Quaderni di Statistica
Vol. 10, 2008

A double imputation method for Data Fusion

Alfonso Piscitelli
Dipartimento di Sociologia, Universitá di Napoli Federico II
E-mail: alfonso.piscitelli@unina.it

Summary: Data fusion consists of merging information coming from two different surveys. The first one is called reference or donor survey while the second is called punctual or receptor survey. Such two independent surveys have a block of common variables that is used as a bridge between them. The aim is to complete the receptor survey exploiting information acquired from the donor one, and file grafting is commonly used for this aim. As file grafting is based on Principal Component Analysis, it does not consider possible dependency structure among the variables. In this work we present a new methodology for data fusion based on the *Constrained Principal Component Analysis* (*CPCA*) technique. The proposal allows to impute the missing information into the second survey taking into account knowledge about the relationship structure among variables.

Keywords: File grafting, Missing values imputation, Non-symmetrical exploratory data analysis.

1. Introduction

In recent years, there is a growing interest in methodologies aiming at combining different sources of information, usually from several surveys. Parallel questionnaires, panel survey, tentatives of enriching basic surveys through specific questionnaires (Santini, 2001) may often require such techniques usually named as data fusion (Aluja-Banet *et al.* 2007).

Data fusion, also known as statistical matching or file grafting, involves the imputation of a complete block of missing variables in independent data sets. It consists of matching two already held surveys in order to make it possible to transfer part of the information contained in one survey to a second one. The first survey is called reference survey (donor matrix); the second is called punctual survey (receptor matrix). Data fusion allows us to treat data coming from the two distinct surveys as a whole. These methods found some applications in media studies (Rius *et al.*, 1999; Lejeune, 2001; Aluja-Banet and Thió, 2001), in web data analysis, and also in national statistical institutes (D'Orazio *et al.*, 2006).

With the aim of determining the complete block of unobserved values of a set of variables included in a first survey but not in a second, data fusion can be approached by means of missing data imputation techniques. Missing data of the receptor matrix will be imputed by exploiting information coming from the donor matrix. To perform such an imputation a set of variables in common to both surveys is required.

Different methodologies have been proposed in literature for data fusion (see e.g. Little and Rubin, 1987; Schulte Nordholt, 1998; Saporta, 2002), and they can be classified in two families. A first group, *explicit model-based estimation methods*, relies on finding a *model* for the variables to be imputed in the donor survey and on applying it for the receptor survey. Explicit models usually exploit regression models and yield good imputations. However, they underestimate the variance of the imputed variables and their correlation coefficients (Shao and Wang, 2002).

The second group includes the so-called *implicit models for imputation*. In such a case, for each statistical unit of the receptor survey, one or more donor units are selected. The values of the donor units are then imputed to the receivers. Among the implicit methods, file grafting, based on Principal Component Analysis (*PCA*), is one of the most largely used. This method aims at defining a common subspace onto which to project the statistical units coming from the two surveys. Such subspace is constructed through a *PCA* performed on the common variables of the reference survey. It is well-known that the *PCA* analyzes the correlation structure, and, in this sense, all the variables play a symmetric role, assuming an interdependence structure among them. However, in sociological and economic theories some relationships are given and well-known, and hence some *a priori* knowledge on dependency structure among the $\mathbf{X}$ and $\mathbf{Y}$ variables is available.

In order to avoid the above-mentioned problems, we propose a file grafting technique that combines the explicit and implicit approaches, and we call it *Non Symmetrical Grafting* (*NSG*). The proposed method exploits the non symmetrical *PCA* to explore the dependency structure of the data, i.e. the *Constrained Principal Component Analysis* (*CPCA*) technique (D'Ambra and Lauro, 1982). The *NSG* algorithm projects individuals belonging to different surveys onto the same subspace, determined through the non symmetrical *PCA*. This projection is made by a linear multiple regression. In such a space, distances among individuals belonging to the different surveys are evaluated, and for each statistical unit of the receiver survey, the "missing values" are imputed using the nearest neighbor donors.

The paper is organized as follows. In Section 2 we present the main ideas about data fusion and models for imputation; in Section 3 the file grafting is described. In Section 4 the proposed Non Symmetrical Grafting procedure along with details concerning the imputation methods. In Section 5 some issues related to validation are discussed. In Section 6 we discuss the main results of a simulation study, and some final remarks conclude the paper.

2. Data fusion and imputation models

Data fusion is generally aimed at combining data coming from several surveys. In this paper, we consider its simplest case, called unilateral fusion, in which there are only two data sets: the donor one that is complete, and the other one with a block of missed variables (recipient data set).

