Comparative analysis of different techniques to impute expenditures into an income data set

Article · December 2020

DOI: 10.34196/IJM.00226

Andr� Decoster
KU Leuven

Jason Loughrey
TEAGASC - The Agriculture and Food Development Authority

Cathal ODonoghue
National University of Ireland, Galway

Comparative analysis of different techniques to impute expenditures into an income data set

André Decoster[1]*, Bram De Rock[1,2], Kris De Swerdt[3], Jason Loughrey[4], Cathal O'Donoghue[5], Dirk Verwerft[3]

[1]University of Leuven, Leuven, Belgium; [2]Université Libre de Bruxelles, Brussels, Belgium; [3]Federal Planning Bureau, Brussels, Belgium; [4]Rural Economy and Development Programme, Teagasc, Carlow, Ireland; [5]National University of Ireland, Galway, Ireland

Abstract Income and budget data seldom are measured in the same dataset. In order to make simulations that need both, one requires a reliable procedure to merge an income and a budget survey into one combined dataset. This paper contains the comparison and evaluation of five different techniques to impute expenditures into income datasets: parametric estimation of Engel curves, nonparametric estimation, both constrained and unconstrained matching using a distance function and grade correspondence. After a detailed description of the methods as well as a comparison of the main pros and cons, their effectiveness is tested upon an artificially split data file. In general, the parametric and non-parametric estimation seem to yield the best results, generating imputed values that are closest to the observed values for the budget shares.

JEL classification: C15, C51, C52, C81, H22, H31

DOI: https://doi.org/10.34196/ijm.00226

1. Introduction

In order to simulate concurrent changes in direct and indirect taxes a dataset which combines income and expenditure data is needed. However, it is unusual to have one data source that contains high quality information on both income and expenditures. A possible solution lies in the creation of a 'new' dataset which merges information of an income and a budget survey by using imputation or matching techniques using the overlapping variables – variables that are held in common by both datasets.

There is a large literature on Statistical Matching in different fields in the microsimulation field (***Cohen, 1991***). ***Sutherland et al. (2002)*** used statistical matching in the UK to combine income and expenditure datasets for indirect tax modelling. ***Decoster et al. (2010)*** also used statistical matching to combine income and expenditure files for indirect tax analysis in different EU countries as does ***Savage (2017)*** for Ireland and ***Donatiello et al. (2014)*** for Italy. ***Peichl and Schaefer (2009)*** utilise statistical matching in the combination of survey and administrative datasets for use in a macro-microsimulation model. ***Abello et al. (2008)*** and ***Von Randow et al. (2012)*** use statistical methods to link surveys in health microsimulation models. ***Cullinan (2010)*** links a spatial microsimulation model with locational data using statistical matching. In the wider inequality literature, ***Borra et al. (2013)*** link time use and income data, while ***Rasner et al. (2013)*** and ***Kum and Masterson (2010)*** look at wealth analysis.

There is a substantial literature, which focuses on combining Official Statistical Sources together (***D'Orazio et al., 2002***; ***D'Orazio et al., 2006a***; ***D'Orazio et al., 2006b***; ***D'Orazio et al., 2012***;

***For correspondence:**
andre.decoster@kuleuven.be

Author Keywords: microsimulation, statistical matching

Leulescu and Agafitei, 2013; ***Serafino and Tonkin, 2017***). Much of the statistical literature focused on techniques for specific methods (***Moriarity and Scheuren, 2001a***; ***Moriarity and Scheuren, 2001b***; ***Moriarity and Scheuren, 2003***; ***Rässler, 2003***). However in general the literature undertakes statistical matching without evaluating the relative performance of different methods, a research gap that this paper aims to undertake.

Specific methods have been evaluated in ***Rodgers (1984)*** and ***Barry (1988)***. However given the range of methods used in the microsimulation modelling and their different strengths there is a relatively sparse literature comparing the performance of different techniques of statistical matching. ***Webber and Tonkin (2013)*** do however undertake a comparison of the statistical matching of the SILC and Household Budget Survey, evaluating the match under a number of different scenarios. They compare the impact on matching variables, mean expenditure by decile using different statistical matching methods and perform an interesting test of conditional independence. ***Rässler (2002)*** compares different imputation and statistical matching methods, but there is no paper that compares the different methods utilised in the microsimulation literature. We will test the distributional assumptions at a disaggregated level relative to these studies. This paper attempts to fill this research gap.

In this paper we evaluate statistical matching algorithms used to link an income dataset and an expenditure dataset in the generation of a dataset to be used to simulate indirect taxation within the EUROMOD model (***Sutherland and Figari, 2013***) using the 2001 and 2002 Belgian Budget Surveys. In the EUROMOD context, the income dataset, on which the direct tax and benefit changes are modelled, cannot be altered. Therefore we designate this income dataset to be the target data set in which expenditure data are to be imputed. The budget dataset then plays the role of the source data set. The purpose of this paper though, is to evaluate an appropriate methodology in which to create a statistically matched dataset rather than to utilise the resulting dataset for a simulation. Therefore, to avoid any issues associated with data definitional issues, differential survey design, differential weights between source and target dataset, we use in this paper donor and recipient data from the same data set, i.e. from the budget survey.

Imputing household expenditure data into income surveys, although not unique, is one of the main uses of statistical matching in microsimulation models (***Sutherland et al., 2002***). Although some tax-benefit microsimulation models use data that contains both income and expenditure, as in the case of earlier models using the former Family Expenditure Survey in the UK or the Household Budget Survey in France (***Bourguignon et al., 1997***), in general the quality of the income variables is weaker in household budget surveys and typically of lower quality and detail required to model income taxation and social transfers (***O'Donoghue et al., 2004***). Similarly the unit of analysis is often at the household level rather than a more disaggregated individual or tax unit level. On the other hand, in most OECD countries, as in the case of the Eurostat European Community Household Panel (ECHP) or the Survey of Income and Living Conditions (SILC), the expenditure data necessary to simulate indirect taxes is missing. The direction of match in some methods such as minimum distance matching is irrelevant as they link both datasets, whilst in explicit methods such as regression based approaches the direction is relevant. In the latter, the income survey is typically taken as the base, because it contains both an appropriate unit of analysis and because of the relative importance of income variables and income based policies relative to expenditure data and expenditure based policies in OECD countries. As a result it is often required as in the case of EUROMOD to use statistical matching to link income and expenditure data.

Five different matching techniques are examined representing the techniques used in the microsimulation literature, which can be divided into two categories. A first category contains the so-called explicit methods that use estimations of Engel curves to impute expenditure information into the income data set.[1] The two techniques that we study in this category can be labelled as parametric (or standard) regression and nonparametric regression. The second category consists of the so-called implicit methods which match to each record in the income survey a record with expenditure information coming from the budget survey. In order to choose the most adequate record in the budget survey two different techniques are used, that is the distance function (both constrained and unconstrained methods) and the grade correspondence. For both techniques there are many variations

[1] An Engel curve is an equation that relates total expenditure to income and other personal characteristics.

possible in the practical application but based on the studies of respectively **Decoster and Van Camp (2002)** and **Taylor et al. (2001)** a limited selection was made.

When it comes to evaluating and comparing the five methods, two criteria will be essential. The first – microscopic – one is the quality of the match, in that one wants to create for each record of the income survey values for a number of new (budget) variables that correspond as well as possible – given the information available – to the true but missing values of that observation. The reason for this is obvious: the primary goal of the matching process is to obtain a dataset with observations that are realistic, in that they represent households that exist in society. A micro-simulation of behavioural change based upon types of behaviour that do not exist in society may not yield very trustworthy results. The second – macroscopic – criterion refers to the fact that the replication of distributions of missing variables is also desirable from the simulation point of view: the observed distribution in the budget survey is considered to be representative and deviations from it may lead to under- or overestimating certain indirect tax change effects (such as distributional effects). Of course, if the distribution of overlapping variables is the same in both datasets, the second criterion is a consequence of the first one: a good individual match will also generate the right marginal distributions for the budget shares. But if this condition is not met, a trade-off between the two criteria will be inevitable, which can best be illustrated by the difference between constrained and unconstrained distance matching (cf. infra).

Section 2 describes the methodology used and the data utilised. Explicit methods are discussed and in the third section the two implicit methods are considered. Both the general strategy and the concrete implementation are discussed. The section makes a theoretical comparison of the different methods. In the next section some evaluation criteria are suggested and the practical performance of the five methods is investigated. Section 4 concludes.

2. Data and methodology

2.1. Explicit methods: imputation by means of Engel curves

As mentioned above the explicit methods use Engel curves to impute for every record in the income survey expenditure information. Theoretically, this expenditure information could be at the most detailed level but in practice this is impossible since this would result in very imprecise estimations of the Engel curves. Consider for instance the influence of the zero expenditures (see e.g. **Pudney, 1989**). This zero expenditure problem illustrates that the reliability of these imputations relies upon an explicit statistical model which can be (slightly) misspecified. It has been assumed that based on the explanatory variables (including disposable income and some demographic characteristics like household size and age of the household head), the behaviour of the dependent variable can be fully captured and, moreover, that (standard) regression issues such as heteroskedasticity and multicollinearity can be adequately dealt with.

Therefore, the application of the explicit techniques in practice boils down to aggregating the expenditure items (in order to avoid zeroes) and then estimating the Engel curves of these aggregates. The quality of these imputation techniques is then completely determined by the quality of the estimation of the Engel curves. Although there exists a large literature on this topic (see for instance **Blundell, 1988**; **Banks et al., 1997**; **Blundell et al., 1998** and references therein), unfortunately in this specific setting the developed machinery cannot be applied fully. For instance, a functional specification has to be determined a priori in the parametric case and the explanatory variables are restricted to the set of overlapping variables. Beside these restrictions, which of course decrease the quality of the estimates, different definitions of the overlapping variables (e.g. income variables) possibly have to be dealt with.[2] Again this could influence the quality of the imputation.

