CESifo Working Papers

7218
2018

August 2018

Linking Tax Morale and Personal Income Tax in Spain

Pilar Rey-del-Castillo, Jaime Villanueva-Garcia

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editors: Clemens Fuest, Oliver Falck, Jasmin Gröschl
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7218*
Category 6: Fiscal Policy, Macroeconomics and Growth

Linking Tax Morale and Personal Income Tax in Spain

Abstract

The paper presents a study of the relationship between the tax morale and the individual payments of personal income tax using the statistical matching of opinion polls with a representative sample of the personal income tax returns in Spain. As an initial step, the method selected to execute the match -imputations using Bayesian Networks- is described. The relationship between a proxy variable of the individual tax morale and other variables in the declared income tax file is later analyzed using the matched files. A first result is that tax morale increases with the level of declared wages, salaries and capital gains, while it has no link with declared business income.

JEL-Codes: C100, C650, H240, H260.

Keywords: statistical matching, opinion polls, personal income tax, tax morale.

*Pilar Rey-del-Castillo**
Institute for Fiscal Studies
Avda. Cardenal Herrera Oria, 378
Spain - 28035 Madrid
mpilar.rey@ief.hacienda.gob.es

Jaime Villanueva-Garcia
Institute for Fiscal Studies
Avda. Cardenal Herrera Oria, 378
Spain - 28035 Madrid
jaime.villanueva@ief.hacienda.gob.es

*corresponding author

1 Introduction

The current literature on tax evasion studies the notion of tax morale as the intrinsic motivation for paying taxes. The traditional economic model explaining the decision of taxpayers on how much income to declare to the tax authority has been improved by including different aspects of moral behaviour (Luttmer and Shingal, 2014). Following this approach, several socio-economic factors such as age, gender, religion, etc. have been identified as having significant impact on people's tax compliance decisions (Dhami, 2017).

Another variable studied in the literature in relation to tax morale is the individual's level of income (Dörrenberg and Peichl, 2013), a feature particularly difficult to measure. What is easier to obtain is the individual's declared income which differs from real income precisely because tax evasion occurs. On the other hand, one of the difficulties with the tax morale concept lies in the way to measure it, being opinion surveys one of the proposals (Alm, 2012).

Although the interesting economic question would be about the relationship between the level of income and tax morale, a possible and easier study is to connect the individual opinion and attitudes collected through opinion polls and related to tax morale, with the individual payments of personal income tax. The resulting analysis could be useful for fiscal authorities, improving the knowledge on the determinants of tax morale and ultimately, tax evasion. Nevertheless, there is not data source containing information on both variables at the same time.

A promising way to cope with the problem is *statistical matching* (Rässler, 2002; D'Orazio et al., 2006), which potentially increases the use and analysis of existing data sources. Statistical matching is a model-based approach for obtaining joint statistical information based on variables which are gathered through different sources. In this way, the proposed jointly analysis of tax morale and personal income tax can be performed using statistical matching, combining the microdata from opinion polls with the declared personal income tax records.

This paper shows the statistical matching of the opinion polls carried out by the Spanish Sociological Research Centre referring to the fiscal issue, with the file prepared as a representative sample of the Personal Income Tax in Spain. The relation between some variables associated with the declared income tax payment –such as the declared income or the family size– and a certain proxy variable of the tax morale are later analyzed using the files obtained through the matching procedure. It should be noted that the use of Bayesian networks implies that the obtained impacts of one variable on another must be understood as predictions based on possible influences and not as exact causal effects (Pearl, 2000).

The paper is organised as follows. The next section describes the original data, that is, the fiscal opinion polls and the file including the declared personal income tax sample. The matching procedure methodology based on Bayesian Networks is detailed in the third section. The forth section presents the results of the analysis of the relation between the selected variables and, finally, a summary and conclusions are presented on section five.

2 Original microdata files

This section presents the original data used for the matching between the opinion polls and the personal income tax sample.

2.1 Annual fiscal opinion poll

The Sociological Research Centre (CIS) is an independent entity assigned to the Ministry of Presidency in Spain. The main mission of the CIS is to contribute to the scientific knowledge of the Spanish society, gathering the necessary data for research in very different fields, from trends in public opinion to applied research. In particular, apart from the results of its polls and studies, the CIS distributes the anonymized microdata files of its surveys to researchers.

Since 1986 the CIS carries out each year one opinion poll gathering information about perceptions and attitudes to fiscal policies whose sample size is close to 2,500 individuals and refers to the population living in Spain. Over last years the survey has included a question on how conscious and responsible is the respondent in fulfilling his/her fiscal duties, with options to reply including "*very conscious and responsible*", "*quite conscious and responsible*", "*little conscious and responsible*" and "*very little conscious and responsible*". There are also questions related to socio-demographic information (sex, age, economic activity, religion...) and questions on the opinion about the fiscal authority in Spain, the fairness of the fiscal laws, etc.

Although there are some doubts about whether the attitudes shown in opinion surveys are indicative of the behaviour in real life (Onu, 2016), this paper tries to measure the tax morale focusing on the first cited question. A variable that has a similar meaning or may be used –in a certain sense– as a proxy for tax morale or tax mentality is created by recoding its options of reply. Thus, the new variable called *"tax responsibility"* is calculated identifying the replies "*very conscious and responsible"* and "*quite conscious and responsible"* with "*Yes"* and "*little conscious and responsible"* and "*very little conscious and responsible"* with "*No"*. This recoding in just two categories seems more appropriate to compare with other measures of tax morale which use binary variables.

Table 1 shows the results for the proposed variable for the years between 2012 and 2017. It can be seen that the values are quite similar, which makes us consider all the surveys jointly as a single one with the aim of supporting more robust inferences. The sample size of this whole survey is 14,546.

Table 1 Distribution of *tax responsibility* on CIS files

Year	***Tax responsibility* (%)**	
	Yes	***No***
2012	90.07	9.92
2013	89.56	10.43
2014	90.47	9.53
2015	91.52	8.48
2016	91.15	8.84
2017	90.92	9.10
ALL	90.62	9.38

The distribution observed in Table 1 suggests that the *tax responsibility* (*TR* hereinafter) could be used as a proxy to tax morale because it is quite similar to the measure of tax morale in Germany through the 11% estimate of participants finding tax evasion acceptable (Dörrenberg and Peichl, 2017). Its joint analysis with others variables in the survey also finds similar results to those described in empirical studies, e.g., elderly people, women or married people are more responsible for their fiscal duties (Cummings et al., 2008; Dörrenberg and Peichl, 2013). This kind of analysis is not the aim of this paper.