More precisely, the donor survey contains information about a set of $p+k = q$ variables observed on n_0 subjects; the recipient survey, contains information about a set of $p + j = z$ variables observed on n_1 subjects. In both surveys a set of p variables $\mathbf{X}$ is in common. We denote with $\mathbf{X}_0$ the set of common variables referred to the donor survey, and with $\mathbf{X}_1$ the other one referred to the receiver survey. Analogously we denote with $\mathbf{Z}_1$ the j specific variables of the receptor survey, with $\mathbf{Y}_0$ the k specific variables of the donor survey and with $\mathbf{Y}_1$ the specific variables to be imputed. The aim is to fill the hyphened part of the second data matrix. We use the donor survey ($\mathbf{X}_0$;$\mathbf{Y}_0$) to impute the set of k variables not observed in the receiver survey (Fig. 1).

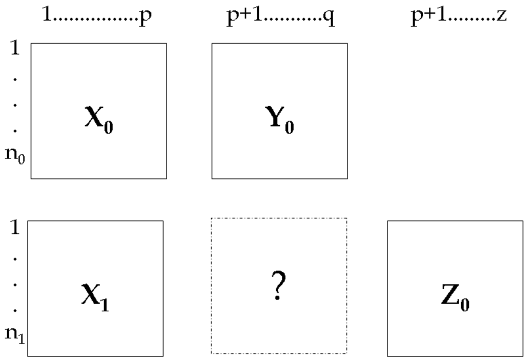

Figure 1. Blocks of shared and unshared information.

Data fusion can be considered as a particular kind of missing data imputation problem. In such a case, the missing values correspond to variables missing by design and a complete block of information should be imputed (Aluja-Banet *et al.* 2007). In this framework, different approaches can be adopted. In the class of explicit models, a very simple imputation procedure relies on a linear regression models of $\mathbf{Y}$ on $\mathbf{X}$, estimated on the available statistical units (namely, the donor data set), in order to impute the missing values through the predicted values $\widehat{\mathbf{Y}}$. For this purpose, the following conditions should be verified: $i)$ regression models should show a good fitting; $ii)$ the relationships among predictors $\mathbf{X}$ and response variables $\mathbf{Y}$ should be constant in both surveys; $iii)$ the partial correlation of $\mathbf{Y}$ given $\mathbf{X}$ and the correlations among the predictors should be equal to zero. More complex regression techniques could be applied for data fusion. Barcena and Tusell (1999) defined a data fusion procedure working with a multiple imputation via classification and regression trees named *forest climbing* algorithm.

However, such methods underestimate the true variance of the variables they are attempting to substitute. When regression models are used for the treatment of missing values, a loss of variability of the genuine values occurs (Little and Rubin, 1987). Indeed, one replaces unknown values scattered around the regression hyperplane through the fitted values lying on the hyperplane (Barcena and Tusell, 1999). Furthermore, with these methods, the correlation structure of the imputed variables is not well-reconstructed (Shao and Wang, 2002).

Another approach to data fusion relies on the use of the *EM algorithm*, that provides an iterative way to maximize the likelihood function of incomplete data (Dempster *et al.*, 1977). In such a case strong assumptions on the likelihood and on the generating mechanism of the data are required. However, data imputed via EM algorithm also suffer same lack of variability with respect to imputed values through regression models. Another method belonging to the explicit models is the *multiple imputation* based on the Bayesian framework (Rubin, 2003), that allows us to simulate the posterior distribution of the missing values by imputing each data with several values according to one or more estimation models. Even if multiple imputation techniques could achieve correct variances, they are really complex and time consuming (Saporta, 2002).

To overcome such problems, on the other hand, implicit model methods for missing data imputation have been developed in literature. A very simple method that does not require assumptions on variable distributions or on relationship structure between the specific and the common variables is the *hot deck imputation* (Ford, 1980). In such a case, the values of some statistical units of the complete survey (donors) are copied and pasted on other incomplete statistical units (receivers). According to some notion of similarity based on the common variables, the best donors are selected. Such methods are data-driven and distribution free; they avoid incoherent estimations since the copied values belong to real observations (Saporta, 2002). The combined use of hot deck imputation through the nearest neighbor principle and of factorial techniques as Multiple Correspondence Analysis or Principal Component Analysis is the base of a reference data fusion procedure known as *file grafting* process (Aluja-Banet *et al.*, 1995).

3. File grafting for data fusion

File grafting technique essentially consists of two steps (Rius *et al.*, 1996): pre-grafting and grafting. The former is aimed at studying the common variables and testing the common space stability in order to ensure the grafting feasibility. In this step, a subset of common variables defining a similar subspace of representation for both data sets is identified. Such variables represent the "bridge" to transfer information from one data set to the other (namely, projecting on it).

