Leunda Iztueta, I., Garmendia Navarro, I., Etxeberria Murgiondo, J. (2017). Statistical Matching en la práctica – Una aplicación a la evaluación del sistema educativo mediante PISA y TALIS. *Revista de Investigación Educativa, 35*(2), 371-388
DOI: http://dx.doi.org/10.6018/rie.35.2.262171

Statistical Matching en la práctica – Una aplicación a la evaluación del sistema educativo mediante PISA y TALIS

Statistical Matching in practice – An application to the evaluation of the education system from PISA and TALIS

Ixiar Leunda Iztueta*, Inés Garmendia Navarro** y Juan Etxeberria Murgiondo*

*EHU-UPV Universidad del País Vasco
**EUSTAT Euskal Estatistika Erakundea

Resumen

Con el nombre Statistical Matching se identifican un conjunto de técnicas que posibilitan integrar información obtenida mediante encuestas independientes con unidades muestrales distintas. El objetivo es obtener un fichero de datos sintético con información plausible para ítems provenientes de distintas fuentes. Método: El Matching parte de la existencia de variables comunes entre los ficheros, usualmente, variables sociodemográficas. Este bloque de información común se emplea para imputar los ítems específicos de las encuestas. Resultados: Explicitamos las fases principales del Statistical Matching y las aplicamos a las encuestas PISA 2012 y TALIS 2013 de España. Proporcionamos pautas para una validación de los resultados. En todas las fases se ha utilizado el software libre R. Conclusiones: La potencialidad de Statistical Matching es enorme en tanto que posibilita enlazar ficheros de origen distinto. Las técnicas de Statistical Matching son accesibles gracias al desarrollo de diversos paquetes de R. Su aplicación en Ciencias Sociales puede ser solución a multitud de problemas metodológicos y contribuir a un mejor conocimiento de la realidad social

Palabras clave: statistical matching; educación; evaluación; Pisa; Talis.

Correspondencia: Juan Etxeberria Murgiondo, juan.etxeberria@ehu.eus, Dpto. MIDE, Facultad F.E.F.A. Universidad del País Vasco. Avenida Tolosa, 70. 20018, Donostia-San Sebastián.

Abstract

Statistical matching methods are aimed at the integration of information collected through multiple sources, usually, surveys drawn from some target population. As opposed to record linkage methods -where we search for identical units-, in statistical matching we search for similar units in order to find statistical relations across databases. Methods: Statistical matching is feasible provided that the independent surveys share a common block of variables. A particular solution is based on imputation methods for missing data: first, the distinct files are concatenated (i.e. rows and columns are joined together to form a unique file); next, empty cells corresponding to non-observed values are interpreted as missing data, and they are imputed according to observed data. Results: The fundamental concepts of statistical matching are shown, and the process is illustrated with the PISA (2012) and TALIS (2013) educational studies with Spain's data. Imputations are carried out using mice package from the free R software. A first validation of the results is performed. Conclusions: Statistical matching offers high potential benefits for the social sciences since it enables to relate information from independent information sources. These techniques can now be applied with relative ease thanks to the development of tools such as R computing environment.

Keywords: statistical matching, education, evaluation, Pisa, Talis.

Introduction

Statistical matching (also known as data fusion data merging or synthetic matching) is a model-based approach for providing joint information on variables or indicators collected through multiple sources, usually, surveys drawn from the same population. The potential benefits of this approach lie in the possibility to enhance the complementary use and analytical potential of existing data sources. In this sense, statistical matching can be viewed as a tool to increase the efficiency of use given the current data collections (Leulescu & Agafitei, 2013).

In the particular case of educational research, independent, large-scale international assessments are made periodically to inspect how aspects of educational institutional arrangements interrelate with each other; the ultimate aim is to ensure equality of educational opportunity. Each study focuses on different aspects. The OECD Program for International Student Assessment (PISA) and the OECD Teaching and Learning International Survey (TALIS) constitute two of the largest ongoing studies, the former focusing on students' performance and the latter on teachers' strategies and practices (Breakspear, 2012; Choi & Jerrim, 2015; Wheater 2013) .

Naturally, policy makers are interested in all three levels of the school system -namely, students, teachers and schools-, in order to fully understand differences in the inputs (e.g. socioeconomic levels), processes (teaching strategies and classroom environment), and outcomes of education (performance levels) (Gustafsson, 2003; Kaplan & Turner, 2013). However, serious limitations arise when trying to extract global conclusions from both studies: particularly, PISA, having questionnaires for students and school principals, is missing teacher-level data; and TALIS, with questionnaires for teachers and school principals, is missing student-level data. Statistical matching could therefore be a fundamental tool with the potential to bridge these gaps between educational studies, and in social science research in general (Taut & Palacios, 2016).

The aim of this article is to show the potential benefits of statistical matching for the social sciences by illustrating the rationale behind the process with a concrete example from educational research, namely, the OECD PISA and TALIS surveys.

The article is organized as follows. First, the fundamental concepts of statistical matching are shown, and a close look is taken at its different steps. The process is illustrated with data from the OECD PISA 2012 and TALIS 2013 studies for the Spain's case (Fernández-Diaz, Rodríguez-Mantilla & Mártinez-Zarzuelo, 2016; González-Such, Sancho-Álvarez & Sánchez-Delgado, 2016). The matching task is tackled by using a multiple imputation method supported by the mice package in the free R software environment. Some relevant results are shown, and after a discussion focusing on the validity of the fused file, the most important conclusions are drawn.

