Data imputation and file merging using the forest climbing algorithm

María Jesús BARCENA* Fernando TUSELL*

Abstract

We address the problem of completing two files with records containing a common subset of variables. The technique investigated involves the use of regression and/or classification trees. An extension (the "forest-climbing" algorithm) is proposed to deal with multivariate response variables. The method is demonstrated on a real problem, in which two surveys are merged, and shown to be feasible and have some desirable properties.

Keywords: file matching; survey linking; imputation; regression trees.

Acknowledgements. We thank for support the Spanish MEC (grant PB95-0346) and UPV/EHU (grant PB95-0346). We gratefully ac-

*Departamento de Estadística y Econometría. Facultad de CC.EE. y Empresariales, Avda. del Lehendakari Aguirre, 83, 48015 BILBAO. E-mail: etptupaf@bs.ehu.es.

knowledge comments from Vicente Nuñez, Eva Ferreira and Karmele Fernández. Any errors and obscurities that remain are our own.

1 Introduction

1.1 The problem

Our starting point are two files, A and B with a total of $N = N_A + N_B$ observations and the following structure:

File A

$X_1, \ldots, X_p$	$Y_1, \ldots, Y_q$	unknown

File B

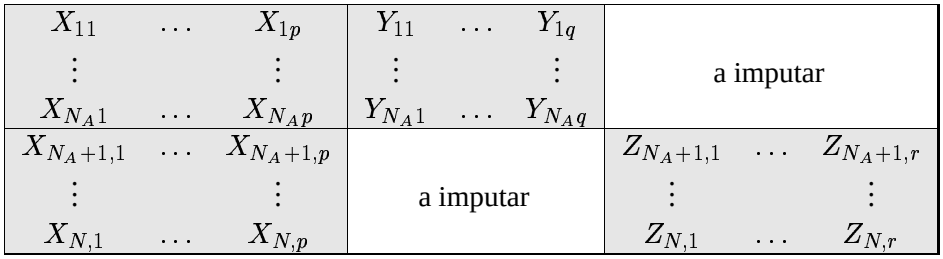
$X_1, \ldots, X_p$	unknown	$Z_1, \ldots, Z_r$

In the problem that motivates this research, this two files contain data from two sample surveys which share a common set of questions. The population sampled is assumed to be the same for both surveys. Variables $X_1, \ldots, X_p$ are thus known for all N cases. Beyond this common set of variables, each file contains also data on a specific set of variables: $Y_1, \ldots, Y_q$ and $Z_1, \ldots, Z_r$ in our notation.

In this paper we deal with the problem of imputing the missing values in the files A y B, using the values of the common variables $X_1, \ldots, X_p$. In other words, if we think of the data arranged in a table such as Figure 1, we address the problem of imputing the non shaded areas. The goal is to create a data set as close as possible to what would have been obtained, had we had the chance to gather complete information on variables (X, Y, Z) for all N cases.

It is clear that we can only hope for limited success, inasmuch as the data set

Figure 1: Data set in a two survey linking problem.



does not provide information on the intrinsic relationship among Y and Z given X. On the other hand, the recovery of the values of Y and Z might be very poor using only the information in X. Nonetheless, the savings realized by creating a complete data set by imputation can be tremendous. In many cases, an existing data set with the structure depicted in Figure 1 is all we have, and the possibility of further sampling complete data on (X, Y, Z) simply does not exist; in those cases, imputation is the only way to go. Either way, the problem is of practical importance, and has received a good deal of attention.

1.2 The application

The EPT-93 (Survey of Time Uses), compiled by the EUSTAT (Basque Institute of Statistics), is a survey comprising responses of 5040 individuals which were asked about the time devoted daily to different purposes, grouped by us in 24 categories. The field work was done in the Fall of 1992 and Spring of 1993; refer to EUSTAT (1997) for a description of the survey.

Ideally, each individual should answer the survey both on working and non working days; but when individuals were asked to compile a daily breakdown of their uses of time for a period one week long, it was found that the quality of the diaries[1]

> " . . . decreases as the week goes on, incomplete diaries proliferate, and the survey cannot be carried out without a substantial monetary reward."

In the end, single day questionnaires were administered, to be completed for a specific day of the week. We therefore have 2521 individuals who answered on working days (in file `trabajo.dat`) and 2519 who answered on weekend days (in file `fiesta.dat`). The information on these two sets of answers compose files A and B in the notation of Section 1.1. In either file we have answers to a common set of characterization variables $X_1, \ldots, X_p$, and then information regarding the use of time in a working day (variables $Y_1, \ldots, Y_q$, in file A) or a weekend day (variables $Z_1, \ldots, Z_r$, in file B).

It is then of interest to create a single data set from files A and B, and the method described in Sections 3 and 4 will be used for that purpose.

1.3 Outline of the paper

Section 2 describes very briefly some of the techniques that have been used for survey imputation or file matching, and provides some pointers to the literature.

[1] Our translation. See EUSTAT (1997), p. XII.

Sections 3 and 4 introduce the proposed method, against the background of the existing techniques. Section 5 demonstrates its feasibility and use. Section 6 shows the performance of trees in simple simulated examples. Some concluding remarks in Section 7 close the paper.

2 Survey linking techniques

A short description of imputation techniques and some pointers to the literature are given next. The interested reader may also refer to Nordholt (1998).

2.1 Regression and the EM algorithm

A quite natural idea is to use the cases with complete values of variables (X, Y) or (X, Z) to fit regressions of Y on X and Z on X, and then use these regressions to impute the missing values by $\hat{Y}$ or $\hat{Z}$. This idea goes back at least to 1960, (see Buck (1960) for an early proponent).

The implied assumptions when using regression in this manner are: i) Constancy of the relationship among the predictors X and the responses in both surveys, and ii) Null partial correlation of Y and Z given X. Obviously, a good fit of the regression models is also required, for the imputation to be any good.

It is important to realize that *even if the above assumptions are justified,* the imputed values will lack the variability of the genuine values: we are replacing unknown values *about* the regression hyperplane by imputed values *on* the hyperplane. This has to be corrected or taken into account in any subsequent analysis.

The EM algorithm advocated in Dempster et al. (1976) (see also Rubin (1991)) provides an easy, iterative way to maximize the likelihood function of incomplete data in a wide variety of settings.

In order to use the EM algorithm we have to specify a likelihood, which requires information or an hypothesis about the generating mechanism of the data. The imputed values are maximum likelihood values given the data, and suffer from the same lack of variability than the regression imputed values —in fact, when the distribution underlying (X, Y, Z) is multivariate normal, the imputed values are linear regression fits.