In the second step, if we consider the case of all continuous variables, the actual graft is performed through a singular value decomposition of $\mathbf{X}_0$, $\mathbf{X}_0 = \mathbf{V_0 \Lambda_0 U_0'}$. The statistical units are represented in the $\mathbf{U}_0$ basis with coordinates $\mathbf{\Psi_0} = \mathbf{X_0 U_0}$, and the elements of the second data set $\mathbf{X}_1$ are *grafted* in the same reference basis $\mathbf{U}_0$. That is, the individuals of $\mathbf{X}_1$ are projected as supplementary points with coordinates $\mathbf{\Psi_1} = \mathbf{X_1 U_0}$.

To perform file grafting the assumption of stability of the relationships among variables is required (Bonnefous *et al.*, 1986). This latter assumption allows us to define a common space on which to represent the whole information of both data sets.

Once all the individuals of the two surveys have been projected on the previously defined subspace, for each individual of the receiver matrix $\mathbf{X}_1$ a donor(s) having the closest profile with respect to the common variables is selected. The *nearest neighbors* to the *i-th* unit of the receptor survey are those individuals of the donor survey having the minimum distance in the common space. In the data fusion original proposal, the *nearest neighbor* (*nn*) algorithm has been applied (Baker *et al.*, 1989); a modified version (Aluja-Banet *et al.*, 2001) exploits and applies the *k-nearest neighbors* (*knn*) algorithm (Fukunaga and Narendra, 1975). Finally, missing data are imputed by hot deck imputation (Ford, 1980).

After the imputation it is necessary to measure the precision of the performed data fusion. One way consists of carrying out a self-imputation of $\mathbf{Y}_0$ variables upon the same individuals $\mathbf{X}_0$. In such a case the observed values can be compared with the imputed ones by the index R_y (Aluja-Banet *et al.*, 2001):

$$R_y = \frac{tr\left[(\mathbf{Y}_0 - \tilde{\mathbf{Y}}_0)'(\mathbf{Y}_0 - \tilde{\mathbf{Y}}_0)\right]}{tr\left[(\mathbf{Y}_0 - \bar{\mathbf{Y}}_0)'(\mathbf{Y}_0 - \bar{\mathbf{Y}}_0)\right]}. \qquad (1)$$

The R_y index is the ratio of the sum of squared errors in the case of file grafting imputation and sum of squared errors when one is imputing by the simple mean of the variable. When the *knn* algorithm is used in the fusion process, the R_y index could be exploited to determine the value of k. Evaluating R_y for the increasing k and plotting R_y as a function of k, the optimum k corresponds to the minimum value of R_y.

4. Non Symmetrical Grafting for data fusion

The descriptive factorial analysis commonly used for file grafting (e.g. PCA, MCA,) do not imply any *a priori* knowledge about the phenomenon under study. However, in many cases of sample survey data, *a priori* information about different roles of the variables may be available or known by the specific literature. In the same survey a dependence structure between two sets of variables often may be reasonably hypothesized (e.g. income and number of the family member affect the consumptions and savings). If a set of variables (dependent variables) depends on another (independent variables) we can use this information to improve the data fusion process. In order to build a common space on which projecting information from the two surveys, we propose the use of the *Constrained Principal Component Analysis* (*CPCA*) technique. *CPCA* consists of carrying out a PCA of the Y's image projected onto the common variables subspace through a suitable orthogonal projection operator.

Let $\mathbf{X}$ and $\mathbf{Y}$ be the two blocks of centered and scaled variables observed on the same n units which identify two sub-sets. The goal of $CPCA$ is the analysis of the relationship of the $\mathbf{Y}$ block with respect to the $\mathbf{X}$ block in terms of principal components associated with the latter block. Let $\Re^q$ be the $p+k$ dimensional vectorial space, and let $\Re^p$ be the vectorial sub-space of $\Re^q$ generated by the columns of $\mathbf{X}$, and consider the image of $\mathbf{Y}$ in the sub-space $\Re^p$:

$$\mathbf{Y}^* = \mathbf{P_X}\mathbf{Y}, \tag{2}$$

i.e $\mathbf{Y}^*$ is the projection of $\mathbf{Y}$ in $\Re^p$ through the orthogonal projection operator $\mathbf{P_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$. The *CPCA* analysis consists on a singular value decomposition of $\mathbf{Y}^*$,

$$\mathbf{Y}^* = \mathbf{V}^*\mathbf{\Lambda}^*\mathbf{U}'^*. \tag{3}$$

In such a case we represent the row elements in the $\mathbf{U}^*$ basis, with coordinate

$$\mathbf{\Psi}^* = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}\mathbf{U}^* = \mathbf{Y}^*\mathbf{U}^* \tag{4}$$

Note that, as $\mathbf{Y}^* = \mathbf{P_X}\mathbf{Y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y} = \mathbf{X}\hat{\beta}$ with the $\hat{\beta}$ the usual OLS estimate, the $CPCA$ is equivalent to the singular value of the predicted value of $\mathbf{Y}$ through the regressors $\mathbf{X}$.

