# ON THE USE OF DATA MINING FOR IMPUTATION

Pilar Rey del Castillo, European Commission, Eurostat

**Abstract**. *Non-response in sample surveys has been a recurrent problem and literature on the topic is widely available, especially on partial or item non-response, where respondents reply to some but not all questions. The state of the art in the theory and practice of handling this type of missing data in surveys is represented by the use of model-based imputations, generated by defining an a priori model for the observed data and making inferences based on the likelihood or posterior distribution under that model. But the use of data models frequently implies strong restrictions and other robust procedures that do not depend on unrealistic model assumptions may aid to extract structures in the data in a reliable way. The use of data mining methods to impute individual missing data signifies a promising approach because these procedures seem to be robust against outliers and easily automatable.*

*This paper presents the results of comparing the imputations of simulated missing numerical data within the files of the EU-SILC in two countries, performed using different data mining methods and classical statistical procedures. Some conclusions are extracted on the behaviour of data mining methods against statistical procedures.*

## 1. Introduction

Non-response in sample surveys has been a recurrent problem and literature on the topic is widely available, especially on partial or item non-response, where respondents reply to some but not all questions.

One of the methods for dealing with partial non-response in surveys once it has occurred is to replace each missing value with an imputation or estimate, which would usually be obtained from the other non-missing variables available. The state of the art in the theory and practice of handling missing data in surveys using imputations involves model-based procedures, generated by defining an *a priori* model for the observed data and making inferences based on the likelihood or posterior distribution under that model. These procedures take into account the uncertainty resulting from incompleteness of data and provide good estimates of the sampling variance of estimators [8].

Multiple Imputation (MI) methods appear to be the most used of the model-based procedures for handling missing data in multivariate analysis at the current time. They involve replacing each missing value with a set of imputations drawn from the assumed model and combining these later in a specific way [13]. One of the suggested advantages of these methods is that only a small number of imputations (between three and five) is needed in order to obtain relatively efficient estimators. Many MI methods have been developed using assumptions from different models for continuous, categorical and mixed continuous and categorical data.

Some problems currently addressed in sample surveys, such as small area estimation or integration of data from different sources through statistical matching, have resulted in imputations being used increasingly widely. In most of these cases, the final aim of the imputation exercise is principally to obtain a microdata file free of missing data rather than to obtain good estimates for any particular parameters of the population. Another important difference between this and the common problem of non-response imputation is that the volume of missing information is usually higher, sometimes close to 80-90% of the data.

The use of data models could lead to significant restrictions on how the imputation can be performed in practice. Other robust procedures that do not depend on unrealistic model assumptions may provide a reliable way of identifying structures in the data. In particular, using data mining methods to impute individual missing data has potential, because these procedures can be robust against outliers and seem to be easier to automate.

As there appears to be no documented experience of using these techniques for non-response imputation, a set of experiments to compare the results of data mining methods and classical statistical procedures have been undertaken. As an initial step, imputations of continuous variables will be tested. The purpose of the experiment is to assess the quality of the individual estimates and the estimates of the mean when some of the variable values have been replaced by imputations. The test will be carried out using simulations of non-response in real-world situations, specifically the anonymised microdata files from the European Union Statistics on Income and Living Conditions (EU-SILC) for two countries.

The remainder of this paper is organised as follows: the next section briefly describes the context of the imputation problem, identifies the data used for the simulations, and reviews the imputation methods to be compared; section 3 then sets out the results of the comparisons; and finally, a number of remarks and conclusions are presented in Section 4.

## 2. Context of the problem and methods of imputation to be compared

### *2.1 The difficulties associated with imputation*

The interest in imputation for official statistics extends beyond its use for replacing non-response in surveys. New requirements for more disaggregated, integrated and consistent data have coincided with a reduction in available resources. This has resulted in significant changes to the systems used for producing statistics. In view of this, the use of imputations could improve efficiency by solving some of the problems currently faced: small area estimation could be performed using mass imputation, which involves providing values for the non-sampled primary units in order to create a complete set of responses for every unit in the universe [9]. Statistical matching also makes use of imputations as a way of providing joint statistical information based on two or more sources.