Although there are other variables in the CIS data, our attention is focused on the *TR* variable. Nevertheless, for the practical implementation of the statistical matching, other socio-demographic variables are also taken into account as candidates for matching variables: *autonomous community of residence, sex, age, marital status* and *housing type*.

2.2 Sample of Personal Income Tax returns 2014

A representative sample of the declared Personal Income Tax (hereinafter PIT) microdata is prepared each year jointly by the State Agency for Tax Administration (AEAT) and the Spanish Institute for Fiscal Studies to be used by researchers. The 2014 sample includes the values of a number of variables from the tax returns of the taxpayers living on the territory under the authority of the AEAT (territory of Spain with the exception of País Vasco and Navarra), corresponding to 2,202,675 returns. A number of variables especially interesting have been chosen for this study:

- Wages and salaries
- Capital gains
- Business income
- Tax exemption for donation
- Family size (computed from other variables; the calculations are described in the corresponding section)

Some of these numeric variables have a big number of zero and/or negative values. For the purpose of using later discrete Bayesian networks, and also to improve the understanding of the results and obtain more robust inferences, they are discretized into categorical variables. The first breakdown in the four first variables is determined by the proportion of negative or zero values, while the "*low*", "*medium*" and "*high*" parts are obtained by trying to split into three sections representing similar proportion of the corresponding variable in the population. The breakdowns in the fifth variable (family size) are directly obtained. The proportions in the global population of the returns for the corresponding categories in PIT 2014 file appear in tables 2 to 6.

Table 2 Distribution of *wages and salaries* in PIT returns

Wages and salaries	% returns
Negative or zero	15.4
Low	28.1
Medium	28.3
High	28.2

Table 3 Distribution of *capital gains* in PIT returns

Capital gains	% returns
Negative or zero	27.1
Low	24.2
Medium	24.1
High	24.6

Table 4 Distribution of *business income* in PIT returns

Business income	% returns
Negative or zero	84.7
Low	5.2
Medium	5.2
High	4.9

Table 5 Distribution of *donation* in PIT returns

Tax exemption for donation	% returns
Zero	85.1
Low	4.9
Medium	4.9
High	5.0

Table 6 Distribution of *family size in* PIT returns

Family size	% returns
One person	66.6
Two people	16.2
More than two people	17.1

Apart from these variables of interest, the socio-demographic variables *autonomous community of residence, sex, age, marital status* and *housing type* −common with the CIS dataset− are considered to perform the statistical matching. Other common variables, e.g., the *number of descendants,* have not been considered as candidates for matching because their definition and measurement are not harmonized in both datasets.

3 The statistical matching of the microdata files

Given the close link between tax morale and tax compliance supported by available literature (Feld and Frey, 2006; Cummings et al., 2008; Dörrenberg and Peichl, 2013; Onu, 2016), a possible exercise to perform is the study of the relationship between real features −such as the declared *wages and salaries* or the declared *capital gains* of taxpayers− with their attitudes to their tax obligations measured by the *tax responsibility*. However, no single survey covers simultaneously these aspects.

The two datasets described in the previous section refer to the same target population with independent and non overlapping samples. One way to cope with the problem is *statistical matching* (also known as *data fusion*) of the two files (Rässler, 2002; D'Orazio et al. , 2006), where the individual records from the two sources are linked on the basis of their similarity on a number of features that are measured in each source. Formally, given Y and Z two random variables, statistical matching is defined as the estimation of the (Y, Z) joint distribution or of some of its parameters when:

- Y and Z are not jointly observed in a survey, but Y is observed in sample A and Z is observed in sample B,
- A set of additional variables X are observed in both A and B,
- A and B are independent and the set of observed units in the two samples do not overlap.

The extension of the definition to the multivariate context −where Y and/or Z are a set of variables− is trivial and straightforward.

The execution of the statistical matching is not a simple operation, requiring technical expertise and raising different methodological questions. In practice, it can be seen as an imputation problem: e.g., of the target variable Y on file B (called *recipient* file) through a model of imputation of Y from X, built from data on file A (called *donor* file). Our focus is on what is sometimes known as *statistical matching on the macro level*, i.e. the study of the relationship between Y (*TR* in our case, just observed in CIS) and Z (PIT variables), and not on the creation of a synthetic dataset including both Y and Z (*statistical matching on the micro level*) (De Waal, 2015).

Usually, apart from the selection of an appropriate matching technique, some key steps must be followed to apply the statistical matching (Leulescu and Agafitei, 2013):

- Harmonisation and reconciliation of sources,
- Analysis of the explanatory power for common variables,
- Selection and application of the matching method,
- Quality assessment.

These steps are described in detail in the next sections.

3.1 Harmonisation and reconciliation of sources

In order to assess whether data files can be compared and integrated, it is necessary to evaluate if they are coherent. Different aspects must be considered for reconciliation: definition of units, reference period, population, variables, classifications, etc.

The population for the CIS poll is the people living in the Spanish territory 18 or more years older, while the PIT data refers to the taxpayers living in Spain with the exception of País Vasco and Navarra. Although the CIS sample refers to all citizens and the income tax data sample just to taxpayers, in a first approach the CIS data can be reconciled with PIT by considering exclusively the subset of units not living in these two regions, resulting in a CIS reduced sample of 13,644 units.

The reference period is 2012-2017 for CIS data and 2014 for PIT data. Nevertheless, it has been considered interesting to use the CIS data of several years to increase the sample data used as basis for the calculations, given the case that the opinions seem to be stable enough through the period.

Another essential point is the existence of a common set of variables that should be homogeneous in their statistical content, that is, the two sample surveys should represent the same population. The common variables considered as candidates are the *autonomous community of residence, sex, age, marital status* and *housing type*. A first step of reconciliation has been done to harmonise the categories of the *housing type*: in CIS the categories were detailed (*property totally paid, property partially paid, property by inheritance or donation, renting, other)*, while in PIT file they were more aggregated (*property, renting, other)*. The different types of property in CIS have been grouped into the category "*property"* to make compatible the two categorizations.

