A Multivariate Approach to Combine Data from Different Sources: Application to Customers Knowledge of Electricité De France

Christian Derquenne
Electricité de France, Research and Development Division
1, av. du Général de Gaulle
92141 Clamart Cedex, FRANCE
christian.derquenne@edf.fr

Nicolas Fischer
Electricité de France, Research and Development Division
1, av. du Général de Gaulle
92141 Clamart Cedex, FRANCE
nicolas.fischer@edf.fr

1. Main issue

A major strategy used by Electricité de France (EDF) has been to develop electric power consumption among residential clients. This has required a good knowledge of customers expectations. The statistical studies have been done (electric consumption, loyalty, new clients, satisfaction surveys, ...). For these ones, EDF owns several data bases on different segments of customers (residential, professional, firm, industries, ...). But, the overall information is not available in a single data base including the same custumers. Indeed, a single data base could provide value added information, then statistical studies could be applied and substential results in marketing objective, for instance, could be obtained. The combining data from different sources is a candidate solution. Then, we have developed a method of generating a sample of artificial individuals involving two main steps based on different areas of statistics : sampling, data analysis and generalized linear models. In this paper, we propose an improving of this method based on a multivariate approach by means the Partial Least Squares regression (PLS) and logistic regression.

2. Multivariate Approach to combine Data from Different Sources

The artificial sample was generated in two steps. The first step, based on a master sample, was a Multiple Correspondence Analysis (MCA) carried out on basic variables. Then “dummy” individuals were generated randomly using the distribution of each “significant” factor in the analysis. Finally, for each individual, a value was generated for each basis variable most closely linked to one of the previous factors. This method ensured that sets of variables were drawn independently. The second step consisted in grafting some other data bases (secondary samples), based on certain property requirements. In the old version of this method, a variable was generated to be added on the basis of its estimated distribution, using a logit model for common variables and those already added. But, the grafting procedure one variable by one variable is not completely satisfactory, because it doesn' t take into account of the correlation structure between variables. Then, PLS regression has been used because it allows the modelling of a block of response variables by a block of "explanatory" variables, including correlation structure. PLS regression has been extented to logistic regression PLS. The same procedure was then used to graft the other samples.

3. Advantages, limits, applications and prospects

This method was applied to the generation of an artificial sample taken from two surveys. The artificial sample that was generated was validated using sample comparison testing. The results were positive, demonstrating the feasibility of this new method in comparison with the previous. Three major advantages of the method are the possibility of generating a sample of containing all information, the taking into account of several variables by new multivariate approach and the positive results obtained. They are two limitations, however, i.e. the fact that the size of samples is generally small, and the complexity of the generation process increases with the number of secondary samples. Moreover, our method will be generalized in terms of longitudinal data (panel data). The proposed method should be applied to other segments of clientele.

REFERENCE

Benzecri, J.-P. et al. (1979). *L'analyse des données,* Tome 1 : *la taxinomie,* Tome 2 : *l'Analyse des correspondances,* 3^{e} éd., Dunod, Paris.

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

Derquenne, C. (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.*

Mc Cullagh, P., and Nelder, J.A. (1990). *Generalized Linear Models,* Monographs on Statistics and Applied Probability **37**, Chapman & Hall.

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

RESUME

La méthode de fusion statistique de données proposée ici est fondée sur la construction d' un échantillon de données artificielles. Elle utilise l' Analyse des Correspondances Multiples et la régression logistique PLS (Partial Least Squares). L' échantillon artificiel généré a été validé statistiquement. Les résultats sont positifs et montrent la faisabilité de cette méthode.