MI methods are currently widely used in multivariate analysis in many countries and for various types of surveys, as previously stated. They are however best suited for computing variances of estimates and do not provide unique individual imputations for each missing value. As a result, they are not appropriate methods for obtaining unique complete datasets, as is required in this case.

The problem this paper addresses is how to impute a numerical variable through its dependence on other numerical and/or categorical variables. The EU-SILC has been chosen as a real-world example for the application of the different imputation techniques. The aim of these statistics is to collect up-to-date and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions. The central part of the data, a very detailed breakdown of the components of income, is mainly collected at an individual level [5].

An important component of income is wages, defined as the payments received by employees in return for work done during the income reference period. Wages is the numerical variable chosen to be imputed using information on other available variables in the microdata file of personal data. This exercise is performed using data for one specific year, 2009, for two

countries, Spain and Austria. The variable is non-zero for those with paid employment, and the samples used to generate simulations are limited to these people.

The same independent numerical and categorical variables are used for all experiments and have been chosen because they have good explanatory power with respect to wages. They are: gender, age, country of birth, marital status, region, degree of urbanisation of the residential area, economic activity, highest level of education, managerial position, occupation, temporary job, part-time job, hours usually worked per week, years of education and years in main job.

Most data mining procedures incorporate a method for dealing with missing data in variables other than the one to be imputed. This gives these procedures a clear advantage over classical statistical methods for imputation, which do not provide imputations when non-response appears in other variables. Nonetheless, the existence of missing values in other variables and their treatment may mask the results of the imputations (of the variable in question). For this paper, to assess the pure imputation results without other effects, the simulations will therefore be conducted only within files with complete information, i.e. without missing values in other variables. The next section describes briefly the methods of imputation being compared.

### *2.2 Procedures for imputing continuous data*

The aim of some of the best known data mining algorithms is classification, i.e. identifying to which of a set of categories a new observation belongs. Procedures which result in classification are called classifiers: they are described as supervised when the classification is based on a training set of data with known outputs, and unsupervised when outputs are not known. Supervised classifiers may be used to impute missing data in categorical variables: every category or value of the variable to be imputed is associated with a class, and the estimation for a missing data input is the classification category.

There are relatively few supervised data mining procedures capable of producing continuous imputations and some of those that do exist are developed by generalising categorical classifiers in such a way that they can be applied to continuous variables. Imputations for non-response are in this case derived directly from the predicted values.

This section briefly presents the procedures used in the experiment to perform imputations of continuous variables from other numerical or categorical variables. The first four are data mining procedures (Least Median Squared Error Regressor, M5P algorithm, Multilayer Perceptron Regressor and Radial Basis Function) and the final two are classical statistical procedures, the first being a simple linear regression, and the second, Predictive Mean Matching, a model-based procedure.

**Least Median Squared Error Regressor imputation (LMS).** Outliers can dramatically affect classical least-squared linear regression because the squared distance accentuates the influence of points which are far away from the regression line. Statistical methods which try to rectify the influence of outliers are called ‘robust’, and one of the ways of performing more robust regressions is to minimise the median (instead of the mean) of the squares of the differences from the regression line. The method of the Least Median Squared Error Regressor repeatedly applies standard linear regression to subsamples of the data, and provides as an output the solution that has the smallest median-squared errors [16].

**M5P algorithm imputation (M5P).** A decision tree is a supervised classifier consisting of nodes and branches connecting the nodes. The nodes located at the bottom of the tree are called leaves while the top node in the tree is called the root. This root contains all the training examples that are to be divided into classes. All nodes except the leaves are called decision

nodes and each of these has a number of children nodes, equal to the number of values that a given feature assumes [3].

The problem of constructing a decision tree can be expressed recursively: once an attribute has been selected to be placed at the root node, one branch is made for each possible value, splitting up the set of examples into subsets, one for every value of the attribute. The process can be repeated recursively for each branch, using only the instances that actually reach the branch and stopping the corresponding part of the tree when all instances at a node have the same classification. Thus, the problem is deciding which attribute to use as the basis for splitting the examples, given a training set of examples with different classes and attributes. The algorithms are easily extended to deal with numerical attributes by placing numerical dividing points between the values.

