A new Method to Match Data Sets Applied to Electric Market

Nicolas Fischer, Christian Derquenne, Gilbert Saporta

A New Method To Match Data Set Applied To Electric Market

Nicolas FISCHER[1], Christian DERQUENNE[1] and Gilbert SAPORTA[2]
[1]*Electricité de France, Research and Development Division*
1, av. du Général de Gaulle
92141 Clamart Cedex, France
nicolas.fischer@edf.fr
christian.derquenne@edf.fr
[2]*Conservatoire National des Arts et Métiers*
Chaire de Statistique Appliquée – CEDRIC
292, rue St Martin
75141 Paris Cedex, France
saporta@cnam.fr

Abstract: The French household electricity market currently dominated by EDF, is going to be opened up to free market economy. In this context, a better understanding of its customers behaviour would be a key feature for the success of the company. To achieve this goal, EDF holds two information sources i.e., a comprehensive customers' invoicing file with few individual data though, as well as survey results performed in regional centres, which contain more data per customer.
We herein present a new method of data fusion based on the generation of virtual individuals. For each of them, a full set of variables will be accessible. This procedure is based on two steps. Firstly, a Multiple Correspondences Analysis (MCA) from an existing master sample is performed on fundamental variables. A sample of virtual individuals (SVI) is randomly selected, based on the distribution of each significant MCA factor. Then, for each virtual individual, a specific value is given for each fundamental variable which is the most correlated to one of the MCA factors. Secondly, the distinct sets of secondary samples are grafted to the previous SVI. A simultaneous estimation of variables distribution is made as based on PLS regression on variables shared by all samples. The use of this method brings about two advantages, namely the possibility to choose SVI size and the avoidance of variance underestimation generally encountered in using the imputation methods.
This process has been so far applied to the treatment of two databases, i.e. two surveys, in order to generate the expected artificial sample. Validation and perspectives will be herein further discussed.

Keywords: Statistical Data Fusion, Survey Data, Multiple Correspondences Analysis, Calibration, Partial Least Squares Regression.

1. Introduction

1.1 Background

A main strategy used by "Electricité de France" has been to develop electric power consumption among residential clients. It has required a knowledge of the relationships between various stakeholders of France's residential electric power market (clients, builders, competitors, etc.). Since the overall information is not available in a single data base including the same customers, a market simulation project has been undertaken. A same problem is raised again when a company wishes to improve its quality of service and or its productivity.

Techniques of data fusion constitute a good approach to solve this problem. Data fusion consists in merging data bases coming from different sources into a single data base when variables are absent or missing in some files. After summarising the main concepts relating to data fusion, we present a new method based on the generation of a sample made of virtual individuals (SVI), drawn from several independent data bases. Each individual is defined by a set of characteristics including socio-demographic, behavioural, equipment items, etc.

1.2 Statistical data fusion [7]

The principle of statistical data fusion, more currently called data fusion, is to obtain a single data-base where all the variables are completed by the union of units. The resulting base may be then analysed with data mining tools. The problem may be formalised in two data files :
the first file contains observations for a set of (p+q) variables taken on n_0 units,
the second file contains observations of only a subset of p variables for n_1 units.

If X stands for the common variables, there is the following scheme:

$\mathbf{X_0}$	$\mathbf{Y_0}$
$\mathbf{X_1}$	**?**

There is thus a blank quarter, our goal is to fill this blank part. In this case, it corresponds to a special type of missing data estimation, where many variables are missing because they have not been collected.
The origins of data fusion are market studies, especially in media and consumption surveys, for which it is impossible to ask all questions because they are too numerous. In order to reduce the questions file, one generally proceeds with two different independent samples, the questions being divided into two parts, with a common set of descriptors.

1.3 Main methods for data fusion [4] and [7]

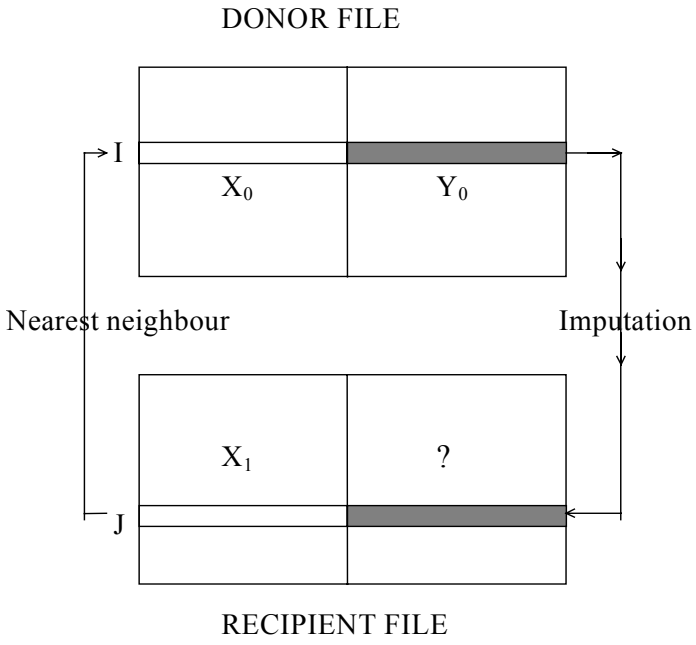
Since the couple (X_0,Y_0) is used to predict the unknown Y part of the second file, the first file is denominated donor-file and the second the recipient-file. Data fusion is a very special case of missing value estimation. Thus, several classical methods can be applied.