The first way to assess the degree of similarity/dissimilarity between different variables in both datasets is to compare their frequency distributions (weighted in case of PIT data). This degree can also be quantified by computing the Hellinger distance (Pollard, 2002) between the corresponding distributions:

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{m} \left(\sqrt{p_j} - \sqrt{q_j}\right)^2}$$

where $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_m)$ are the proportions of the observed values that fall into the corresponding m categories or intervals for each of the two distributions. Although it is set up on an arbitrary basis, it is a generally accepted rule for univariate distributions to consider two distributions close if the Hellinger distance is not greater than 0.05.

Table 7 shows the frequency distributions and the Hellinger distance computed between the CIS and the PIT data distributions for the five variables considered. For this purpose, a harmonization procedure has been previously applied to the *age* variable, recoding it in different categories than those usually employed in CIS in order to improve the similarity of the distributions.

Table 7 Comparison of distributions of variables in CIS and PIT files

Variable	CIS (%)	PIT (%)	Hellinger distance
Autonomous community			0.0328
Andalucía	19.1	16.8	
Aragón	3.1	3.5	
Asturias	2.7	2.6	
Baleares	2.4	2.5	
Canarias	4.6	4.0	
Cantabria	1.4	1.4	
Castilla La Mancha	4.7	4.6	
Castilla León	6.1	6.3	
Cataluña	16.9	18.0	
Ceuta	0.1	0.1	
Extremadura	2.7	2.4	
Galicia	6.9	6.6	
La Rioja	0.7	0.8	
Madrid	14.1	16.0	
Melilla	0.1	0.1	
Murcia	3.1	3.0	
Valencia	11.4	11.1	
Sex			0.0512
Man	48.7	55.9	
Woman	51.3	44.1	
Age			0.0628
From_35_to_49	29.5	35.8	
From_50_to_64	23.9	25.8	
From_65_onwards	22.2	19.8	
Up_to_34	24.3	18.6	
Marital status			0.0238
Divorced	6.9	7.2	
Married	53.9	56.4	
Single	31.5	30.1	
Widow	7.6	6.3	
Housing property			0.0583
Other	18.6	22.9	
Property	75.4	68.3	
Renting	5.9	8.8	

The average Hellinger distance for the five variables is 4.6%. Overall, both data show similar distributions and there are no grounds for rejecting the hypothesis of coherence of variables. Thus, it can be established that the common variables meet the pre-conditions for their selection as matching variables in the matching process.

3.2 Analysis of the explanatory power for common variables

The validity of a matching exercise depends largely on the explanatory power of the matching variables to predict the specific information to be transferred from the donor to the recipient file. For the matching it is recommended to include all the common variables in both files that show some significant relation in the donor file with the variables to be imputed, the *TR* in our case.

Table 8 Association of matching variables with *tax responsibility* in CIS file

Variable	Adjusted Pearson coefficient	Cramer's V coefficient
Autonomous community	0.101	0.084
Sex	0.085	0.060
Age	0.184	0.145
Marital status	0.157	0.126
Housing type	0.068	0.052

A summary overview of the global explanatory power of each common variable with the *TR* in CIS is provided in Table 8 through two traditional measures of association for categorical variables, the adjusted Pearson contingency coefficient and the Cramer's V coefficient (Agresti, 1996) that are scaled between 0 and 1. Results of standard tests are not provided because –as CIS dataset are a large sample– the *p-values* produced are always near zero, resulting in any minor effect becoming statistically significant (Lin et al., 2013). From the table, all five variables show a certain association with *TR*, being *age* and *marital status* the variables with the strongest association. Although the individual explanatory powers are not strong, considering all together may improve their predictive performance.

3.3 Selection and application of the matching method

Many different algorithms and procedures have been used over the past years to perform statistical matching. A traditional way to cope with the problem is to assume the *Conditional Independence Assumption*, that is, that *Y* and *Z* variables are probabilistically independent conditionally on the common matching variables *X*. But in several situations this assumption may not be justified and it cannot be tested from the data sets. Due to the fact that *Y* and *Z* are never jointly observed, there is an intrinsic uncertainty related to the inferences, and therefore there are no unique estimates for the parameters measuring the association between these variables. What can be done in practice is to draw inferences considering all the compatible distributions. The advantage of the method here proposed is that a number of different matched files can be generated through simulations, making it possible to obtain point and interval estimates.

The traditional non-parametric method of statistical matching to impute a non-observed variable for a record in *Z* –*random hot deck imputation*– randomly chooses a donor record from the group of similar records in *Y*, groups which are developed by exploiting the information of the common matching variables.

In order to impute to each PIT unit the *TR* variable, Bayesian Networks (Pearl, 1988) are proposed in this paper. This kind of networks are particularly useful for reducing the complexity of a phenomenon by representing joint relationships between a set of variables through conditional relationships between subsets of these variables. An important attribute of the imputations based on Bayesian Networks is that they preserve the joint relationships between variables (Di Zio et al., 2004; Hruschka et al., 2007). And the advantage of Bayesian networks over random hot deck imputations is that the grouping of similar records based on assumptions of relationships and dependences between variables are substituted by a rigorous probabilistic model.

A Bayesian network is a directed acyclic graph in which nodes represent random variables of interest, and directed edges between nodes (also called *arrows* and *arcs*) indicate conditional dependency relationships. The structure of the network has the potential to explicitly handle complex relationships not easily captured by simpler functions. This structure can be learned from data using a search algorithm through the space of all possible structures, and a score metric to assess its suitability can be computed. The joint probability distribution of a set of variables using the compact and structured representation provided by Bayesian networks can be estimated if a large enough dataset is given including the corresponding variables. Once the network parameters –that is, the structure of dependences and the probabilities for each node– have been learned from a set of training samples, the network can be used for inference, e. g., to impute the value of any subset of variables given evidence or conditioned to the values of any other subset (Heckerman, 1998). Many algorithms have been developed –both exact and approximate– to perform the probabilistic calculations (Pear, 1988; Cowel et al., 1999).