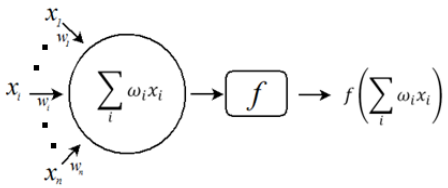
All algorithms for decision trees are based on a 'divide and conquer' strategy. This means that at each node they find the attribute that produces the purest daughter nodes using a certain measure of the purity of each node. The first decision tree algorithm –developed by Quinlan [11] – uses Shannon's entropy as a criterion to select the most discriminatory feature. Quinlan has developed other decision trees (C4.5, C5.0) as successors to the first one [12]. Trees used for numerical prediction are called model trees. They are similar to ordinary decision trees, except that at each leaf they store a linear regression model that predicts the value for instances reaching that leaf. Instead of maximising the information gain, in the case of numerical prediction the intra-subset variation in the values occurring along each branch is minimised.

The algorithm M5P (M5') is a reconstruction of Quinlan's algorithms for inducing trees of regression models. Its operation is fully described in [15].

**Multilayer Perceptron Regressor imputation (MLP)**. The computational architecture of artificial neural networks is based on the neural structure of the brain and seeks to model the information processing capabilities of nervous systems. They are called 'adaptive' because they can learn to estimate the parameters of a population using a number of examples.

Neural networks are essentially built from simple units called neurons which are linked by a set of weighted connections. These neurons are usually organised into several layers, the first being called the input layer and the last the output layer. When there are intermediate layers these are called hidden layers. The neurons in the input layer correspond to variables or features in the input data set. The information to be analysed is fed into the neurons in the first layer and then propagated to the neurons in the second layer for further processing, the result of this processing then being propagated to the next layer and so on until the last layer is reached. Each neuron or unit receives information, either from other units or from the external world through some sort of device, and processes it, thus producing the output of that unit [1].

**Figure 1. Processing of information in a basic neuron**




The objective of the network is to learn some structure or association between input and output. Learning is typically performed by adjusting the connection weights between neurons. From a statistical perspective, the connections could be seen as parameters to be estimated using the

training data, while the learning process functions as an algorithm by which to arrive at the estimates. Figure 1 shows the processing of information from the input through to the output response in a basic neural unit: firstly, the activation of the neuron is computed as the weighted sum of its inputs; secondly, the activation is transformed into the output by using a transfer function. Any function whose domain is the real numbers may be used as a transfer function, the logistic function $f(x) = 1/(1+e^{-x})$ being one of the more common. It maps the real numbers onto the interval [-1,1], and its derivative –needed for the learning process– is easily computed as $f'(x) = 1/(1+e^{-x})$.

The behaviour of the network is specified completely by the structure of layers and neurons, together with the transfer functions and the weights. The learning rules determine the way the weights change as a function of their performance. The 'gradient descent' or 'delta rule' with back-propagation is the most widely used supervised learning rule, using the difference between the real and the expected output as the error signal. In back-propagation the error signal is propagated from the output end to the input on a layer-by-layer basis. During back-propagation, the values of the weights are adjusted by the error feedback and the continuous modification of the weights and the offsets is applied to make the real output of the network closer to the expected one.

Multilayer perceptron is one of the most popular types of neural network. It belongs to a class called feed-forward networks, which are named as such because they contain no loops or cycles within the same layer and the output depends only on the current input instance. The network has one hidden layer, uses the delta rule as its learning algorithm and the logistic function as the transfer function for the neurons in the hidden layer. The output layer predicts a numerical variable and therefore has only one node or neuron with linear activation, whose output value is compared to the original value in the input to compute the error. Given the error function $E(w_{ij})$ and the learning rate $\eta$, the modification to the weight $\Delta w_{ij} = -\eta\, \delta E(w_{ij})/\delta w_{ij}$ is applied to each weight $w_{ij}$ for each training instance until the network error function is small enough.