In the rest of this section, let y_h denote the disposable income of household h, E_{jh} the expenditures of the household on the aggregate j, and $\mathbf{O}_h$ the vector of overlapping variables between the datasets (excluding y_h). For the first (standard) method the imputation is carried out by estimating the Engel curves of the budget shares:

$$w_{jh} : = \frac{E_{jh}}{y_h} = f\left(y_h, \mathbf{O}_h\right), \tag{1}$$

[2] **Decoster et al. (2007)** deals in detail with this issue.

using ordinary least squares regression on the budget dataset. Note that savings are treated in the same way as the budget categories in that it is also modeled by a regression equation. This points out why disposable income appears in the denominator rather than total expenditure: the budget and saving shares sum up to one. The explanatory variables are, as stated above, chosen out of the set of overlapping variables. In this way, the obtained model can be used to predict budget shares for the observations in the income survey.

In practice, the construction of the model is performed using the QUAIDS specification. The independent variables thus span the logarithm of the disposable income up to the second degree as well as the other overlapping variables:

$$w_{jh} = \alpha_j + \beta_j log\left(y_h\right) + \lambda_j log^2\left(y_h\right) + \boldsymbol{\delta_j g}\left(\mathbf{O}_h\right) + \varepsilon_{jh}, \tag{2}$$

where α_j, β_j, λ_j and δ_j are the parameters to be estimated and ε_{jh} is the error term. The function $\mathbf{g}$ is included so as to allow squared values and cross effects of demographic variables to be taken into account (e.g. age as in ***O'Donoghue et al., 2004***). Note also that the condition that the predicted budget shares have to sum up to one for each household needs no explicit restriction, since by the properties of the least squares estimators, the OLS performs this task automatically (see e.g. ***Deaton and Muellbauer, 1980, p. 19***):

$$\sum_j \alpha_j = 1 \quad \sum_j \beta_j = \sum_j \lambda_j = \sum_j \delta_{ij} = 0 \quad 1 \leq i \leq m \tag{3}$$

m being the dimension of the image of **g**. The regression equations derived by this procedure can then be applied to the observations of the income dataset, generating new variables w_j (possibly with an error term to randomize the results to some extent). An important remark in this respect is that the marginal distributions of the variables w_j will not necessarily be the same in the source and target dataset, except when the multivariate distribution of the overlapping variables is identical. The differences between the distributions of the overlapping variables in the source and target dataset are described in ***Decoster et al. (2007)***.

The nonparametric method starts from the same idea as the parametric method: to find a function that relates the budget shares to the overlapping variables in the household survey and in the next step apply this function to the observations in the income dataset. The difference lies in the fact that the parametric method starts from a functional specification while the non-parametric does not. In this way a misspecification of the Engel curves is avoided and much more flexibility is obtained for estimating the relation between the explanatory variables and the dependent variable. The nonparametric procedure consists of estimating density functions directly. In the univariate case, this can be visualised intuitively by a histogram, being the empirical density function:

$$\hat{f}\left(t\right) = \frac{1}{N}\sum_{i=1}^{N} \#\left\{t_j, j = 1, ..., H|t_j \in \left[a_i, b_i\right)\right\} \mathbf{1}_{\left[a_i, b_i\right)}\left(t\right) \tag{4}$$

where {[ai, bi), i = 1, ..., N} is a partition of the domain of *f*(*t*), *H* is the number of observations and $\mathbf{1}_A$ is the indicator function of a set *A*. Note that so far, no regression has yet been performed. In most cases, the result will be a highly discontinuous function (which can be thought of as caused by the fact that the sample was finite). For continuous random variables, the question also arises which partition should be chosen to represent the data. Both problems are tackled at the same time by the use of a density kernel estimator K:

$$\hat{f}\left(t\right) = \frac{1}{Hb}\sum_{k=1}^{H} K\left(\frac{t_k - t}{b}\right). \tag{5}$$

K represents a continuous function that integrates to one and acts as a smoothing device: $f\left(t\right)$ will indeed be continuous as a finite sum of continuous functions, and will integrate to one as one expects from a density function. Here the standard normal density function has been chosen to play the part of *K*, but the choice of *K* has been reported not to be of major importance (***Decoster et al., 2004***). The parameter *b* on the other hand, is a measure for the width of the intervals and is much more influential. If *b* is small, only those t_k 's close to *t* will have a significant impact on $f\left(t\right)$ (in the case of the standard normal density), and hence the bandwidth is smaller. A higher bandwidth has a more smoothening effect, while a smaller bandwidth will keep closer to the observed data, and the choice of b is therefore a trade off between variance and bias. A proposed optimal value for b that has been adopted here (see ***Deaton, 1997***) is:

$$b = 1.06 * min\left(\sigma, 0.75 * IQR\right) H^{-\frac{1}{5}}, \tag{6}$$

using H for the number of households, σ for the sample standard deviation and IQR for the sample interquartile range.

The method to estimate density functions can be used in this context since the Engel curve can be formulated as follows:

$$E\left(w_i|y, \mathbf{O}\right) = \int w_i f\left(w_i|y, \mathbf{O}\right) dw_i = \int w_i \frac{f\left(y, \mathbf{O}, w_i\right)}{f\left(y, \mathbf{O}\right)} dw_i = \frac{\int w_i f\left(y, \mathbf{O}, w_i\right) dw_i}{\int f\left(y, \mathbf{O}, w_i\right) dw_i} \tag{7}$$

Discretisation of the last expression (see ***Decoster et al., 2004***) yields the following nonparametric estimator:

$$\hat{E}\left(w_i|y, \mathbf{O}\right) = \frac{\sum_{k=1}^{H} w_i K\left(y_k - y, \mathbf{O}_k - \mathbf{O}\right)}{\sum_{k=1}^{H} K\left(y_k - y, \mathbf{O}_k - \mathbf{O}\right)}. \tag{8}$$

In this expression, K is a function on a more-dimensional space, which can be easily implemented by using e.g. the multivariate standard normal density function. There is, however, a problem when the number of dimensions becomes too large: in order to estimate a functional relationship adequately, one typically needs a lot of observations, but the required number of data increases with the number of dimensions ("the curse of dimensionality"). A possible solution consists in limiting the set of independent variables that enter nonparametrically, and use a standard (multiple) regression method for estimation of the other explanatory variables. This is exactly what is done in semiparametric models. In this application, the variables y and age are taken up in the nonparametric part, as in ***Decoster et al. (2004)***, while the effect of the other independent variables $\tilde{\mathbf{O}}$ is estimated by least squares. The resulting equation takes the form:

$$E\left(w_j|y, age, \tilde{\mathbf{O}}\right) = \beta_j E\left(\tilde{\mathbf{O}}|y, age\right) + F_j\left(y, age\right), \tag{9}$$

and subtracting this from the model equation $w_j = \beta_j \tilde{\mathbf{O}} + F_j\left(y, age\right) + \varepsilon_j$ yields:

$$w_j - E\left(w_j|y, age, \tilde{\mathbf{O}}\right) = \beta_j\left(\tilde{\mathbf{O}} - E\left(\tilde{\mathbf{O}}|y, age\right)\right) + \varepsilon_j. \tag{10}$$

The expectation values on the right and on the left can be estimated nonparametrically as before, and what remains of the equation is a model that can be estimated using least squares regression. Note that the estimated w_j's again sum up to one, as in the parametric case. See ***Blundell et al. (1998)***, and ***Decoster et al. (2004)*** for more details and an application of these semiparametric techniques.

We briefly compare both regression techniques. It is obvious that theoretically the semiparametric models are at least as good as the parametric method. Indeed, if the functional specification in a parametric method is the correct one, then the semiparametric method will result in similar estimates. But clearly the opposite does not hold. See for instance ***Härdle and Mammen (1993)***, for a comparison of both methods. In practice however estimating semiparametric methods can be very time consuming while estimation of parametric models can be done by using well known standard procedures. Finally, the parametric method has the advantage that a regression model estimated upon the budget data can be obtained in countries where the data themselves are inaccessible due to legal restrictions (see e.g. ***O'Donoghue et al., 2004***).

2.2. Implicit methods: imputing complete records

The implicit methods avoid the (theoretical) assumptions and their implications by using as little theory as possible, meaning that they do not rely on an explicit statistical model to impute the expenditure information. These methods try to concatenate expenditure information to observations in the income dataset by using the values of an observation in the expenditure survey that is as similar as possible to the target observation. Similarity is expressed mathematically as a distance function which has to be minimized and which can take the form of a numerical value or of belonging to the same categories and having the same rank within these categories (cf. infra). To find a similar record, the overlapping variables in both surveys are used. Although this is a very simple idea (without theoretical assumptions), the performance crucially depends upon the available overlapping variables and the method used to find the matching records. To give a hypothetical example, suppose that one of the overlapping variables is a (unique) identification number and that in both surveys the same households are present. Then one can of course match to every record of the income survey a unique record of the budget survey based on this number. In reality, however, no such precise overlapping variables are available. What is more, the observations in both surveys are not the same. Finding an exact match is therefore impossible. Before describing the two implicit methods used in this application to find the best possible match, two remarks are given that apply to both.