2.2 Nearest neighbours and deck replacement

Another common idea is to replace the missing values in one case by those of another case in some sense "close" to it, according to a predefined notion of "closeness" in the space of common variables X. For instance, to impute $\mathbf{Y_j} = (Y_{j,1}, \ldots, Y_{j,q})^T$ for $j \in \{N_A + 1, \ldots, N\}$ we could use $\hat{\mathbf{Y}}_\mathbf{j} = \mathbf{Y_i}$ for some $i \in \{1, \ldots, N_A\}$ such that $\mathbf{X_i} \simeq \mathbf{X_j}$. This gives rise to a variety of flavours of the nearest neighbour idea, depending on how we define proximity in the space $\mathcal{X}$ of the X variables.

Sometimes, "closeness" means "close in the card deck", reflecting the practice of replacing the missing values in one case with those of the case next to it in the computer card deck —a reasonable procedure if the order in the deck reflects geographical contiguity or otherwise similarity among cases. See for instance (Little and Rubin, 1987, p. 60).

In spite of their simplicity, deck replacement methods have advantages: the imputed values do not suffer from the lack of variability that afflict the regression imputed values. Also, the imputed values belong to some other case in the sample, hence are realistic and internally coherent. We will turn to this issue later.

2.3 Factorial techniques

There have been various proposals of methods to uncover the relationship among Y and Z in settings like ours. Aluja and Rius (1994) and Aluja et al. (1995) show how to project the information in one survey onto the factorial planes obtained from the other, using techniques such as multiple correspondence analysis and principal components analysis. While these techniques can conceivably be used to obtain imputed values from the projections, we think their main advantage is their ability to provide visual access to the structure of the problem at hand.

Similar and closely related methods are Dear's principal components method and Krzanowski's singular value decomposition method: see Bello (1993) for a short description and references, and a simulation study comparing their performance.

2.4 Neural networks

Artificial Neural Networks have shown great usefulness in many problems, as universal approximators. They are ideally suited to model complex relationships when there is no clear choice of a parsimonious model. Useful monographs are Ripley

(1996) and Bishop (1996). Nordbotten (1996) and Villagarcía and Muñoz (1997) are examples of uses in survey imputation.

2.5 Multiple imputation

Not an imputation technique, is rather a method that can improve many of them: Little and Rubin (1987) convincingly shows its rationale and benefits. See also Rubin (1986).

The idea is to generate not one but several complete data sets by imputation. In our setting, we would create several matrices such as the one in Figure 1, sharing the shaded areas but with different imputations for the missing data. We can then perform several classical, complete data analysis, and compare them to have an idea on how much the results vary due to random fluctuation in the imputation.

3 Imputation using regression or classification trees

We propose the use of regression and/or classification trees to impute missing values. Using trees has a number of advantages: it gives a unified treatment of continuous and categorical variables, provides useful byproducts to assess the goodness of fit and makes multiple imputation easy. Trees also have well known advantages: flexibility, few assumptions, relative insensitivity to outliers, etc. The seminal book Breiman et al. (1984) describes these advantages.

3.1 Univariate Y and Z

Consider the simplest possible case in which we have p common variables X and $q = r = 1$, i.e. there is only one specific variable to impute in each survey (refer to Figure 1, p. 3). The case (usually more relevant in practice) $q > 1$, $r > 1$ is taken up in Section 3.2, and a method is proposed in Section 4.

Let $\mathcal{X}$ be the space of all possible values of X. The idea is to build two partitions of $\mathcal{X}$ such that in each class of the first (respectively, second) we have like values of Y (respectively, Z). Since no restrictions are imposed on the kind and distributions of the variables, the CART methodology described in Breiman et al. (1984) seems a good way to build such partitions, growing one tree for each specific variable. Then, to impute a case we drop it down the relevant tree and look at the leave where it ends. This is formalized in Algorithm 1, pág. 10.

Notice that the method described lends itself quite well to multiple imputation, since each case will normally end in a terminal node which contains cases with more than one value of Z (or Y). Thus, we can sample among them at random to create multiple imputations. We can also impute using the mean, median, etc., or mode in the case of categorical variables.

It is worth mentioning that Algorithm 1 can be seen as a regression method —only a tree replaces familiar linear or nonlinear regression, with both advantages and disadvantages. It can also be seen as a nearest neighbour method. But, while a nearest neighbour method using, for instance, Mahalanobis distance in the space $\mathcal{X}$, would disregard the values of the Y (or Z), here a different notion of proximity

Algorithm 1 – Univariate imputation using trees.

1. Build a tree $\mathcal{Y}_X$ "regressing" Y on the X, using cross validation and observations $i = 1, \ldots, N_A$. Let $\mathcal{Y}_1, \ldots, \mathcal{Y}_a$ the leaves of said tree, and $\mathcal{Y}$ the partition they form.

2. Build a tree $\mathcal{Z}_X$ "regressing" Z on the X, using cross validation and observations $i = N_A + 1, \ldots, N$. Let $\mathcal{Z}_1, \ldots, \mathcal{Z}_b$ be the leaves of said tree and $\mathcal{Z}$ the partition they form.

3. To impute the value of Z for a case with $i \in \{1, \ldots, N_A\}$, drop it down the tree $\mathcal{Z}_X$. If it falls in the leave $\mathcal{Z}_{\delta(i)}$, we impute Z as a function of the values Z observed in that leave.

   Do likewise to impute Y for the cases in which such variable is missing ($i \in \{N_A + 1, \ldots, N\}$).

is used. A case is “near” another if both happen to fall in the same leave when dropped down the relevant tree. Thus, the notion of proximity used does take into account the response variable. This is quite important, and is further discussed in Bárcena and Tusell (1998).

3.2 Multivariate Y and Z

When we attempt to generalize the method to multivariate Y and Z (q>1 and r>1), we stumble upon a pitfall. We would like a method to construct trees partitioning the $\mathcal{X}$ space in such a way that each class contains like values of the (multivariate) response: but there is no unique way to define likeness in a multidimensional space. A possibility is to use Kullback-Leibler or a similar measure of discrepancy as in Ciampi (1991), but this requires a model for the distribution of the response variables.

One obvious way out is to use different trees to impute each of the variables in Y (or Z), effectively turning a multivariate problem into q (or r) univariate ones. This is clearly undesirable, for it disregards relationships that may exists among components of Y (or Z). Eventually, nonsensical imputations might be produced which fail to comply logical or arithmetic constraints that we know must hold. In the application shown below (EPT - Survey on the Uses of Time) we might produce for some cases imputations which do not add up to 24 hours, as they should.

To circumvent such problem, it is desirable to impute all variables for each case i at once, taking the values of an observed “similar” case. This automatically

guarantees consistency of the imputed values, and is a commonly accepted way to proceed (see Lejeune (1995), pág. 140 and Lebart and Lejeune (1995) in this connection.) Section 4 describes the method we propose for multivariate imputation.