- Explicit model based estimation

Every missing value is estimated with classical techniques, such as regression, the general linear model for numerical Y, or logistic regression if Y is categorical : each variable of Y_0 is modelled with X_0 as predictors, and the model is applied to the recipient file. These techniques suffer from several drawbacks. Estimations are made variable by variable, not taking into account their correlation and, thus could lead to inconsistent results. Maximum likelihood estimation may be applied but this method does not prevent incoherent results [5]. Another drawback of estimation methods consists in the fact that two units having the same values of the predictors will have the same estimate of their Y variable, hence a loss of variability. Multiple imputation techniques, based on a bayesian framework [6] allow to simulate the posterior distribution of the missing values by imputing each data with several values according to one or more estimation models. One can thus recover correct variances with multiple imputations. However, these techniques are very complex and time consuming for large data sets.
To sum up, explicit estimation methods seem fitted more to sparse missing values than to the estimation of blocks of thousands of missing data like in data fusion.

- Imputation with implicit models

Much more simpler methods than the previous ones consist in giving to the variables called Y of a receiver the whole vector of variables of a donor : copy and paste.
We note the integer i the receiver. The basic idea is to look for a donor called j having a close profile with the variables X : a double if all the variables are identical or a nearest neighbour such as an appropriate distance, called d(i,j) in the P^P space of common variables, is minimal.
This method avoids incoherent estimations since the copied values belong to real observations. Furthermore, in order to avoid loss of variability, one can use a penalty function such that a same donor cannot be used too many times.



2. A generation of a sample of virtual individuals (SVI)

C. Derquenne [3] provided a method of generating a sample of artificial individuals involving two main steps based on different areas of statistics i.e., sampling, data analysis and generalised linear models.

In this work, we are introducing a multivariate approach using the Partial Least Square regression (PLS regression). The method herein presented, is based on two steps like the method of C. Derquenne.
- The first step called (i) is to generate the *first artificial sample* based on primary sample called χ.
- The second step called (ii) is to statistically graft a secondary sample $\mathbf{Y}^1$ onto the first artificial sample to obtain the second artificial sample.
The second step is repeated to provide the *final artificial sample* or Sample of Virtual Individuals (SVI) by progressively grafting other secondary samples $\mathbf{Y}^2$,...$\mathbf{Y}^K$.

3. Original Data

Two sets of survey samples are available: *primary sample* χ and *secondary samples* $\Upsilon^1,...,\Upsilon^K$, respectively :
- the primary sample includes the variables of the sample design taken from the survey design called X_{MP} and also the measured variables called X_M.
- the secondary samples have some variables shared with the X_{MP} variables, called $Y_{MP}^{(k)}$, and others measured variables which are in part common and for another part non-common to X_M, called $Y_M^{(k)}$.

This method is applied to the generation of a sample of 10,000 virtual individuals for all variables of the surveys described below. The primary sample is linked to the 1990 CREDOC survey on the living conditions and aspirations of the French population, while the secondary sample was taken from a 1990 SOFRES survey on home heating. The following table shows the variables of the sample design and the measured variables:

	CREDOC 1990 (2.000 persons)	***SOFRES 1990 (8.000 clients)***
Sampling variables X_{MP} and Y_{MP}	- gender×age - occupation - size of town	- age - occupation - size of town
Measured variables X_M and Y_M	- dwelling characteristics - household characteristics - principal heat source - opinion regarding nuclear power stations - opinion regarding the environment	- dwelling characteristics - household characteristics - principal heat source - level of satisfaction with respect to heat source (cost, safety, thermal comfort,etc.)

4. Generation of artificial data samples

The principle used to generate the sample of artificial individuals involved two main steps, described below.

4.1 Generating a first artificial sample S_0 based on primary sample ?

Firstly, we have generated the first artificial sample. This procedure uses the primary sample fixed previously. We have taken a vector called $X_{MP,}$ the size of which is equal to Q, composed by $X_{MP(1)}$,...,$X_{MP(Q)}$, i.e. these variables represent the sampling variables which are all qualitative. We have then introduced an other vector X_M of size R, composed by $X_{M(1)},...,X_{M(R)}$. These variables denote the measured variables which are nominal, ordinal or discretized.