Hence, exploiting the properties of $CPCA$ and the characteristics of file grafting procedure we propose a three-step procedure that we call Non Symmetrical Grafting. In the first step, as in the imputation by regression models, a subset of the $\mathbf{X}$ must be selected; in the second step the file grafting through $CPCA$ is performed, while in the third step the missing variables are imputed via hot deck imputation.

4.1. Building the basic matrix

In order to use $a\ priori$ information for grafting, we should identify the common variables influencing the specific variables to be imputed. In other words, a subset of the $\mathbf{X}$ variables on which we will perform the *CPCA* has to be selected. We propose to use the *backward elimination* criterion in regression analysis. Considering only the complete survey, we fit a regression model for each variable belonging to $\mathbf{Y_0}$ block on $\mathbf{X_0}$, and we select the predictors through the backward elimination. Then, those selected predictors, in common to both surveys, will be used to build the $\mathbf{X}$ matrix to be analyzed through *CPCA*. This subset of variables will be the basis of the common space onto which the incomplete survey will be grafted.

4.2. Graft in CPCA

Once the common space is built, in order to jointly represent the two data clouds the grafting process consists of projecting the whole information in such a common space. Generally, it is possible to perform the projection of additional individuals which are described by the matrix $[\mathbf{Y}_s|\mathbf{X}_s]$, starting from the singular value decomposition of the *CPCA*. The coordinates of the supplementary individuals will be:

$$\mathbf{\Psi}^*_\mathbf{s} = \mathbf{X_s}(\mathbf{X'_sX_s})^{-\mathbf{1}}\mathbf{X'_sY_sU^*} = \mathbf{Y^*_sU^*}. \tag{5}$$

However, in our case, the individuals to be supplementary projected are lines of the receptor matrix, and hence the $\mathbf{Y}_s$ values we need in (5) are missed. To overcome such a problem we propose to estimate them by means of a regression model for each variable, starting from the reference survey's data. The usual OLS estimate $\hat{\beta}_\mathbf{0} = (\mathbf{X'_0X_0})^{-\mathbf{1}}\mathbf{X'_0Y_0}$ is exploited to perform a first imputation of the specific variables $\hat{\mathbf{Y}}_1$ in the punctual survey, with

$$\hat{\mathbf{Y}}_1 = \mathbf{X}_1\hat{\beta}_\mathbf{0}. \tag{6}$$

Then, for each individual of the receiver matrix $\mathbf{Y}_s$, defined in (5), is replaced by the $\hat{\mathbf{Y}}_1$ values obtained from the application of the estimated regression models. Hence, in the case of Non Symmetrical Grafting, the coordinates of the supplementary points of the receptor matrix will be the following:

$$\mathbf{\Psi}^*_\mathbf{1} = \mathbf{X_1}(\mathbf{X'_1X_1})^{-\mathbf{1}}\mathbf{X'_1\hat{Y}_1U^*_0} = \mathbf{P_{X_1}\hat{Y}_1U^*_0} \tag{7}$$

where $\mathbf{U^*_0}$ is the basis of the *CPCA* for $\mathbf{Y^*_0} = \mathbf{P_{X_0}Y_0}$. This result solves the problem of the projection in supplementary of the receptor matrix individuals.

4.3. Imputation

Once all the individuals of the two surveys are projected in the same subspace constructed through the *CPCA* technique, for each unit of the receptor matrix we calculate the distances from the individuals of the donor matrix. Adopting the Euclidean metric, among the donors we selected the *nearest neighbor* (or the group of *k-nearest neighbor*) of each receiver statistical unit. Then, for the imputation we exploit the *hot deck imputation*.

In the case of just one *nearest neighbor*, i.e. $k = 1$, the imputation consists of copying the donor survey specific variable values given by the best donor and imputing (pasting) them to the corresponding receiver. On the other hand, to obtain a smoother imputation the k *nearest neighbor* algorithm can be adopted. In such a case, we calculate the average on the specific variable values given by the optimal *knn* donors and impute it to the considered receptor.

