Strictly unemployed	2.536	415	321	2.311	**5.584**
Non-work activity (household chores, studying, military service)	16.426	7.238	7.056	13.743	**44.462**
Strictly retired and others	1.899	1.051	155	993	**4.099**
Total	**104.042**	**72.587**	**62.116**	**116.599**	**355.344**

Row %	1- Spanish only	2- Spanish and others	3- Spanish and Basque	4- Spanish, Basque and others	Total
Employed individuals	27,0%	21,0%	17,9%	34,1%	100%
Non-work activity (looking for work)	36,7%	24,2%	21,2%	17,9%	100%
Strictly unemployed	45,4%	7,4%	5,8%	41,4%	100%
Non-work activity (household chores, studying, military service)	36,9%	16,3%	15,9%	30,9%	100%
Strictly retired and others	46,3%	25,6%	3,8%	24,2%	100%
Total	*29,3%*	*20,4%*	*17,5%*	*32,8%*	*100%*

Column %	1- Spanish only	2- Spanish and others	3- Spanish and Basque	4- Spanish, Basque and others	Total
Employed individuals	73,1%	81,6%	81,3%	82,4%	*79,3%*
Non-work activity (looking for work)	6,8%	6,4%	6,6%	3,0%	*5,4%*
Strictly unemployed	2,4%	0,6%	0,5%	2,0%	*1,6%*
Non-work activity (household chores, studying, military service)	15,8%	10,0%	11,4%	11,8%	*12,5%*
Strictly retired and others	1,8%	1,4%	0,3%	0,9%	*1,2%*
Total	*100%*	*100%*	*100%*	*100%*	*100%*

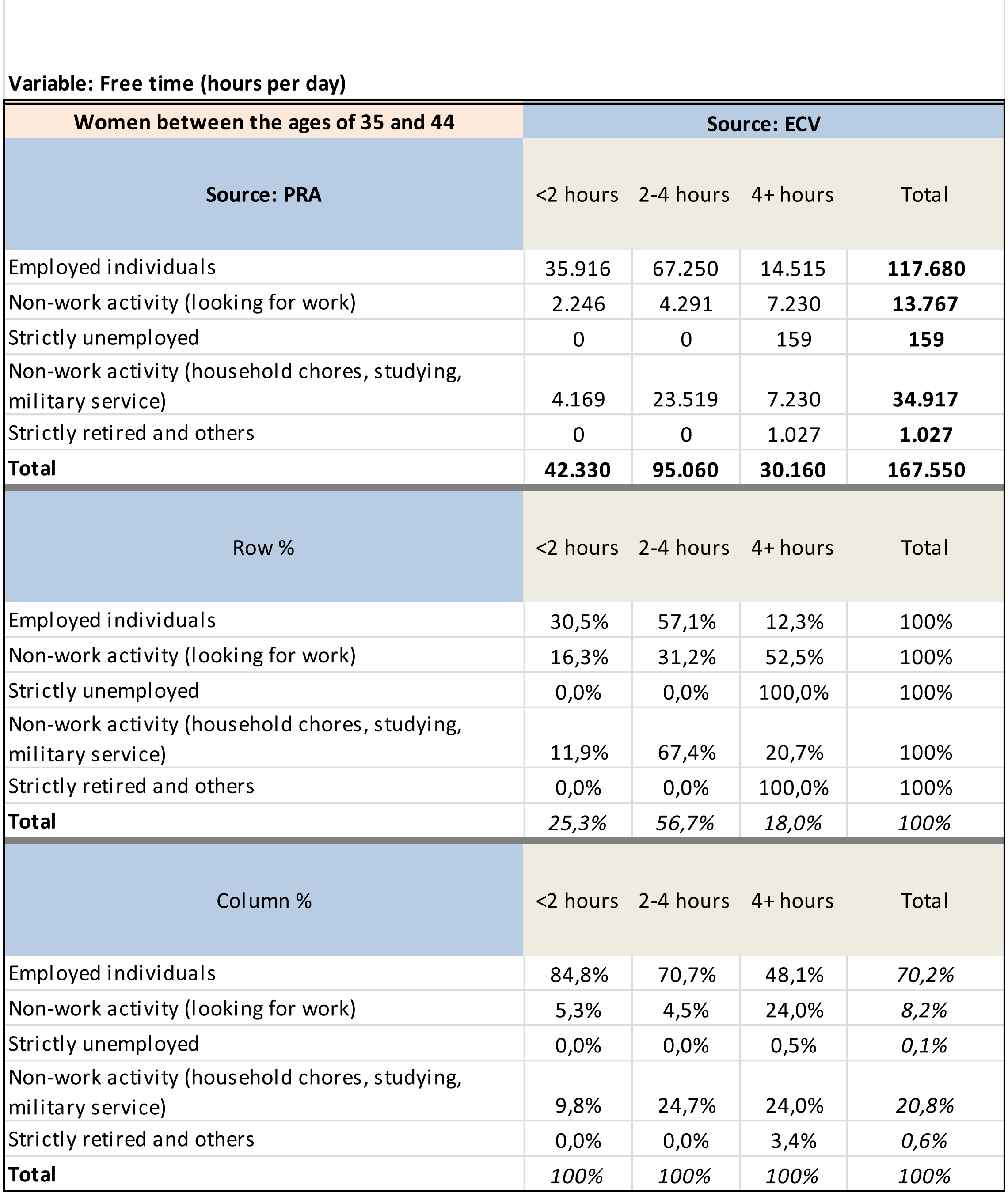
Result 2-2. *Knowledge of Languages (Source: ECV) vs PRA Division (Source: PRA) . Stratum selected: Aged between 35 and 44 (inclusive). Totals and percentages of rows and columns*



Variable: Free time (hours per day)

	Source: ECV			
Source: PRA	<2 hours	2-4 hours	4+ hours	Total
Employed individuals	171.098	551.052	238.162	**960.312**
Non-work activity (looking for work)	8.430	28.436	28.963	**65.830**
Strictly unemployed	1.161	17.341	19.131	**37.633**
Non-work activity (household chores, studying, military service)	48.494	247.203	284.977	**580.673**
Strictly retired and others	11.415	48.415	148.713	**208.543**
Total	**240.598**	**892.447**	**719.946**	**1.852.991**
Row %	<2 hours	2-4 hours	4+ hours	Total
Employed individuals	17,8%	57,4%	24,8%	100%
Non-work activity (looking for work)	12,8%	43,2%	44,0%	100%
Strictly unemployed	3,1%	46,1%	50,8%	100%
Non-work activity (household chores, studying, military service)	8,4%	42,6%	49,1%	100%
Strictly retired and others	5,5%	23,2%	71,3%	100%
Total	*13,0%*	*48,2%*	*38,9%*	*100%*
Column %	<2 hours	2-4 hours	4+ hours	Total
Employed individuals	71,1%	61,7%	33,1%	*51,8%*
Non-work activity (looking for work)	3,5%	3,2%	4,0%	*3,6%*
Strictly unemployed	0,5%	1,9%	2,7%	*2,0%*
Non-work activity (household chores, studying, military service)	20,2%	27,7%	39,6%	*31,3%*
Strictly retired and others	4,7%	5,4%	20,7%	*11,3%*
Total	*100%*	*100%*	*100%*	*100%*