Our goal is to provide a reduced-dimension space of initial variables, which allows to keep the greatest quantity of well-separated, available information. In this step, we have applied a Multiple Correspondence Analysis (MCA) [1] to the sampling variables called $X_{MP(1)},...,X_{MP(Q)}$ in order to generate the principal components. These sampling variables are used as active variables to generate the principal component, whereas the measured variables called $X_{M(1)},...,X_{M®}$ are used as supplementary variables. The Multiple Correspondence Analysis is used to provide reduced dimension as well as new, uncorrelated variables with principal components.

We have chosen a vector designated $Z = (Z_1, ...,Z_T)$, which its principal components (numerical) have been derived by the Multiple Correspondence Analysis. A subset Z* of the set Z corresponding to "significant" eigenvalues called λ_t has then be selected. The subset Z* is equal to $(Z_1 , ...,Z_{T*})$, in which the integer called T* represents the number of eigenvalues higher than 1/MP(Q). A space of N_0 individuals has thus been obtained on the T^* principal components from the Multiple Correspondence Analysis. The principal components which depend on the initial variables should be defined. Therefore, we have generated a few groups of variables which best correlated with the principal components. To do so, the correlation ratio η^2 between sampling variables and "significant" principal components was calculated. The maximum of the correlation ratio η^2 was established such that :

$$\eta^2 (X_{MP (q)} , Z_u) = \max_{t=1,T^*} \eta^2 (X_{MP (q)} , Zt),$$

where $\eta^2 (X_{MP (q)} , Z_t) = R^2 (Z_t ; X_{MP(q)}^{(1)},\ldots, X_{MP(q)}^{(mq)})$

Thus, one gets: $X_{MP}^{(t)} = \{X_{MP (q)} ; \eta^2 (X_{MP (q)} , Z_u), \forall t, t=u\}$ which is a group of variables correlated with the principal Z_t component, included in the subset called Z^*. The same procedure was followed with the measured variables to obtain supplementary variable groups: $X_M^{(t)}$.

How artificial individuals can be drawn?

A discrete space is substituted to the continuous space as based on the principal components in order to determinate a empirical distribution. To do so, a paving space is generated.
In the first step, each principal component of the subset called Z^* is discretized into k_t intervals so as to generate a paving space such that :

$$k_t = \left[(n_d)^{1/T^*} \times \lambda_t \Big/ \prod_{t=1}^{T^*} \lambda_t \right]$$

where n_d is equal to n/s, s previously fixed.

Finally, $\widetilde{K} = \prod_{t=1}^{T^*} k_t \leq n_d$ is the number of windows (w_l) in the paving space.

We put $f_l = n_l / n$, denoting the observed distribution in the paving space, where $n_l = \sum_{i=1}^{n} 1_{(i \in w_l)}$ for all integer l = 1 to $\widetilde{K}$. Thus, *N* artificial individuals are drawn from this distribution, where the integer *N* represents the generated sample size. It should be higher than that of the primary sample. These artificial individuals are called "dummy individuals":

$\widetilde{z}_{\widetilde{i}}^{\,1} = \left(\widetilde{z}_{\widetilde{i}(1)}^{\,1}, \ldots, \widetilde{z}_{\widetilde{i}(T^*)}^{\,1}\right)$, for all integer $\widetilde{i}$ = 1 to *N*.

As a second step, each set of variables ($X_{MP(q)}^{(t)}$; $X_{M(r)}^{(t)}$) also involves a distribution observed in each window of the paving space. In this case, the empirical distribution $f^{MP(q)}_{l(1)}$ is equal to $n^{MP(q)}_{l(1)} / n_l$. *N* artificial individuals are drawn from the distribution observed of $X_{MP(q)}^{(t)}$ and $X_{M(r)}^{(t)}$ knowing $\widetilde{z}_{\widetilde{i}}^{\,1}$. These new artificial individuals are called "first replicates": $\widetilde{x}_{\widetilde{i}}^{\,\widetilde{z}} = \left(\widetilde{x}_{\widetilde{i}(1)}^{\,\widetilde{z}}, \ldots\ldots, \widetilde{x}_{\widetilde{i}(MP(Q)+M(R))}^{\,\widetilde{z}}\right)$.

4.2 Statistical graft based on secondary samples using the logistic PLS regression

The first sample is chosen within the set of secondary samples $\mathbf{Y^1}, \ldots, \mathbf{Y^K}$. This choice is normally based on the number of sampling variables common to the primary sample and the variables of the secondary sample determined as important for the study.