Fundamentals of statistical matching

The statistical matching task begins with two or more independent survey samples from the same population of interest, each of which produces measures regarding specific questions (for example, living styles and wages), but sharing a block of common variables (usually sociodemographic variables such as the age, sex, or social status), see Figure 1. The basic assumption is that the number of individuals or units appearing in both samples (i.e., the overlap) is negligible. In this respect, the fundamental difference with respect to other methods such as record linkage is that, in the latter, we have identical units that we want to match exactly, while in statistical matching we know the units are different, but we wish to find similar ones.

Observations	Common variables (Sources #1 & #2)	Specific variables (Sources #1)	Specific variables (Sources #2)
Obs_1 Obs_2 : Obs_k	Z	X	
Obs_{k+1} Obs_{k+1} : Obs_{k+1} Obs_{k+n}			Y

Figure 1. Starting point of statistical matching of two independent data sources sharing a block of common variables. Source #1 is composed of k observations and Z+X variables; Source #2 is composed of n observations and Z+Y variables. After concatenating the files we get k+n observations and Z+X+Y variables

In order to extract conclusions about specific variables in distinct files, statistical matching takes advantage of the fact that, the more they are determined by common

characteristics of individuals (e.g. income is determined by sex, age, educational level… to a certain extent), the more the relation between specific variables will be determined. Accordingly, statistical matching can be viewed as a problem of effectively using these kind of relations in databases in order to enrich the possibilities of statistical analysis.

Two main approaches can be delineated in terms of statistical matching goals (D'Orazio, Di Zio & Scanu, 2006). In the macro approach the source files are used in order to have a direct estimation of certain characteristics of the specific variables, such as joint distributions, marginal distributions or correlation matrices. In the micro approach the aim is the construction of a synthetic file which is complete, e.g., a file that contains complete records on specific variables that were originally measured in separate files.

In the micro case, the term synthetic refers to the circumstance that the file is not a product of direct observation, but it is obtained from the independent source files by some statistical transformation. In an ideal situation, any analysis based on the statistically matched (or fused) file may be performed as though the matched file had been obtained as a random sample from the underlying population (Rässler, 2002).

The statistical matching task can be formulated as a large missing data imputation problem, in which large blocks of variables are missing by design (i.e. because they were not measured). By adopting this view, well-known statistical techniques used to handle missing value problems can be applied here.

The statistical matching process

Regardless of the method used to infer the synthetic file (in the micro case) or certain parameters for the relation between specific variables (in the macro case), the application of statistical matching techniques implies a series of phases closely related to the stages of a survey process. It is important to keep in mind that the selection of an appropriate matching technique is only one of these steps, and often not the most important.

First, the objectives of the matching task have to be specified (see Figure 2). Specifically, the concrete specific variables to relate as well as the desired type of result (e.g. micro or macro results) must be specified.

The second stage comprises two main steps. First, coherence between the survey samples to be matched has to be studied in detail. This effort must comprehend an assessment of the degree of harmonisation and reconciliation between the sources, including (but not limited to) the following aspects (D'Orazio et al., 2006):

(i) harmonisation on the definition of units and the reference period, and the completion of population (e.g., assuring that the survey samples refer to the same population of interest),
(ii) harmonisation of (common) variables and classifications
(iii) adjustment of measurement errors (accuracy)
(iv) adjustment for missing data
(v) derivation of variables

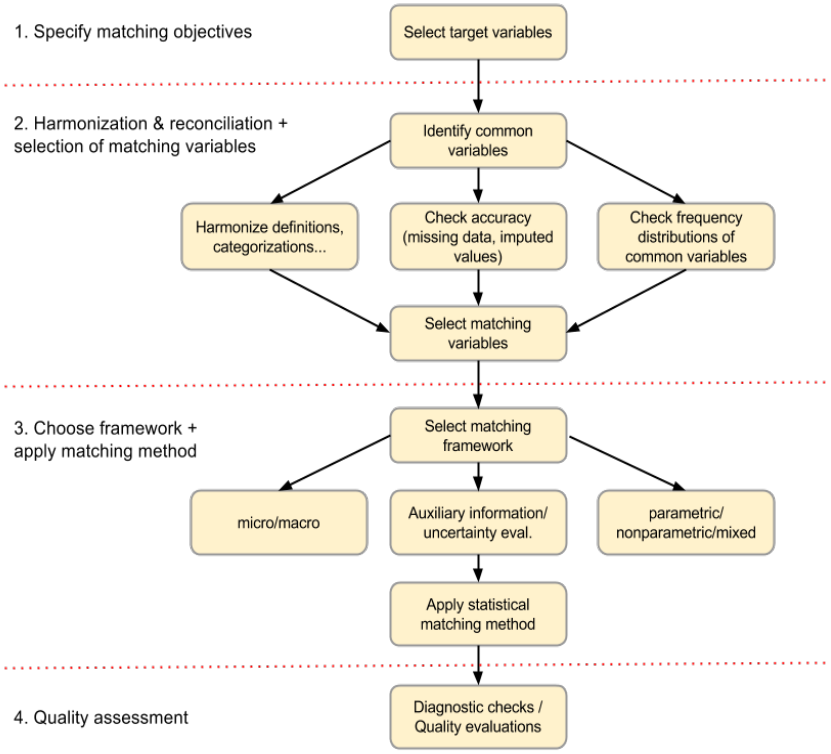

Figure 2. Stages of statistical matching. Adaptado de Eurostat (2008). ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data.

Discrepancies will usually emerge at different points, mainly in the data collection (different unit definitions, variables measured with different categories...), but also due to posterior statistical treatments such as calibrations or the derivation of different indicators. The analyst must proceed to identify a set of common variables between the sources that are as homogeneous as possible in their statistical content. That is: both samples should estimate, for example, the same age, sex... distributions.