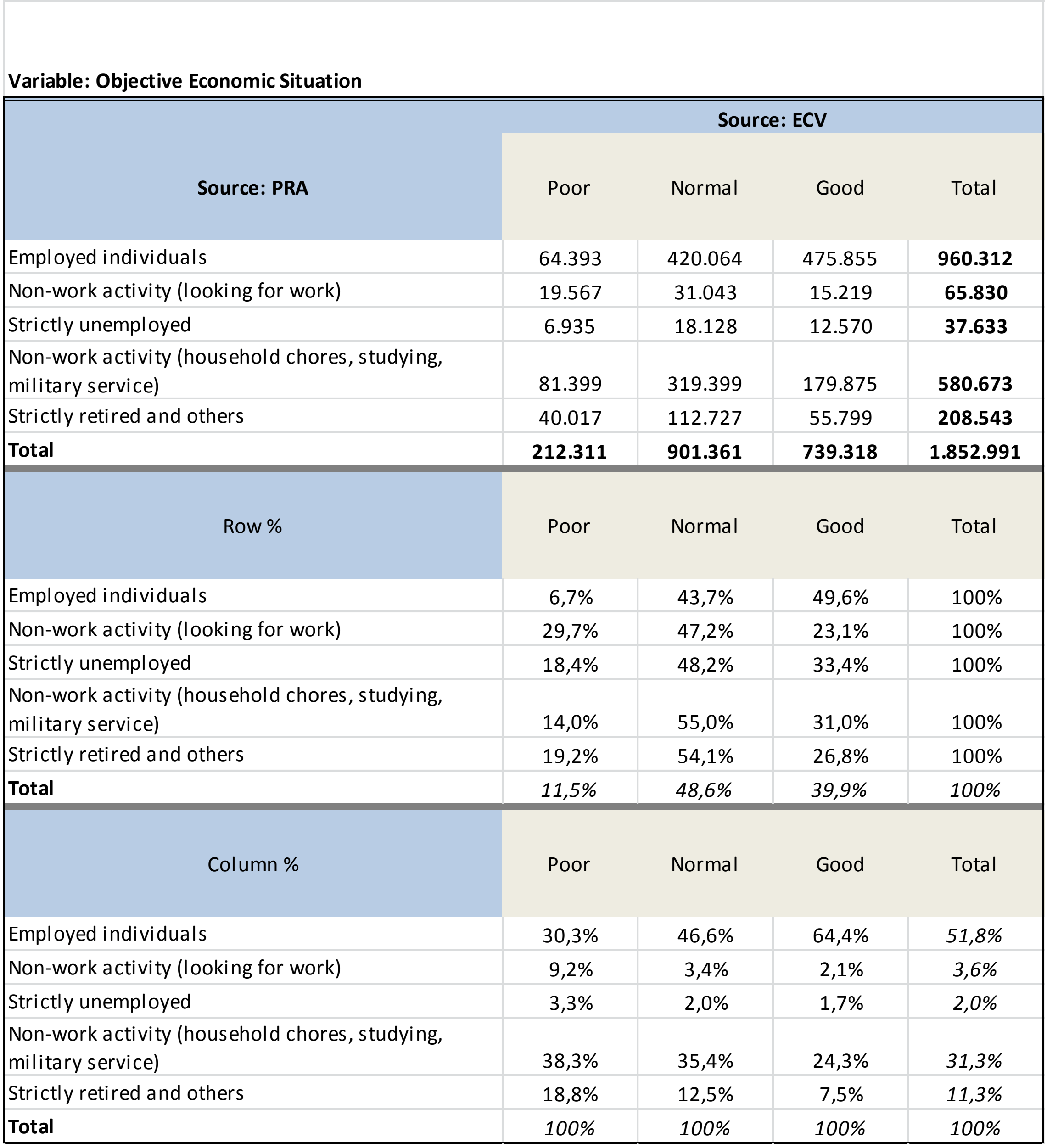
Result 3-1. *Free time per day (Source: ECV) vs PRA Division (Source: PRA). Totals for the population as a whole, and row and column percentages.*



Variable: Free time (hours per day)

Women between the ages of 35 and 44	**Source: ECV**			
Source: PRA	<2 hours	2-4 hours	4+ hours	Total
Employed individuals	35.916	67.250	14.515	**117.680**
Non-work activity (looking for work)	2.246	4.291	7.230	**13.767**
Strictly unemployed	0	0	159	**159**
Non-work activity (household chores, studying, military service)	4.169	23.519	7.230	**34.917**
Strictly retired and others	0	0	1.027	**1.027**
Total	**42.330**	**95.060**	**30.160**	**167.550**
Row %	<2 hours	2-4 hours	4+ hours	Total
Employed individuals	30,5%	57,1%	12,3%	100%
Non-work activity (looking for work)	16,3%	31,2%	52,5%	100%
Strictly unemployed	0,0%	0,0%	100,0%	100%
Non-work activity (household chores, studying, military service)	11,9%	67,4%	20,7%	100%
Strictly retired and others	0,0%	0,0%	100,0%	100%
Total	*25,3%*	*56,7%*	*18,0%*	*100%*
Column %	<2 hours	2-4 hours	4+ hours	Total
Employed individuals	84,8%	70,7%	48,1%	*70,2%*
Non-work activity (looking for work)	5,3%	4,5%	24,0%	*8,2%*
Strictly unemployed	0,0%	0,0%	0,5%	*0,1%*
Non-work activity (household chores, studying, military service)	9,8%	24,7%	24,0%	*20,8%*
Strictly retired and others	0,0%	0,0%	3,4%	*0,6%*
Total	*100%*	*100%*	*100%*	*100%*

Result 3-2. *Free time (Source: ECV) vs PRA Division (Source: PRA). Stratum selected: Women aged between 35 and 44 (inclusive). Totals and percentages of rows and columns*