Overall, two strategies are possible, which have both been implemented in this application: *unconstrained matching* allows replacement of already chosen records in the source dataset, while *constrained matching* forbids replacement. By construction the unconstrained technique will yield the lowest total distance, but with constrained matching it is possible to replicate the marginal distribution of the variables w_j in the target dataset. A necessary condition for this to happen is that the number of observations in the source and the target dataset is the same. Since in most datasets the "number" of observations is represented by means of a weight variable (which gives the weight of the observation in the entire population), this prerequisite of an equal number of observations is realized by some procedure of "reweighting" the data via a duplication mechanism in the source set.[3] We sum up the weights of all the observations in the source dataset, the result being the number of households in the country. Dividing each weight by this sum, multiplying by the number of observations in the target dataset, and rounding the result (reweighting) gives the number of times an observation has to be duplicated (or deweighted) to get a source dataset with the same number of observations as the target dataset.[4]

A second choice concerns the weights that will be assigned to the different overlapping variables in the distance function. Indeed, not every overlapping variable has to be equally important in defining the distance. In this paper we consider two different applications of this weighting procedure: one with finite weights and one in which some variables get weight infinity.

The most basic implementation of implicit methods uses distance functions with finite weights for the overlapping variables. To be precise, for a given record in the income survey, the distance in the (selected) overlapping variables to every record in the budget survey is calculated. This could for instance be the difference in the number of children, the difference in disposable income, the difference in household size, etc. Then the weighted sum of these differences is calculated and finally the record of the budget survey which has the smallest weighted sum is picked out. If there are several records which result in the same minimum distance, one of these records is chosen at random.

In this case, there are no variables that are deemed so important that matching is forced within their categories. Of course, this does not mean that all overlapping variables are of equal importance: assigning a finite weight to each variable can make it relatively more or less influential in determining the distance (with the special case of putting the weight equal to zero for variables that will not be considered). The strategy adopted here consists of calculating the Mahalanobis distance. Let $\mathbf{t}_i$ be the realisation of overlapping variables of observation i in the target dataset and $\mathbf{t}_j$ that of observation j in the source set, then the Mahalanobis distance is defined as:

$$d\left(\mathbf{t}_i, \mathbf{t}_j\right) = \sqrt{\left(\mathbf{t}_i - \mathbf{t}_j\right)' \cdot \Sigma^{-1} \cdot \left(\mathbf{t}_i - \mathbf{t}_j\right)}, \qquad (11)$$

where Σ stands for the covariance matrix of the overlapping variables in the source dataset. Intuitively, one can keep in mind what this means for the uni- and bivariate case. If there is only one variable, the Mahalanobis distance equals the usual, Euclidean distance divided by the standard deviation. This introduces a correction which considers the same absolute distance as less important when the variable

[3] As stated before, the number of observations in the income dataset cannot be modified within the EUROMOD context, so only the source set can be adjusted.

[4] In cases where there were large differences in size between surveys or within groups, we replicated the datasets according to their weights and matched on the replicated sample.

under consideration has a high variance, than in case it is more concentrated. With several overlapping variables, also the correlations between the variables enter the scene (the off diagonal elements of matrix Σ^{-1}). Compared to the Euclidean distance ($\Sigma = \mathbf{1}$),the expression under the square root will be lower if there are two variables that are highly correlated (which means they have a high covariance). This is in line with intuition. Since highly covariating overlapping variables essentially capture the same information, we do want to decrease the weight of these variables in the distance function.

The Mahalanobis distance thus accounts for differences in variation of and correlation between the overlapping variables. Yet it does not allow for making qualitative distinctions between those variables (e.g. it is more important to put together households with the same income level than with the same education level) other than putting the weight of one variable equal to 0 (which means leaving one variable out of consideration). Also from empirical studies it seems that 'subjective weights' perform better (see e.g. ***Moriarity and Scheuren, 2001b***, and ***Decoster and Van Camp, 2002***). These subjective weights are mainly based on the quality of the overlapping variable (for instance the definition of the overlapping variable is the same in both surveys) and on the explanatory power. A possible way to tackle this problem is to determine weights by using a stepwise linear regression. This concept points to a collection of algorithms that try to find the most efficient regression equation given a set of explanatory variables. In a number of consecutive steps, a model is tested leaving out a variable or adding one. If the explanatory effect of this variable is significant, the variable is retained, otherwise it is dropped. Consider the model that comes out of an algorithm like this. The magnitude of the regression coefficients is a measure for the influence of the respective regressors on the dependent variable. Therefore, these magnitudes can be used as weights for a distance function, setting the weights for variables that were left out equal to zero. The distance between observations in one variable can be taken to be the absolute value of the difference. This method clearly accounts for differences in explanatory power of the common variables in that the more influential a variable is, the more weight it will get. Variance and correlation effects of the independent variables are also taken into account via the regression model.[5]

The grade correspondence technique consists in first clustering the observations in both datasets according to some overlapping variables (see ***Taylor, 2000*** and ***Taylor et al., 2001***). In a way, one sets the weights of these variables equal to infinity, because no matter how large the difference in the other variables within a cluster, the model does not allow matching across clusters. The division into clusters can be done based on experience (which variables are the most important?) or formal clustering procedures can be used (see the above references for a discussion of such procedures). Then, in a second step, a distance function is applied within each cluster, in the same way as before. So, one can again choose between a constrained and an unconstrained matching, and for the former method a reweighting/deweighting procedure can be put in place to obtain the same number of records in the source dataset as in the target dataset.

In this paper, the grade correspondence method is implemented using 18 a priori defined clusters: the observations are assigned to a cluster according to the age of the household head (below 40, between 40 and 60, and above 60), the profession of the household head (not working, blue or white collar worker) and whether children are present or not. These broadly correspond to the categories chosen in ***O'Donoghue et al. (2004)***.[6] Then the clusters in both surveys are made equally large by reweighting/deweighting. The distance between observations is determined by the rank of disposable income within each cluster. So the record with the smallest disposable income in cluster A in the budget survey will be matched to the record with the smallest disposable income in the corresponding cluster in the income dataset.

An important remark is that one has to avoid small clusters since this could lead to bad matching results. For instance, instead of using the exact number of children as a variable for the clustering, one can use the fact that there are children or not to avoid small clusters. On the other hand, ***Taylor et al. (2001)*** show that clustering significantly improves their results and that their results are similar for different sets of clusters. These statements are mainly based on elementary statistics concerning the deciles and on the performance when dealing with different tax simulations.

[5] We utilised unweighted regression coefficients in this study.

[6] It is true that the choice of cluster may make a difference in outcomes. However in this paper, we have tried to keep this issue relatively simple, using similar variables for different methods. It may be worth testing the sensitivity of the conclusions in this paper to the choice of cluster, but given the existing length required to explain and test the existing methods with a single classification system, this is left for future exploration.

We end this section by briefly comparing the two implicit methods. Theoretically, grade correspondence can more or less be considered as a special case of the method based on distance functions. Note moreover that the clustering idea can also be used to improve the results of the matching by distance functions (which actually implies that some weights are infinite or very high and so these variables can no longer be used in the distance function). On the other hand there is also a subtle difference. Since in grade correspondence we use the ranking information of the income variable, this method is less sensitive to difference in the income distribution in both surveys. This robustness can be a real advantage in situations where the measurement of the disposable income is not entirely reliable. Finally, in practice, applying the grade correspondence technique is straightforward while choosing the optimal weights for the distance functions can be quite cumbersome.

2.3. Prior comparison of explicit versus implicit methods

In the next section we will evaluate the *empirical performance* of the different matching techniques. Yet, it is also worth the while considering the prior theoretical arguments for choosing the ‘best’ matching technique, as well as the arguments and intuitions stemming from the considered literature. Note however that in the literature there are hardly any comparisons of the different techniques.

Purely theoretically, one is tempted to favour the implicit methods, since they do not rely on theoretical assumptions and they avoid many of the problems of the explicit methods. There are three types of problems associated with the latter.

1. A first problem concerns the influence of zero expenditures on the estimation of the Engel curves. From empirical studies it is clear that this highly influences the results.
2. Secondly, it is unfeasible to estimate Engel curves for hundreds of commodities. If one uses Engel curves, one first has to construct expenditure aggregates. This evidently also constrains the imputation of expenditure information to these aggregates. Since the aggregates are fixed before the matching procedure takes place, this deprives EUROMOD users of the possibility to define other expenditure aggregates in a later stage. Implicit methods allow for more flexibility in that the records matched will be the same regardless of the number and the magnitude of the aggregates (since the overlapping variables of the records stay the same). So one can anticipate user manipulation by using many small aggregates. For the explicit methods, this would decrease the quality of the match, e.g. because of the zero expenditure problem.
3. The third problem is the variability in the imputed expenditure information. If estimated Engel curves are used to impute expenditures, one actually imputes ‘averages’. To increase the variability in the matched dataset, one could draw random errors from a normal distribution with mean zero and variance equal to the mean square error of the model, or draw errors randomly from the error terms in the budget dataset. But this again induces problems, such as negative expenditures. Again this is not an issue when using the implicit methods.

In this theoretical scenario, we assume that the implicit methods can be applied at full strength. But this is not the case in this application. Recall that in EUROMOD the direction of the matching is fixed (because it is useless to impute income information into the budget survey), and secondly that the income survey cannot be modified (for instance to duplicate observations). The latter implication means that we have to use either unconstrained matching methods, which implies that we possibly do not use all the information of the budget survey (since unconstrained matching might use only part of the source dataset records), or either forms of constrained matching which do not duplicate observations in the target dataset.

A final note pertains to the way possible tax and benefit changes will be evaluated and simulated. In simulating the effects, a change in the behaviour of the households could be incorporated. With explicit methods the estimated Engel curves can be used to simulate these behavioural reactions as far as real income changes are concerned. The implicit methods on the other hand have not modelled this.