To determine the optimal value of k in the nearest neighbor algorithm, we proceed to the auto-imputation of the variables $\mathbf{Y}_0$ on $\mathbf{X}_0$ itself, in order to be able to measure the produced error, and to evaluate the R_y index (Aluja-Banet *et al.*, 2001).

5. Validation of imputation

Once the imputation is performed, it is necessary to validate the imputed data. In this respect, we have three validation levels to measure the imputation quality. The first consists of a global statistics comparison. We perform an hypothesis testing for differences between the means of the block of the imputed variables $\tilde{\mathbf{Y}}_1$ and the block of the donor matrix specific variables $\mathbf{Y}_0$.

The second validation level tends to assess the homogeneity of imputations evaluating internal and external coherency of the imputed variables. The former is based on comparison between the correlation coefficient matrix of $\tilde{\mathbf{Y}}_1$ and the corresponding correlation coefficient matrix of $\mathbf{Y}_0$. The latter tends to verify the homogeneity of the two cross-correlation matrix of $\mathbf{X}_0$ with $\mathbf{Y}_0$ and of $\mathbf{X}_1$ with $\tilde{\mathbf{Y}}_1$.

In order to evaluate both internal and external coherency, we use the Fisher transformation of the correlation coefficient and we perform a set of significant tests based on the Z distribution to verify the pairwise correlation coefficient's homogeneity. Hence, the imputed variables are coherent when a reasonable number of tests on the differences among the correlation coefficients is not significant for a given p-value.

Finally, the third level of the validation process considers the accuracy of the imputation, where the term accuracy denotes the agreement between the imputed values and the "real values". The accuracy can be measured evaluating the root mean square error ($RMSE$) among the imputed values $\tilde{\mathbf{Y}}_1$ and the "real values" $\mathbf{Y}_1$, that we should have had if we observed those variables in the punctual survey:

$$RMSE = \sqrt{n_1^{-1} tr\left[(\tilde{\mathbf{Y}}_1 - \mathbf{Y}_1)'(\tilde{\mathbf{Y}}_1 - \mathbf{Y}_1)\right]}. \tag{8}$$

Obviously, it is possible to perform such a validation only in the case of simulation studies, when the real values are known.

6. Simulation study

In this section we present the results of a simulation study performed to compare the proposed *NSG* algorithm with respect to both the classical file grafting methodology based on *PCA* and the multiple regression imputation. We generate 1000 observations from two multivariate normal distributions; we generate a $10 - dimensional$ standard normal as $\mathbf{X}$ variables and a 5-$dimensional$ standard normal as $\mathbf{Y}$. In the simulation study we consider eight different covariance patterns. In the first case, all the correlations have been set equal to zero ($Sim.1$), while in the last case the correlation of $\mathbf{X}$, of $\mathbf{Y}$ and the cross correlations of $\mathbf{X}$ and $\mathbf{Y}$ vary between 0.2 and 0.6 ($Sim.8$). All the other intermediate cases consider in turns different combinations of independent/dependent $\mathbf{X}$, independent/dependent $\mathbf{Y}$, and independent/dependent $\mathbf{Y}$ on $\mathbf{X}$.

Furthermore, we adopt the *missing data at random* (MAR) approach, i.e. we randomly delete 500 observations in the $\mathbf{Y}$ matrix. For the PCA file grafting and for the $CPCA$ file grafting we consider also the case of both $k = 1$ and $k > 1$ for the nearest neighbor hot deck imputation, and the possibility of using a limited number of eigenvectors depending on the scree plot. For the three methods we performed tests to compare the "true" variances with the variances of the imputed variables, the "true" correlation coefficients with the ones obtained on the imputed variables (internal and external coherences), and finally we evaluated the $RMSE$ value for the five variables in $\mathbf{Y}$.

Looking at whole results we observed the presence of a trade-off between the external and internal coherence and the $RMSE$. Generally, the better performances in terms of $RMSE$ are achieved by the knn hot deck imputation (through both PCA and $CPCA$) and by the regression. This suggest us the idea that the hot deck imputation when k is quite large is a sort of nonparametric and local regression. On the other hand, the regression shows the poorest performances in terms of external and internal coherence, while our proposed procedure shows the best results.