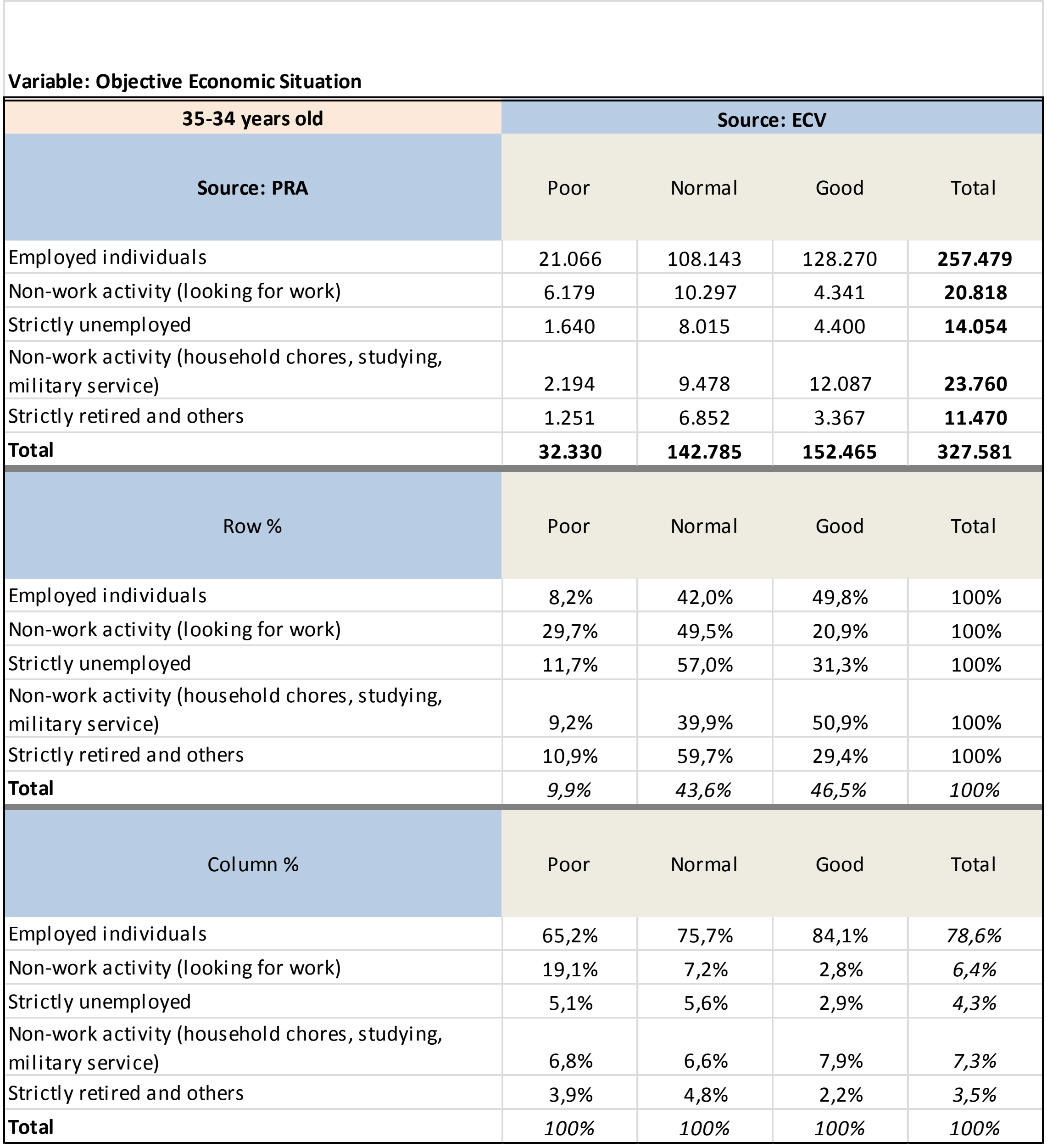
Variable: Objective Economic Situation

Source: PRA	**Source: ECV**			
	Poor	Normal	Good	Total
Employed individuals	64.393	420.064	475.855	**960.312**
Non-work activity (looking for work)	19.567	31.043	15.219	**65.830**
Strictly unemployed	6.935	18.128	12.570	**37.633**
Non-work activity (household chores, studying, military service)	81.399	319.399	179.875	**580.673**
Strictly retired and others	40.017	112.727	55.799	**208.543**
Total	**212.311**	**901.361**	**739.318**	**1.852.991**

Row %	Poor	Normal	Good	Total
Employed individuals	6,7%	43,7%	49,6%	100%
Non-work activity (looking for work)	29,7%	47,2%	23,1%	100%
Strictly unemployed	18,4%	48,2%	33,4%	100%
Non-work activity (household chores, studying, military service)	14,0%	55,0%	31,0%	100%
Strictly retired and others	19,2%	54,1%	26,8%	100%
Total	*11,5%*	*48,6%*	*39,9%*	*100%*

Column %	Poor	Normal	Good	Total
Employed individuals	30,3%	46,6%	64,4%	*51,8%*
Non-work activity (looking for work)	9,2%	3,4%	2,1%	*3,6%*
Strictly unemployed	3,3%	2,0%	1,7%	*2,0%*
Non-work activity (household chores, studying, military service)	38,3%	35,4%	24,3%	*31,3%*
Strictly retired and others	18,8%	12,5%	7,5%	*11,3%*
Total	*100%*	*100%*	*100%*	*100%*

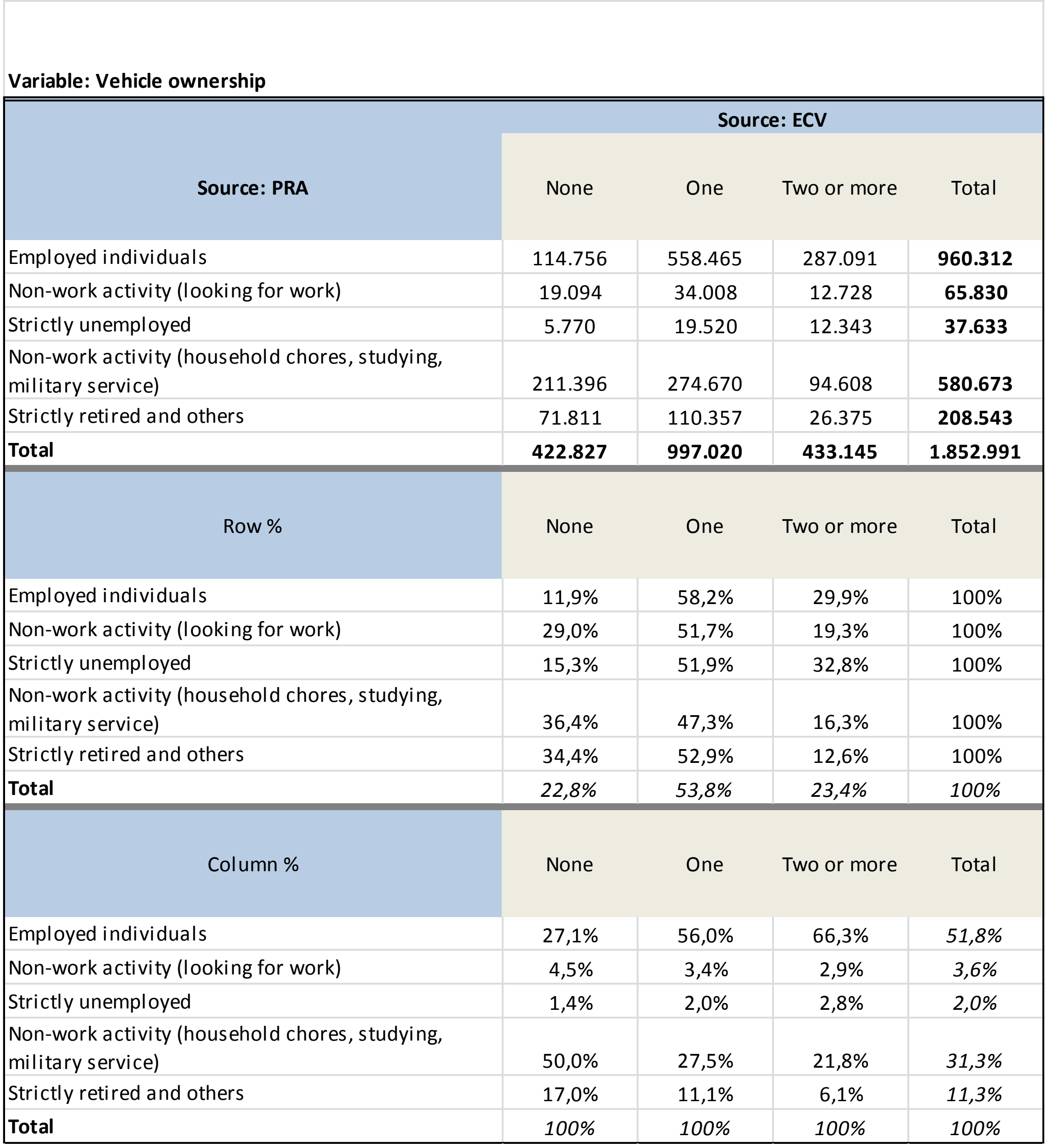
Result 4-1. *Objective economic situation (Source: ECV) vs PRA Division (Source: PRA). Totals for the population as a whole, and row and column percentages.*