**Radial Basis Function imputation (RBF)**. This is another popular type of feed-forward neural network differing from the multilayer perceptron in the way the hidden unit performs computations. Each hidden unit represents a prototype in the input space and its activation or output for a particular input point depends on the distance between the prototype and the input point, the activation being stronger when the two points are closer. This is achieved by using a non-linear transformation function to convert the distance into a similarity measure through a *Gaussian* activation function. The hidden units are called RBFs because the points in the input space which produce the same activation form a hypersphere or a hyperellipsoid. The output layer is the same as that of a multilayer perceptron. It takes a linear combination of the outputs of the hidden units and pipes it through the sigmoid function. The parameters to be learnt in such a network are the centres, i.e. the prototypes defining the hidden units, and the weights used to form the linear combination of the outputs obtained from the hidden layer. The first set of parameters may be determined by clustering. The second set of parameters can then be learnt by keeping the first fixed [16].

**Regression imputation (REG).** Imputation by regression is performed by simply computing for each input of covariate variables the regression forecast using the parameters of the regression estimated from the training set. Since some of the covariates are in this case categorical, they are previously treated by constructing appropriate dummy variables for each category or class (less one, because its value can be derived from the values of the others). This imputation is obtained as a baseline for comparisons only: it is well known that the variance from the file imputed in this way is underestimated, this being the reason for introducing model-based imputation procedures.

**Predictive Mean Matching imputation (PMM)**. The predictive mean matching method works in a similar way to the regression method, the difference being that, for each missing value, it imputes a value randomly from the set of observed values having the closest predicted values to the predicted value obtained from the simulated regression model [14]. Pre-treatment of the categorical covariates is also performed so as to introduce them as inputs in the regression model. This particular MI method has been included in the testing because it was identified as performing the best imputations on the proposed files [7].

## 3. Comparison of imputation methods

The experiment consists of testing the imputation methods proposed in the previous section using the 2009 EU-SILC personal microdata files of two countries, Spain and Austria. These files have been anonymised to prevent individuals being identified. As the purpose is to simulate missing wages values, only the instances referring to individuals performing a paid job, with information on wages and with no missing values in the fifteen covariate variables listed in Section 2, have been collected for testing. This provides a sample of 9 810 individuals in Spain and 4 560 in Austria.

Simulations are always carried out using the same set of files for all steps, methods and countries. This set of files is generated as follows: the two original sample files (Spain and Austria) are randomly ordered ten times, and each time the file is then split into two parts. For each of the ten orders, the imputation is performed in two ways:

1. using the first half of the data as training data and the second half as the test data to be imputed; and
2. using the second half of the data as training data and the first half as the test data to be imputed.

This method provides, for each of the countries, twenty possibilities for performing imputations which can then be compared to the original known individual values, assuming 50% missing data in the wages variable, i.e. it makes it possible to perform ten two-fold cross-validations within the data set for each country. This means that, for each version of the Spanish file, 4 905 cases are imputed from the other 4 905 which are used as the training set each time. Similarly for Austria, 2 280 cases are imputed from the other 2 280. A proportion of 50% non-response has been chosen so as to try to assess the performance of the methods when non-response is high and also for practical reasons.

Version 3.7.9 of the WEKA data mining software has been used to compute the results of the data mining procedures [6]. This is an open source project containing a collection of machine learning algorithms and tools that make it easy to test and compare different procedures. SAS/STATS software, Version 9.2 of the SAS System for Windows, Copyright © 2002-2008 by SAS Institute Inc., Cary, NC, USA has also been used for other types of processing.