The second part of this stage should assess the predictive power of the common variables (concluded to be comparable from the previous study) in both files with respect to the specific variables (e.g. the one selected as the target of the matching task, for example, wages on the one hand, and occupational status on the other). Ideally, an appropriate subset of common variables should be identified containing all association existing between the specific, not jointly observed, variables. In this point, care must be taken not to include variables with low predictive value, since doing so may have negative impact on the computational procedure. In order to choose an optimal subset of predictors, multivariate techniques such as stepwise regression or factor analyses may be used. The derivation of new common variables from the original ones is also possible.

The third step focusses on selecting an appropriate matching method and applying it to the data. The selection will strongly depend on the matching objectives (micro/macro) and on what kind of information we have. For instance, if auxiliary information is available (possibly in the form of a third, complete file with observed values for all the variables of interest), that information can be advantageously used as an integral part of the matching method. Also, the choice of a parametric versus a non-parametric setting has to be considered: in the first case, a specific model is assumed for the joint distribution of variables (typically, the normal multivariate in the case of continuous variables, or multinomial in the case of categorical variables). In the second case, the method will be free of distributional assumptions.

The selection of the matching method is not straightforward. Since the first matching efforts in the fields of official statistics (in the United States and in Canada) in the early 1970's and in market research (in Europe) in the 1960's many different statistical techniques have been applied for the matching task, (see Rässler, 2002, for a brief history on statistical matching). According to Leulescu and Agafitei (2013), four main groups of techniques can be identified:

1. Hot deck methods: the most popular matching techniques by far, they are non-parametric in nature. Basically, for each record in one of the files (named as the recipient file, say file A), the method searches for a similar record in the other file (known as the donor file, say file B), and the observed values in that record (in file B) are used back to impute the values of the initial record (in file A).
   Regression-based methods: Grounded on a parametric framework, they use maximum likelihood to estimate the joint distribution of the variables as a product of conditional and marginal likelihood functions derived from both files, but they have serious limitations like regression towards the mean.
2. Mixed methods: These methods effectively combine the advantages of the parametric and non-parametric frameworks. One of these methods is predictive mean matching, which, in a first step, performs a regression of the specific variables versus the common ones. Next, for each record a nearest record is searched based on the predicted values from the regression equation.
3. Multiple imputation methods: First proposed by Rubin (1987), these may be used to overcome the inherent uncertainty in the matching task. The idea is to impute more than one value (that is, m>1 values, usually between 3 and 5) for each missing value. In this way, m fused files are obtained rather than one, and the uncertainty about which value to impute is reflected.

Finally, a quality assessment has to be performed. For this purpose a process approach has to be adopted, since each of the steps (the quality and coherence of the original sources, assumptions on distributional features or conditional correlations, the matching method itself) can have a potential impact on the quality of results.

Rässler (2002) established four levels of validity for the evaluation of statistical matching results. The first level is the easiest to check, and measures to what extent the

marginal and joint distributions of variables in the donor sample are correctly reproduced in the fused file (that is, the recipient file completed with imputed columns).

The second level (preserving correlation structures and higher moments of variables) and the third level (preserving the true -but unknown- joint distribution of all variables) would be the most interesting to attain, since their fulfillment would assure that secondary analysis (such as regression models) based on the fused file would be grounded on reliable estimations of the missing data, and therefore correctly reflecting the true, unobserved, relations in the population.

Finally, the fourth level deals with preserving individual values, that is, it measures at what extent the matching procedure is capable of producing true (but unknown) values when completing the recipient file. This level is never assessed outside simulation studies (e.g. matching exercises simulating the A and B files by splitting an original survey, and checking if the true values are reproduced after matching). In any case, since statistical analyses based on the synthetic file will not refer to the individual values themselves –the primary reason why we want a fused file–, this level of validity is not of real interest in itself, and will not be regarded generally.

Software implementations

Focusing on the R free statistical computing environment, many statistical matching methods as well as methods related to other steps of the statistical matching process (such as checking coherence of sources or validity levels) are currently available to the social research community through several packages, some related to official statistics (such as StatMatch), and others to the missing data research (such as mice, Amelia, or BaBooN).

StatMatch (version by D'Orazio, 2013), developed within the framework of two ESSnet projects on data integration, implements nonparametric hot deck imputation methods (random, rank and nearest neighbour donor) that can be used to derive a synthetic data set; it also implements some mixed procedures based on predictive mean matching and methods to deal with data from complex sample surveys.

MICE (version by van Buuren, 2014) stands for multiple imputation via chained equations, and implements a powerful methodology to impute missing data that can be used for the statistical matching task. Each variable has its own imputation model and built-in imputation models are provided for continuous data (predictive mean matching, normal), binary data (logistic regression), unordered categorical data (polytomous logistic regression) and ordered categorical data (proportional odds). Also, various diagnostic plots are available to inspect the quality of imputations.

Case study: Application of statistical matching to the PISA and TALIS studies

Kaplan and Turner (2013) performed an experimental evaluation of a representative group of data fusion methods to link variables from the OECD PISA 2009 and TALIS 2008 studies using data from Iceland. Iceland is the only OECD country to implement both studies to all members (school principals, students, and teachers) of the relevant populations. For this reason, since all PISA and TALIS variables are measured for the

same units, Iceland data are ideal to assess the performance of statistical matching methods. In the words of his authors, the Iceland data study provided a proof of concept of statistical matching methodology applied to educational research, concluding that "statistical matching PISA and TALIS might be a reasonable option for countries that are unable to administer both surveys to the same sample schools". This is the case for Spain.