As an example we present the main results of the two extreme cases denoted by $Sim.1$ and $Sim.8$. Moreover, in the following tables $\tilde{\mathbf{Y}}^1_{PCA}$ and $\tilde{\mathbf{Y}}^1_{CPCA}$ are the imputed matrices through the usual file grafting and our file grafting using one nearest neighbor and all eigenvectors, respectively. In the same way, $\tilde{\mathbf{Y}}^k_{PCA}$ and $\tilde{\mathbf{Y}}^k_{CPCA}$ denote the imputed values using the k nearest neighbors and all the eigenvectors, while $\tilde{\mathbf{Y}}_{REG}$ is the imputation through multiple regression. Finally, when the scree plot suggested a reduced number of eigenvectors, we added this number to the abbreviations PCA and $CPCA$.

Table 1. Simulation study 1: Number of P-values less than 0.05 (over 5 tests) for the variance ratio tests $H_0 : Var(\mathbf{Y}_1) = Var(\tilde{\mathbf{Y}}.)$

For 5 variables	$P < 0,05$
$\tilde{\mathbf{Y}}^1_{PCA}$	4
$\tilde{\mathbf{Y}}^1_{CPCA}$	2
$\tilde{\mathbf{Y}}^1_{CPCA(2)}$	1
$\tilde{\mathbf{Y}}^k_{PCA}$	5
$\tilde{\mathbf{Y}}^k_{CPCA}$	5
$\tilde{\mathbf{Y}}^k_{CPCA(2)}$	5
$\tilde{\mathbf{Y}}_{REG}$	5

In the first case (Tables 1–4), we note that the grafting methods with the knn algorithm and the multiple regression perform in a similar way in terms of $RMSE$ (Table 4). However, multiple regression presents the

Table 2. Simulation study 1: Number of P-values less than 0.05 (over 10 tests) of tests for the homogeneity of the real correlation coefficients among the $\mathbf{Y}_1$ *and the correlation coefficients among the imputed* $\tilde{\mathbf{Y}}$*..*

For 10 couples	$P < 0,05$
$\tilde{\mathbf{Y}}^1_{PCA}$	0
$\tilde{\mathbf{Y}}^1_{CPCA}$	2
$\tilde{\mathbf{Y}}^1_{CPCA(2)}$	0
$\tilde{\mathbf{Y}}^k_{PCA}$	3
$\tilde{\mathbf{Y}}^k_{CPCA}$	2
$\tilde{\mathbf{Y}}^k_{CPCA(2)}$	4
$\tilde{\mathbf{Y}}_{REG}$	9

Table 3. Simulation study 1: Number of P-values less than 0.05 (over 10 tests) of tests for the homogeneity of the real cross-correlation coefficients among $\mathbf{X}_1$ *and* $\mathbf{Y}_1$*, and the cross-correlation coefficients among* $\mathbf{X}_1$ *and the imputed* $\tilde{\mathbf{Y}}$*..*

For 50 couples	$P < 0,05$
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}^1_{PCA}]$	4
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}^1_{CPCA}]$	2
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}^1_{CPCA(2)}]$	3
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}^k_{PCA}]$	19
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}^k_{CPCA}]$	17
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}^k_{CPCA(2)}]$	22
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}_{REG}]$	30

Table 4. Simulation study 1: $RMSE$ *values for different types of imputation* $\tilde{\mathbf{Y}}$*..*

	$\mathbf{Y_1}$	$\mathbf{Y_2}$	$\mathbf{Y_3}$	$\mathbf{Y_4}$	$\mathbf{Y_5}$
$\tilde{\mathbf{Y}}^1_{PCA}$	1.4334	1.3627	1.3584	1.3950	1.3343
$\tilde{\mathbf{Y}}^1_{CPCA}$	1.4454	1.4096	1.4393	1.4966	1.4043
$\tilde{\mathbf{Y}}^1_{CPCA(2)}$	1.4297	1.3372	1.3835	1.4307	1.4271
$\tilde{\mathbf{Y}}^k_{PCA}$	1.0629	0.9583	1.0530	1.0342	1.0484
$\tilde{\mathbf{Y}}^k_{CPCA}$	1.0668	0.9516	1.0288	1.0178	1.0665
$\tilde{\mathbf{Y}}^k_{CPCA(2)}$	1.0683	0.9589	1.0411	1.0248	1.0762
$\tilde{\mathbf{Y}}_{REG}$	1.0657	0.9527	1.0247	1.0267	1.0583

worst results in terms of both internal and external coherence (Tables 2 and 3). Note that, as all the correlations are equal to zero the PCA and the $CPCA$ perform similarly. Looking at tests for the variance comparison between the real value of the variances of $\mathbf{Y}_1$ and the variances of the imputed values (Table 1), both *knn* algorithm and multiple regression imputation underestimate the original variances as all tests rejects the null hypothesis of equal variances.