The first assessment of the 40 sets of imputations (two countries x two halves x 10 ordered files), is performed based on various measures of the closeness to the original non-imputed values. Let $X_i$ , $i=1,\ldots, n$ be the original non-imputed value, $\hat{X}_i, i=1,\ldots,n$ the corresponding imputations obtained using the other half as training set, and $w_i$ , $i=1, \ldots, n$ the corresponding sample weights. The following measures are computed:

(a) Correlation coefficient: $\rho = \left( \sum_{i=1}^{n} w_i \left(X_i - \overline{X}\right)\left(\hat{X}_i - \overline{\hat{X}}\right) \right) \Big/ \left( \sqrt{\sum_{i=1}^{n} w_i \left(X_i - \overline{X}\right)^2 \left(\hat{X}_i - \overline{\hat{X}}\right)^2} \right)$

(b) Mean Absolute Error: $MAE = \sum_{i=1}^{n} w_i \left| X_i - \hat{X}_i \right| \Big/ \sum_{i=1}^{n} w_i$

(c) Root Mean Squared Error: $RMSE = \sqrt{\sum_{i=1}^{n} w_i \left(X_i - \hat{X}_i\right)^2 \Big/ \sum_{i=1}^{n} w_i}$

(d) Relative Absolute Error: $RAE = \left( \sum_{i=1}^{n} w_i \left| X_i - \hat{X}_i \right| \Big/ \sum_{i=1}^{n} w_i \left| X_i - \overline{X} \right| \right).100$

(e) Root Relative Squared Error: $RRSE = \sqrt{\sum_{i=1}^{n} w_i \left| X_i - \hat{X}_i \right| \Big/ \sum_{i=1}^{n} w_i \left| X_i - \overline{X} \right|}.100$

These five measures are commonly used to assess the performance of data mining predictors for numerical values [16]: (a) is simply the correlation between the original and the imputed data; (b) and (c) represent absolute figures (in this case, wages measured in euros); and (d) and (e) are relative values, comparing the average difference between original and imputed data against the average deviation of the original values (these figures can be greater or less than 100).

**Table 1**

| COUNTRY | METHOD | Correlation | MAE | RMSE | RAE | RRSE |
|---|---|---|---|---|---|---|
| ES | LMS | 0.74 | 435.8 | 708.3 | 59.1 | 68.0 |
| ES | M5P | 0.75 | 431.3 | 694.5 | 58.4 | 66.7 |
| ES | MLP | 0.73 | 449.6 | 718.8 | 60.9 | 69.0 |
| ES | PMM | 0.55 | 634.8 | 982.5 | 86.0 | 94.3 |
| ES | RBF | 0.75 | 430.0 | 696.2 | 58.3 | 66.8 |
| ES | REG | 0.73 | 443.8 | 716.9 | 60.1 | 68.8 |
| AT | LMS | 0.53 | 648.5 | 1551.6 | 63.6 | 84.6 |
| AT | M5P | 0.55 | 636.3 | 1529.1 | 62.4 | 83.2 |
| AT | MLP | 0.44 | 751.7 | 1733.7 | 73.7 | 96.1 |
| AT | PMM | 0.33 | 944.5 | 2067.1 | 92.7 | 116.6 |
| AT | RBF | 0.53 | 643.7 | 1543.1 | 63.1 | 84.0 |
| AT | REG | 0.52 | 655.7 | 1561.9 | 64.3 | 85.2 |

Table 1 shows the average values of the five measures for the test data sets for each country, for each of the imputation methods considered, and using the same set of 15 variables as inputs.
It is clear that the best overall result when all five measures are taken into consideration (i.e. higher correlation and smaller errors) for each country is achieved using the M5P decision tree, a method whose processing time is much lower than that of the other data mining methods. Also, the worst result, by a significant margin, is obtained using the PMM method. The results of the imputations by regression (REG) are of similar order to that of the worst data mining method, the MLP. In summary, it can be noted that the data mining methods for producing individual imputations tested are more successful in reproducing the original data than the classical statistical procedures. They offer significantly superior results to those produced by the PMM MI method, and also represent an improvement, albeit a smaller one, on the imputations by regression.