Figure 1

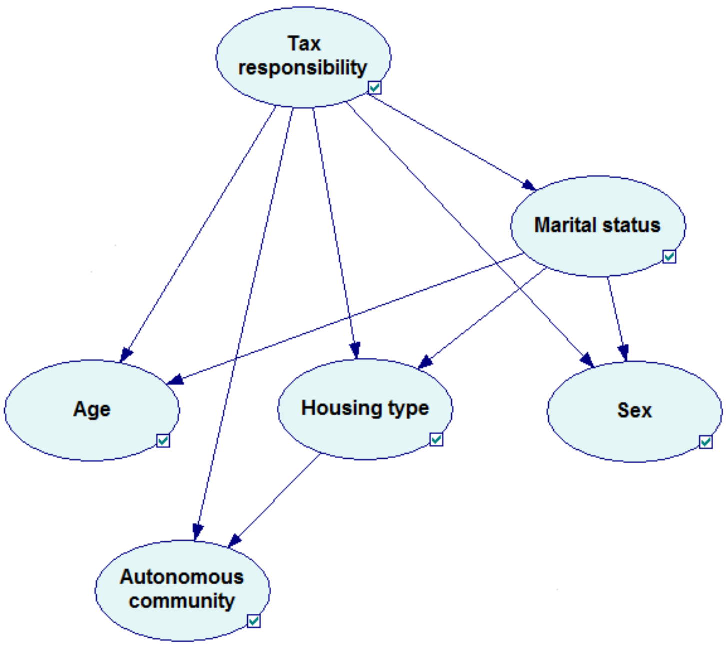


Figure 1 shows the structure of a Bayesian network linking the matching variables and *TR*, learned from the CIS data using the *Tree Augmented Naïve Bayes* search algorithm (Friedman et al., 1997). It must be stressed that similar results to the ones here presented may be obtained using other possible network structures computed using other search algorithms, whenever these structures of dependences show direct arcs between tax responsibility and the rest of the nodes. The reason is that, in this case, the conditional probabilities are computed directly from the corresponding observed frequencies.

It can be shown that the network –including all possible marginal and conditional distributions– is completely determined by this structure of dependences plus the tables of probability of each node conditioned to the values of its parents, that is, to the nodes from which there is an edge directed to the given node. These conditional probabilities are obtained as maximum likelihood estimates (Cower et al., 1999; Di Zio et al., 2004).

Once the network has been learnt from CIS data, it can be computed, for each unit in PIT dataset with known values for the five matching variables, the probability that *TR* would take the value "*Yes*" conditioned to these values. From these probabilities, it is possible for each record in PIT file to effectively impute *TR* by randomly drawing a value from the distribution of *TR* conditioned to the actual values of the matching variables. This technique obtains imputations that preserve as much as possible the original relationships between variables (Di Zio et al., 2007) and –a particularly interesting attribute– allows for obtaining different versions of the synthetic dataset. Using the Python software (Van Rossum, 1995), 1,000 independent simulations of the PIT completed dataset –started with different random seeds– have been generated. The analysis of these datasets allows for obtaining estimates for the parameters of interest, as will be shown in following sections.

3.4 Quality assessment

The first quality assessment in a matching exercise is traditionally focused on the comparison of marginal and joint distributions in the donor and matched datasets. In this case, our matched files are the simulated datasets that include the imputed *TR* for the PIT records.

Table 9 Hellinger distance for bivariate joint distributions in CIS and PIT files

Variables	Autonomous community	Sex	Age	Marital status	Housing type
Autonomous community	.	0.063	0.075	0.049	0.098
Sex	0.063	.	0.084	0.070	0.078
Age	0.075	0.084	.	0.070	0.130
Marital status	0.049	0.070	0.070	.	0.095
Housing type	0.098	0.078	0.130	0.095	.

The assessment of the similarity between the marginal distributions of the common variables –that is, the five matching variables– in the original CIS and the matched files has been done when performing the reconciliation of sources. Thus, what should be assessed now is the similarity between the joint (bivariate) distributions of each matching variables jointly with *TR,* in the CIS data and the imputed PIT datasets. Table 9 provides a summary overview of the similarity between the bivariate distributions generated by the matching variables in CIS and PIT datasets through the corresponding Hellinger distances. The values are not large, with maybe the exception of the (*age, housing type*) joint distribution having a Hellinger distance of 0.130. The average for these bivariate distributions is a Hellinger distance of 0.081.

To complete the previous assessment, what remains to be checked is the similarity between the marginal distribution of *TR* and the joint distributions of *TR* with the matching variables in the original CIS and the imputed PIT files.

To start with, the 9.38% proportion of negative *TR* units in CIS (Table 1) can be compared in Table 10 with the 9.18% average proportion in the 1,000 simulated PIT datasets. It can be seen that the corresponding 0.001 average Hellinger distance is really small, confirming that the used procedure preserves the distribution of *TR* in the matched datasets.

Table 10 Results of simulations for *tax responsibility*

	Mean	Std	Min	Max
Percentage of *TR* = "*No*" in PIT matched files	9.18	0.10	8.87	9.52
Hellinger distance to *TR* distribution in CIS	0.001	0.001	0.000	0.005

The connection between the different simulations poses an interesting question. Are they so similar in order to always produce imputations reproducing so well the original CIS distribution? Taking the first simulation produced as a baseline, Table 11 shows the average result of the comparison between its imputed values for *TR* and the ones imputed for the 999 resting simulations. As the simulations are random and independently produced, this can be an estimate for the coincidence/non-coincidence in the imputations.

Table 11 Average result for simulations
Number of imputations (%)

Value of imputation	*TR* = "*Yes*" in other simulations	*TR* = "*No*" in other simulations
***TR* = "*Yes*" in first simulation**	81.9	8.3
***TR* = "*No*" in first simulation**	8.3	1.3

It can be seen in Table 11 that there are, on average, 16.6% of non-coincident imputations, and only 1.3% are coincidently imputed as negative *TR*. Given that the categories are imbalanced, the simulations are different enough, while they preserve the original distribution of *TR* in CIS dataset.

For assessing now the joint distributions of *TR* with the matching variables, Table 12 shows the averages of the Hellinger distances between the bivariate distributions of each matching variable jointly with *TR* in CIS and PIT for the simulated files. These distances can be considered acceptable in all cases.