2.4. Conditional independence

It is appropriate to underline that all matching techniques rely on the conditional independence assumption. In order to believe that the simulations with the ‘new’ data set are reliable, one has to be convinced that this conditional independence assumption holds. To recall the assumption, let us label the variables in the income survey by (X, Y) and the ones in the budget survey by (X, Z), meaning that we call the overlapping variables X and the non-overlapping variables Y and Z. The conditional independence assumption then states that given X, Y and Z should be independent, or equivalently, that all the correlation between Y and Z has to be explained by X. Note that this can be a heavy assumption in the case of budget and income data. Consider, for instance, two people with the same disposable income and the same socio-demographic profile (and so with the same values of X). Suppose they both have a car, but one of them has bought an energy-economical car so as to get an income tax reduction. In that way there can be a positive correlation between the height of the income tax, which belongs to Y, and the height of the private transportation costs, which belongs to Z.[7]

2.5. Data

In the project that funded this research, we undertook this evaluation using data from different countries. In this paper we have chosen to select a particular country, grounded in our familiarity with the data, rather than for any specific country reason. The country we have chosen is Belgium. Also, in oder to avoid any issues associated with data definitional issues, differential survey design, differential weights etc., we in this study take the donor and receiving data from the same data set. The data we use is the Belgium Budget Survey from 2000 and 2001, collected by the Belgian National Institute for Statistics (NIS) containing 3,550 households.

Since the budget surveys only contain net or disposable household income (after taxes) and not gross income, we first used the micro-simulation model described in the next paragraph to reconstruct gross incomes from net earnings. This backward calculation was based on the fiscal and parafiscal regulations of the year of the survey itself.

The unit of analysis of incomes is mostly individual, excluding housing allowances, social assistance, rental income and inheritance/lottery winning, whilst the period of collection is mostly monthly income together with the number of months received during the reference year. Household level cross-sectional weights (shared weights) and individual level longitudinal weights are created that take into account of adjustment for sample attrition and external checks on population structure (demographic/socio-economic/social welfare)

2.6. Summary

In summary, the different methods take the same overlapping variables and try to generate variables (expenditure and shares) from the target dataset to introduce into the source dataset. Parametric and non-parametric methods generate an estimate of each variable, conditional on the match variables. For budget shares, we do not utilise error terms assigning the same shares to the simulated expenditure (which incorporates an error term). This is because of the computational challenge of sampling from a multi-dimensional error distribution. Of course, it is possible to generate univariate distributions for the error term. However we believe that the outcome which would make conditional distributions independent of each other to be a more serious issue and would result in budget shares not summing to one. The matching methods retain the inter-variable correlations as an observation in one dataset is linked to another, avoiding this problem. The grade correspondence method matches on the rank of a single variable, while the other two match on the Mahalanobis distance which contains more information. Nevertheless as datasets with different means and structures, they may generate different conditional means than in the regression based methods. They also come at an increased computation cost, albeit the grade correspondence method is much quicker.

3. Results: empirical evaluation of the different techniques

A number of tests can be used to determine the comparative matching strength between the different methods described above. However, the main assumption that underlies all methods, the conditional independence assumption, cannot be tested in an exact sense. There are a number of papers where this assumption is further investigated (e.g. ***Ingram et al., 2000***; ***Black and Smith, 2004***). For this application one has to keep in mind that it is possible that the CIA does not hold, and this can have negative effects on the matching methods' efficiency. But since all methods will be affected by it, it

[7] In this particular example, an overlapping variable containing the type of car owned would solve the problem. But other examples might be found since the set of overlapping variables is finite. See further Rodgers (1984) and ***Ingram et al. (2000)*** for a discussion and test of the conditional independence assumption.

seems reasonable to compare the methods relative to each other.

Further, the differences in distributions of the overlapping variables of the two datasets can distort the outcome of the matching process: the marginal distributions of the budget variables for instance will not be reproduced in this case unless constrained matching is applied.

Finally, some care has to be given to the fact that the overlapping variables have to be defined in the same way in both datasets, to avoid misidentification. Therefore, in what follows, the methods will be tested upon a dataset which was split artificially and randomly, so that firstly the problems of different distributions and differing variable definitions of the overlapping variables do not occur and secondly the imputed values can be compared to the observed values on an individual level. For the testing, the Belgian Budget Surveys of the years 2001 and 2002 were used. The datasets were concatenated so as to have more observations and the resulting dataset was split randomly in two equally large datasets, which acted as source and target dataset. Next, the budget shares were imputed from the source into the target dataset using the five different methods.

The relative matching quality was then evaluated by means of two criteria: a goodness of fit measure, and tests of the equality of the distributions of the imputed and the observed budget shares.

For the goodness of fit measure we calculated the differences between the observed and the imputed values for each budget share and took the root of the mean sum of squares of these differences, in short the root mean squared error (RMSE). The RMSE can be interpreted as a measure for the performance of the methods at the household level, in that it gives the expected deviation between the imputed and observed budget shares per household.

To test the equality of the distribution of observed and imputed budget shares, we want to take into account differences in the distribution of overlapping variables between the source and target dataset.[8] We therefore perform tests on the conditional distributions. Ideally, this conditionality should be implemented for all overlapping variables simultaneously. Yet, due to lack of data, we only performed the tests conditional upon some important overlapping variables: for each income decile, for different household types (single with or without children, cohabiting with or without children), for different age groups of the household head (younger than

Table 1. RMSE per method and per expenditure category

RMSE	FOOD, NON-ALCOHOLIC BEVERAGES	ALCOHOLIC BEVERAGES	TOBACCO	CLOTHING AND FOOTWEAR	HOME FUELS AND ELECTRICITY	RENTS	HOUSEHOLD SERVICES	HEALTH	PRIVATE TRANSPORT
Parametric	0.1597	0.0284	0.0249	0.0632	0.0790	0.2113	0.1538	0.0830	0.1156
Kernel	0.1607	0.0286	0.0252	0.0637	0.0817	0.2220	0.1529	0.0836	0.1137
Distance	0.2231	0.0382	0.0338	0.2074	0.1324	0.3567	0.1938	0.1309	0.2232
Constrained	0.2242	0.0394	0.0348	0.2048	0.1813	0.3556	0.1980	0.1336	0.2396
Grade Corr.	0.4926	0.2005	0.0684	0.2314	0.1812	0.4314	0.3804	0.2247	0.3722

RMSE	PUBLIC TRANSPORT	COMMUNI-CATION	RECREATION AND CULTURE	EDUCATION	RESTAURANTS AND HOTELS	OTHER GOODS AND SERVICES	DURABLES	SAVINGS	AVERAGE	WEIGHTED AVERAGE
Parametric	0.0415	0.0897	0.1259	0.0155	0.1241	0.1464	0.2818	0.8323	0.1431	0.2274
Kernel	0.0430	0.0896	0.1279	0.0157	0.1229	0.1470	0.2826	0.8288	0.1439	0.2278
Distance	0.0461	0.1023	0.1994	0.0239	0.1900	0.2115	0.4133	1.5169	0.2357	0.3819
Constrained	0.0477	0.1024	0.1977	0.0231	0.1788	0.2370	0.3939	1.5244	0.2398	0.3865
Grade Corr.	0.0497	0.0690	0.1574	0.0300	0.3155	7.2507	0.4445	11.024	1.2180	2.2701

[8] The procedure of imputing within the same budget survey already minimizes this possible difference in distributions of the overlapping variables. Yet, they are still present due to sampling errors.

30, between 30 and 50, between 50 and 65 and older than 65) and for different professional statuses (not employed, (self-) employed, retired or other). Three tests are carried out.

1. To compare the equality of distributions, the *Kolmogorov- Smirnov* test was used. This is a non-parametric test: since the distribution of the imputed values is not known or assumed a priori for the implicit methods, parametric tests are not adequate here. The Kolmogorov-Smirnov test compares the distribution functions by using the maximal distance between them as a test statistic. Note that this may disadvantage the explicit methods since they will create degenerate distributions conditional upon the overlapping variables by construction.
2. Two other non-parametric tests take a somewhat intermediate position: they test the equality of the conditional distributions of imputed and observed budget shares, but at the same time recognize that the budget shares are paired: every observation has a value for the imputed and for the observed share. Both tests calculate the differences between imputed and observed values and test whether the median of the resulting distribution is equal to zero.

- The *sign test* takes the number of positive values (which should be around half of the total number of observations) as a test statistic.
- The *signed rank test* also takes the magnitude of the differences into account: all observations get a rank number according to the magnitude of the difference between observed and imputed value and afterwards the ranks of the positive differences are summed. This sum should be around one half of the total sum of ranks.

3.1. Goodness of fit of the five different methods

The results for the RMSE are summarized in ***Table 1*** for all expenditure groups separately, and by means of an unweighted and a weighted average of the RMSE's, in which we use the shares in disposable income as weights. The conclusion is that overall, the explicit methods have a lower RMSE, and so the quality of these imputations is better than that of the ones created by the implicit methods. Among the explicit methods, the parametric and the non-parametric case yield almost the same RMSE. At first sight, it is surprising that the non parametric kernel regression does not have a lower RMSE than the parametric Engel curve. Note however that the Engel curve used here is in fact not fully non parametric, but only semi-parametric. Only income and age are treated non parametrically. The other factors are treated parametrically. Our results suggest that the QUAIDS specification, with sufficient cross effects built in, is flexible enough to capture all curvatures captured by the semi-parametric specification.