A vector $Y_{MP}^{(1)} = \{Y_{MP(1)}^{(1)}; \ldots; Y_{MP(Q1)}^{(1)}\}$ denotes the sampling variables common to the primary sample, and let $Y_M^{(1)} = \{Y_{M(1)}^{(1)}; \ldots; Y_{M(R1)}^{(1)}\}$ denotes the other measured variables, G_1 represents the number of variables not shared with the primary sample to be grafted onto the first artificial sample S_0. The second artificial sample was then generated in two steps.

As a first step, the secondary sample was adjusted in relation to the sample design of the primary sample, using a marginal calibration method [2].

As a second step, G_1 variables were grafted one by one from the secondary sample into the first artificial sample S_0. However, the grafting variable one by one is not completely satisfactory, because it doesn't take into account the correlation structure between variables. Therefore, PLS regression has been used since it allows the modelling of a block of response variables by a block of "explanatory" variables, including correlation structure.

Let G_1 variables denoted by $y_{M^*}^{(1)} = \{y_{M^*(1)}^{(1)}; \ldots; y_{M^*(G1)}^{(1)}\}$ to graft. These variables can either be Boolean, nominal, or ordinal. Let P_1 variables shared by the primary and the second samples, these variables are denoted by $y_C^{(1)} = \{y_{C(1)}^{(1)}; \ldots; y_{C(P1)}^{(1)}\}$. The goal is to estimate the distribution of each variable to graft.

To predict the variables to be explained with the candidate "explanatory" variables, the Partial Least Square regression method, such as PCA methods, aims to extract some components with the help of the candidate explanatory variables. The number of PLS components is determined by cross-validation. These components enable to estimate response variables.
Here, $y_C^{(1)}$ were the "explanatory" variables. Thus a logistic PLS regression was performed, an extension of PLS regression, in order to obtain the estimated distribution of $y_{M^*}^{(1)}$. The complete algorithm of logistic PLS regression is given by M. Tenenhaus [8]. Then *N* artificial individuals $\widetilde{y}_{M^*}^{(1)}$ are drawn from the estimated distribution of $y_{M^*}^{(1)}$, the characteristics of the significant "explanatory" variables in the first artificial sample being known. The generation of other artificial samples follows the same procedure until the final sample of virtual individuals is obtained.

5. Validation, advantages, limits and prospects

Which are the quality indicators? Recovery of the values at an individual level seems appealing but too demanding in most cases. Users are not generally interested in individual predictions and may be satisfied with predictions that are correct on average for groups of units. It is however not enough to recover marginal distributions or mean values, since a random sampling could do this adequately ! The main problem is to keep the covariance structure, or for categorical data to have some correct cross-tabulations between variables of interest. Therefore, tests are performed, comparing for example the marginal distributions for an existing survey and the generated sample or comparing correlations between two generated variables and two observed variables taken from the same survey. A more complete statistical validation of the final SVI is proposed by C. Derquenne [3] involving six tests of increasing complexity. In all circumstances, in order to have satisfactory results, it is necessary to have a large enough number of common variables and high correlations between the block of common variables and variables to be grafted.
Finally, the main advantages of this new method consist in the possibility to achieve to generate a sample of variable size, in the positive results obtained, in the possibility to take into account the correlations between variables, given the multivariate approach (PLS regression) and in the generation of artificial individuals using an a priori knowledge. There are two limitations, however, the fact that the size of the survey samples is generally small and the complexity of the generation process increases with the number of grafted variables and secondary samples. Nevertheless, our method will be generalized in terms of longitudinal data (panel data). The proposed method will be used for the study of other segments of EDF customers' group.

References

[1] J.-P. Benzecri et al. (1979). L'Analyse des Données, Tome 1: la Taxinomie, Tome 2: l'Analyse des correspondances, 3e éd., Dunod, Paris.

[2] W.E. Deming and F.F. Stephan (1940). On a least square adjustment of sampled frequency table when the expected marginal total are known, Annals of Mathematical Statistics, Vol. 11, 427-444.

[3] C. Derquenne (1999). A Method of Generating a Sample of Artificial Data from several existing data tables : Application Based on the Residential Electric Power Market, Proceedings of Statistics Canada Symposium 99, Combining Data from Different Sources.

[4] N. Fischer and G. Saporta (2000) Fusion et Greffes de Données, In Proceedings 5èmes Journées Modulad : Data Mining des Données Clientèle, Clamart.

[5] R.J.A. Little, D.B. Rubin (1987). Statistical analysis with missing data, Wiley, New-York.

[6] D.B. Rubin (1987). Multiple Imputation for Nonresponse in Surveys, Wiley, New- York.

[7] G. Saporta (2000) Data Fusion and Data Grafting, In Proceedings NMDM2000, International meeting on Non-linear Methods and Data Mining, Rome.

[8] M. Tenenhaus (2000). La Régression Logistique PLS, Journées d'Etudes en Statistique, Modèles Statistique pour données Qualitatives.