In the second simulation we report, both $\mathbf{X}$ and $\mathbf{Y}$ present a correlation structure and there is also a cross-correlation structure between the $\mathbf{Y}$ block and the $\mathbf{X}$. Our proposed procedure with one nearest neighbor works better than the other methods, especially in terms of internal and external coherence and in terms of variance estimation. Note that with respect to the same criterion, the worst imputation is obtained by multiple regression, especially in terms of both homogeneity (external and internal) and variance reconstruction. Our method with k nearest neighbor works similarly to regression. This is probably due to the large value of k that transforms our method in a local regression. Being the simulated relationships linear, these two methods provide similar results. Further investigations should be done in case of nonlinear relationships, or to find a way to limit the value of k.

Table 5. Simulation study 8 Number of P-values less than 0.05 (over 5 tests) for the variance ratio tests $H_0 : Var(\mathbf{Y}_1) = Var(\tilde{\mathbf{Y}}.)$

For 5 variables	$P < 0,05$
$\tilde{\mathbf{Y}}^1_{PCA}$	3
$\tilde{\mathbf{Y}}^1_{PCA(3)}$	3
$\tilde{\mathbf{Y}}^1_{CPCA}$	4
$\tilde{\mathbf{Y}}^k_{PCA}$	5
$\tilde{\mathbf{Y}}^k_{PCA(3)}$	5
$\tilde{\mathbf{Y}}^k_{CPCA}$	5
$\tilde{\mathbf{Y}}_{REG}$	5

Table 6. Simulation study 8: Number of P-values less than 0.05 (over 10 tests) of tests for the homogeneity of the real correlation coefficients among the $\mathbf{Y}_1$ *and the correlation coefficients among the imputed* $\tilde{\mathbf{Y}}$*..*

For 10 couples	$P < 0,05$
$\tilde{\mathbf{Y}}^1_{PCA}$	0
$\tilde{\mathbf{Y}}^1_{PCA(3)}$	4
$\tilde{\mathbf{Y}}^1_{CPCA}$	3
$\tilde{\mathbf{Y}}^k_{PCA}$	8
$\tilde{\mathbf{Y}}^k_{PCA(3)}$	8
$\tilde{\mathbf{Y}}^k_{CPCA}$	10
$\tilde{\mathbf{Y}}_{REG}$	10

Table 7. Simulation study 8: Number of P-values less than 0.05 (over 10 tests) of tests for the homogeneity of the real cross-correlation coefficients among $\mathbf{X}_1$ *and* $\mathbf{Y}_1$*, and the cross-correlation coefficients among* $\mathbf{X}_1$ *and the imputed* $\tilde{\mathbf{Y}}$*..*

For 50 couples	$P < 0,05$
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}^1_{PCA}]$	7
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}^1_{PCA(3)}]$	10
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}^1_{CPCA}]$	3
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}^k_{PCA}]$	25
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}^k_{PCA(3)}]$	35
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}^k_{CPCA}]$	15
$[\mathbf{X}_1, \mathbf{Y}_1] vs [\mathbf{X}_1, \tilde{\mathbf{Y}}_{REG}]$	15

Table 8. Simulation study 8: $RMSE$ *values for different types of imputation* $\tilde{\mathbf{Y}}$*..*

	RMSE				
	$\mathbf{Y_1}$	$\mathbf{Y_2}$	$\mathbf{Y_3}$	$\mathbf{Y_4}$	$\mathbf{Y_5}$
$\tilde{\mathbf{Y}}^1_{PCA}$	1.5002	1.5520	1.5203	1.5214	1.4517
$\tilde{\mathbf{Y}}^1_{PCA(3)}$	1.6046	1.5638	1.6094	1.5704	1.5881
$\tilde{\mathbf{Y}}^1_{CPCA}$	1.4434	1.3656	1.4138	1.4402	1.4368
$\tilde{\mathbf{Y}}^k_{PCA}$	1.1755	1.1358	1.1524	1.1773	1.1129
$\tilde{\mathbf{Y}}^k_{PCA(3)}$	1.2515	1.1784	1.2111	1.2260	1.1706
$\tilde{\mathbf{Y}}^k_{CPCA}$	1.1242	1.0695	1.0883	1.1164	1.0593
$\tilde{\mathbf{Y}}_{REG}$	1.0657	1.0287	1.0441	1.0686	1.0526

7. Final Remarks

Data fusion can be considered a particular kind of missing data imputation problem and, hence can be treat through several methodologies depending on both the nature of the data and the correlation structure.