4 A method for multivariate imputation

4.1 Notation and description of the method

To simultaneously impute $\mathbf{Y_i} = (Y_{i1}, \ldots, Y_{iq})$ we use the univariate trees $\mathcal{Y}_X^{(j)}$ constructed for each of the variables Y_j, $j = 1, \ldots, q$, as described next and formalized in Algorithm 2. In the interest of brevity we address the problem of imputing the variables in Y; to impute Z we proceed likewise.

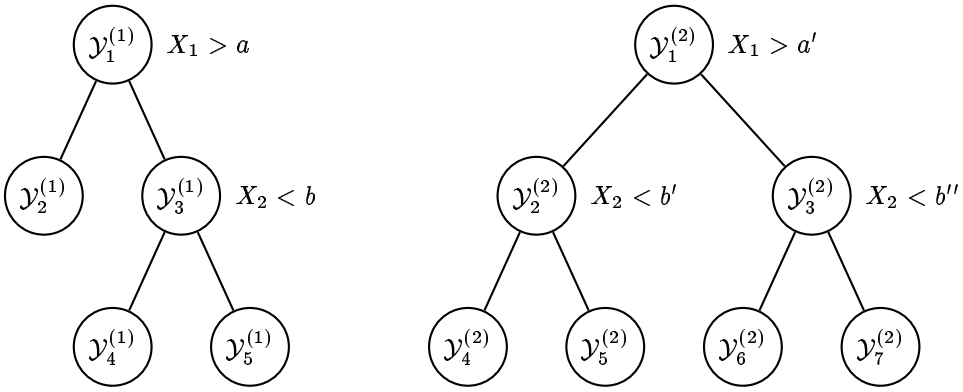
Let the nodes of each tree be numbered, and let $\mathcal{Y}_k^{(j)}$ be the k-th node of tree $\mathcal{Y}_X^{(j)}$. We use $\mathcal{Y}_k^{(j)}$ to denote the node, the subset of cases ending in, or going through, that node, and the corresponding region of $\mathcal{X}$. For instance, let $q = 2$ (i.e., there are two variables Y_1 e Y_2 in survey A) and let the trees $\mathcal{Y}_X^{(1)}$ and $\mathcal{Y}_X^{(2)}$ have the simple form depicted in Figure 2. Then, all cases in the training sample with $X_1 \leq a$ will end up in node $\mathcal{Y}_2^{(1)}$ when dropped down the tree constructed for variable Y_1; the corresponding region $\mathcal{Y}_2^{(1)}$ of $\mathcal{X}$ is shown in Figure 3.

For each q-tuple $(\alpha_1, \ldots, \alpha_q)$ such that α_j ($j \in \{1, \ldots, q\}$) is the label of a node in tree $\mathcal{Y}_X^{(j)}$, we define

$$\mathcal{C}_{\alpha_1,\ldots,\alpha_q} \quad = \quad \mathcal{Y}_{\alpha_1}^{(1)} \cap \mathcal{Y}_{\alpha_2}^{(2)} \cap \ldots \cap \mathcal{Y}_{\alpha_q}^{(q)}. \tag{1}$$

Finally, let node $\mathcal{Y}_{(\uparrow\alpha_k)}^{(j)}$ be the "father" of node $\mathcal{Y}_{\alpha_k}^{(j)}$ in tree $\mathcal{Y}_X^{(j)}$. When there is no

Figure 2: Trees $\mathcal{Y}_X^{(1)}$ and $\mathcal{Y}_X^{(2)}$. Next to each non terminal node is the condition whose fulfillment sends a case through the right son.



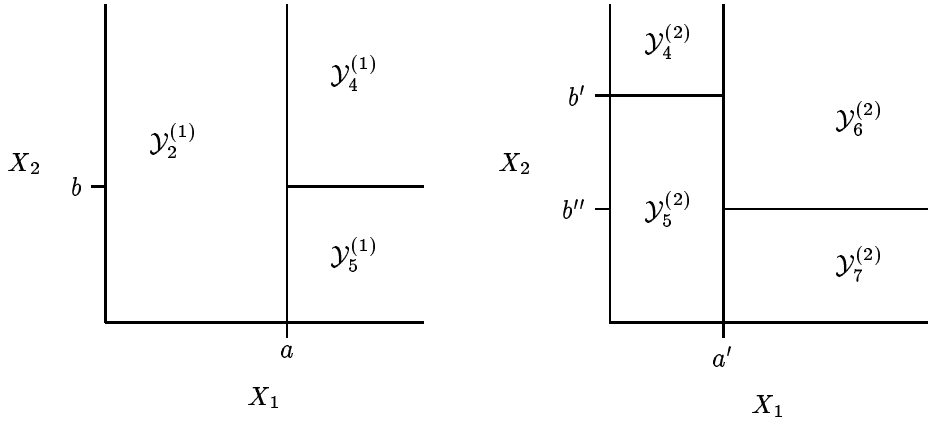
ambiguity about the tree referred to, we will simply refer to nodes $(\uparrow \alpha_k)$ and α_k.

Consider now case i, ($i \in \{N_A + 1, \ldots, N\}$), for which an imputation of $\mathbf{Y_i}$ is sought. Assume that when dropping that case through the trees built for each of the variables in Y, it ends in the leaves $\mathcal{Y}_{i_1}^{(1)}, \ldots, \mathcal{Y}_{i_q}^{(q)}$ and hence belongs to $\mathcal{C}_{i_1,\ldots,i_q}$. The simple idea in our method is to impute $\mathbf{Y_i}$ as a function of the values $\mathbf{Y}$ from cases in the training sample (file A) which also belong to $\mathcal{C}_{i_1,\ldots,i_q}$. Those cases have values for each variable $Y_1, \ldots, Y_q$ which, as far as the relevant trees can ascertain, are indistinguishable from the ones of the case to impute.

In the previous example, consider a case to impute i such that $a' < X_1 < a$ and $X_2 < b''$; it will end in leaves $\mathcal{Y}_2^{(1)}$ and $\mathcal{Y}_7^{(2)}$ when dropped down the trees $\mathcal{Y}_X^{(1)}$ and $\mathcal{Y}_X^{(2)}$. The intersection of those leaves,

$$\mathcal{C}_{2,7} = \mathcal{Y}_2^{(1)} \cap \mathcal{Y}_7^{(2)}, \tag{2}$$

Figure 3: Partitions of the $\mathcal{X}$ space induced by trees $\mathcal{Y}_X^{(1)}$ and $\mathcal{Y}_X^{(2)}$.



is shown in Figure 4. We propose to impute $\mathbf{Y_i}$ using the values of $\mathbf{Y}$ observed for cases in the training sample that also fall in $\mathcal{C}_{2,7}$.