Beyond illustrating the statistical matching process itself, another reason for our study was to check if statistical matching could be effectively performed in a situation such as the Spain's PISA and TALIS studies, in order to extract combined conclusions from partial data. Specifically, we were interested in seeing what kind of achievements could be made, and what kind of problems do arise in practice, when using this methodology for combined statistical analysis of educational studies.

Method

Data

For the Spain's case, the PISA 2012 study surveyed 889 schools, containing responses for 889 school principals and 25313 students. TALIS 2013 surveyed 192 schools, containing responses for 192 school principals and 3339 teachers.

The matching process

First, we specified the objective of producing a synthetic file containing all the records in both studies (PISA and TALIS), with variables from both sources, in a similar fashion as in (Kaplan & Turner, 2012). That is, a micro approach was adopted. We focused our study to relate performance in mathematics, measured in PISA, to classroom processes between teachers and students, as measured in TALIS. As the Index of socioeconomic status (an important indicator of PISA) is a well known determinant of students' performance, this variable from PISA was also included as a target for our matching exercise.

In order to achieve a common level of analysis, data had to be aggregated to school level in both files. This circumstance implied an additional set of transformations, for example, computing the average (mean value) of plausible values in math for PISA schools and the average job satisfaction (which is, originally, a categorical ordinal variable with 4 categories) for TALIS schools. All the variables finally considered for the matching task, along with the corresponding name in the codebooks, can be seen in Table 1.

Next, considerable effort was made to identify and harmonise the common variables between the studies: for each variable (question) present in both files, the concordance of univariate distributions was assessed, both visually and by using specific statistical measures. Variables with low concordance between the files were discarded, and in some cases (such as the type of community where schools are located), some levels had to be aggregated to achieve comparable distributions.

Table 1

Matching variables in the PISA-2012 and TALIS-2013 codebooks

VARIABLES	SOURCE	CODE BOOK
COMMON VARIABLES	Sector: type of school, 2 levels: public, private.	PISA: SC02Q01 TALIS: BCG08
	Community: type of community where school is located, 2 levels (town/small town/village, city/ large city)	PISA: SC04Q01 TALIS: BCG10
	School size: number of student enrolled.	PISA: SC10Q01 TALIS: BCG12
	Student/teacher ratio: no. students divided by no teachers.	PISA: SCQ-STRATIO TALIS: BCG-STRATIO
	Shortage library materials: subjective measure of school´s material shortage by school principal (4 levels -degree of agreement that there is shortage-: not at all, very little, to some extent, a lot).*	PISA: SC11Q12 TALIS: BCG29H
	Disciplinary climate: subjejctive measure of disciplinary climate – by students in PISA and by teachers in TALIS (z-scores)	PISA: STQ - DISCLIMA TALIS: BTG-CCLIMATE
	Student-teacher relations: subjective measure of student/teacher relations by students in PISA and by teachers in TALIS (z-scores)	PISA: STQ-STUDREL TALIS: BTG- TSRELAT
	Teacher´s respons budget*: responsably of teacher´s with respect to allocating budget.*	PISA: SCQ- Q.24(f2) TALIS: BCG-q.31(e)
PISA-SPECIFIC VARIABLES	PV1MATH: first plausible value mathematics	PISA: STQ - PV1MATH
	PV2MATH: second plausible value mathematics	PISA: STQ - PV2MATH
	PV3MATH: third plausible value mathematics	PISA: STQ - PV3MATH
	PV4MATH: fourth plausible value mathematics	PISA: STQ - PV4MATH
	PV5MATH: fifth plausible value mathematics	PISA: STQ - PV5MATH
	ESCS: Index of socioeconomic and cultural status (mean value of school)	PISA: STQ - ESCS
	Metasum: summarising skills (meta-cognition)	PISA: STQ - METASUM
	Undrem: understanding and remembering skills (meta-cognition)	PISA: STQ - UNDREM
	Memor: use of memorisation strategies	PISA: STQ - MEMOR

TALIS- SPECIFIC VARIABLES	BTG31A: Teacher´s degree of job satisfaction	TALIS: BTG-q.31(a)
	SELFEF: measure of teacher´s self-efficacy (synthetic factor, centered to mean value 0 for all OECD countries)	TALIS: BTG-SELFEF. Syn. Factor of 4 items: BTG-q.31(b)-(c)-(d)-(e)
	TPSTRUC: Classroom teaching practice: structuring	TALIS: BTG-TPSTRUC. Syn. Factor of 5 items: BTG-q.42(b-h-i-m),q.30(c)
	TPSTUD: Classroom teaching practice: student-oriented	TALIS: BTG-TPSTUD Syn. Factor of 4 items: BTG-q.42(d-e-f-n)
	TPACTIIV: Classroom teaching practice: enhanced activities	TALIS: BTG-TPACTIIV Syn. Factor of 4 items: BTG-q.42(j-o-q-s)

Here, it is important to point out that we looked for all types of common information: beyond the usual characteristics such as school sector or the number of students, subjective measures of key aspects such as disciplinary climate or student-teacher relations were also considered for inclusion. As we sill see in the discussion, this is important to ensure that the uncertainty remaining after the matching process regarding the relation between specific variables is minimized.

The predictive power of the common variables was assessed with respect to the specific ones. This was done by inspecting bivariate relations in each of the school-level aggregated files, taking each common variable as an independent variable and each specific variable as dependent. Those variables with no predictive power were discarded (among them, teachers' responsibility in choosing textbook, and similar ones). Ultimately a set of common variables was selected by including only one variable for each set of clearly redundant variables (such as students' absenteeism or class disruption, corresponding to the same block of questions in the principal questionnaires). This is generally advisable in order to avoid potential computational problems in the matching step.