Within the group of implicit methods there is a lot of variation in performance. In general the constrained and unconstrained distance function seem to give the same result when it comes to expected deviation from the observed values. The grade correspondence technique performs worse for most of the budget shares, except for "communication" and "recreation and culture". For some categories, especially for "saving", all the methods perform very badly.

Note that this goodness of fit measure obviously omits a possible important criterion for selection of the best method. It only looks at the best fit for each expenditure category separately, but does not assess how well the methods replicate or preserve the covariance between the different expenditure categories. We will try to integrate this criterion of assessment in our future research.

3.2. Are the distributions of imputed and observed budget shares different?

A second important issue to be assessed is whether the distribution of both the observed and imputed budget shares conditional upon the overlapping variables is the same. As already mentioned, conditionality upon all the overlapping variables is not an option, since this would require a lot more data then available to get significant results. Therefore, in what follows, the three tests will be carried out conditionally upon four important variables (cf. supra). Since this results in four tables with three sub- tables per budget share (one table for each test), a selection of budget shares has been made. ***Table 2*** present the p-values for the tests for the "food and non-alcoholic beverages"-category, ***Table 3*** for "clothing and footwear", ***Table 4*** for "private transport" and ***Table 5*** for "saving" categorised by income category. We perform a similar analysis for age group, employment status and family status in the appendix. For each budget share, the first table gives the p-values for the three tests per income decile, the second per age group of

Table 2. P-values for the Statistical Tests (Kolmogorov Smirnov, Sign Test, Signed Rank Test) for Food & Non-alcoholic Beverages by Income decile

	INCOME DECILE 1	INCOME DECILE 2	INCOME DECILE 3	INCOME DECILE 4	INCOME DECILE 5	INCOME DECILE 6	INCOME DECILE 7	INCOME DECILE 8	INCOME DECILE 9	INCOME DECILE 10
KOLM. SMIRNOV P-VALUE										
Parametric	0.000	0.000	0.000	0.000	0.000	0.000	0.045	0.002	0.213	0.003
Kernel	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.071	0.058
Distance	0.019	0.937	0.226	0.524	0.640	0.536	0.084	0.019	0.572	0.408
Constrained	0.008	0.624	0.444	0.977	0.405	0.849	0.002	0.605	0.572	0.304
Grade Corr.	0.000	0.000	0.000	0.002	0.022	0.569	0.009	0.013	0.000	0.000
SIGN TEST P-VALUE										
Parametric	0.0001	0.0000	0.0074	0.3093	0.3068	0.4292	0.7453	0.8685	0.2816	0.0055
Kernel	0.0913	0.0016	0.0172	0.0109	0.1252	0.0397	0.0511	0.0311	0.1615	0.3181
Distance	0.0073	0.3239	0.1974	0.4767	0.9186	0.1876	0.3297	0.0174	0.1061	0.3181
Constrained	0.0006	0.5540	0.6918	1.0000	0.2609	0.8744	0.0092	0.1674	0.1615	0.3748
Grade Corr.	0.0000	0.0000	0.0003	0.1201	0.7616	0.6435	1.0000	0.1725	0.0000	0.0000
SIGNED RANK TEST P-VALUE										
Parametric	0.0086	0.0000	0.0206	0.9100	0.6720	0.5854	0.4199	0.9981	0.7140	0.0118
Kernel	0.7695	0.0092	0.0775	0.3235	0.0927	0.2790	0.0871	0.0355	0.2895	0.2565
Distance	0.0012	0.3908	0.3347	0.4168	0.8198	0.1974	0.3532	0.0812	0.4333	0.1810
Constrained	0.0001	0.9497	0.2668	0.9551	0.2932	0.9222	0.0019	0.3668	0.0741	0.0764
Grade Corr.	0.0000	0.0000	0.0000	0.0009	0.4365	0.6466	0.6104	0.0185	0.0000	0.0000

Table 3. P-values for the Statistical Tests (Kolmogorov Smirnov, Sign Test, Signed Rank Test) for Clothing & Footwear by Income decile

	INCOME DECILE 1	INCOME DECILE 2	INCOME DECILE 3	INCOME DECILE 4	INCOME DECILE 5	INCOME DECILE 6	INCOME DECILE 7	INCOME DECILE 8	INCOME DECILE 9	INCOME DECILE 10
KOLM. SMIRNOV P-VALUE										
Parametric	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Kernel	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Distance	0.228	0.451	0.101	0.049	0.703	0.792	0.102	0.605	0.152	0.598
Constrained	0.547	0.307	0.500	0.771	0.309	0.665	0.440	0.863	0.447	0.532
Grade Corr.	0.006	0.067	0.021	0.010	0.240	0.260	0.152	0.754	0.000	0.000
SIGN TEST P-VALUE										
Parametric	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000
Kernel	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000
Distance	0.0742	0.1581	0.3233	0.1746	0.9178	0.9145	0.3254	1.0000	0.7869	0.7393
Constrained	0.9171	0.0386	0.2559	0.2521	0.3768	0.7488	0.8265	0.8664	0.5882	0.5044
Grade Corr.	0.0205	0.1301	0.0807	0.0057	0.9589	0.6399	0.1910	0.9562	0.0157	0.0000
SIGNED RANK TEST P-VALUE										
Parametric	0.0001	0.0001	0.0000	0.0000	0.0048	0.0003	0.0658	0.0056	0.0158	0.0001
Kernel	0.0006	0.0024	0.0000	0.0000	0.0204	0.0000	0.0120	0.0004	0.0047	0.0017
Distance	0.5840	0.3338	0.2254	0.0556	0.6569	0.9276	0.0646	0.5447	0.5943	0.5390
Constrained	0.7142	0.0521	0.1738	0.3413	0.2139	0.8735	0.2613	0.5648	0.5366	0.4289
Grade Corr.	0.0192	0.0345	0.0113	0.0007	0.3995	0.2746	0.1999	0.8189	0.0012	0.0000

Table 4. P-values for the Statistical Tests (Kolmogorov Smirnov, Sign Test, Signed Rank Test) for Private Transport by Income decile

	INCOME DECILE 1	INCOME DECILE 2	INCOME DECILE 3	INCOME DECILE 4	INCOME DECILE 5	INCOME DECILE 6	INCOME DECILE 7	INCOME DECILE 8	INCOME DECILE 9	INCOME DECILE 10
KOLM. SMIRNOV P-VALUE										
Parametric	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Kernel	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Distance	0.001	0.028	0.393	0.710	0.405	0.024	0.244	0.264	0.152	0.259
Constrained	0.000	0.123	0.982	0.361	0.230	0.082	0.207	0.359	0.883	0.800
Grade Corr.	0.000	0.001	0.056	0.603	0.832	0.115	0.071	0.430	0.032	0.001
SIGN TEST P-VALUE										
Parametric	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Kernel	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Distance	0.0071	0.0286	0.0895	0.6435	0.6426	0.0793	0.3566	0.2687	0.6276	0.6573
Constrained	0.0134	0.0750	0.1793	0.2801	0.7580	0.0173	0.7043	0.2941	0.1178	0.7393
Grade Corr.	0.0001	0.0080	0.0997	0.1885	0.4136	0.0382	0.0177	0.4783	0.4214	0.0071
SIGNED RANK TEST P-VALUE										
Parametric	0.0000	0.0000	0.0000	0.0068	0.0005	0.0000	0.1635	0.0150	0.0093	0.0002
Kernel	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0505	0.0124	0.0070	0.0374
Distance	0.0365	0.0278	0.2695	0.7604	0.6489	0.0135	0.1925	0.0949	0.2934	0.8728
Constrained	0.0243	0.0652	0.4758	0.2550	0.9930	0.0270	0.5431	0.1092	0.1507	0.9040
Grade Corr.	0.0000	0.0003	0.0989	0.3547	0.3325	0.0104	0.0463	0.2263	0.2372	0.0005

Table 5. P-values for the Statistical Tests (Kolmogorov Smirnov, Sign Test, Signed Rank Test) for Saving by Income decile

	INCOME DECILE 1	INCOME DECILE 2	INCOME DECILE 3	INCOME DECILE 4	INCOME DECILE 5	INCOME DECILE 6	INCOME DECILE 7	INCOME DECILE 8	INCOME DECILE 9	INCOME DECILE 10
KOLM. SMIRNOV P-VALUE										
Parametric	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Kernel	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Distance	0.722	0.050	0.226	0.771	0.405	0.234	0.018	0.673	0.829	0.011
Constrained	0.920	0.199	0.982	0.977	0.640	0.418	0.630	0.912	0.508	0.030
Grade Corr.	0.040	0.247	0.067	0.078	0.027	0.003	0.213	0.053	0.000	0.000
SIGN TEST P-VALUE										
Parametric	0.0000	0.0000	0.0000	0.0044	0.0000	0.0000	0.1590	0.0001	0.0002	0.0000
Kernel	0.0000	0.0000	0.0000	0.0005	0.0000	0.0000	0.0171	0.0000	0.0001	0.0000
Distance	0.3892	0.0607	0.4278	0.7604	1.0000	0.3703	0.1041	0.6193	0.5904	0.1491
Constrained	0.1037	0.0102	0.3723	0.4767	0.5398	0.8744	0.5882	0.2040	0.5183	0.0959
Grade Corr.	1.0000	0.3562	0.4961	0.2925	0.3118	0.1357	0.7467	0.0560	0.3084	0.0000
SIGNED RANK TEST P-VALUE										
Parametric	0.0000	0.0000	0.0010	0.0084	0.0010	0.0155	0.9880	0.0011	0.0264	0.0000
Kernel	0.0001	0.0000	0.0048	0.0010	0.0003	0.0001	0.1471	0.0000	0.0146	0.0000
Distance	0.8222	0.0934	0.4229	0.9461	0.3304	0.2561	0.0876	0.6433	0.8904	0.0281
Constrained	0.5855	0.0650	0.6691	0.6501	0.8823	0.9395	0.3458	0.4177	0.3587	0.0085
Grade Corr.	0.9726	0.6301	0.1146	0.0371	0.0456	0.4947	0.3127	0.1129	0.2783	0.0000

the household head, the third per professional status of the household head and the fourth per household type. In each table, the first sub-table gives the results for the Kolmogorov-Smirnov test, the second one for the sign test and the third one for the signed rank test.