Variable: Objective Economic Situation

35-34 years old	**Source: ECV**			
Source: PRA	Poor	Normal	Good	Total
Employed individuals	21.066	108.143	128.270	**257.479**
Non-work activity (looking for work)	6.179	10.297	4.341	**20.818**
Strictly unemployed	1.640	8.015	4.400	**14.054**
Non-work activity (household chores, studying, military service)	2.194	9.478	12.087	**23.760**
Strictly retired and others	1.251	6.852	3.367	**11.470**
Total	**32.330**	**142.785**	**152.465**	**327.581**
Row %	Poor	Normal	Good	Total
Employed individuals	8,2%	42,0%	49,8%	100%
Non-work activity (looking for work)	29,7%	49,5%	20,9%	100%
Strictly unemployed	11,7%	57,0%	31,3%	100%
Non-work activity (household chores, studying, military service)	9,2%	39,9%	50,9%	100%
Strictly retired and others	10,9%	59,7%	29,4%	100%
Total	*9,9%*	*43,6%*	*46,5%*	*100%*
Column %	Poor	Normal	Good	Total
Employed individuals	65,2%	75,7%	84,1%	*78,6%*
Non-work activity (looking for work)	19,1%	7,2%	2,8%	*6,4%*
Strictly unemployed	5,1%	5,6%	2,9%	*4,3%*
Non-work activity (household chores, studying, military service)	6,8%	6,6%	7,9%	*7,3%*
Strictly retired and others	3,9%	4,8%	2,2%	*3,5%*
Total	*100%*	*100%*	*100%*	*100%*

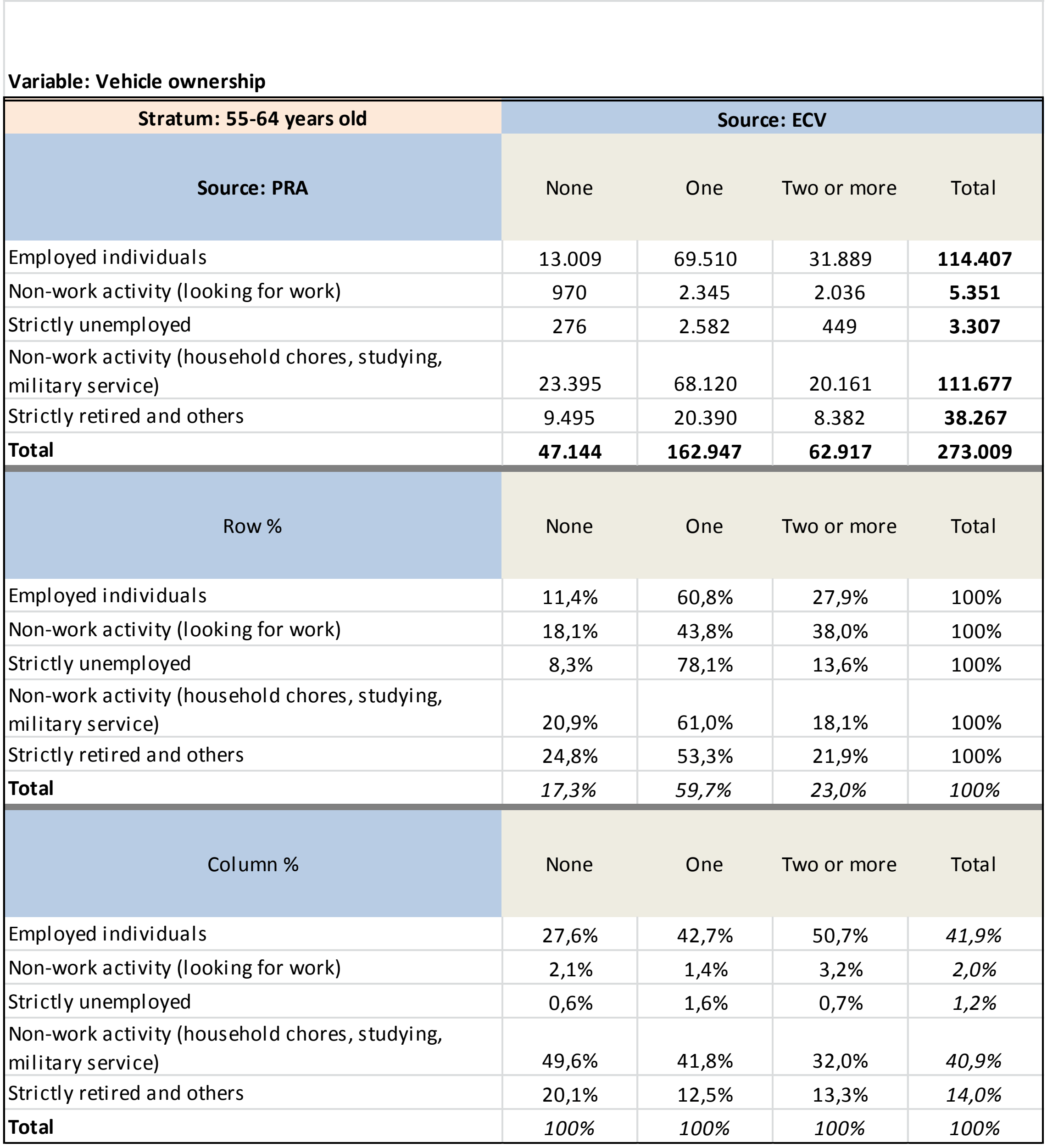
Result 4-2. *Objective economic situation (Source: ECV) vs PRA Division (Source: PRA). Stratum selected: Individuals aged between 35 and 44 (inclusive). Totals and percentages of rows and columns*



Variable: Vehicle ownership

	Source: ECV			
Source: PRA	None	One	Two or more	Total
Employed individuals	114.756	558.465	287.091	**960.312**
Non-work activity (looking for work)	19.094	34.008	12.728	**65.830**
Strictly unemployed	5.770	19.520	12.343	**37.633**
Non-work activity (household chores, studying, military service)	211.396	274.670	94.608	**580.673**
Strictly retired and others	71.811	110.357	26.375	**208.543**
Total	**422.827**	**997.020**	**433.145**	**1.852.991**
Row %	None	One	Two or more	Total
Employed individuals	11,9%	58,2%	29,9%	100%
Non-work activity (looking for work)	29,0%	51,7%	19,3%	100%
Strictly unemployed	15,3%	51,9%	32,8%	100%
Non-work activity (household chores, studying, military service)	36,4%	47,3%	16,3%	100%
Strictly retired and others	34,4%	52,9%	12,6%	100%
Total	*22,8%*	*53,8%*	*23,4%*	*100%*
Column %	None	One	Two or more	Total
Employed individuals	27,1%	56,0%	66,3%	*51,8%*
Non-work activity (looking for work)	4,5%	3,4%	2,9%	*3,6%*
Strictly unemployed	1,4%	2,0%	2,8%	*2,0%*
Non-work activity (household chores, studying, military service)	50,0%	27,5%	21,8%	*31,3%*
Strictly retired and others	17,0%	11,1%	6,1%	*11,3%*
Total	*100%*	*100%*	*100%*	*100%*