A further assessment of the imputation methods involves comparing the output for the mean and other related parameter estimates. The closer these estimates are to the originals obtained without simulated missing data, the better the imputation method. There are 20 files for each

country, each of which can be used for producing new estimates. Each file contains 50% original data and 50% imputed data. For instance, if the first half of the data is imputed, the mean estimate is computed as:

$$\overline{X} = \frac{\sum_{i=1}^{n/2} w_i \hat{X}_i + \sum_{i=n/2+1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

Table 2 shows the original values (from the files with no imputations) and the averages in the corresponding 50%-imputed files of the estimates of some typical statistics (mean, mode, median and standard deviation), for each method and each country.

**Table 2**

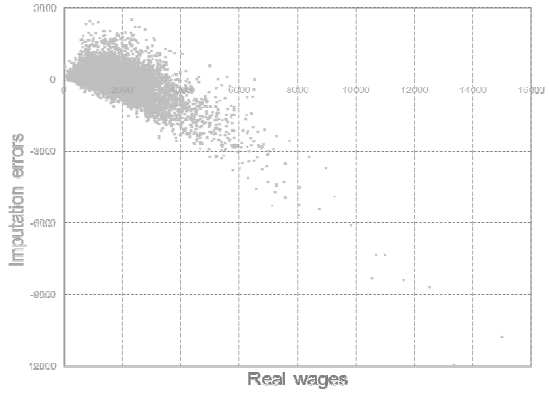
| COUNTRY | METHOD | Mean | Mode | Median | STD |
|---|---|---|---|---|---|
| ES | ORIGINAL | 1820 | 1400 | 1575 | 10.5 |
| ES | LMS | 1780 | 1400 | 1595 | 8.9 |
| ES | M5P | 1777 | 1400 | 1592 | 9.2 |
| ES | MLP | 1782 | 1400 | 1587 | 9.3 |
| ES | PMM | 1819 | 1305 | 1572 | 10.5 |
| ES | RBF | 1776 | 1400 | 1586 | 9.2 |
| ES | REG | 1775 | 1400 | 1605 | 9.0 |
| AT | ORIGINAL | 2287 | 1800 | 1955 | 27.3 |
| AT | LMS | 2228 | 1915 | 1998 | 21.8 |
| AT | M5P | 2209 | 1915 | 1993 | 21.5 |
| AT | MLP | 2238 | 1915 | 1989 | 23.1 |
| AT | PMM | 2288 | 1500 | 1968 | 25.9 |
| AT | RBF | 2214 | 1915 | 1994 | 21.8 |
| AT | REG | 2205 | 1915 | 1997 | 21.6 |

It can be noted that, although the individually imputed values are much closer to the original when data mining methods are used for imputation, the mean estimates from these methods are further away from the original. For both countries, the mean estimates obtained from data mining methods are systematically smaller than the original mean. The same is true of imputations by regression but not of the PMM procedure, where the means are almost identical.

The median and standard deviation estimates show a similar pattern, with the estimates obtained using the PMM MI procedure being closer to the original values. In contrast, it appears that much better estimates of the mode are obtained using data mining and regression imputation methods. More detailed analysis shows the reason for this to be that imputations performed using data mining procedures are not centred, i.e. the forecast errors are not randomly distributed around the original values and the imputations show a certain shrinkage towards the mean [4].

This can be seen in Figure 2 which shows how, when using the M5P method of imputation on one version of the Spanish file, the imputation errors corresponding to the values of the original wages change as we move away from the mean value. To the left of the mean value of 1 820, most of the imputation errors are positive while they become increasingly negative towards the right. This phenomenon can also be seen with similar magnitude when the imputations are computed using other data mining methods or using simple regression. PMM imputations show this shrinkage to a lesser extent, in contrast, resulting in the estimates of the mean and standard deviation being more centred.

**Figure 2. Imputation errors corresponding to the original wages variable**




An alternative way of assessing the effects of the imputation methods on the possible inferences from the 50%-imputed files is to evaluate the similarity between the empirical original distribution of the wages and the values obtained with 50%-imputed and 50%-original values.An appropriate measure to determine the similarity between empirical distributions when dealing with an ordered variable is the Kolmogorov-Smirnov statistic [2]:

$$KS(P,Q) = \max_{1 \le k \le m} \left| \sum_{j=1}^{k} p_j - \sum_{j=1}^{k} q_j \right|$$

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 ordered categories or intervals for each of the two distributions. Another independent measure which could be considered is the Hellinger distance [10] between distributions, although this is more suitable for categorical non-ordered variables:

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

The averages of these two distances between the original and the 50%-imputed distributions, calculated using intervals constructed based on the deciles of the original distribution of wages are shown in Table 3.