Take for instance ***Table 2*** For the fifth income decile, both the parametric and non-parametric methods yield a p-value of zero for the Kolmogorov- Smirnov test, which means that the null hypothesis of equality of distributions is rejected at a significance level of, for instance, 0.05. For the unconstrained and constrained distance function (p-values of 0.640 and 0.405) the null hypothesis can clearly not be rejected in this decile, whereas for the grade correspondence (p-value of 0.022) the null is rejected at a significance level of 0.05 but not at a level of 0.01.

Overall, the implicit methods seem to replicate the conditional distributions, whereas this is not the case for the explicit methods. The bad performance of the explicit methods for the Kolmogorov- Smirnov test can perhaps be explained by the fact that the conditional distributions of their imputed values are degenerate: if the overlapping values are the same, they predict only one share, without variation in the results. The Kolmogorov-Smirnov test is by construction very sensitive when it comes to comparing a degenerate to a nondegenerate distribution. If this is indeed the explanation, doping the imputed values of the explicit methods with random error terms as described above may improve the test results. This is planned for future research. However, also for the sign and signed rank test, the results of the explicit methods are worse than those of the implicit methods. The fact that the conditionality implemented here is only partial because of too few observations may explain the bad performance of the explicit methods, although this shortcoming is also present for the other methods.

4. Conclusion

This paper tried to formulate a solution to the fact that there often exists no single dataset in which both income and budget variables are present. The solution consists of a matching procedure, in which two datasets are merged using variables that are common to both sets. Many different methods are utilised within the literature, but there is no strong consensus as to the appropriate method to use.

Five different matching procedures were investigated in this paper: the parametric and non-parametric estimation of Engel curves, the use of an unconstrained or constrained distance function and grade correspondence. The first two generate a model estimated on the budget set that predicts expenses based on the overlapping variables, and then apply this model to the income set. The other methods attach to each observation of the income dataset the values for the budget variables of an observation in the budget set that is most similar to the original record. A difference in the (mathematical) definition of similarity leads to the three methods discussed above.

We applied the five procedures to the 2001 and 2002 Belgian Budget Surveys in order to test their quality. Overall, the parametric and non-parametric methods seem to generate the best fit of the imputed values with respect to the observed values, which was demonstrated by lower root mean square errors. Concerning the distribution of budget shares conditional upon disposable income, age and professional status of the household head and household type, the distance functions seem to yield the best result, whereas the parametric and non-parametric methods do not reproduce the same distribution. This result can be biased, however, by the fact that estimation procedures yield degenerate conditional distributions by construction.

Future research will be to see whether the above conclusions are robust with respect to the introduction of more variation after the imputation by means of the explicit methods (adding error terms), and/or with respect to inserting an additional criterion of assessment, to wit: how well are the covariances between the budget shares preserved under the different methods? While this study uses the same donor and receiver dataset, it would be of interest to test the robustness of the conclusions to datasets that were different.

We hope in this study to have provided some guidance to microsimulation model builders who wish utilise statistical matching. As in the case of ***Webber and Tonkin (2013)***, there are pros and cons with different methods. Ultimately, minimum distance measures produce better distributions both within variables and between variables, but weaker means than the parametric or non-parametric methods, but come at a significant computational requirement.

As to what to trust, as in the case of all data preparation for microsimulation models, it requires detailed validation that matched or imputed variables broadly follow the distributions and means from the matched dataset and that additional corrections occur when there are discrepancies. It should however be noted that one is generating a model for microsimulation purposes. They are by definition wrong, but hopefully useful. As microsimulation models typically are based upon differences between baseline and simulated distributions, the bar is not quite as high as when one is generating merely a base distribution as some of the differences cancel out. Nevertheless discipline norms and high standards of validation and verification remain essential.

Given the importance of inter-variable relationships in distributional analysis, our preference is to use minimum distance methods where possible, perhaps correcting means if necessary. However in the case of the original EUROMOD analysis, where micro data sets were not available (***O'Donoghue et al., 2004***), or in the case of very large datasets such as the base data for dynamic microsimulation models or spatial microsimulation models, we are willing to sacrifice the improved distributional precision for a lower computational cost and use parametric methods.

ORCID iDs

André Decoster https://orcid.org/0000-0002-0521-8610
Bram De Rock https://orcid.org/0000-0002-5114-1386
Jason Loughrey https://orcid.org/0000-0001-9658-0801
Cathal O'Donoghue https://orcid.org/0000-0003-3713-5366

Acknowledgements

We are grateful for helpful comments made by three anonymous referees, and are indebted to the many people who have contributed to the development of EUROMOD and participants at several EUROMOD workshops, where this paper has been presented. EUROMOD has originally been developed and maintained by the Institute for Social and Economic Research (ISER) in Essex in collaboration with national teams from the EU countries. Since 2021 EUROMOD is maintained, developed and managed by the Joint Research Centre (JRC) of the European Commission, in collaboration with EUROSTAT and the national teams. The results and their interpretation are the authors' responsibility.

Funding

This research was carried out under European Commission FP6 contract no. 028412, "Accurate Income Measurement for the Assessment of Public Policies" (AIM-AP).

Conflict of Interest

No competing interests reported.

Data and Code availability

The authors are willing to share code with model builders (contact: cathal.odonoghue@nuigalway.ie). As a code written for the authors' and project's research purposes, it is not set up for generic matching, however it can and has been adapted for statistically matching other datasets. The considerable lag between when the AIM-AP project was funded and this publication made the code quite 'aged'. However, we felt that it is useful for the microsimulation community for this work to be showcased, given the importance of the techniques in the field.

References

Abello A, Lymer S, Brown L, Harding A, Phillips B. 2008. Enhancing the Australian national health survey data for use in a microsimulation model of pharmaceutical drug usage and cost. *Journal of Artificial Societies and Social Simulation* **11**:2.

Banks J, Blundell R, Lewbel A. 1997. Quadratic Engel curves and consumer demand. *Review of Economics and Statistics* **79**:527–539. DOI: https://doi.org/10.1162/003465397557015

Barry JT. 1988. An investigation of statistical matching. *Journal of Applied Statistics* **15**:275–283. DOI: https://doi.org/10.1080/02664768800000038

Black DA, Smith JA. 2004. How robust is the evidence on the effects of college quality? *Evidence from matching, Journal of Econometrics* **121**:99–124.

Blundell R. 1988. Consumer Behaviour: Theory and Empirical Evidence - A Survey. *The Economic Journal* **98**:16–65. DOI: https://doi.org/10.2307/2233510

Blundell R, Duncan A, Pendakur K. 1998. Semiparametric estimation and consumer demand. *Journal of Applied Econometrics* **13**:435–461. DOI: https://doi.org/10.1002/(SICI)1099-1255(1998090)13:5<435::AID-JAE506>3.0.CO;2-K

Borra C, Sevilla A, Gershuny J. 2013. Calibrating time-use estimates for the British household panel survey. *Social Indicators Research* **114**:1211–1224. DOI: https://doi.org/10.1007/s11205-012-0198-2

Bourguignon F, O'Donoghue C, Sastre-Descals J, Spadaro A, Utili F. 1997. Eur 3: A prototype European Tax-Benefit model. *DELTA Working Papers 97-30, DELTA*.

Cohen ML. 1991. Statistical matching and microsimulation models. In Improving Information for Social Policy Decisions: The Uses of Microsimulation Modelling, Volume II. In:Citro CF, Hanushek EA (editors). *Technical Papers*. Washington, DC: National Academy Press.

Cullinan J. 2010. Developing a continuous space representation of a simulated population. *Spatial Economic Analysis* **5**:317–338. DOI: https://doi.org/10.1080/17421772.2010.493954

D'Orazio M, Di Zio M, Scanu M. 2002. Statistical matching and official statistics. *Rivista di statistica ufficiale*.

D'Orazio M, Di Zio M, Scanu M. 2006a. Statistical matching for categorical data: Displaying uncertainty and using logical constraints. *Journal Of Official Statistics-Stockholm* **22**:137.

D'Orazio M, Di ZioM, Scanu M. 2006b. *Statistical matching: Theory and practice*. John Wiley & Sons.

D'Orazio M, Di Zio M, Scanu M. 2012. Statistical matching of data from complex sample surveys. In:*Proceedings of the European Conference on Quality in Official Statistics-Q2012*. **29**.

Deaton A. 1997. *The analysis of household surveys: A microeconometric approach to development policy*. Baltimore, MD: Johns Hopkins University Press.

Deaton A, Muellbauer J. 1980. *Economics and consumer behavior*. Cambridge NY: Cambridge University Press.

Decoster A, De Rock B, De Swerdt K, Flannery D, Loughrey J, O'Donoghue C, Verwerft D. 2007. Comparative analysis of different techniques to impute expenditures into an income data set, work package 3.4 of accurate income measurement for the assessment of public policies (AIM-AP contract no 028412), Leuven.

Decoster A, De Swerdt K, Van Camp G. 2004. Matching of income and expenditure data by means of nonparametric estimation of Engel curves, report of the D.W.T.C. project AG/01/079.