Next, the two school-level aggregated files were concatenated (e.g., records from both files were stacked one below the other, and an extra variable identifying the source -PISA or TALIS-, was added, see Figure 3). There were some missing values among the selected common variables; after exploring the missing data patterns (e.g. percentage of cases with missing values, missingness conditioned on the most important variables),

missing values were imputed by using the information from all the other variables by using mice package. The final file, with no other missing data rather than non-observed values, contains 871 data rows corresponding to PISA and 182 corresponding to TALIS. As an illustration, some records in this file are shown in Figure 3.

school ID	source	weight	sector	community	school size	student / teacher ratio	shortage library materials	1st p. value Maths	ESCS Index	job satisfaction	self-efficacy
97	pisa	2,01	Public	City/Large city	492	8,13	Very little	553,26	0,28	NA	NA
98	pisa	2,52	Public	town/small town/village	380	5,80	Very little	497,28	-0,47	NA	NA
99	pisa	3,02	Private	City/Large city	936	16,14	Very little	545,44	0,21	NA	NA
1001	talis	15,35	Public	town/small town/village	241	5,48	Not at all	NA	NA	3,00	-0,76
1002	talis	71,35	Private	town/small town/village	338	14,08	Not at all	NA	NA	3,17	-0,20
1003	talis	67,38	Private	town/small town/village	315	14,32	Not at all	NA	NA	3,26	-0,34

Figure 3. Six records from the concatenated file. This file was obtained by keeping only variables chosen for the matching task, choosing uniform, easy-to-use names, and finally stacking all the records in both files.

Finally the concatenated file was processed by function mice() in the mice package version 2.21, to perform predictive mean matching with m=5 imputations. The coding instructions were adapted from (Kaplan et al., 2013), and they are included here as an Annex. As a result we obtained 5 synthetic files with 1053 complete observations for all the specific variables included in Table 1. Figure. 4 gives a view of the first records in that file, for imputation number 1.

school ID	source	weight	sector	community	school size	student / teacher ratio	shortage library materials	1st p. value Maths	ESCS Index	job satisfaction	self-efficacy
97	pisa	2,01	Public	City/Large city	492	8,13	Very little	553,26	0,28	3,14	-0,29
98	pisa	2,52	Public	town/small town/village	380	5,80	Very little	497,28	-0,47	3,18	-0,43
99	pisa	3,02	Private	City/Large city	936	16,14	Very little	545,44	0,21	3,32	-0,21
1001	talis	15,35	Public	town/small town/village	241	5,48	Not at all	455,41	-0,74	3,00	-0,76
1002	talis	71,35	Private	town/small town/village	338	14,08	Not at all	491,74	-0,05	3,17	-0,20
1003	talis	67,38	Private	town/small town/village	315	14,32	Not at all	455,41	-0,66	3,26	-0,34

Figure 4. Six records of the synthetic file corresponding to imputation #1 after processing the concatenated file in Figure 3 with mice() function in the mice package. The R code is given in Annex 1.

It should be noted that in most cases Statistical Matching reproduces univariate distributions for specific variables as well as bivariate distributions between each specific and common variable.

The quality of results was assessed with respect to the first level of validity, via graphical analysis (density plots and quantile-quantile plots for univariate distributions, and boxplots for bivariate distributions) and by means of descriptive statistics measures comparing each pair of original and imputed distributions.

Results

Density plots for specific variables in PISA are shown in Figure 5; those in TALIS are shown in Figure 6. It can be seen that some variables are better reproduced than others. For instance, the TALIS variables TPSTUD (Classroom teaching practice: student-oriented) and TPACTIIV (Classroom teaching practice: enhanced activities) have lower performance. That is because these variables are not sufficiently determinated and explained by the common variables of the two databases. Moreower, it must be taken into account that the TALIS sample size is much smaller than the PISA. In order to illustrate how secondary analysis based on the multiply-imputed file may be carried out in mice we included some instructions for performing a multivariate linear regression in the Annex.