Table 12 Results of simulations for Hellinger distances of joint distributions of matching variables and *tax responsibility* in CIS and PIT

Variable	Mean	Std	Min	Max
Autonomous community	0.034	0.0002	0.033	0.035
Sex	0.052	0.0002	0.051	0.052
Age	0.063	0.0001	0.063	0.063
Marital status	0.026	0.0004	0.024	0.027
Housing type	0.060	0.0003	0.059	0.061

Finally, the figures in Table 13 provide additional information on the quality of the match, showing the corresponding adjusted Pearson and Cramer's V coefficients. The values of the association are of similar order of magnitude to those on the original CIS dataset in Table 8, with a certain increase for *housing type*. But this does not seem grave enough to critically jeopardize the success of the matching.

Table 13 Association of matching variables with *tax responsibility* in PIT matched files
Results of simulations

Variable	Adjusted Pearson coefficient				Cramer's V coefficient			
	Mean	**Std**	**Min**	**Max**	**Mean**	**Std**	**Min**	**Max**
Autonomous community	0.108	0.002	0.102	0.114	0.089	0.002	0.085	0.095
Sex	0.057	0.005	0.044	0.073	0.041	0.004	0.031	0.051
Age	0.170	0.005	0.156	0.186	0.134	0.004	0.123	0.147
Marital status	0.125	0.005	0.111	0.146	0.098	0.004	0.087	0.115
Housing type	0.114	0.005	0.098	0.129	0.087	0.004	0.075	0.098

4 Relationship of tax responsibility to other variables

From the results in previous section, in a first approximate approach the matching procedure may be considered acceptable. Recognising that tax morale serves as an explanation for tax compliance (Feld and Frey, 2006; Cummings et al., 2008), it is now interesting to explore the determinants of tax morale. For this purpose, the relationship between *TR* −as a proxy to tax morale− and some variables in PIT file will be analysed.

The parameter describing the behaviour of interest through the following paragraphs will be the percentage of units with negative *TR*, whose distribution on the simulated datasets was shown in Table 10.

4.1 Declared income

A first analysis gives attention to the impact of the declared income on tax morale. What has been more frequently studied is the relationship of tax morale and real income. As noted by Dörrenberg and Peichl (2013), the theoretical effects of real income on tax morale are not clear, finding ambiguous empirical evidence in the literature. There are reports referring to different countries that find a negative relationship while others do not discover significant effects. The models currently used for the economic analysis of an individual taxpayer's compliance decision assume that declared income may differ from the real one just because tax evasion occurs (Hashimzade et al., 2010). From this approach, the relationship of *tax responsibility* and declared income is unknown and may take place, or not, in any direction.

In the case of this study, the matched and simulated files allow for studying separately the impact of different types of declared income: *wages and salaries*, *capital gains* and *business income*. The analysis will focus on the recoded categorical variables as defined in section 2.2. As an initial assessment, Table 14 shows the measures of association between the income and the imputed *TR*.

Table 14 Association of *income* with *tax responsibility* in PIT matched files
Results of simulations

Variable	Adjusted Pearson coefficient				Cramer's V coefficient			
	Mean	**Std**	**Min**	**Max**	**Mean**	**Std**	**Min**	**Max**
Wages and salaries	0.026	0.005	0.013	0.042	0.021	0.004	0.011	0.033
Capital gains	0.081	0.004	0.063	0.093	0.064	0.004	0.050	0.073
Business income	0.012	0.003	0.002	0.021	0.009	0.002	0.002	0.016

It can be seen through the coefficients that a certain relationship between the different income and *tax responsibility* exist. The strongest would be the one corresponding to *capital gains*, while the one to *business income* is almost nonexistent.

Another perspective confirming this assessment is presented in Figure 2 and Tables 15, 16 and 17, which show summary results of each source of income for the 1,000 simulated datasets. Results of traditional tests of difference of groups are not provided because of the problems with using *p-values* in very large samples (Lin et al., 2013). Nevertheless, some conclusions can be extracted from the tables.

Figure 2. Declared income: negative tax responsibility (%)

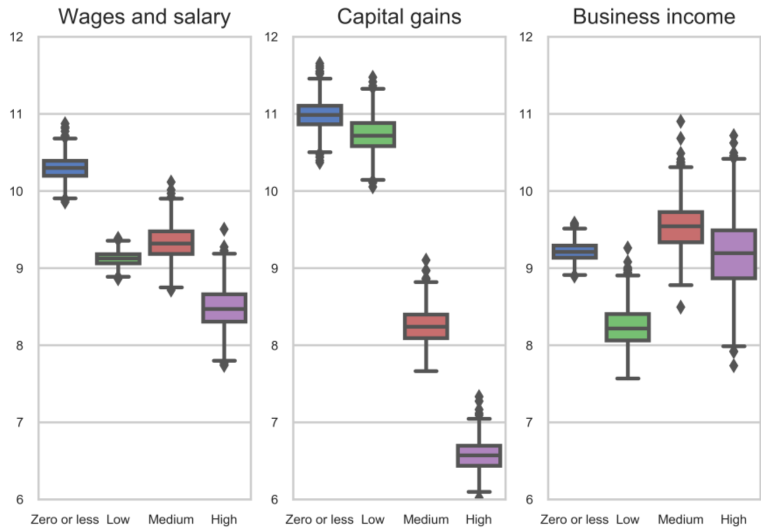


Table 15 Percentage of *tax responsibility* = *"No"*
Results of simulations for *Wages and salaries*

Category	Mean	Std	Min	Max
Negative or zero	10.3	0.15	9.86	10.87
Low	9.12	0.09	8.86	9.39
Medium	9.33	0.21	8.72	10.12
High	8.48	0.27	7.74	9.51

First of all, the figures for *wages and salaries* in Table 15 show a clear breakdown in the level of tax responsibility between the category *negative or zero* on the one hand, and the rest on the other (although not of a big magnitude), while there is no drop between *low* and *medium* categories. Thus, grouping low and medium categories, it may be said that the *wages and salaries* have a positive impact on *TR*, or, as is directly seen in the table, increases in its value are linked to decreases in the percentage of taxpayers with a "*No*" value for *TR*. The lowest *tax morale* is found in the group with negative or zero declared *wages and salaries.* In other words, the most important levels of tax evasion (as opposed to tax morale) arise at negative or zero employment declared income, and the lower in the group with high *wages and salaries*. This result is somehow consistent with one study in Italy that estimates tax evasion (as opposite to tax morale) comparing declared income with replies to a survey (Fiorio and D'Amuri, 2005).