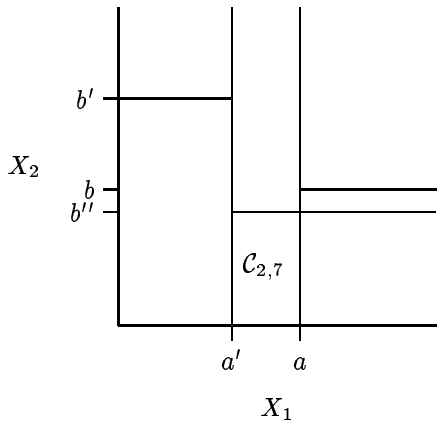
As in the univariate Algorithm 1, a variety of options exist: to impute using a $\mathbf{Y}$ sampled randomly from $\mathcal{C}_{i_1,\ldots,i_q}$, using the mean, the median, or any other suitable function.

4.2 Details of the implementation

If there is a large number of variables, substantial savings can be realized by performing first a principal component analysis, and building trees only for a reduced number of sufficiently descriptive components.

A problem may arise if case i to be imputed belongs to an intersection $\mathcal{C}_{i_1,\ldots,i_q}$ which is empty; no cases in the training sample belong to that particular intersection. When this happens, the intersection needs to be gradually enlarged to a non

Figure 4: Overlay of partitions of $\mathcal{X}$ induced respectively by trees $\mathcal{Y}_X^{(1)}$ and $\mathcal{Y}_X^{(2)}$, and intersection $\mathcal{C}_{2,7}$.



empty set: starting from the leaves $\mathcal{Y}_{i_1}^{(1)}, \ldots, \mathcal{Y}_{i_q}^{(q)}$ where i ended, our algorithm "climbs" the trees, replacing one node at a time by its "father". In doing so, we have at each step a choice of q trees that we may climb —hence the "forest climbing" name. The goal is to choose at each step in such a way that the quality of the imputation suffers least.

Let us see the heuristics implemented in our algorithm, which is one possible way of doing it. We refer in the following to imputation of variables Y (the Z's are treated likewise). Continuous variables Y are assumed, but the idea can be generalized.

In the construction of trees, nodes are divided for as long this improves the fit in terms of *deviance* —for regression trees, usually the sum of squares is used; see Breiman et al. (1984), Cap. 3—. Let $R(t)$ be the deviance at node t and $R(T)$ the

total deviance of tree T, defined as

$$R(t) = \sum_{i \in t} (y_i - \overline{y}_t)^2 \tag{3}$$

$$R(T) = \sum_{t \in \tilde{T}} R(t), \tag{4}$$

where $\tilde{T}$ is the set of "leaves" or terminal nodes of tree T and $\overline{y}_t$ is the arithmetic mean of values of the response variable for the cases in node t.

For any of the trees $\mathcal{Y}_X^{(j)}$, $j = 1, \ldots, q$, the cost of climbing from node t_h to its father node t_p can be evaluated by

$$c^{(j)}(t_h) = \frac{\sum_{i=1}^{N_p} (y_{ij} - \overline{y}_{j,t_p})^2}{N_p} - \frac{\sum_{i=1}^{N_h} (y_{ij} - \overline{y}_{j,t_h})^2}{N_h} \tag{5}$$

$$= \hat{R}(t_p)/N_p - \hat{R}(t_h)/N_h \qquad \forall j = 1, .., q, \tag{6}$$

where $\hat{R}(t_h)$ and $\hat{R}(t_p)$ are resubstitution estimates of the deviance in node t_h (from which we consider climbing) and its father node t_p, N_p and N_h are the number of cases in nodes t_p and t_h respectively, and $\overline{y}_{j,t_p}$, $\overline{y}_{j,t_h}$ the means of variable Y_j for said nodes.

With the previous notation, we can specify Algorithm 2. A few comments are worth making. First, Algorithm 2 can be applied both to the original variables or to any suitable transformations. The motivation behind using principal components is to reduce the number of trees to construct: this speeds up the process of finding an intersection. But both the first and last step in the algorithm are optional. Principal components have been used in the illustration of Section 5.

Second, whether we use the original variables, principal components or any other suitable transformations, the criterion to choose which tree must be climbed

Algorithm 2 – Multivariate imputation using trees.

(optionally) Compute the principal components of the variables Y to impute using the training sample.

Construct trees $\mathcal{Y}_X^{(1)}, \ldots, \mathcal{Y}_X^{(q)}$.

for $i \in$ {Cases to impute} **do**

  Drop case i down the trees, and determine the intersection $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$ of the leaves $\mathcal{Y}_{\alpha_1}^{(1)}, \ldots, \mathcal{Y}_{\alpha_q}^{(q)}$ where it falls.

  if $\mathcal{C}_{\alpha_1,\ldots,\alpha_q} \neq \emptyset$ **then**

    break

  else

    while $\mathcal{C}_{\alpha_1,\ldots,\alpha_q} = \emptyset$ **do**

      Compute the costs $c^{(1)}(\alpha_1), \ldots, c^{(q)}(\alpha_q)$ of climbing from the current nodes.

      Select k such that climbing from node α_k is of minimal cost.

      $\alpha_k \leftarrow \quad (\uparrow \alpha_k)$; replace node α_k by its father.

    end while

  end if

  Impute i from $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$.

end for

(if required) Reconstruct the original variables from the imputed principal components

is scale dependent. Therefore, we may want to scale the variables to have common variance, or variances which reflect their importance.

5 Imputation in the Survey of Uses of Time (EPT-93)

In the following, we illustrate the concepts above linking the files `trabajo.dat` and `fiesta.dat` referred to in Section 1.2. Linking those two files can be seen as a particular case of the problem referred to in subsection 1.1, with p=5, $q = r$ =24, $N_A = 2521$ y $N_B = 2519$. Each file contains,

- Common or characterization variables X, the description of which is given in Table 1.
- Specific variables: daily time in minutes allocated to activities in Table 2.

The two files have been linked using Algorithm 2. The computation has been performed with functions programmed using the S-PLUS language and primitives: see Becker et al. (1988) and Chambers and Hastie (1992) for a description of that package.

We chose to transform the variables to principal components and grow trees on these. As a descriptive measure of the goodness of fit we computed the relative mean quadratic error of each tree.

The relative mean quadratic error $RE(T^*)$ was computed by cross validation (see Breiman et al. (1984) for details), dividing the sum of squares of deviations with respect to the mean of each terminal node by the total "sum of squares" of the

Table 1: Common or characterization variables X, along with their categories.