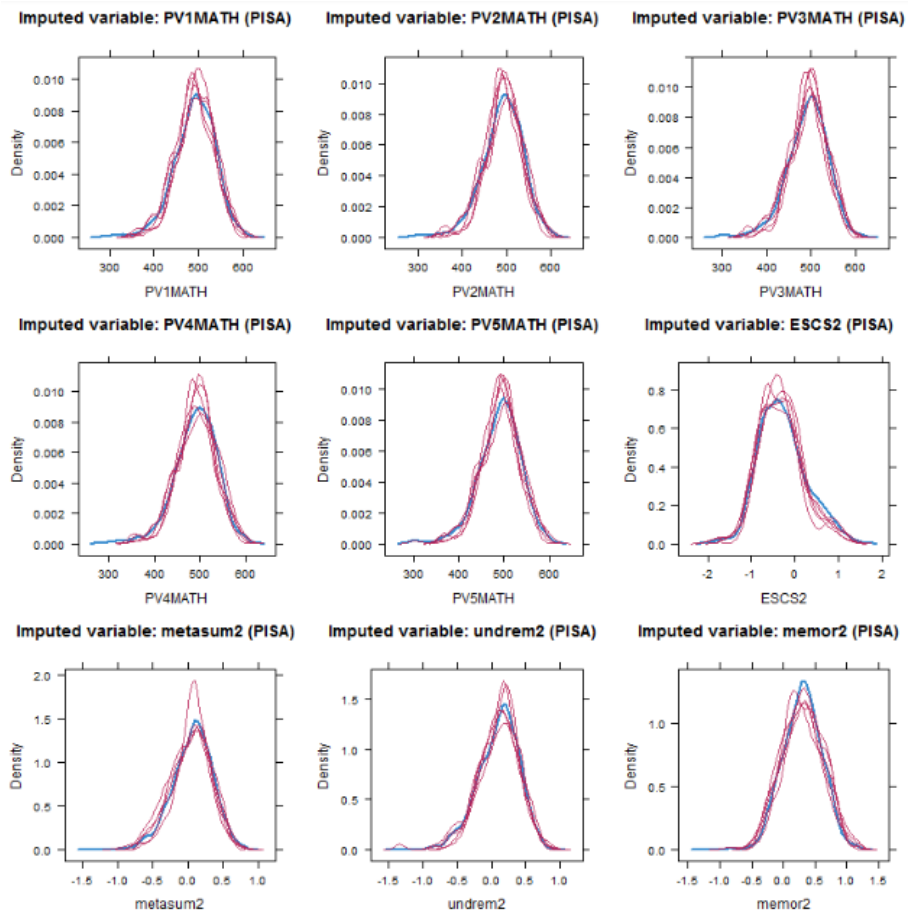

Figure 5. Density plots for PISA variables. Blue line showing original variable; red lines showing imputed variables accross m=5 imputations

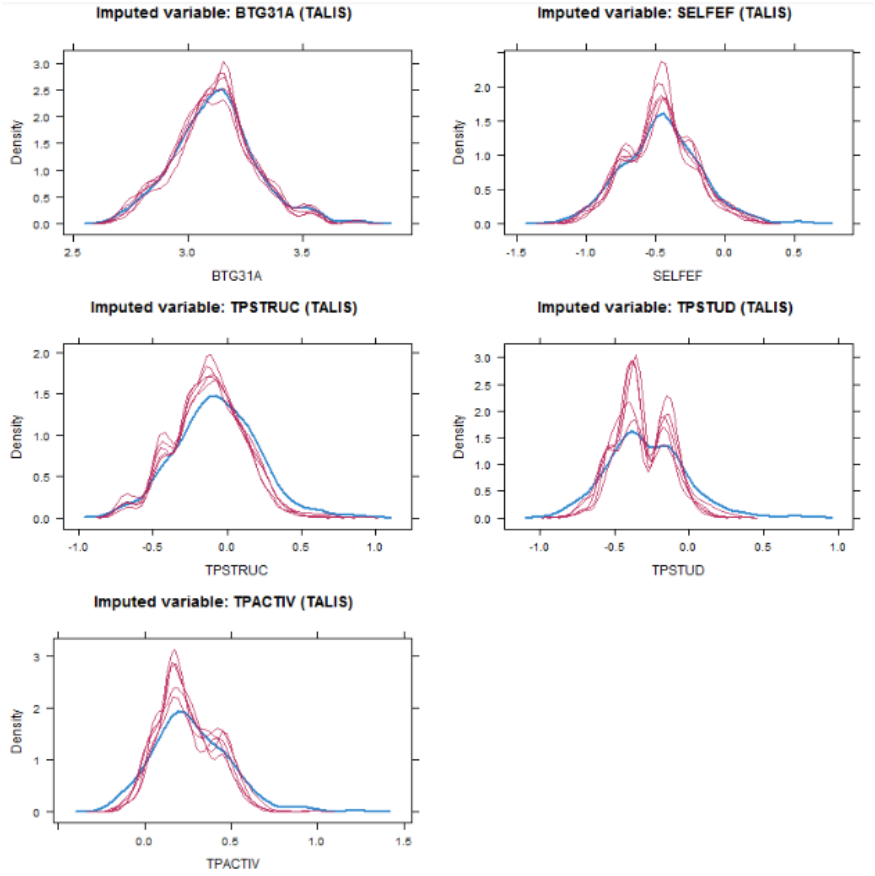

Figure 6. Density plots for TALIS variables. Blue line showing original variable; red lines showing imputed variables across m=5 imputations

Discusion

Our study for statistically matching PISA and TALIS data followed recommendations given in Leulescu and Agafitei (2013) and showed that fusing this kind of data is a feasible task provided that adequate software is used and the stages of the process are properly tackled. For Spain's data, a multiply-imputed synthetic file was obtained that correctly reproduced (in most of the cases) univariate distributions for specific variables as well as bivariate distributions (not shown here) between each specific and common variable.

The process of statistical matching comprehends a thorough search for all kind of common, comparable or harmonizable information between the files. This is important in order to make sure that uncertainty about the non-observed relation between specific variables is minimized. This key aspect of the process has been pointed out

by many authors (Jolani, Frank & van Buuren, 2014; Jong, van Buuren & Spiess, 2014). For example, in Leulescu and Agafitei (2013) it is said that "As no sample containing joint information on our target variables is (usually) available, the only possible solution would be to use proxy variables that can improve estimations if included in the imputation model." (p. 36). In our study, the proxy variables used to match PISA and TALIS were subjective measures of disciplinary climate and student-teacher relations (given by students in the former study, and by teachers in the latter).

PISA and TALIS are based in two independent survey samples, which are representative of each OECD member country. Different variables are measured in each survey; the level of the teachers and the schools in TALIS and the level of the students and the schools in PISA. Information units are different for each sample.

Statistical matching offers a potential bridge-tool between PISA and TALIS. The matching aggregates the PISA and TALIS files based on the school level, which is present in both files.