**Table 3**

| Country | Method | Hellinger distance | Kolmogorov-Smirnov distance |
|---|---|---|---|
| ES | LMS | 0.0504 | 0.0305 |
| ES | M5P | 0.0428 | 0.0280 |
| ES | MLP | 0.0359 | 0.0233 |
| ES | PMM | 0.0148 | 0.0085 |
| ES | RBF | 0.0410 | 0.0268 |
| ES | REG | 0.0518 | 0.0354 |
| AT | LMS | 0.0488 | 0.0284 |
| AT | M5P | 0.0502 | 0.0301 |
| AT | MLP | 0.0362 | 0.0216 |
| AT | PMM | 0.0180 | 0.0119 |
| AT | RBF | 0.0445 | 0.0260 |
| AT | REG | 0.0504 | 0.0302 |

The Kolmogorov-Smirnov distances obtained imply that the averages of the differences between the cumulative distribution functions in each decile are always less than 4%. Irrespective of which country or measure of distance is considered, the PMM method exhibits a significantly greater similarity with the original data distribution than any other method. It can also be seen that, based on this assessment measure, the distribution of the 50%-imputed data obtained using imputations by regression seems to be slightly worse than that obtained using data mining procedures.

Although the previous results are valuable, the purpose of the imputation exercise should also be considered. Thus, for example, when imputations are performed with the purpose of obtaining complete files free of missing data, the results at a more detailed level of disaggregation can be reversed. The same simulations have revealed that, descending to the regional level, the comparative advantages of data mining procedures seem to hold while those of the PMM method can disappear. An example of this may be seen in Tables 4, 5 and 6 which show the resulting comparisons for the region of Extremadura in Spain.

**Table 4**

| METHOD | Correlation | MAE | RMSE | RAE | RRSE |
|---|---|---|---|---|---|
| LMS | 0.85 | 317.2 | 489.9 | 54.8 | 57.1 |
| M5P | 0.83 | 313.5 | 489.4 | 54.2 | 57.1 |
| MLP | 0.80 | 337.8 | 521.8 | 58.3 | 60.8 |
| PMM | 0.66 | 504.8 | 731.8 | 87.5 | 86.0 |
| RBF | 0.84 | 314.8 | 480.1 | 54.4 | 56.0 |
| REG | 0.82 | 339.1 | 504.7 | 58.6 | 58.9 |

**Table 5**

| METHOD | Mean | Mode | Median | STD |
|---|---|---|---|---|
| LMS | 1477 | 1372 | 1348 | 5.7 |
| M5P | 1471 | 1372 | 1337 | 6.0 |
| MLP | 1476 | 1372 | 1340 | 6.1 |
| ORI | 1492 | 1400 | 1317 | 6.8 |
| PMM | 1557 | 1373 | 1374 | 6.9 |
| RBF | 1467 | 1372 | 1323 | 6.0 |
| REG | 1519 | 1372 | 1393 | 5.9 |

**Table 6**

| METHOD | Hellinger distance | Kolmogorov-Smirnov distance |
|---|---|---|
| LMS | 0.083 | 0.055 |
| M5P | 0.068 | 0.045 |
| MLP | 0.067 | 0.044 |
| PMM | 0.076 | 0.063 |
| RBF | 0.062 | 0.038 |
| REG | 0.088 | 0.086 |

## 4. Final remarks

The previous section drew comparisons between the results of data mining and statistical methods for imputation. The main findings can be summarised as follows:

- Data mining procedures provide imputations which reproduce the original individual values significantly better than the PMM imputation procedure.
- When imputation methods are used to produce estimates of statistical parameters such as the mean and standard deviation, the PMM method produces significantly better estimates than data mining methods.
- The results of imputation by regression are slightly worse than those of most data mining imputation procedures, both in terms of reproducing original individual values and estimating typical statistical parameters.