In our opinion, if a dependency structure among the common variables and the ones to be imputed is present, in *implicit model imputation* it is not sufficient to evaluate the closeness of donors only on the common variable. In order to define the best donor(s) it is necessary to consider also the relationship structure among variables. In this sense, our *NSG* algorithm and multiple regression imputation take into account such a dependency structure, in contrast with the usual file grafting. In addition, with respect to the imputation by regression models our proposal works better in reconstructing variances and covariances of the imputed variables. Indeed, the simulation study shows that *NSG* algorithm performs better than multiple regression in terms of both homogeneity (internal and external) and variance reconstruction.

More simulation studies will be performed to analyze how our proposal performs when nonlinear relationship structure are present, and further works will be done in order to find a new criterion to select the number k of nearest neighbor donors.

References

Aluja-Banet T., Nonell R., Rius R., Martínez M. (1995), File Grafting, in F. Mola and A. Morineau (eds) *Actes du IIIme Congrés International d'Analyses Multidimensionnelles des Données, NGUS'95,* Centre Int. de Statistique et d'Informatique Appliquées, CISIA-CERESTA, 23–32.

Aluja-Banet T., Thio S. (2001), Survey Data Fusion, *Bulletin of Sociological Methodology,* 72, 20–36.

Aluja-Banet T., Daunis-i-Estadella J., Pellicer D. (2007), GRAFT, a Complete System for Data Fusion, *Computational statistics and data analysis,* 52, 635–649.

Baker K., Harris P., O'Brien J. (1989), Data Fusion: an Appraisal and Experimental Evaluation, *Journal of the Market Research Society,* 31, 153–212.

Barcena M.J., Tusell F. (1999), Enlace de encuestas: una propuesta metodológica y aplicación a la Encuesta de Presupuestos de Tempo, *Qüestiio*, 23, 297–320.

Bonnefous S., Brenot J., Pages J.P. (1986), Methode de la greffe et communication entre enquetes, in E. Diday *et al., Data Analysis and Informatics IV,* North Holland, 603–617.

D'Ambra L., Lauro N.C. (1982), Analisi in componenti principali in rapporto a un sottospazio di riferimento, *Rivista di Statistica applicata,* 15, 51–67.

Dempster A.P., Laird N.M., Rubin, D.B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of Royal Statistical Society,* Ser. B 39, 1–38.

D'Orazio M., Di Zio M., Scanu M. (2006), *Statistical Matching. Theory and Practice,* Wiley, New York.

Ford B. (1980), An Overview of Hot-Deck Procedures, in Madow W., et. al. (eds), *Incomplete Data in Sample Surveys,* 2, Academic Press, New York, 185–206.

Fukunaga K., Narendra P.M., (1975), A Branch and Bound Algorithm for Computing k-nearest neighbors. *IEEE Trans. Computers,* C-24, 7, 750–753.

Lejeune M. (2001), *Traitements des fichiers d'enquêtes,* Presses Universitaires de Grenoble.

Little J.A., Rubin D.B. (1987), *Statistical Analysis with Missing Data,* Wiley & Sons, New York.

Rius R., Nonell R., Aluja-Banet T. (1996), File Grafting: a Data Sets Communication Tool, *COMPSTAT '96,* Physica Verlag, 417–22.

Rius R., Aluja-Banet T., Nonell R. (1999), File Grafting in Market Research, *Applied Stochastic Models in Business and Industry,* 15, 451–60.

Rubin D.B. (2003), Discussion on Multiple Imputation. *International Statistical Review* 71, 3, 619–625.

Santini G. (2001), Méthode de fusion procustéenne, in Lejeune M., *Traitements des fichiers d'enquêtes,* Presses Universitaires de Grenoble, 75–82.

Saporta G., Co V. (1999), Fusion de fichiers: une nouvelle méthode basée sur l'analyse homogéne, in G. Brossier and A.M. Dussaix (eds) *Enquêtes et sondages,* Dunod, Paris, 81–93.

Saporta G. (2002), Data Fusion and Data Grafting, *Computational Statistics and Data Analysis,* 38, 465–473.

Schafer J.L., (1997), *Analysis of Incomplete Multivariate Data,* Chapman & Hall, London.

Schafer J.L., Olsen M.K. (1998), Multiple Imputation for Multivariate Missing-Data Problems: a Data Analysts Perspective, *Multivariate Behavior Re-*

search, 33, 545–571.

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

Shao J., Wang H. (2002), Sample Correlation Coefficients Based on Survey Data Under Regression Imputation, *Journal of the American Statistical Association,* 97, 544–552.

Winkler W.E. (1995), Matching and Record Linkage, in B. G. Cox (eds), *Business Survey Methods,* John Wiley, New York, 355–384.