Decoster A, Loughrey J, O'Donoghue C, Verwerft D. 2010. How regressive are indirect taxes? A microsimulation analysis for five European countries. *Journal of Policy Analysis and Management* **29**:326–350. DOI: https://doi.org/10.1002/pam.20494

Decoster A, Van Camp G. 2002. De constructie van één samengesteld bestand op basis van twee bestanden: koppeling van de budgetenquete 1997-98 en het fiscaal bestand 1999 (inkomstens 1998) [i.e. Match the expenditure survey of 1997-98 to the income survey of 1999].

Donatiello G, D'Orazio M, Frattarola D, Rizzi A, Scanu M, Spaziani M. 2014. Statistical matching of income and consumption expenditures. *International Journal of Economic Sciences* **3**:50.

Härdle W, Mammen E. 1993. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics* **21**:1926–1947. DOI: https://doi.org/10.1214/aos/1176349403

Ingram D, O'Hare J, Scheuren F, Turek J. 2000. Statistical matching: A new validation case study. *proceedings American Statistical association*.

Kum H, Masterson TN. 2010. Statistical matching using propensity scores: Theory and application to the analysis of the distribution of income and wealth. *Journal of Economic and Social Measurement* **35**:177–196. DOI: https://doi.org/10.3233/JEM-2010-0332

Leulescu A, Agafitei M. 2013. Statistical matching: A model based approach for data integration. *Eurostat-Methodologies and Working papers*.

Moriarity C, Scheuren F. 2001a. Statistical matching: Pitfalls of current procedures. *Proceedings of the Annual Meeting of the American Statistical Association,* August 5-9.

Moriarity C, Scheuren F. 2001b. Statistical matching: A paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics* **17**:407.

Moriarity C, Scheuren F. 2003. A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics* **21**:65–73. DOI: https://doi.org/10.1198/073500102288618766

O'Donoghue C, Baldini M, Montovani D. 2004. Modelling the redistributive impact of indirect taxes in Europe: An application of EUROMOD. *EUROMOD Working Paper No. EM7/01*.

Peichl A, Schaefer T. 2009. FiFoSiM - An Integrated Tax Benefit Microsimulation and CGE Model for Germany. *International Journal of Microsimulation* **2**:1–15. DOI: https://doi.org/10.34196/ijm.00008

Pudney S. 1989. *Modelling individual choice: The econometrics of corners, kinks and holes*. Oxford: Blackwell. ISBN:0-631-14589-3

Rasner A, Frick JR, Grabka MM. 2013. Statistical matching of administrative and survey data: An application to wealth inequality analysis. *Sociological Methods & Research* **42**:192–224.

Rässler S. 2002. *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. **168**. Springer Science & Business Media.

Rässler S. 2003. A Non-Iterative Bayesian approach to statistical matching. *Statistica Neerlandica* **57**:58–74. DOI: https://doi.org/10.1111/1467-9574.00221

Rodgers W. 1984. An evaluation of statistical matching. *Journal of Business and Economic Statistics* **2**:91–102.

Savage M. 2017. Integrated modelling of the impact of direct and indirect taxes using complementary datasets. *The Economic and Social Review* **48**:171–205.

Serafino P, Tonkin R. 2017. Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey. *Eurostat Statistical Working Papers.Simar, L., 2004, An Invitation to the Bootstrap: Panacea for statistical inference?, course handout.*

Sutherland H, Figari F. 2013. EUROMOD: the European Union tax-benefit microsimulation model. *International Journal of Microsimulation* **6**:4–26. DOI: https://doi.org/10.34196/ijm.00075

Sutherland H, Taylor R, Gomulka J. 2002. Combining household income and expenditure data in policy simulations. *Review of Income and Wealth* **48**:517–536. DOI: https://doi.org/10.1111/1475-4991.00066

Taylor R. 2000. Guidelines for identifying clusters using grade correspondence analysis: Practical and technical issues. *Microsimulation unit research note MU/RN/39*.

Taylor R, Sutherland H, Gomulka J. 2001. Using POLIMOD to evaluate alternative methods of expenditure imputation. *Microsimulation unit research note MU/RN38*.

von Randow M, Davis P, Lay-Yee R, Pearson J. 2012. Data matching to allocate doctors to patients in a microsimulation model of the primary care process in New Zealand. *Social Science Computer Review* **30**:358–368. DOI: https://doi.org/10.1177/0894439311417153

Webber D, Tonkin R. 2013. Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation. Net SILC2 project working paper. https://ec.europa.eu/eurostat/documents/3888793/5857145/KS-RA-13-007-EN.PDF/37d4ffcc-e9fc-42bc-8d4f-fc89c65ff6b1

Appendix A

Table A1 P-values for Food & Non-alcoholic Beverages by Age of ReferencePerson

	AGE < 30	30 <= AGE < 50	50 <= AGE < 65	65 <= AGE
KOLM. SMIRNOV P-VALUE				
Parametric	0.000	0.000	0.000	0.000
Kernel	0.000	0.000	0.000	0.000
Distance	0.040	0.975	0.143	0.408
Constrained	0.049	0.774	0.737	0.886
Grade Corr.	0.005	0.000	0.000	0.001
SIGN TEST P-VALUE				
Parametric	0.0000	0.0290	0.0000	0.9159
Kernel	0.0000	0.0001	0.0050	0.0843
Distance	0.0419	0.4176	0.2738	0.8053
Constrained	0.0324	0.3010	0.8884	0.9159
Grade Corr.	0.0241	0.8906	0.6175	0.2359
SIGNED RANK TEST P-VALUE				
Parametric	0.0000	0.4845	0.0007	0.5176
Kernel	0.0000	0.0181	0.0324	0.5517
Distance	0.0094	0.9159	0.7293	0.9754
Constrained	0.0120	0.5883	0.7856	0.7631
Grade Corr.	0.0016	0.0211	0.0045	0.0176

Table A2 P-values for Food and Non-alcoholic Beverages by Professional Status

	(SELF-) EMPLOYED	UNEMPLOYED	(EARLY) RETIRED	OTHER
KOLM. SMIRNOV P-VALUE				
Parametric	0.000	0.000	0.000	0.003
Kernel	0.000	0.000	0.000	0.003
Distance	0.407	0.087	0.036	0.000
Constrained	0.772	0.033	0.017	0.002
Grade Corr.	0.000	0.000	0.001	0.008
SIGN TEST P-VALUE				
Parametric	0.0011	0.3409	0.0000	0.0639
Kernel	0.0001	0.1440	0.0000	0.2480
Distance	0.1775	0.2276	0.0086	0.0639
Constrained	0.5839	0.0042	0.0016	0.0370
Grade Corr.	0.2880	0.1033	0.5616	0.6530
SIGNED RANK TEST P-VALUE				
Parametric	0.0297	0.0216	0.0000	0.0151
Kernel	0.0049	0.0026	0.0000	0.0296
Distance	0.9349	0.0608	0.0051	0.0057
Constrained	0.4898	0.0087	0.0021	0.0160
Grade Corr.	0.0003	0.0025	0.0407	0.0619

Table A3 P-values for Foods and Non-alcoholic Beverages by Household Type

	SINGLE WITHOUT CHILDREN	SINGLE WITH CHILDREN	COHABITING WITHOUT CHILDREN	COHABITING WITH CHILDREN
KOLM. SMIRNOV P-VALUE				
Parametric	0.000	0.019	0.000	0.000
Kernel	0.000	0.008	0.000	0.000
Distance	0.365	0.180	0.393	0.056
Constrained	0.781	0.424	0.601	0.784
Grade Corr.	0.000	0.152	0.000	0.000
SIGN TEST P-VALUE				
Parametric	0.0000	0.8439	0.0000	0.0000
Kernel	0.0000	0.6937	0.0111	0.0000
Distance	0.6006	0.0756	0.4393	0.0873
Constrained	0.8752	0.4307	0.0400	0.7434
Grade Corr.	0.0000	0.8482	0.0000	0.0000
SIGNED RANK TEST P-VALUE				
Parametric	0.0000	0.9659	0.0000	0.0000
Kernel	0.0000	0.7772	0.0000	0.0000
Distance	0.9581	0.1487	0.8952	0.4990
Constrained	0.9793	0.3869	0.0580	0.8262
Grade Corr.	0.0000	0.6513	0.0000	0.0000

Table A4 P-values for Clothing and Footwear by Age of Reference Person

	AGE < 30	30 <= AGE < 50	50 <= AGE < 65	65 <= AGE
KOLM. SMIRNOV P-VALUE				
Parametric	0.000	0.000	0.000	0.000
Kernel	0.000	0.000	0.000	0.000
Distance	0.060	0.266	0.027	0.445
Constrained	0.246	0.606	0.440	0.280
Grade Corr.	0.513	0.467	0.021	0.292
SIGNTEST P-VALUE				
Parametric	0.0000	0.0000	0.0000	0.0000
Kernel	0.0000	0.0000	0.0000	0.0000
Distance	0.6225	0.6440	0.0966	0.9132
Constrained	0.5109	0.3863	0.4384	0.5614
Grade Corr.	0.6676	0.9104	0.8883	0.9432
SIGNED RANK TEST P-VALUE				
Parametric	0.0000	0.0000	0.0000	0.0006
Kernel	0.0000	0.0000	0.0000	0.0014
Distance	0.5804	0.4255	0.0816	0.3289
Constrained	0.3702	0.5941	0.6191	0.1008
Grade Corr.	0.4759	0.3834	0.1766	0.6787