Variable	**Description**	**Code**	**Categories**
X_1	Age	EDA1	Up to 34 years.
		EDA2	Between 35 and 59 years.
		EDA3	60 years and more.
X_2	Sex	VARO	Male
		MUJE	Female
X_3	Marital status	SOLT	Single.
		CASD	Married.
		REST	Other.
X_4	Education level	PRIM	Primary school.
		MEDI	Secondary school.
		SUPE	University.
X_5	Activity	SRMI	Military service.
		OCUP	Working.
		PARA	Unemployed.
		JUBI	Retired.
		ESTD	Studying.
		LAHO	Household chores.
		OTRS	Others.

principal component, i.e., n times the associated eigenvalue λ:

$$RE(T^*) = \frac{R(T^*)}{n\lambda}. \tag{7}$$

Thus, $RE(T^*)$ is very much alike $1 - R^2$ in ordinary regression. The results for the Y variables (similar results were obtained for the Z's) are shown in Table 3.

We may notice that the $RE(T^*)$ are in general quite high, as could be expected from the nature of the variables analysed. On the other hand, the better fits are obtained for the first principal components, also as expected. For two principal components (15 and 24) the CART methodology produced no subdivisions, and hence the corresponding trees were dropped from the analysis.

In this application we opted for single imputation using one case taken at random from $\mathcal{C}_{\alpha_1,\ldots\alpha_q}$, the first non empty intersection. We kept track of how many times climbing was necessary, because the intersection of leaves reached in the first instance was empty. Only 33 cases (1.31% of the total) required climbing when imputing the variables Y, and 48 cases (1.91% of the total) when imputing the Z's. Climbing was mainly up the trees built for the last principal components —the less costly. In a few cases, climbing all the way up to the root of one tree was required.

It is usual practice to compare the distributions of the imputed and observed variables (see Lebart and Lejeune (1995)). If all assumptions that make imputation feasible and worthwhile hold, they should not differ by much. Usually, univariate measures of location and scale are compared. In our case, results of the imputation

Table 2: Specific variables measuring uses of time. Variables $Y_1, \ldots, Y_{24}$ correspond to time use in working days and $Z_1, \ldots, Z_{24}$ are the homologous variables for holidays.

Variables	**Description**
Y_1, Z_1	Sleeping.
Y_2, Z_2	Toilet and personal care.
Y_3, Z_3	Meals.
Y_4, Z_4	Private or undescribed activities.
Y_5, Z_5	Work.
Y_6, Z_6	Learning and education.
Y_7, Z_7	Household chores.
Y_8, Z_8	Shopping.
Y_9, Z_9	Bureaucratic or administrative steps.
Y_{10}, Z_{10}	Partly leisure activities (sewing, painting, sculpture, fixing things in the household, gardening, pet care...)
Y_{11}, Z_{11}	Adults and children care.
Y_{12}, Z_{12}	Family and social meetings (meals, funerals, weddings, hospital visits,...)
Y_{13}, Z_{13}	Time spent with friends, drinking, talking,
Y_{14}, Z_{14}	Involvement in political or religious activities.
Y_{15}, Z_{15}	Sport and gymnastics.
Y_{16}, Z_{16}	Promenades and outings.
Y_{17}, Z_{17}	Leisure at home (watching TV, listening to music,..)
Y_{18}, Z_{18}	Leisure out of home (movies, theater, concerts, museums and exhibitions, attendance to sportive events,...)
Y_{19}, Z_{19}	Other leisure activities (micro computers, photography, playing cards or other games, solving crosswords,...)
Y_{20}, Z_{20}	Commuting to the place of work or education.
Y_{21}, Z_{21}	Spending time with others.
Y_{22}, Z_{22}	Waiting time either at work or education.
Y_{23}, Z_{23}	Waiting for medical or administrative attention.
Y_{24}, Z_{24}	Other waiting times.

Table 3: Relative mean quadratic error $RE(T^*)$ for the trees grown on the principal components of the Y's.

Principal component							
1	2	3	4	5	6	7	8
0.346	0.695	0.597	0.993	0.802	0.985	0.974	1.000

Principal component							
9	10	11	12	13	14	15	16
0.929	1.000	0.985	0.991	0.982	1.000	NA	0.957

Principal component							
17	18	19	20	21	22	23	24
0.998	0.946	1.000	0.978	0.972	0.978	0.971	NA

Table 4: Imputation of uses of time in working days. Location and scale statistics for observed data and imputed data of the variables Y. Starred variables are those for which the mean of the imputed values differed from the mean of observed values by more than two standard deviations.

	Observed values						Imputed values					
	Min.	q_1	Me	$\overline{x}$	q_3	Max.	Min.	q_1	Me	$\overline{y}$	q_3	Max.
Y_1	179	450	499	511.30	559	1214	179	450	495	508.50	555	975
Y_2	0	20	35	42.23	55	295	0	20	35	41.90	55	295
Y_3	0	75	105	109.90	135	374	0	75	100	109.90	135	374
Y_4	0	0	0	0.33	0	105	0	0	0	0.25	0	75
Y_5	0	0	0	196.50	455	990	0	0	0	194.70	460	929
Y_6^*	0	0	0	23.84	0	760	0	0	0	56.25	0	760
Y_7^*	0	0	75	133.10	235	770	0	0	45	107.60	185	630
Y_8^*	0	0	0	30.22	55	355	0	0	0	25.34	45	330
Y_9	0	0	0	2.73	0	415	0	0	0	3.05	0	240
Y_{10}^*	0	0	0	16.90	0	634	0	0	0	13.81	0	634
Y_{11}^*	0	0	0	20.53	0	735	0	0	0	15.71	0	625
Y_{12}^*	0	0	0	7.32	0	630	0	0	0	8.90	0	630
Y_{13}^*	0	0	0	40.43	60	510	0	0	5	47.35	75	510
Y_{14}	0	0	0	4.64	0	365	0	0	0	5.12	0	350
Y_{15}^*	0	0	0	6.91	0	315	0	0	0	8.39	0	230
Y_{16}	0	0	0	54.83	90	600	0	0	0	53.92	90	600
Y_{17}^*	0	74	140	163.00	225	964	0	65	135	157.70	224	730
Y_{18}^*	0	0	0	1.48	0	239	0	0	0	2.09	0	239
Y_{19}	0	0	0	10.76	0	405	0	0	0	9.59	0	350
Y_{20}	0	0	0	20.48	30	340	0	0	0	21.14	30	340
Y_{21}^*	0	0	10	33.32	50	355	0	0	15	39.25	60	355
Y_{22}	0	0	0	1.49	0	165	0	0	0	1.81	0	165
Y_{23}	0	0	0	0.50	0	165	0	0	0	0.56	0	165
Y_{24}	0	0	0	0.79	0	75	0	0	0	0.94	0	75

Note: q_1, q_3 are the first and third quartiles; Me is the median.

for the variables Y (similar results are available for the Z's) are displayed in Table 4. The imputed values for Y reproduce reasonably well the distribution of the observed values, the only exceptions that catch the eye being Y_7 ("Household chores") and Y_{13} ("Meetings with friends").