Table 16 Percentage of *tax responsibility* = "No"
Results of simulations for *Capital gains*

Category	Mean	Std	Min	Max
Negative or zero	10.99	0.20	10.37	11.65
Low	10.73	0.23	10.05	11.48
Medium	8.25	0.22	7.66	9.10
High	6.57	0.19	6.02	7.33

The case of the *capital gains* income source in Table 16 is similar, but the breakdown on the level of *TR* between categories is always on the same direction and of higher magnitude, that is, capital gains have a positive impact on *tax responsibility*. Consequently, the biggest proportion of possible tax evaders is found on taxpayers with negative or low capital gains.

Table 17 Percentage of *tax responsibility* = "No"
Results of simulations for *Business income*

Category	Mean	Std	Max	Min
Negative or zero	9.21	0.11	8.90	9.58
Low	8.24	0.26	7.57	9.26
Medium	9.55	0.31	8.49	10.90
High	9.20	0.47	7.73	10.72

Lastly, there is not a clear result for the relationship between *business income* and *TR* in Table 17, because there is no apparent breakdown in the *TR* level through the categories. It must be noted that this type of income is easier to conceal and only a small proportion of the returns have a greater-than-zero value (16% of the PIT returns from Table 4), in contrast to 84% of returns for *wages and salaries* and 73% for *capital gains*. Another consideration to be taken into account is that taxpayers can obtain income from more than one income source, which may somehow distort the obtained figures.

As summary result, the *wages, salaries and capital gains* sources of declared income have a clear impact on tax morale, while the *business income* does not seem to show significant relationship. These results are in apparent contradiction with the cited results about the real income, but being different measures of income, the straightforward comparisons are not completely fair.

4.2 Donations

There is some literature recommending the use of charitable contributions instead of declared income for studying the relationship between income and tax evasion. The reason is that this type of donations may be related to the level of real income, and people can be more likely to declare charitable contributions while concealing real earnings (Feldman and Slemrod, 2007).

Table 18 Association of *donations* with *tax responsibility* in PIT files
Results of simulations

Adjusted Pearson coefficient				Cramer's V coefficient			
Mean	**Std**	**Min**	**Max**	**Mean**	**Std**	**Min**	**Max**
0.023	0.005	0.009	0.038	0.018	0.004	0.007	0.030

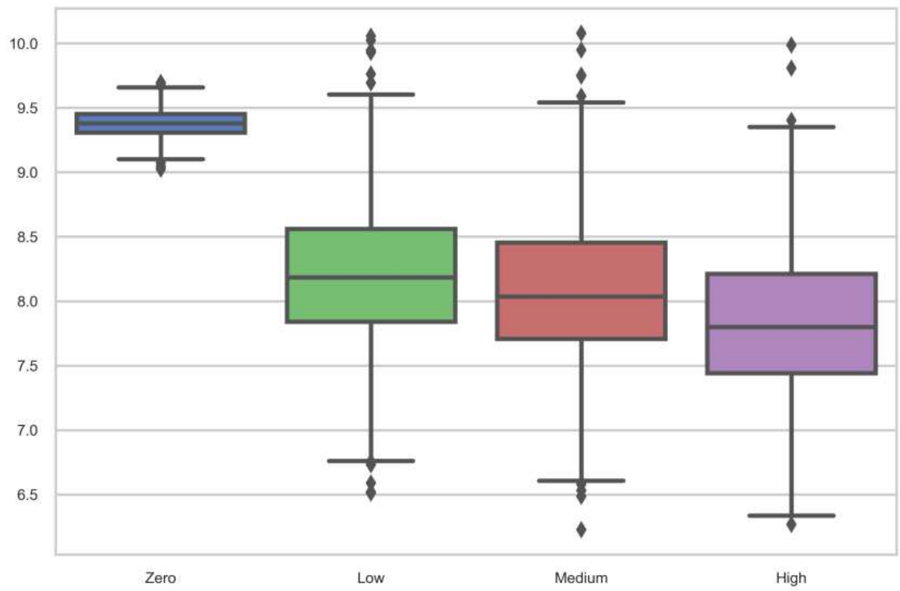
As a proxy analysis using the PIT dataset, what can be studied is the relationship between *tax exemptions for donations* and *TR*. Measures of the association appear in Table 18, where the existence of a minimal relationship can be seen.

Other figures confirming this outcome are shown in Table 19 and graphically in Figure 3 through the mean results for the 1,000 simulated datasets. There seems to be a positive relationship between tax responsibility and donations, although weak in intensity. A possible reason is that donations do not distinguish between taxpayers with different sources of income.

Table 19 Percentage of *tax responsibility = "No" in PIT files*
Results of simulations for *tax exemptions for donations*

Category	Mean	Std	Min	Max
Zero	9.38	0.11	9.02	9.70
Low	8.20	0.54	6.51	10.06
Medium	8.08	0.56	6.23	10.08
High	7.83	0.57	6.27	9.99

Figure 3. Donations: negative tax responsibility (%)



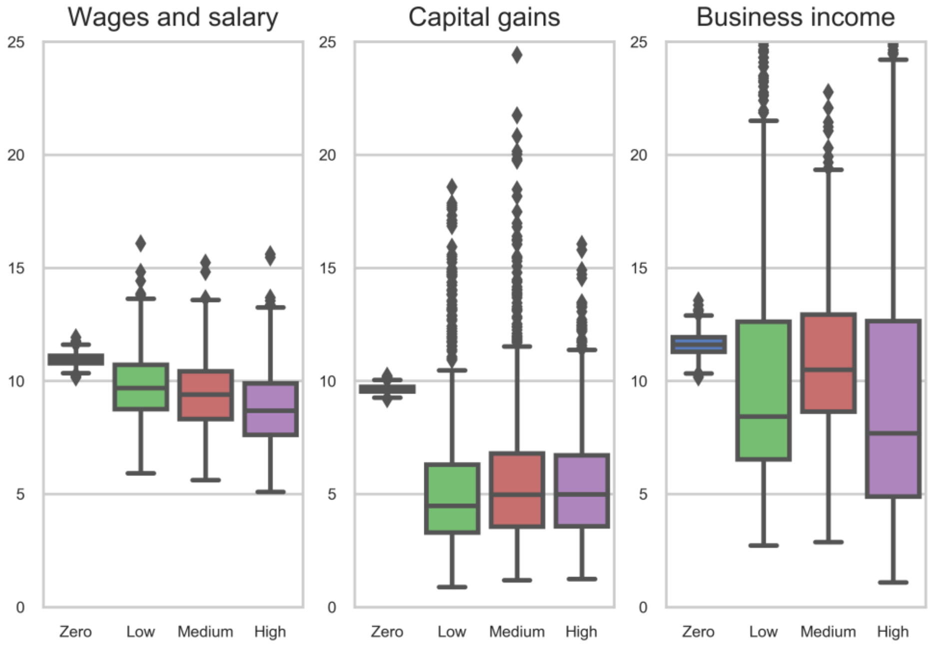
A deeper analysis by sources can be done in Table 20 and Figure 4, where, in each simulation, the proportion having negative *TR* has been computed using taxpayers with greater-than-zero values exclusively in one of the income sources, *wages and salaries*, *capital gains* or *business income* (13%, 11% and 2% of PIT sample, respectively).