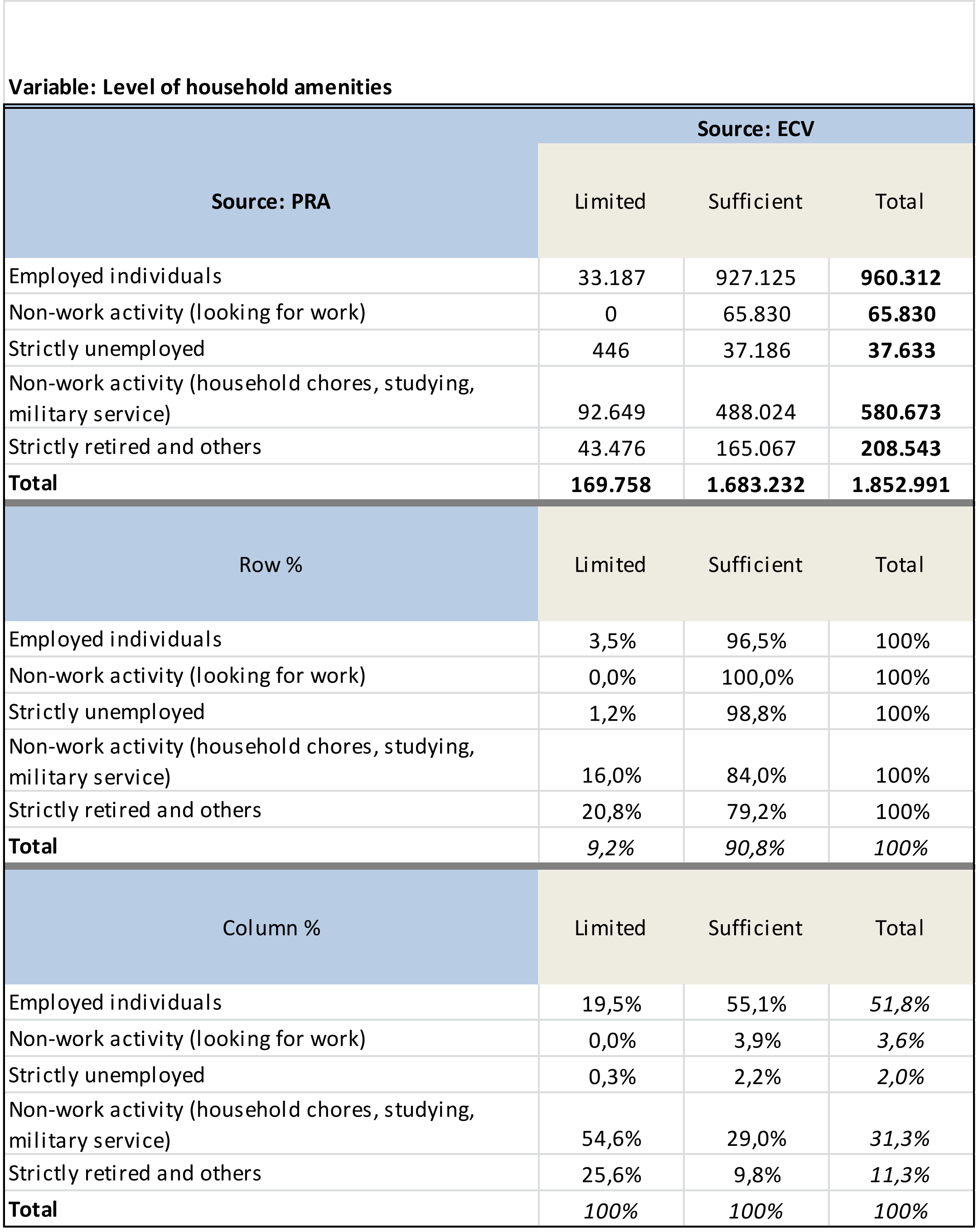
Result 5-1. *Vehicle ownership (Source: ECV) vs PRA Division (Source: PRA). Totals for the population as a whole, and row and column percentages.*



Variable: Vehicle ownership

Stratum: 55-64 years old	**Source: ECV**			
Source: PRA	None	One	Two or more	Total
Employed individuals	13.009	69.510	31.889	**114.407**
Non-work activity (looking for work)	970	2.345	2.036	**5.351**
Strictly unemployed	276	2.582	449	**3.307**
Non-work activity (household chores, studying, military service)	23.395	68.120	20.161	**111.677**
Strictly retired and others	9.495	20.390	8.382	**38.267**
Total	**47.144**	**162.947**	**62.917**	**273.009**
Row %	None	One	Two or more	Total
Employed individuals	11,4%	60,8%	27,9%	100%
Non-work activity (looking for work)	18,1%	43,8%	38,0%	100%
Strictly unemployed	8,3%	78,1%	13,6%	100%
Non-work activity (household chores, studying, military service)	20,9%	61,0%	18,1%	100%
Strictly retired and others	24,8%	53,3%	21,9%	100%
Total	*17,3%*	*59,7%*	*23,0%*	*100%*
Column %	None	One	Two or more	Total
Employed individuals	27,6%	42,7%	50,7%	*41,9%*
Non-work activity (looking for work)	2,1%	1,4%	3,2%	*2,0%*
Strictly unemployed	0,6%	1,6%	0,7%	*1,2%*
Non-work activity (household chores, studying, military service)	49,6%	41,8%	32,0%	*40,9%*
Strictly retired and others	20,1%	12,5%	13,3%	*14,0%*
Total	*100%*	*100%*	*100%*	*100%*