Table A5 P-values for Clothing and Footwear by Professional Status

	(SELF-) EMPLOYED	UNEMPLOYED	(EARLY) RETIRED	OTHER
KOLM. SMIRNOV P-VALUE				
Parametric	0.000	0.000	0.000	0.000
Kernel	0.000	0.000	0.000	0.000
Distance	0.292	0.418	0.040	0.328
Constrained	0.772	0.167	0.569	0.732
Grade Corr.	0.009	0.439	0.830	0.762
SIGN TEST P-VALUE				
Parametric	0.0000	0.0002	0.0000	0.0106
Kernel	0.0000	0.0000	0.0000	0.0106
Distance	0.1116	0.7870	0.3112	0.9022
Constrained	0.5905	0.3074	0.6233	1.0000
Grade Corr.	0.7841	0.9474	0.9047	0.9050
SIGNED RANK TEST P-VALUE				
Parametric	0.0000	0.2529	0.0000	0.1108
Kernel	0.0000	0.1388	0.0000	0.2108
Distance	0.2560	0.8363	0.2228	0.7173
Constrained	0.8673	0.2142	0.9046	0.8845
Grade Corr.	0.0694	0.5889	0.8375	0.8468

Table A6 P-values for Clothing and Footwear by Household Type

	SINGLE WITHOUT CHILDREN	SINGLE WITH CHILDREN	COHABITING WITHOUT CHILDREN	COHABITING WITH CHILDREN
KOLM. SMIRNOV P-VALUE				
Parametric	0.000	0.000	0.000	0.000
Kernel	0.000	0.000	0.000	0.000
Distance	0.111	0.424	0.507	0.980
Constrained	0.562	0.245	0.689	0.923
Grade Corr.	0.000	0.576	0.001	0.000
SIGN TEST P-VALUE				
Parametric	0.0000	0.1145	0.0000	0.0000
Kernel	0.0000	0.0482	0.0000	0.0000
Distance	0.4529	0.2276	0.9784	0.2743
Constrained	0.9111	0.1933	0.8290	0.8267
Grade Corr.	0.0000	0.2853	0.0090	0.0000
SIGNED RANK TEST P-VALUE				
Parametric	0.0000	0.5362	0.0000	0.0005
Kernel	0.0000	0.6380	0.0000	0.0004
Distance	0.6935	0.3204	0.8651	0.7866
Constrained	0.8083	0.1432	0.9476	0.8488
Grade Corr.	0.0000	0.5147	0.0771	0.0000

Table A7 P-values for Private Transport by Age of Reference Person

	AGE < 30	30 <= AGE < 50	50 <= AGE < 65	65 <= AGE
KOLM. SMIRNOV P-VALUE				
Parametric	0.000	0.000	0.000	0.000
Kernel	0.000	0.000	0.000	0.000
Distance	0.803	0.208	0.027	0.184
Constrained	0.179	0.888	0.024	0.253
Grade Corr.	0.117	0.437	0.078	0.321
SIGN TEST P-VALUE				
Parametric	0.0000	0.0000	0.0000	0.0000
Kernel	0.0000	0.0000	0.0000	0.0000
Distance	1.0000	0.7100	0.2313	0.9139
Constrained	0.1707	0.5658	0.2922	0.4492
Grade Corr.	0.0123	0.7783	0.0526	0.5238
SIGNED RANK TEST P-VALUE				
Parametric	0.0000	0.0000	0.0000	0.0140
Kernel	0.0000	0.0000	0.0000	0.0107
Distance	0.7543	0.5907	0.1387	0.3946
Constrained	0.0849	0.5879	0.1294	0.1398
Grade Corr.	0.0064	0.7072	0.0205	0.3618

Table A8 P-values for Private Transport by Professional Status

	(SELF-) EMPLOYED	UNEMPLOYED	(EARLY) RETIRED	OTHER
KOLM. SMIRNOV P-VALUE				
Parametric	0.000	0.000	0.000	0.000
Kernel	0.000	0.000	0.000	0.000
Distance	0.292	0.418	0.071	0.732
Constrained	0.698	0.863	0.017	0.234
Grade Corr.	0.711	0.003	0.057	0.262
SIGN TEST P-VALUE				
Parametric	0.0000	0.0000	0.0000	0.0805
Kernel	0.0000	0.0000	0.0000	0.0021
Distance	0.9831	0.1704	0.6200	0.3916
Constrained	0.4439	0.6797	0.2659	0.1048
Grade Corr.	0.8337	0.0037	0.0736	0.3742
SIGNED RANK TEST P-VALUE				
Parametric	0.0000	0.0018	0.0000	0.2079
Kernel	0.0000	0.0004	0.0000	0.1102
Distance	0.5281	0.2536	0.2885	0.2036
Constrained	0.2410	0.9629	0.0380	0.1353
Grade Corr.	0.6898	0.0017	0.0074	0.3010

Table A9 P-values for Private Transport by Household Type

	SINGLE WITHOUT CHILDREN	SINGLE WITH CHILDREN	COHABITING WITHOUT CHILDREN	COHABITING WITH CHILDREN
KOLM. SMIRNOV P-VALUE				
Parametric	0.000	0.000	0.000	0.000
Kernel	0.000	0.000	0.000	0.000
Distance	0.022	0.130	0.091	0.221
Constrained	0.042	0.959	0.919	0.404
Grade Corr.	0.000	0.576	0.001	0.030
SIGN TEST P-VALUE				
Parametric	0.0000	0.0176	0.0000	0.0000
Kernel	0.0000	0.0015	0.0000	0.0000
Distance	0.3470	0.0375	0.1241	0.2141
Constrained	0.1562	0.4752	0.2821	1.0000
Grade Corr.	0.0000	0.7688	0.0049	0.0101
SIGNED RANK TEST P-VALUE				
Parametric	0.0000	0.1487	0.0000	0.0000
Kernel	0.0000	0.0527	0.0000	0.0000
Distance	0.4570	0.0383	0.1483	0.4078
Constrained	0.1728	0.6356	0.5067	0.8123
Grade Corr.	0.0000	0.7509	0.0022	0.0141

Table A10 P-values for Saving by Age of Reference Person

	AGE < 30	30 <= AGE < 50	50 <= AGE < 65	65 <= AGE
KOLM. SMIRNOV P-VALUE				
Parametric	0.000	0.000	0.000	0.000
Kernel	0.000	0.000	0.000	0.000
Distance	0.021	0.416	0.703	0.853
Constrained	0.006	0.476	0.383	0.694
Grade Corr.	0.000	0.381	0.008	0.016
SIGN TEST P-VALUE				
Parametric	0.0000	0.0000	0.0000	0.0001
Kernel	0.0000	0.0000	0.0000	0.0000
Distance	0.1060	0.1242	0.6334	0.6984
Constrained	0.0105	0.1242	0.7153	0.4179
Grade Corr.	0.0814	0.0990	0.2439	0.7804
SIGNED RANK TEST P-VALUE				
Parametric	0.0000	0.0000	0.0000	0.0097
Kernel	0.0000	0.0000	0.0000	0.0022
Distance	0.0473	0.2626	0.5934	0.9932
Constrained	0.0031	0.2073	0.9326	0.4939
Grade Corr.	0.0033	0.9675	0.7188	0.1205

Table A11 P-values for Saving by Professional Status

	(SELF-) EMPLOYED	UNEMPLOYED	(EARLY) RETIRED	OTHER
KOLM. SMIRNOV P-VALUE				
Parametric	0.000	0.000	0.000	0.328
Kernel	0.000	0.000	0.000	0.072
Distance	0.571	0.167	0.005	0.162
Constrained	0.698	0.719	0.000	0.046
Grade Corr.	0.000	0.048	0.001	0.361
SIGN TEST P-VALUE				
Parametric	0.0000	0.0359	0.0000	0.8176
Kernel	0.0000	0.0488	0.0000	0.8176
Distance	0.6131	0.1823	0.0007	0.2480
Constrained	0.8331	0.1122	0.0002	0.0639
Grade Corr.	0.9834	0.2880	0.0014	1.0000
SIGNED RANK TEST P-VALUE				
Parametric	0.0000	0.2288	0.0000	0.7038
Kernel	0.0000	0.3944	0.0000	0.8534
Distance	0.6747	0.5829	0.0035	0.2225
Constrained	0.6992	0.5988	0.0001	0.0559
Grade Corr.	0.0311	0.5618	0.0055	0.5543

Table A12 P-values for Saving by Household Type

	SINGLE WITHOUT CHILDREN	SINGLE WITH CHILDREN	COHABITING WITHOUT CHILDREN	COHABITING WITH CHILDREN
KOLM. SMIRNOV P-VALUE				
Parametric	0.000	0.019	0.000	0.000
Kernel	0.000	0.012	0.000	0.000
Distance	0.070	0.326	0.091	0.001
Constrained	0.202	0.424	0.991	0.139
Grade Corr.	0.000	0.697	0.001	0.000
SIGN TEST P-VALUE				
Parametric	0.0000	0.1145	0.0000	0.0000
Kernel	0.0000	0.0482	0.0000	0.0000
Distance	0.3734	0.4307	0.2742	0.0454
Constrained	0.2285	1.0000	0.7693	0.2030
Grade Corr.	0.0090	0.8482	0.0238	0.0000
SIGNED RANK TEST P-VALUE				
Parametric	0.0000	0.6009	0.0017	0.0000
Kernel	0.0000	0.4572	0.0000	0.0000
Distance	0.3145	0.6833	0.4703	0.0598
Constrained	0.2334	0.7697	0.5726	0.4227
Grade Corr.	0.0003	0.1839	0.2573	0.0000