The result is consistent with what has previously been found for declared income, that is, the tax morale (the opposite to tax responsibility = "*No*") increases with the level of donations for people having wages, salaries and capital gains, while it has not a clear link for people with business income. Although the figures are not totally conclusive, what seems to be true is that – whatever their income source is– the proportion of taxpayers with negative *TR* is greater for people having zero donations.

Table 20 Percentage of *tax responsibility = "No"* in PIT files
Results of simulations *for tax exemptions for donations* by type of income

Income source	Donation category	Mean	Std	Min	Max
Wages and salaries	Zero	10.96	0.25	10.16	11.93
	Low	9.79	1.47	5.93	16.08
	Medium	9.47	1.52	5.62	15.24
	High	8.83	1.65	5.10	15.60
Capital gains	Zero	9.65	0.16	9.20	10.22
	Low	5.32	3.12	0.89	18.58
	Medium	5.73	3.21	1.20	24.41
	High	5.42	2.52	1.25	16.06
Business income	Zero	11.63	0.49	10.15	13.55
	Low	10.19	5.20	2.73	32.77
	Medium	10.84	3.16	2.88	22.78
	High	9.83	6.75	1.09	38.73

Figure 4. Donations by income source: negative tax responsibility (%)



This is, theoretically, an expected connection, because donations can be linked to a certain ethical behaviour, the same that happens with the *tax responsibility* concept. Consequently, people with zero donations may feel less prone to tax compliance, that is, may have less tax morale.

4.3 Family size

There is in the literature some evidence suggesting that income under-reporting increases with family size (Erard, 1997; Matsaganis and Flevotomou, 2010). To verify whether this assumption holds true for declared income, a new variable to measure the family size has been calculated in PIT file. Using other variables appearing in PIT, its expression is:

$$Family\ size = Number\ of\ ancestors + Number\ of\ descendants + \phi(Marital\ status)$$
$$\text{where } \phi(x) = \begin{cases} 1 & if\ x = \text{'}Married\text{'} \\ 0 & in\ any\ other\ case \end{cases}$$

Now the relationship between *tax responsibility* and *family size* can be analyzed. The magnitude of the association measured by the adjusted Pearson and Cramer's V coefficients seems to be minimal in Table 21 (it must be taken into account that the family size definition is just a proxy of the real measure that may be distorted by other people who are not declared but are cohabiting with the taxpayer, such as unmarried couples, sisters, and others).

. **Table 21** Association of *family size* with *tax responsibility* in PIT files
Results of simulations

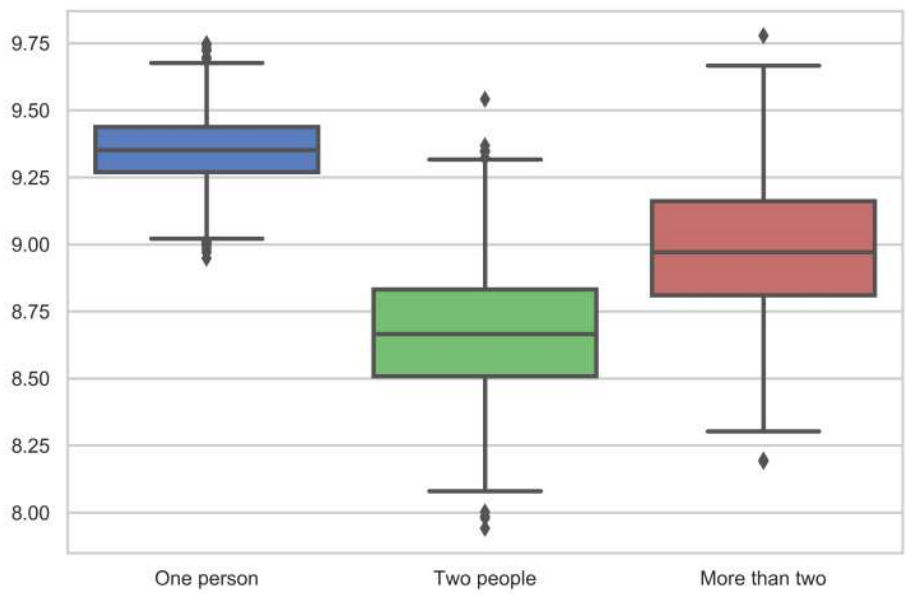
Adjusted Pearson coefficient				**Cramer's V coefficient**			
Mean	**Std**	**Min**	**Max**	**Mean**	**Std**	**Min**	**Max**
0.013	0.004	0.000	0.028	0.010	0.003	0.000	0.021

The result is confirmed in Table 22 and Figure 5 which show –through their joint distribution in the simulations– that a relationship between both variables cannot be found.

Table 22 Percentage of *tax responsibility = "No" in PIT files*
Results of simulations for *family size*

Category	**Mean**	**Std**	**Min**	**Max**
One person	9.35	0.13	8.95	9.75
Two people	8.67	0.24	7.94	9.54
More than two people	8.99	0.25	8.19	9.78

Figure 5. Family size: negative tax responsibility (%)



5 Final remarks

The availability of matched files including tax responsibility and other variables from a sample of personal income tax returns in Spain has allowed for analyzing the possible relationships between them.

The first thing to remark is that the selected procedure to execute the match –imputations using Bayesian Networks– has shown to be particularly suitable because it has two important features: it preserves the distributions of the variables in the original and matched files, and it allows for producing simulations that may be used in subsequent inferences. Its advantage over other methods for statistical matching resides in that imputations are based on a rigorous probabilistic model.