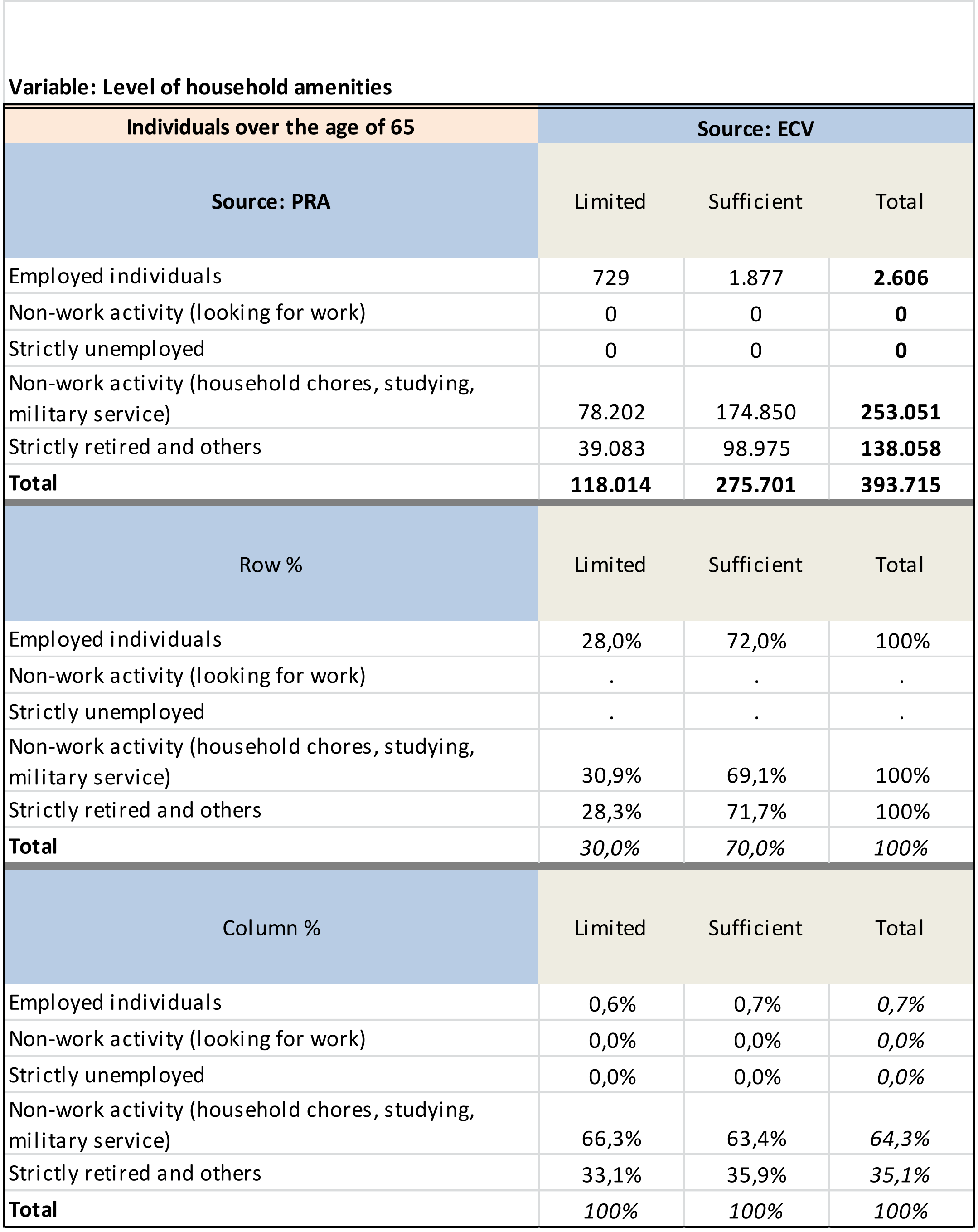
Result 5-2. *Vehicle ownership (Source: ECV) vs PRA Division (Source: PRA). Stratum selected: Individuals aged between 55 and 64 (inclusive). Totals and percentages of rows and columns*



Variable: Level of household amenities

Source: PRA	**Source: ECV**		
	Limited	Sufficient	Total
Employed individuals	33.187	927.125	**960.312**
Non-work activity (looking for work)	0	65.830	**65.830**
Strictly unemployed	446	37.186	**37.633**
Non-work activity (household chores, studying, military service)	92.649	488.024	**580.673**
Strictly retired and others	43.476	165.067	**208.543**
Total	**169.758**	**1.683.232**	**1.852.991**
Row %	Limited	Sufficient	Total
Employed individuals	3,5%	96,5%	100%
Non-work activity (looking for work)	0,0%	100,0%	100%
Strictly unemployed	1,2%	98,8%	100%
Non-work activity (household chores, studying, military service)	16,0%	84,0%	100%
Strictly retired and others	20,8%	79,2%	100%
Total	*9,2%*	*90,8%*	*100%*
Column %	Limited	Sufficient	Total
Employed individuals	19,5%	55,1%	*51,8%*
Non-work activity (looking for work)	0,0%	3,9%	*3,6%*
Strictly unemployed	0,3%	2,2%	*2,0%*
Non-work activity (household chores, studying, military service)	54,6%	29,0%	*31,3%*
Strictly retired and others	25,6%	9,8%	*11,3%*
Total	*100%*	*100%*	*100%*

Result 6-1. *Level of household (Source: ECV) vs PRA Division (Source: PRA). Totals for the population as a whole, and row and column percentages.*



Variable: Level of household amenities

Individuals over the age of 65	Source: ECV		
Source: PRA	Limited	Sufficient	Total
Employed individuals	729	1.877	**2.606**
Non-work activity (looking for work)	0	0	**0**
Strictly unemployed	0	0	**0**
Non-work activity (household chores, studying, military service)	78.202	174.850	**253.051**
Strictly retired and others	39.083	98.975	**138.058**
Total	**118.014**	**275.701**	**393.715**
Row %	Limited	Sufficient	Total
Employed individuals	28,0%	72,0%	100%
Non-work activity (looking for work)	.	.	.
Strictly unemployed	.	.	.
Non-work activity (household chores, studying, military service)	30,9%	69,1%	100%
Strictly retired and others	28,3%	71,7%	100%
Total	*30,0%*	*70,0%*	*100%*
Column %	Limited	Sufficient	Total
Employed individuals	0,6%	0,7%	*0,7%*
Non-work activity (looking for work)	0,0%	0,0%	*0,0%*
Strictly unemployed	0,0%	0,0%	*0,0%*
Non-work activity (household chores, studying, military service)	66,3%	63,4%	*64,3%*
Strictly retired and others	33,1%	35,9%	*35,1%*
Total	*100%*	*100%*	*100%*

Result 6-2. *Level of household amenities (Source: ECV) vs PRA Division (Source: PRA). Stratum selected: Individuals over the age of 65. Totals and percentages of rows and columns*

Chapter
5

Development of a suitable R package

During the process of matching the ECV and PRA surveys, a series of functions were created to speed up the calculations in different phases of the process: variables selection, imputation by strata and validation of the results. These functions, based on tested and contrasted packages (see the third chapter Software), were encapsulated in a suitable R package called `micromatch` with the goal that these functions would be distributed together with this technical manual.

The suitable R packages (i.e. linked to a project) present a number of advantages (Chambers, J,, 2008):

- They articulate the R functions created for the project as an integrated whole, in such a way that the calculations are more efficient and reliable

- They let you create a simple documentation type that is highly useful, for the user too

- They allow you to generalise the calculations and other similar or connected problems

- They allow you to easily distribute the code of other potential users

The ECV-PRA matching project already enjoys these advantages.

In a phase subsequent to this project, the possibility was recognised of generalising the `micromatch` package to generically address any statistical matching of independent surveys. To take this step, a structure of classes and methods in system S4 for *object-oriented programming* was designed (Chambers, J., 2008). The idea of the **classes** is that they can conceive complex concepts as objects, such as "a survey to match" (in `micromatch`, the class `filetomatch`). Then, **methods** are designed that act on these objects, for instance: "compare all the variables with respect to another survey to match".

Unlike other packages, the `micromatch` package does not provide suitable methods for statistical matching. Instead, it is based on packages that have already been tested and contrasted, and implements a generic solution that covers (and speeds up) all the phases of a match.

Thus, the idea of `micromatch` is to create an environment where the user can use and test any type of matching method in a simple way. Ultimately, the package does the following:

- It offers an environment where the calculations involved in a statistical match can be sped up

- It offers an effective, robust computation environment where, thanks to the system of classes, all calculations can be efficiently interconnected

- It can disseminate survey methodology by inserting a wide ensemble of techniques currently in separate packages into one single package

- It can spread the work carried out during this grant, thereby offering an environment where the user can reproduce the ECV-PRA calculations.

This package was presented during the VI-grade R User Conference, held in Santiago de Compostela in October 2014. It is also available on Eustat's website.

Chapter
6

Conclusions

Statistical matching allows us to make more efficient use of information acquired from independent surveys referring to the same population, and we can do this by obtaining integrated indicators and statistics.