Ideally, the full, multivariate distribution of imputed and observed values should be compared. It is not possible to go that far, but one step in that direction is to

compare the correlation matrices of both the imputed and observed values. This has been done for the variables Y (again, similar results are available for the Z's). The largest discrepancy among correlations is 0.1124, and occurs for the correlation between Y_5 and Y_6. Rather than embark in a tedious comparison one by one, we can see the likeness of both correlation matrices by looking at Figure 5, in which correlations of both observed and imputed values are displayed with different shades.

6 Simulations

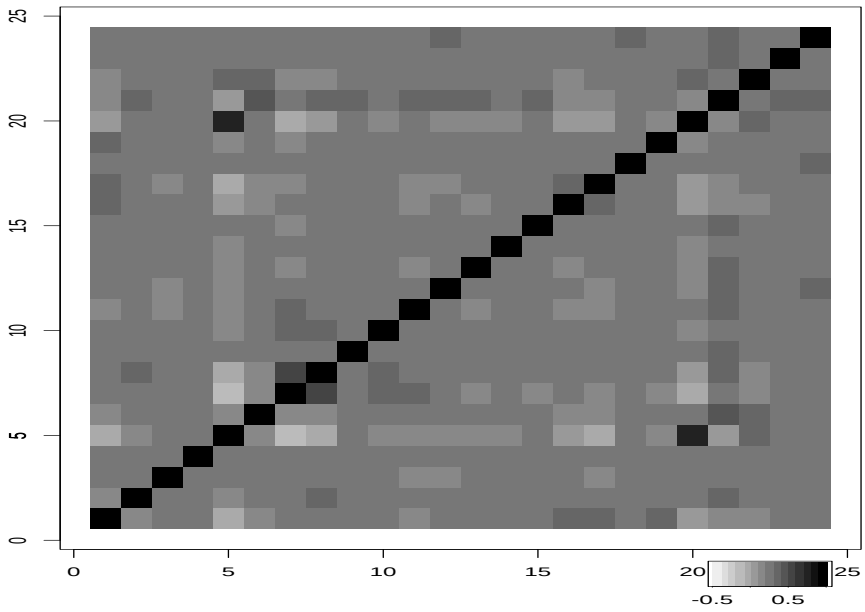
We tried to gain some insight on the performance of the method carrying out some simulations. We report below the results obtained in three simple situations, which give a flavour of the rest.

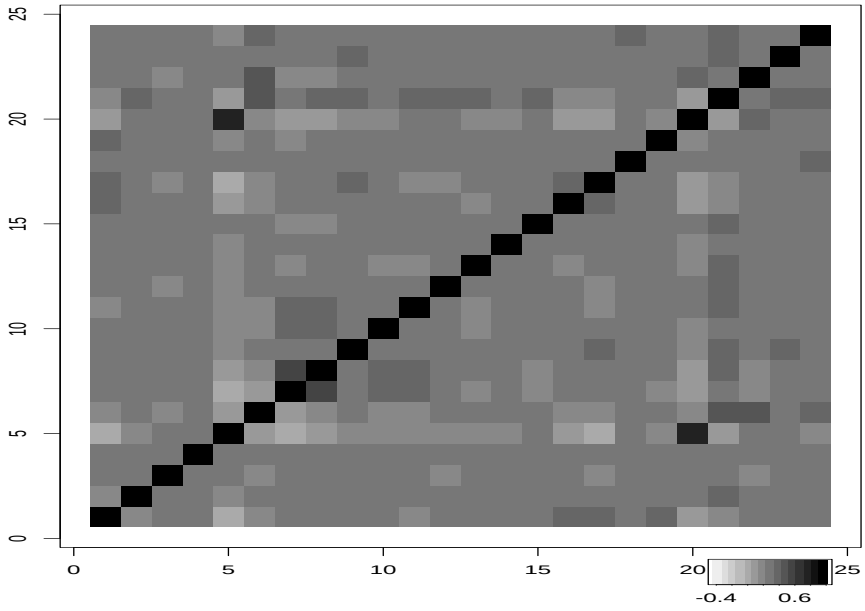
6.1 Setup common to all problems

Since the application that motivated this research required the merging of two files, the program was written with a data layout such as Figure 1 in mind. Therefore, each table shown below gives results on the imputation of variables Y and Z, which could be reported independently.

A few elements are common to all simulations reported. The number of common variables X was set at $p = 5$. We considered the univariate case ($q = r = 1$), and set the size of each file to merge at $N_A = N_B = 1000$. Each simulation run consisted of one thousand replications of the same problem, recreating the sample,

Figure 5: Pictorial representation of the correlation matrices of observed (above) and imputed (below) Y's. Notice the unconventional position of the main diagonal, and the similarity of both matrices.





growing the trees and performing the merging of the two files each time, using Algorithm 1.

The specific variables Y and Z were made to depend on some of the X's in various ways, which included situations in which a tree may find hard the detection of structure. The specific variables contained also noise, which was taken to be Gaussian with different variances. To be precise, we simulated

$$Y = 10 + (X_1 - 4)(X_4 - 2) + \epsilon \quad (8)$$

$$Z = 16 - X_1^2 + X_4^2 + \nu \quad (9)$$

with ϵ and ν zero mean Gaussian variables. The part of Y which can be recovered as a function of the X's is shaped like a saddle, a difficult surface to approximate by a tree with splits based on thresholds of a single variable X at a time ("Is $X_i < a$?").

We generated complete data for both Y and Z, then deleted N_B and N_A values and performed the imputation. This allows us to compare the results that would have been obtained with complete data, and those obtained with partially imputed data. Of particular interest is the comparison among the estimated correlation using full data (r_{yz}) and partially imputed data ($\hat{r}_{yz}$).

6.2 Problem 1: discrete X, many levels

We considered a relatively large number of levels in the variables, requiring a potentially large number of terminal nodes. The X's were generated as follows:

$$X_i \quad = \quad \begin{cases} b(p = 0.5, n = 8) & \text{for } i = 1, 2, 3, \\ b(p = 0.5, n = 4) & \text{for } i = 4, 5. \end{cases} \tag{10}$$

Two variances for ϵ and ν in (8) and (9) have been used: 1 and 25.

There are two variables relevant in the approximation of both Y and Z, having 9 and 5 levels; an ideal tree would partition the $\mathcal{X}$ space in 45 cells corresponding to each possible combination of X_1 and X_4. In practice, trees with fewer nodes are chosen by the CART methodology. Table 5 displays the number of terminal nodes of the trees for Y and Z ($|\tilde{\mathcal{Y}}_X|$ and $|\tilde{\mathcal{Z}}_X|$, respectively), and the mean squared residual, computed by cross validation ($R^{cv}_{\mathcal{Y}}$ and $R^{cv}_{\mathcal{Z}}$) and using an independently generated test sample ($R^{ts}_{\mathcal{Y}}$ and $R^{ts}_{\mathcal{Z}}$). We also report the estimated correlations among Y and Z using complete data and partly imputed data.