Although some of the assumptions for a perfect statistical matching are hardly accomplished, some preliminary results can be established. The relationships analyzed have tried to link the tax responsibility with some variables –wages and salaries, capital gains, business income, exemptions for donations and family size– from the sample of personal income tax returns. Given the connection between tax morale and the tax responsibility used in this paper, the remarks given refer to tax morale.

As a general comment, all the associations found do not have a high magnitude.

From the perspective of the declared income, the relationship depends on the source. Thus, for the ones of two different sources –corresponding to employment and capital sources of income– the tax morale has a positive dependence, increasing with the level of income, while it has no link with business income.

What has been found about the donations is that the proportion of taxpayers with no tax morale is greater for people giving zero donations. Lastly, the proportion of tax non-responsibility does not seem to be related to the family size.

References

Agresti, A. (1996), *Introduction to categorical data analysis*, New York, John Wiley and Sons.

Alm, J. (2012), Measuring, Explaining, and Controlling Tax Evasion: Lessons from Theory, Experiments, and Field Studies, *International Tax and Public Finance*, Vol. 19, pp. 54-77.

Cowell, R.G., Dawid, A.P., Lauritzen, S.L., and Spiegelhalter, D. J. (1999), *Probabilistic Networks and Expert Systems*, New York, Springer.

Cummings, R.G., Martinez-Vazquez, J., McKee, M., Torgler, B. (2008), Tax Morale Affects Tax Compliance: Evidence from Surveys and an Artefactual Field Experiment, *Journal of Economic Behavior and Organization*, doi:10.1016/j.jebo.2008.02.010.

De Waal, T. (2015), Statistical matching: Experimental results and future research questions, *Discussion Paper* 2015/19, Statistics Netherlands.

Dhami, S. (2017), Human Ethics and Virtues: Rethinking the Homo-Economicus Model, *CESifo Working papers* No. 6836, Category 13: Behavioural Economics.

Di Zio, M., Scanu, M., Coppola, L., Luzi, O. and Ponti, A. (2004), Bayesian Networks for Imputation, *Journal of the Royal Statistical Society* Vol. 167, part 2, pp. 309-322.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006), *Statistical Matching: Theory and Practice*, John Wiley & Sons, Ltd, ISBN: 0-470-02353-8.

D'Orazio, M., Di Zio, M. and Scanu, M. (2017), The use of uncertainty to choose matching variables in statistical matching, *International Journal of Approximate Reasoning* Vol. 90, pp. 433–440.

Dörrenberg, P. and Peichl, A. (2013), Progressive Taxation and Tax Morale, *Public Choice*, Vol. 155, No. 3-4, pp. 293-316.

Dörrenberg, P. and Peichl, A. (2017), Tax morale and the role of social norms and reciprocity, *ifo Working Papers*, No. 242, ifo Institute – Leibniz Institute for Economic Research at the University of Munich.

Erard, B. (1997), A Critical Review of the Empirical Research on Canadian Tax Compliance, *Working Paper*, No. 97-6, Carleton University.

Feld, L. P, and Frey, B.S. (2006), Tax Compliance as the Result of a Psychological Tax Contract: The Role of Incentives and Responsive Regulation, *Working Paper* 2006/10, Center for Research in Economics, Management and the Arts (CREMA).

Feldman, N., and Slemrod, J. (2007), Estimating Tax Noncompliance with Evidence from Unaudited Tax Returns, *Economic Journal*, Vol. 117, No. 518, pp. 327-352, available from: http://doi.org/10.1111/j.1468-0297.2007.02020.x.

Fiorio, C.F. and D'Amuri, F. (2005), Workers' Tax Evasion in Italy, *Giornale degli Economisti e Annali di Economia*, Vol. 64, No. 2/3, pp. 247-270.

Friedman, N., Geiger, D. and Goldszmidt, M. (1997), Bayesian Network Classifiers, *Machine Learning*, Vol. 29, pp. 131--163.

Hashimzade, N., Myles, G.D., and Tran-Nam, B. (2010), *New Approaches to the Economics of Tax Evasion*, University of Exeter, Exeter, United Kingdom, available from: http://people.exeter.ac.uk/gdmyles/papers/pdfs/Non-euEva.pdf

Heckerman, D., (1998), Bayesian networks for data mining, *Data Mining and Knowledge Discovery*, Vol. 1, No. 1, pp. 79-119.

Hruschka Jr., E.R., Hruschka, E.R., and Ebecken, N.F.F. (2007), Bayesian networks for imputation in classification problems, *Journal of Intelligent Information systems*, Vol. 29, pp. 231-252.

Leulescu, A. and Agafitei, M. (2013), Statistical matching: a model based approach for data integration, *Methodologies and Working Papers*, Eurostat, Luxembourg, Publications Office of the European Union.

Lin, M., Lucas, H., and Shmueli, G. (2013), Too Big to Fail: Large Samples and the p-Value Problem, *Information Systems Research*, Vol. 24, No. 4, pp. 906-917.

Luttmer, E. F. P., and Singhal, M. (2014), Tax Morale, *Journal of Economic Perspectives*, Vol. 28, No. 4, pp. 149-168.

Matsaganis, M. and Flevotomou, M. (2010), Distributional implications of tax evasion in Greece, *Hellenic Observatory Papers on Greece and Southeast E*urope GreeSE Paper no. 31, The Hellenic Observatory, LSE, London, UK.

Onu, D. (2016), Measuring Tax Compliance Attitudes: What Surveys Can Tell Us about Tax Compliance Behaviour, in John Hasseldine (ed.), *Advances in Taxation*, Vol. 23, Emerald Group Publishing Limited, pp. 173 – 190.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems*, San Francisco, CA, Morgan Kaufmann.

Pearl, J. (2000), *Causality: Models, Reasoning and Inference*, Cambridge University Press, England.

Pollard, D. E. (2002), *A user's guide to measure theoretic probability*, Cambridge, UK, Cambridge University Press, ISBN 0-521-00289-3.

Rässler, S. (2002), Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches, *Lecture Notes in Statistics*, Vol. 168, Springer, New York.

Van Rossum, G. (1995*),* Python tutorial, *Technical Report CS-R9526*, Centrum voor Wiskunde en Informatica (CWI), Amsterdam.