The practical case developed in this technical manual, i.e. matching two independent EUSTAT surveys: The Survey on Living Conditions and The Survey on Population in Relation to Activity, have both allowed us to appreciate the importance of various elements in every procedure in a match:

- An adequate selection of **common variables**, which requires an in-depth meta-analysis of the questionnaires to be performed in order to identify information that is comparable with regards to the definition and the empirical distributions observed in the data files.

- The use of **stratum variables**, (in the ECV-PRA surveys, age and gender groups) which will generally be closely related to the specific variables (living conditions: free time, social relations... in the ECV survey, and the labour market in the PRA survey)

- The need to **validate the results**, which requires the algorithm to produce marginal distributions that are comparable to the ones observed. However, always bear in mind that there might be more sources of uncertainty involved

More generally, the adoption of the statistical matching methodology turns out to be beneficial seeing as it compels us to view the surveys not as independent instruments, but as an integrated whole. In this respect, A. Leulescu y M. Agafitei (2013) make a number of recommendations:

- **Standardise** the questionnaires as much as possible, formulating questions in a comparable fashion in order to guarantee their consistency

- If possible, **include a small, common module** for all the surveys, gathering determined, "specific" but basic aspects, such as income or health self-perception. When it comes to matching the surveys, these variables will play a key role in improving the quality (i.e. reduction of uncertainty) of the match results.

In the specific case of the ECV-PRA match, it would be desirable to have harmonised variables for income or the level of education, which will help to reduce uncertainty and thus improve the match results.

As to the **implementation** of these techniques, the development of the free R software is increasingly providing more tools, and perhaps in the future, we will see better coordination between all the packages. During the project, some steps have been made in this direction thanks to the development of the `micromatch` package.

Bibliography

General references

Agresti (2014). Categorical data analysis. John Wiley & Sons.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. Journal of statistical software, 45(3).

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). Statistical matching: Theory and practice. John Wiley & Sons.

D'Orazio, M., Di Zio, M., & Scanu, M. (2010). Old and new approaches in statistical matching when samples are drawn with complex survey designs. Proceedings of the 45th "Riunione Scientifica della Societa'Italiana di Statistica", Padova, 16-18.

Leulescu, A. & Agafitei, M. (2013). Statistical matching: a model based approach for data integration. Eurostat Methodologies and working papers.

Rässler, S. (2002). Statistical matching. Springer.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business & Economic Statistics, 4(1), 87-94.

European projects

Data Integration: http://www.cros-portal.eu/content/data-integration-finished

ISAD Integration of Survey and Administrative Data. http://www.cros-portal.eu/content/isad-finished

Report WP2 ESS-net, Statistical Methodology Project on Integration of Surveys and Administrative Data. Recommendations on the use of methodologies for the integration of surveys and administrative data.

Internal EUSTAT documents

EUSTAT, methodology file for the Survey on Population in Relation to Activity (PRA) http://es.eustat.es/document/poblact_c.html

EUSTAT, methodology index for the Survey on Living Conditions (ECV), http://es.eustat.es/document/ecvida_c.html#axzz37QhmZ17l

Software

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Chambers, J. (2008). Software for data analysis: programming with R. Springer.

D'Orazio, M. (2013). StatMatch: Statistical Matching (aka data fusion). http://CRAN.R-project.org/package=StatMatch

Lumley, T. (2012) survey: analysis of complex survey samples. http://CRAN.R-project.org/package=survey

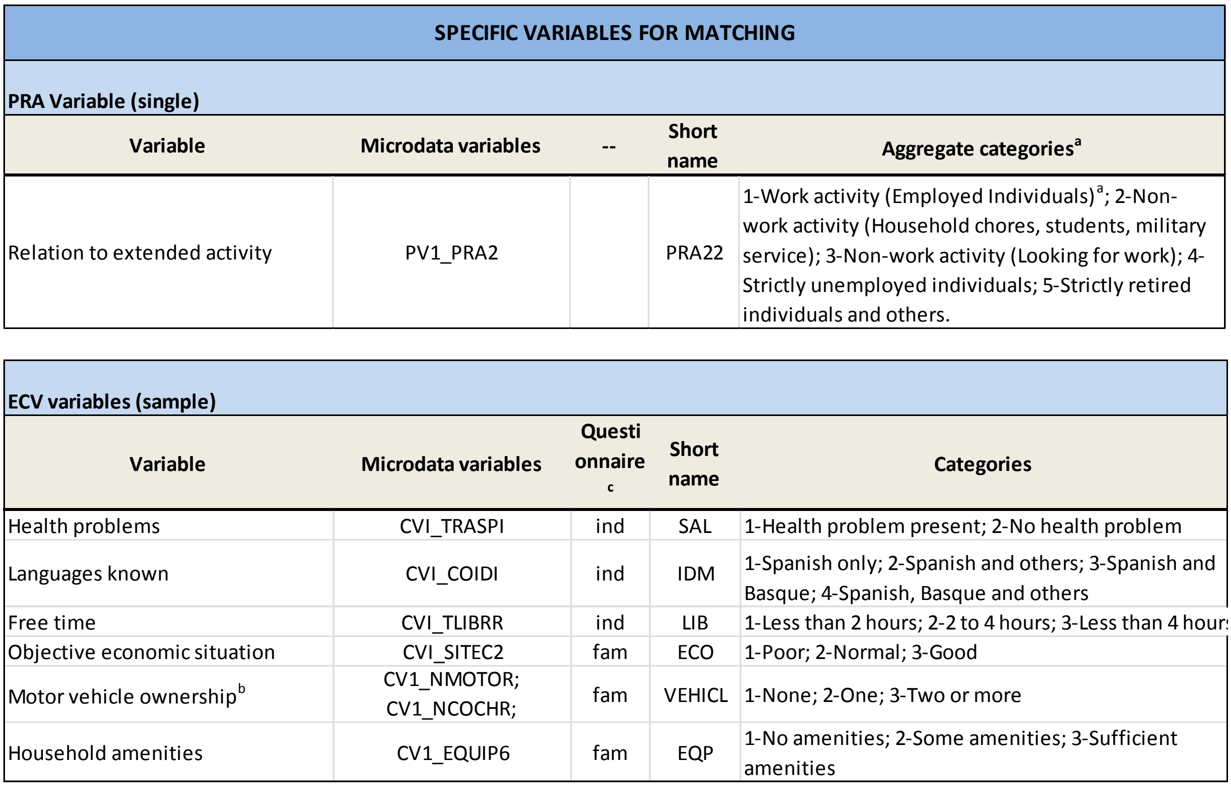
Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL http://www.jstatsoft.org/v45/i03/

Harrell Jr, F. E. (2014). Hmisc: Harrell Miscellaneous. R package version 3.14-3. http://CRAN.R-project.org/package=Hmisc

Honaker, J.,King, G. & Blackwell, M. (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), 1-47. URL http://www.jstatsoft.org/v45/i07/.

Meinfelder, F. (2011). BaBooN: Bayesian Bootstrap Predictive Mean Matching – Multiple and single imputation for discrete data. http://CRAN.R-project.org/package=BaBooN

Annexes



SPECIFIC VARIABLES FOR MATCHING				
PRA Variable (single)				
Variable	**Microdata variables**	**--**	**Short name**	**Aggregate categories**[a]
Relation to extended activity	PV1_PRA2		PRA22	1-Work activity (Employed Individuals)[a]; 2-Non-work activity (Household chores, students, military service); 3-Non-work activity (Looking for work); 4-Strictly unemployed individuals; 5-Strictly retired individuals and others.