The results are shown for the two different variances, $\sigma^2_\epsilon = \sigma^2_\nu = 1$ and $\sigma^2_\epsilon = \sigma^2_\nu = 25$, representative of low and moderate noise. For each statistic, the mean ($\overline{x}$), median (Me), first and third quartiles (q_1 and q_3) and extremes are reported. The mean squared residuals ($R_{\mathcal{Y}}$ or $R_{\mathcal{Z}}$, estimated either by cross validation or using a test sample) can be compared to the variance of ϵ and ν to appraise the goodness of fit.

It is interesting to note that as the variance of the noise increases from $\sigma^2_\epsilon = 1$ to $\sigma^2_\epsilon = 25$, the ability of the tree to pick up structure degrades dramatically: the

number of leaves of $\mathcal{Y}_X$ drops on the average from 20.8 to 10.3. The opposite seems to be true for the $\mathcal{Z}_X$ tree. While the fit of the $\mathcal{Y}_X$ is quite good, with the estimated mean squared residual ($R^{cv}_{\mathcal{Y}}$ and $R^{cv}_{\mathcal{Z}}$) only about 15% above σ^2_ϵ, this is not so for the Z_X tree. However, r_{yz} is still acceptably recovered by $\hat{r}_{yz}$ estimated from partially imputed data.

6.3 Problem 2: discrete X, few levels

Next we consider a relatively small number of levels in the predictors X. The X's were generated as follows:

$$X_i \quad = \quad \begin{cases} b(p = 0.5, n = 3) & \text{for } i = 1, 2, 3, \\ b(p = 0.5, n = 2) & \text{for } i = 4, 5. \end{cases} \tag{11}$$

The results are shown in Table 6. The ideal tree would require 12 nodes, since there are 12 different combinations of values of X_1 and X_4. The trees constructed are indeed smaller than in Problem 1 above. The same patterns recur, though, in that an increase in the variance of the noise brings about a decrease in the number of nodes for $\mathcal{Y}_X$, unlike in the case of $\mathcal{Z}_X$. The correlation between both variables is again acceptably recovered.

Table 5: Results in one thousand replications of Problem 1.

Statistic	**Min.**	$\mathbf{q_1}$	**Me**	$\overline{\mathbf{x}}$	$\mathbf{q_3}$	**Max.**
	$\sigma_\epsilon^2 = \sigma_\nu^2 = 1$					
$\lvert\tilde{\mathcal{Y}}_X\rvert$	16	19	20	20.8	22	35
$\lvert\tilde{\mathcal{Z}}_X\rvert$	10	11	11	11.38	12	13
$R_{\mathcal{Y}}^{cv}$	1.01	1.12	1.16	1.16	1.20	1.41
$R_{\mathcal{Y}}^{ts}$	0.954	1.10	1.14	1.15	1.19	1.48
$R_{\mathcal{Z}}^{cv}$	11.05	13.00	13.55	13.56	14.08	15.84
$R_{\mathcal{Z}}^{ts}$	10.25	12.82	13.43	13.52	14.15	22.56
r_{yz}	-0.4400	-0.3821	-0.3652	-0.3652	-0.3483	-0.2930
$\hat{r}_{yz}$	-0.4906	-0.4320	-0.4129	-0.4128	-0.3938	-0.3221
	$\sigma_\epsilon^2 = \sigma_\nu^2 = 25$					
$\lvert\tilde{\mathcal{Y}}_X\rvert$	3.0	7.0	11.0	10.3	13.0	28.0
$\lvert\tilde{\mathcal{Z}}_X\rvert$	14.00	17.00	18.00	18.22	19.00	22.00
$R_{\mathcal{Y}}^{cv}$	23.29	25.38	26.26	26.33	27.19	30.42
$R_{\mathcal{Y}}^{ts}$	21.76	25.47	26.24	26.30	27.08	31.32
$R_{\mathcal{Z}}^{cv}$	26.96	30.34	31.37	31.40	32.47	37.67
$R_{\mathcal{Z}}^{ts}$	26.24	29.90	31.08	31.21	32.54	38.89
r_{yz}	-0.22740	-0.17270	-0.15710	-0.15650	-0.14030	-0.07153
$\hat{r}_{yz}$	-0.34330	-0.23680	-0.20880	-0.20910	-0.18190	-0.06886

Table 6: Results in one thousand replications of Problem 2.

Statistic	**Min.**	$\mathbf{q_1}$	**Me**	$\overline{\mathbf{x}}$	$\mathbf{q_3}$	**Max.**
	$\sigma_\epsilon^2 = \sigma_\nu^2 = 1$					
$\lvert\tilde{\mathcal{Y}}_X\rvert$	7	9	10	10.5	11	20
$\lvert\tilde{\mathcal{Z}}_X\rvert$	9	11	11	11	11	12
$R_{\mathcal{Y}}^{cv}$	0.8742	0.9878	1.0190	1.0190	1.0510	1.1570
$R_{\mathcal{Y}}^{ts}$	0.8727	0.9901	1.0190	1.0210	1.0510	1.2750
$R_{\mathcal{Z}}^{cv}$	0.9451	1.1540	1.2130	1.2180	1.2870	1.5370
$R_{\mathcal{Z}}^{ts}$	0.8622	1.1220	1.2030	1.2100	1.2840	1.8300
r_{yz}	-0.4212	-0.3680	-0.3545	-0.3545	-0.3420	-0.2888
$\hat{r}_{yz}$	-0.4807	-0.4160	-0.4008	-0.4000	-0.3832	-0.3115
	$\sigma_\epsilon^2 = \sigma_\nu^2 = 25$					
$\lvert\tilde{\mathcal{Y}}_X\rvert$	2.00	4.00	5.00	5.42	7.00	20.00
$\lvert\tilde{\mathcal{Z}}_X\rvert$	7.00	11.00	13.00	12.96	14.00	31.00
$R_{\mathcal{Y}}^{cv}$	21.84	24.69	25.41	25.45	26.23	28.85
$R_{\mathcal{Y}}^{ts}$	21.96	24.79	25.48	25.54	26.28	31.34
$R_{\mathcal{Z}}^{cv}$	21.39	24.96	25.76	25.76	26.54	30.05
$R_{\mathcal{Z}}^{ts}$	21.45	24.95	25.72	25.77	26.58	30.20
r_{yz}	-0.14270	-0.09637	-0.08099	-0.08046	-0.06597	-0.01067
$\hat{r}_{yz}$	-0.23150	-0.14810	-0.12400	-0.12380	-0.09825	-0.00281