In addition to these general conclusions, the following comments and questions merit particular mention:

(a) Given the original non-imputed population of wages values, it may at first seem contradictory to find one imputed-population having, at the same time, more similar individual values and a more divergent statistical distribution than another differently imputed population. This is however clearly explained by the bias produced by the shrinkage to the mean seen in the imputed values obtained from data mining methods.
(b) The PMM being a multiple imputation procedure, it produces random rather than fixed imputations that are specifically designed to improve the estimates of the standard deviation. This objective is achieved and the procedure also gives good estimates for the mean. It seems however that the objective is achieved at the cost of closeness to the individual values of the original file, because the simple imputations obtained by regression, which form the basis for the PMM method, are similarly closer than those obtained using data mining methods.
(c) The results of data mining methods in producing estimates from the partially-imputed files can be improved in a number of ways. There is a training set where the discrepancies between the original and the imputed data can be learned and data mining procedures can be used to correct the first imputations. Other possibilities include correcting the initial imputations in the opposite way as the 'shrinkage estimator' [4] does, or randomising the imputations in a similar way as the PMM method does.
(d) It may also be worth considering the individual one-to-one likeness when assessing the similarity between empirical populations. This criterion may be of particular interest when the aim is to obtain a complete microdata file, as is the case in imputations produced for small area estimation purposes.
(e) The advantage of the PMM imputation method in the estimates of the mean and other parameters at lower levels of disaggregation can disappear, while the advantage of data mining methods in terms of individual one-to-one likeness seems to hold.

**References**

[1] Abdi, H., Valentin, D., and Edelman, B. (1999), Neural Networks, Sage University, Thousand Oaks, CA.

[2] Carruth, J., Tygert, M., and Ward, R. (2012), A comparison of the discrete Kolmogorov-Smirnov statistic and the Euclidean distance, Eprint arXiv:1206.6367.

[3] Cios, K. J., Pedrycz, W., Swiniarski, R. W., and Kurgan, L. A. (2007), Data Mining: A Knowledge Discovery Approach, Springer, USA.

[4] Copas, J. B. (1983), Regression, Prediction and Shrinkage, Journal of the Royal Statistical Society, series B, vol. 45 (3), 311-354.

[5] Eurostat (2013), European Union Statistics on Income and Living Conditions, http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc (Accessed 30 November 2013).

[6] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009), The WEKA Data Mining Software: An Update, SIGKDD Explorations vol. 11 (1).

[7] Leulescu. A., Agafitei. M. (2012), A quality framework for matching EU social surveys, Dissertation, European Conference on Quality in Official Statistics.

[8] Little R. J., and Rubin, D. B. (2002), Statistical Analysis with Missing Data, 2nd ed., John Wiley and Sons, New York.

[9] Moore, R., and Robbins, N. (2004), A Study of Mass Imputation in Small-area Estimation, Joint Statistical Meeting, Toronto, Canada.

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

[11] Quinlan, J. R. (1986), Induction of Decision Trees, Journal of Machine Learning vol.1 (1), 81–106.

[12] Quinlan, J. R. (1996), Improved use of continuous attributes in C4.5, Journal of Artificial Intelligence Research vol. 4, 77-90.

[13] Rubin, D.B. (1978), Multiple Imputations in Sample Surveys, Proceedings of the Section on Survey Research Methods, 20-34.

[14] Schenker, N., Taylor, J. M. G. (1996), Partially Parametric Techniques for Multiple Imputation, Computational Statistics and Data Analysis vol. 22 (4), 425–446.

[15] Wang, Y., and Witten, I. H. (1996), Induction of Model Trees for Predicting Continuous Classes, Working Paper Series 96/23, University of Waikato, New Zealand.

[16] Witten, I.H., Frank, E., and Hall, M.A. (2011), Data mining: practical machine learning tools and techniques, 3rd ed. Morgan Kaufmann Series in Data Management Systems, Elsevier Inc.