The main limitation of this case study is the reduced sample size of TALIS (192 schools) in comparison to PISA (889 schools), both in relative and absolute terms. As a consequence of the small sample, some sampling errors are spread to the whole study when the matching method is applied.

Any other approximation of multivariate analysis must firstly consider the existence of qualitative and quantitative variables. This means that the original data needs to be aggregated, which causes a substantial reduction of information; e.g. Multiple Factor Analysis between common and specific PISA and TALIS variables departing from a multiple contingence table.

Nevertheless, the development of useful indicators of the validity of statistical matching results still remains an open question. Also, Bayesian methods properly reflecting uncertainty in the matching process could be an interesting tool to carry out sensitivity analyses, that is, to see how the imputation results change when prior assumptions are made about the (non-observed) relation between the specific variables, conditioned to the common ones. This approach is presented in both Rässler (2002) and Kaplan and Turner (2012).

Mice was chosen because of the easy-to-use, powerful, well-documented techniques it contains. However, other choices would be equally valid. So far, SPSS can perform multiple imputations for missing disperse values of a unique file, by using linear or logistic regressions. The Statistical Matching method is not automaticaly implemented in SPSS. Nevertheless, last IBM-SPSS versions enable simultaneous work with R, using the SPSS databases. Thus, the R code must be activated from the SPSS application and results of the imputation come back in SPSS.

Conclusions

In this article we presented the statistical matching fundamentals as well as the main steps of the process. The potential benefits for the social sciences, particularly for educational studies, were shown. The concepts were illustrated by matching the PISA and TALIS studies for Spain's data files.

After indicating the way each step was solved, the main results were discussed. The exact variables used for matching were indicated in Table 1, with inclusion of their correspondence in the codebooks. R coding instructions were included for the chosen statistical matching method (predictive mean matching via chained equations) as well as for performing simple secondary analyses such as multivariate linear regressions based on the multiply-imputed file.

Although obtaining a fused file is relatively easy if all the stages are adequately tackled and proper software is used, assessing the validity of the fused file -understood as its capability to reflect the real correlations between the specific variables of interest- is not straightforward.

A possible application of this method is to analyse the influence that leadership has in the students' mathematics skills. Leadership is measured by two TALIS variables: BTG31A, teacher's degree of job satisfaction and FCSGCD, school's goals and the curricular development. The students' maths skills are measured by the following PISA variable: mPVMATH, mean of plausible value mathematics.

Based on the pooled imputed synthetic PISA & TALIS file, a general analyse could be run to study the influence of the considered TALIS variables on the PISA academic results.

Nowadays powerful software such as mice package in the free R software exists facilitating that kinds of analyses. We think our work may be valuable beyond educational studies, for any social science researcher interested in the efficient use of information from various independent sources.

References

Breakspear, S. (2012). The policy impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance. *OECD Journals, 71*, 1–32. doi:http://dx.doi.org/10.1787/19939019

Choi, A., & Jerrim, J. (2016). The Use (and Misuse) of PISA in Guiding Policy Reform: The Case of Spain. Comparative education. *56*(2), 230–245. doi:http://dx.doi.org/10.1080/03050068.2016.1142739

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical Matching: Theory and Practice.* NJ: Wiley.

D'Orazio, M. (2012). *StatMach: Statistical Matching. R package version 1.2.0*. Recuperado de http://CRAN.R-project.org/package=StatMatch

D'Orazio, M. (2013). *Statistical Matching: Metodological issues and practice with R-StatMatch (or. 69).* XXVI. Seminario Internacional de Estadística. Eustat. Recuperado de http://www.eustat.es/prodserv/seminario_i.html#axzz2sF9JV1rV

Eurostat (2008). *Recommendations on the use of methodologies for the integration of surveys and administrative data.* Recuperado de http://www.cros-portal.eu/sites/default/files//Report_of_WP2.doc

Fernández-Díaz, M. J.; Rodríguez-Mantilla J. M., & Martínez-Zarzuelo, A. (2016). PISA y TALIS ¿congruencia o discrepancia? *RELIEVE*, 22(1), art. M6. doi:http://dx.doi.org/10.7203/relieve.22.1.8247

González-Such, J., Sancho-Álvarez, C., & Sánchez-Delgado, P. (2016). Cuestionarios de contexto pisa: Un estudio sobre los indicadores de evaluación. *RELIEVE,* 22(1), art. M7. doi:http://dx.doi.org/10.7203/relieve.22.1.8274

Gustafsson, J. E. (2003). What Do We Know About Effects of School Resources on Educational Results? *Swedish Economic Policy Review,* 10, 77-110.

Jolani S, Frank L.E., & van Buuren S (2014). Dual imputation model for incomplete longitudinal data. *British Journal of Mathematical and Statistical Psychology*, 67(2), 197-212. doi: 10.1111/bmsp.12021

Jong R., van Buuren S., & Spiess M. (2014). Multiple imputation of predictor variables using generalized additive models. *Communications in Statistics - Simulation and Computation, 45*(3), 1-18. doi:http://dx.doi.org/10.1080/03610918.2014.911894

Kaplan, D., & Turner, A. (2012). *Statistical Matching of PISA 2009 and TALIS 2008 Data in Iceland (OECD Education Working Papers).* Paris: Organisation for Economic Co-operation and Development. Recuperado de http://www.oecd-ilibrary.org/;jsessionid=2ah7v0n0eg9ce.x-oecd-live-02content/workingpaper/ 5k97g3zzvg30-en

Kaplan, D., & Turner, A. (2013). *Data fusion with international large scale assessments: a case study using the OECD PISA and TALIS Surveys.* Springer-Verlag. Recuperado de http://link.springer.com/article/10.1186%2F2196-0739-1-6/fulltext.html

Leulescu, A., & Agafitei, M. (2013). *Statistical matching: A model based approach for data integration.* Luxembourg: European Commission, Eurostat. Publications Office.