ECV variables (sample)				
Variable	**Microdata variables**	**Questionnaire**[c]	**Short name**	**Categories**
Health problems	CVI_TRASPI	ind	SAL	1-Health problem present; 2-No health problem
Languages known	CVI_COIDI	ind	IDM	1-Spanish only; 2-Spanish and others; 3-Spanish and Basque; 4-Spanish, Basque and others
Free time	CVI_TLIBRR	ind	LIB	1-Less than 2 hours; 2-2 to 4 hours; 3-Less than 4 hour
Objective economic situation	CVI_SITEC2	fam	ECO	1-Poor; 2-Normal; 3-Good

Motor vehicle ownership[b]	CV1_NMOTOR;
	CV1_NCOCHR;
		fam	VEHICL	1-None; 2-One; 3-Two or more

Household amenities	CV1_EQUIP6	fam	EQP	1-No amenities; 2-Some amenities; 3-Sufficient amenities

[a] Within the PRA category for employed individuals, there are subcategories depending on the rate of employment. In this study, they have been combined into one single category: 'Employed individuals'

[b] Variable generated from: CV1_NMOTOR 'Number of 50cc + motorbikes', CV1_NCOCHR 'Number of cars', CV1_NFURGR 'Number of vans', CV1_OTRVEHICR 'Other vehicles'

[c] Source questionnaire: individual (ind) or family (fam).

Index A: Specific variables for matching independent EUSTAT surveys, ECV and PRA.

	PRA-2009 4T		ECV-2009			
Variable	Name[a]	Questionnaire question number[b]	Name[a]	Source file[c]	Questionnair e question number[b]	Aggregate levels[c]
Social-demographic variables						
Gender	PV1_SEXO	p12	ind: CV1_SEXOI	ind	--	M-Man; W-Woman
Age	PV1_EDAD	p10	ind: CV1_EDADIR	ind		01-"<=15 years old"; 02-"16-24 years old"; 03-"25-34 years old"; 04-"35-44 years old"; 05-"45-54 years old"; 06-"55-64 years old"; 07-">=65 years old"
Family size	TAMAÑO_FAM	--	fam: CV1_TFAMR	fam	--	1, 2, 3+ members
Level of education						
Undertaking accredited studies[e]	PV1_ENRE (!D)[d]	p43	CV1_SITES (B)	ind	pI2	1-Studying; 0-Not studying
Undertaking distance learning Y/N	PV1_ENRE (C)	p43	CVI_SISTE (F)	ind	pI4	1-Yes; 0-No
Illiterate Y/N	PV1_LEES	p34	CVI_ANALF	ind	pI15	1-Yes; 0-No
D. Variables related to employment status						
Relation to activity - ILO	PV1_PRA1	--	CV1_RELA1	ind	--	Employed individuals; Unemployed Individuals; Inactive Individuals
Looking for work Y/N	PV1_BUSQ	p140	CVI_BUSQ	ind	pT23	1-Yes; 0-No
Hours worked per week (Employe	PV1_HTRA[g]	p107	CVI_HOTAT	ind	pT22	Numerical
Devotion to household chores	PV1_SILH	p55	CV1_TDOME1	ind	Synthetic indicator[f]	Does household chores; Does not do household chores

[a] Variable name in the microdata file.

[b] Question number in the questionnaire.

[c] Source file for the ECV survey: indicated in the source microdata file (individual: ind; family: fam)

[d] "!" symbol: all categories accepted except the one indicated.

[e] Hereinafter 'Student Y/N'.

[f] Indicator derived from 4 items in question pT27: COMPR2-Buy food; COMID2-Prepare food; FREG2-Wash the dishes; ROPA2-Prepare clothes; LIMPC2-Clean the house.

[g] Segment OIT=Employed individuals

Index B: Meta-analysis of the variables in common between the ECV and PRA surveys.

Common values		Coherence	Predictive value						
Variable	Short name	Hellinger Distance	Cramer's V (global)						
			PRA: "PRA"	**ECV:** "SAL"	**ECV:** "IDM"	**ECV:** "LIB"	**ECV:** "ECO"	**ECV:** "VEH"	**ECV:** "EQP"
Age	ED	0,002	0,372	0,304	0,267	0,230	0,152	0,289	0,396
Gender	Y	7,68E-06	0,409	0,014	0,027	0,075	0,051	0,123	0,054
Family size	TF	0,001	0,192	0,169	0,099	0,167	0,194	0,285	0,275
Student Y/N	EST	0,009	0,329	0,097	0,333	0,138	0,028	0,084	0,089
Unemployed Y/N	OCP	0,002	1,000	0,275	0,291	0,402	0,345	0,373	0,285
Unemployed Y/N	PAR	0,012	0,998	0,021	0,046	0,105	0,142	0,029	0,073
Inactive Y/N	INA	0,007	0,999	0,288	0,309	0,380	0,285	0,366	0,322
Looking for work Y/N	BUSQ	0,025	0,850	0,067	0,118	0,037	0,073	0,035	0,087
Household chores	DOM	0,054	0,548	0,054	0,061	0,105	0,032	0,096	0,002

Table C1. *Global measurements.*

Stratum	Common variables selected	Sample size		Coherence	Predictive value
(Short name)	Variables selected by stratum	ECV (recipient)	PRA (donor)	Hellinger distances. (by variable)	Cramer's V Dependent variable: "PRA"
H.15-24	EST, BUSQ	157	512	EST: 0.048 - BUSQ: 0.052	EST: 0.887 - BUSQ: 0.907
W. 15-24	EST, BUSQ	166	491	EST: 0.09 - BUSQ: 0.003	EST: 0.835 - BUSQ: 0.848
M. 25-34	BUSQ, DOM	359	746	BUSQ: 0.002 - DOM: 0.064	BUSQ: 0.895 - DOM: 0.389
W. 25-34	PAR, DOM	372	740	PAR: 0.004 - DOM: 0.049	PAR: 1.000 - DOM: 0.378
M. 35-44	PAR, FS2	426	920	PAR: 0.011 - FS2: 0.047	PAR: 1.000 - FS2: 0.117
W. 35-44	PAR, OCP	386	957	PAR: 0.019 - OCP: 0.051	PAR: 1.000 - OCP: 1.000
M. 45-54	PAR, BUSQ	377	966	PAR: 0.004 - BUSQ: 0.034	PAR: 1.000 - BUSQ: 0.870
W. 45-54	PAR, BUSQ	403	1034	PAR: 0.032 - BUSQ: 0.004	PAR: 1.000 - BUSQ: 0.730
M. 55-64	INA, BUSQ	349	843	INA: 0.033 - BUSQ: 0.023	INA: 1.000 - BUSQ: 0.848
W. 55-64	INA	391	887	INA: 0.027	INA: 1.000
M. 65+	TF2	540	1158	TF2: 0.018	TF2: 0.295
W. 65+	TF2	823	1611	TF2: 0.030	TF2: 0.111

Table C2. *Measurements by Age and Gender strata for the variables selected.*

Index C: Variables selection by stratum. Empirical coherence measurements (Hellinger distances) and predictive value measurements (Cramer's V). Global measurements (Table C1) and by stratum (C2).

Organismo Autónomo del

EUSKO JAURLARITZA
GOBIERNO VASCO

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA
www.eustat.eus

www.eustat.eus