6.4 Problem 3: mixed X, correlated predictors

Next we consider predictors X either discrete with a moderate number of levels (nine) or continuous and correlated, generated as follows:

$$X_i = \begin{cases} b(p = 0.5, n = 8) & \text{for } i = 1, 2, 3, \\ \begin{pmatrix} X_4 \\ \\ X_5 \end{pmatrix} \sim N(\vec{0}, \Sigma) & \text{with } \Sigma = \begin{pmatrix} 30 & 20 \\ & \\ & 40 \end{pmatrix} \end{cases} \tag{12}$$

This is a situation particularly difficult to handle by a tree, and indeed the results in Table 7 show substantially degraded performance. The mean squared error nowhere approaches the variance of the noise; in the case of the $\mathcal{Z}_X$ tree is much larger. Interestingly enough, r_{xy} is again recovered quite acceptably, which illustrates the fact that if Y and Z depend on the X's and have no partial correlation between them, the correlation can be well approximated even if the variables are not.

7 Summary and conclusions

A new method for imputation has been presented and its feasibility demonstrated on a real data set. It can cope with a large variety of problems, because of the generality of the tool used for approximation —classification or regression trees. It makes few assumptions, is computationally feasible, and appears to give good results. Simulations seem to confirm this; the method works well whenever the common variables X are good predictors for the Y's and Z's and the functional relationship among predictors and responses can be reasonably well approximated

Table 7: Results in one thousand replications of Problem 3.

Statistic	**Min.**	$\mathbf{q_1}$	**Me**	$\overline{\mathbf{x}}$	$\mathbf{q_3}$	**Max.**
	$\sigma_\epsilon^2 = \sigma_\nu^2 = 1$					
$\lvert\tilde{\mathcal{Y}}_X\rvert$	16.00	19.00	20.00	19.77	21.00	25.00
$\lvert\tilde{\mathcal{Z}}_X\rvert$	13.00	16.00	17.00	16.82	18.00	21.00
$R_{\mathcal{Y}}^{cv}$	2.869	3.889	4.385	4.588	5.043	10.340
$R_{\mathcal{Y}}^{ts}$	2.688	3.806	4.271	4.439	4.849	17.880
$R_{\mathcal{Z}}^{cv}$	31.46	60.48	72.57	78.57	89.37	322.70
$R_{\mathcal{Z}}^{ts}$	29.36	55.20	67.90	73.73	86.12	275.60
r_{yz}	0.1567	0.2947	0.3303	0.3304	0.3698	0.4709
$\hat{r}_{yz}$	0.08066	0.30430	0.35040	0.34820	0.39660	0.53280
	$\sigma_\epsilon^2 = \sigma_\nu^2 = 25$					
$\lvert\tilde{\mathcal{Y}}_X\rvert$	11.00	18.00	21.00	21.36	24.00	39.00
$\lvert\tilde{\mathcal{Z}}_X\rvert$	15.00	21.00	22.00	22.29	24.00	30.00
$R_{\mathcal{Y}}^{cv}$	25.48	28.83	29.97	30.10	31.21	39.59
$R_{\mathcal{Y}}^{ts}$	25.06	28.64	29.68	29.87	30.95	43.92
$R_{\mathcal{Z}}^{cv}$	49.06	77.45	89.22	95.56	106.40	335.20
$R_{\mathcal{Z}}^{ts}$	47.51	71.57	85.14	90.57	104.10	294.10
r_{yz}	0.1130	0.2122	0.2382	0.2384	0.2671	0.3668
$\hat{r}_{yz}$	0.02346	0.23830	0.28010	0.27580	0.31520	0.45500

by a tree.

We see room for improvement, specially in the climbing strategy, at the expense of increased complexity and computational burden. Our work proceeds along this line. Further work is also required in comparing our method to other flexible, all-purpose methods of imputation, like those using neural networks.

References

Aluja, T., Nonell, R., Rius, R., and Martínez, M. (1995). File grafting. In F. Mola and A. Morineau, editors, *Actes du IIIème Congrès International d'Analyses Multidimensionnelles des Données - NGUS'95*, pp. 23–32, Centre Int. de Statistique et d'Informatique Appliquées, CISIA-CERESTA.

Aluja, T. and Rius, R. (1994). Inserción de datos de encuesta mediante análisis de componentes principales. *Presented at the XXI Congreso nacional de Estadística e Investigación Operativa (SEIO). Calella.*

Becker, R., Chambers, J., and Wilks, A. (1988). *The New S Language. A Programming Environment for Data Analysis and Graphics.* Pacific Grove, California: Wadsworth & Brooks/Cole.

Bello, A. (1993). Choosing among imputation techniques for incomplete multivariate data: a simulation study. *Communications in Statistics - Theory and Methods*, 22, 853–877.

Bishop, C. (1996). *Neural Networks for Pattern Recognition.* Oxford: Clarendon Press.

Bárcena, M. and Tusell, F. (1998). Enlace de encuestas: una propuesta metodológica y aplicación a la Encuesta de Presupuestos de Tiempo. Technical Report 98.07, Departamento de Econometría y Estadística.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth.

Buck, S. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Ser. B*, 22, 302–306.

Chambers, J. and Hastie, T. (1992). *Statistical Models in S*. Pacific Grove, Ca.: Wadsworth & Brooks/Cole.

Ciampi, A. (1991). Generalized regression trees. *Computational Statistics and Data Analysis*, 12, 57–78.

Dempster, A., Laird, N., and Rubin, D. (1976). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, 39, 1–38.

EUSTAT (1997). *Análisis de Tipologías de Jornadas Laborales*. Vitoria/Gazteiz: Instituto Vasco de Estadística, (EUSTAT).

Lebart, L. and Lejeune, M. (1995). Assessment of Data Fusions and Injections. In *Encuentro Internacional AIMC sobre Investigación de Medios*, pp. 1–18, Madrid.

Lejeune, M. (1995). De l'usage des fusions de données dans les études de marché. In *Proceedings of the IASS Meeting, Beijing.*

Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley and Sons.

Nordbotten, S. (1996). Neural Network Imputation Applied to the Norwegian 1990 Population Census Data. *Journal of Official Statistics*, 12, 385–401.

Nordholt, E. (1998). Imputation: Methods, Simulation Experiments and Practical Examples. *International Statistical Review*, 66, 157–180.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rubin, D. (1986). Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations. *Journal of Business and Economic Statistics*, 4, 87–94.

Rubin, D. (1991). EM and beyond. *Psychometrika*, 56, 241–254.

Villagarcía, T. and Muñoz, A. (1997). Imputación de datos censurados mediante redes neuronales: una aplicación a la EPA. *Cuadernos Económicos de I.C.E.*, pp. 193–204.