OECD (2012). Pisa 2012. Recuperado de http://www.oecd.org/pisa/keyfindings/pisa-2012-results.htm

OECD (2013). *Talis 2013*. Recuperado de http://www.oecd.org/edu/school/talis.htm

Rässler, S. (2002). *Statistical matching. A frequentist theory, practical applications, and alternative Bayesian approaches*. New York: Springer.

Rubin. D.B. (1987). *An overview on multiple imputation*. Recuperado de www.amstat.org/sections/srms/Proceedings/papers/1988_016.pdf

Taut, S., & Palacios, D. (2016). Interpretaciones no intencionadas e intencionadas y usos de los resultados de PISA: Una perspectiva de validez consecuencial. *RELIEVE*, 22(1), art. M8. doi:http://dx.doi.org/10.7203/relieve22.1.8294

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software,* 45(3), 1–67.

van Buuren, S. (2014). *Mice. Imputation by random forests*. Recuperado de http://www.inside-r. org/ packages/cran/mice/docs/mice.impute.rf

Wheater, R. (2013). Achievement of 15 year olds in England: PISA 2012 national report. *OECD Programme for International Student Assessment.* Recuperado de https://www.nfer.ac.uk/publications/PQUK02/PQUK02.pdf

Annex I: R Code

R instructions for applying the predictive mean matching for statistically matching PISA and TALIS files by means of the mice package.

Need to have mice properly installed and loaded into the workspace. A data frame must be first prepared that includes concatenated PISA and TALIS records, with student-level and teacher-level variables aggregated to school-level and containing both common and specific variables.

1st step: select specific variables in PISA and TALIS

```
# PISA variables
varEspPisa <- c("PV1MATH", #STQ – PV1MATH
                "PV2MATH", #STQ – PV1MATH
                "PV3MATH", #STQ – PV1MATH
                "PV4MATH", #STQ – PV1MATH
                "PV5MATH", #STQ – PV1MATH
                "ESCS2",   #STQ – ESCS
                "metasum2",#STQ - METASUM
                "undrem2", #STQ - METASUM
                "memor2")  #STQ - MEMOR
# TALIS variables
varEspTalis <- c("BTG31A",  #BTG - q.31(a)
                 "SELFEF",  #BTG – SELFEF
                 "TPSTRUC", #BTG – TPSTRUC
                 "TPSTUD",  #BTG – TPSTUD
                 "TPACTIV") #BTG – TPACTIV
## 2nd step: select common variables

varCom <- c("sector",    #PISA: SC02Q01; TALIS: BCG08
      "community", #PISA: SC04Q01; TALIS: BCG10 (2 levels)
      "size",      #PISA: SC10Q01; TALIS: BCG12 (imputed missing values)
      "stratio",   #PISA: SCQ-STRATIO;TALIS:BCG-STRATIO (imputed m. values)
      "disclima",  #PISA: STQ-DISCLIMA;TALIS: BTG-CCLIMATE (z-scores)
      "studtea")   #PISA: STQ – STUDREL; TALIS: BTG- TSRELAT (z-scores)
## 3rd step: impute values

# Select columns (common & specific variables)

pisatalis.mice <- pisatalis[,c(varCom,varEspPisa,varEspTalis)]

# Prepare data to apply mice() – dry run.

ini <- mice(pisatalis.mice, max=0, pri=F)
```

```
# Force specific variables to take values between (min,max) of original variables. (Use
min(), max() functions).
#PISA variables
post["ESCS2"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-1.885,1.508))"
post["PV1MATH"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(288.1,614.6))"
post["PV2MATH"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(282.4,609.1))"
post["PV3MATH"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(291.0,617.2))"
post["PV4MATH"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(290.1,608.6))"
post["PV5MATH"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(294.2, 605.2))"
post["metasum2"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-1.381 , 0.796))"
post["undrem2"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c( -1.333,0.849))"
post["memor2"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-1.235,1.181))"
#TALIS variables.

post["BTG31A"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(2.700,3.733 ))"
post["SELFEF"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-1.177,0.518 ))"
post["TPSTRUC"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-0.715, 0.855 ))"
post["TPSTUD"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-0.863, 0.727 ))"
post["TPACTIV"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-0.201,  1.220 ))"
# Run mice with m=5 (five imputations) and 5000 iterations.

pisatalis.pmm <- mice(pisatalis.mice, m=5, maxit=5000, post=post)
# Obtain synthetic file
pisatalis.fused <- complete(x=pisatalis.pmm, "long")
## Perform secondary analyses with mice
# Multivariate linear regression
# Simple linear regression with original variables in PISA.
fit1 <- lm(PV1MATH ~ ESCS2, data= pisatalis) #Stores the model in fit1
summary(fit1) #Summary statistics for the model
# Include variables from (imputed) TALIS
# Need to have object of class “mice” (pisatalis.pmm)
fit2 <- with(pisatalis.pmm, lm(PV1MATH ~ ESCS2 + BTG31A))
summary(fit2) # Summary statistics for the model, for each of the imputed files
> summary(pool(fit2)) # Pooled results for the m imputed files
```

Fecha de recepción: 27 de junio de 2016
Fecha de revisión: 27 de junio de 2016
Fecha de aceptación: 06 de noviembre de 2016