

# **LA COMBINAISON DE DONNÉES DE SOURCES DIFFÉRENTES : APPARIEMENT STATISTIQUE *vs* MODÈLES DE PRÉDICTION**

Christian Derquenne  
EDF R&D - 1, avenue du Général de Gaulle- 92141 Clamart Cedex - France

## **RÉSUMÉ**

Ce papier fait le point sur différentes méthodes de fusion statistique de données utilisées dans le cadre de l'enrichissement de base de données clientèle EDF. Le principe de la fusion statistique est de greffer des informations issues d'un fichier donneur (par exemple, une enquête de satisfaction auprès de la clientèle EDF) sur un fichier receveur (par exemple, une base de facturation d'un Centre EDF). Les différentes méthodes de fusion statistique que nous avons développées et utilisées ont été appliquées à un jeu de données test. Les résultats obtenus sont comparés. D'autre part, les méthodes sont confrontées de façon plus théorique. L'utilisation adéquate de ces méthodes dépend des objectifs visés et donc des atouts et des faiblesses de chacune.

## **1. PROBLÉMATIQUE DE LA FUSION STATISTIQUE DE DONNÉES**

Dans le contexte actuel de développement du Groupe EDF et dans un environnement concurrentiel, une meilleure connaissance des attentes et des besoins de ses clients est primordiale. En effet, cette connaissance représente un enjeu essentiel permettant la construction de stratégies et d'actions décisives pour placer le Groupe EDF aux premiers rangs des opérateurs européens. Il est alors très intéressant pour les experts clientèle dans les agences commerciales de disposer d'informations supplémentaires (satisfaction, attentes, services rendus, ...), leur permettant d'avoir une aide à la décision pour construire leurs cibles et mettre en œuvre des actions marketing. Cependant, l'ensemble de ces informations n'est pas disponible dans une seule et même base de données regroupant les mêmes clients. Une des solutions possibles pour résoudre ce problème est la combinaison de données de sources différentes et plus particulièrement, la fusion statistique de données. Celle-ci, introduite fin des années 80, permet de greffer des informations d'un ou plusieurs fichiers de données (enquêtes, recensement, ...) nommés fichiers donneurs, sur une base de données, dit fichier receveur, (par exemple un fichier de facturation d'un Centre EDF). L'opération de greffe ne peut être réalisée que s'il y a des informations communes dans les fichiers donneur et receveur. Signalons que si l'ensemble des méthodes de fusion statistique permet d'obtenir plus d'informations sur les individus, celles-ci ne sont que des prédictions auxquelles sont associées des erreurs. L'objectif de ce papier est de comparer les méthodes de fusion statistique développées et appliquées à EDF R&D à des fins d'enrichissement de données.

## **2. MÉTHODES DE FUSION STATISTIQUE MISES EN ŒUVRE**

### **2.1 Deux types de méthodes : appariement statistique *vs* modèles de prédiction**

La fusion statistique de données est un cas particulier du traitement des données manquantes où des variables, dites "spécifiques" sont complètement absentes dans le fichier receveur. Il y a alors deux types de méthodes pour estimer ces variables absentes. Le premier consiste à mimer les observations présentes du fichier donneur dans le fichier receveur, car elles se ressemblent sur les variables communes. En d'autres termes, cela permet de copier des valeurs issues des variables spécifiques du premier fichier dans le second, au moyen de la similitude des valeurs des variables communes entre ces deux fichiers. Ce principe est souvent nommé "*appariement statistique*" et repose généralement sur la technique des plus proches voisins. Le deuxième type de méthodes, revient plutôt à raisonner de façon macro. L'idée directrice est de prédire l'information, en tenant compte des informations (les variables communes) qui ont

une influence sur les valeurs des variables spécifiques. Ce principe repose sur les "modèles de prédiction" : les variables communes du fichier donneur vont servir de variables candidates à l'explication, alors que les variables spécifiques (à greffer) représenteront les variables réponses à expliquer.

## 2.2 Modélisation des données : trois approches (univariée, séquentielle, multivariée)

Soient  $(Y_1, \dots, Y_q, \dots, Y_Q)$  des variables réponses qualitatives (à greffer) ayant chacune  $R_1, \dots, R_Q$  modalités, que l'on cherche à expliquer par un ensemble  $(X_1, \dots, X_p, \dots, X_p)$  de variables qualitatives candidates à l'explication avec  $m_1, \dots, m_j, \dots, m_p$  catégories, respectivement (ces notations sont également valables pour l'appariement).

Le principe de *l'approche univariée* est de construire autant de modèles que de variables réponses à greffer. La méthode, nommée **LOGIT**, revient à modéliser une par une chacune des variables spécifiques par régression logistique, à l'aide de la méthode du maximum de vraisemblance classique. Mais l'approche univariée ne permet pas de tenir compte des relations (corrélations) souvent existantes entre les variables à expliquer.

*L'approche séquentielle (SEQ)* [Derquenne, 1999] permet en partie de pallier le problème de non prise en compte des corrélations. En s'appuyant toujours sur la régression logistique, l'idée est de tenir compte dans la modélisation des liens existant entre les différentes variables  $Y$ . Au départ, la démarche est la même que précédemment, chaque  $Y$  est estimée à l'aide d'un modèle de régression logistique sur les  $X$  candidates à l'explication. Mais au lieu de greffer chacune des  $Y$  indépendamment, on ne conserve que le modèle donnant le meilleur taux de bien classés pour la variable considérée, et seule l'estimation de celle-ci sera greffée dans le fichier receveur. La seconde étape consiste à reprendre les  $Q-1$  variables  $Y$  restant à estimer et à reconstruire pour chacune un nouveau modèle de régression mais en intégrant cette fois la variable estimée dans les variables candidates à l'explication de l'échantillon donneur. Ne sera retenue à nouveau que la variable obtenant le meilleur taux de bien classés lors de cette étape. Ce processus continue jusqu'à épuisement des variables à greffer. Enfin, cette approche est dépendante de l'ordre d'entrée des variables  $Y$ .

*L'approche multivariée* a pour objectif de prédire un ensemble de variable à expliquer (de différents types) en une seule fois. Dans le cas où il n'y a que des variables numériques, la régression PLS2 (Partial Least Squares) peut être utilisée [Wold & al., 1983], car elle tient compte de la relation entre les variables à expliquer. Malheureusement, cette méthode n'a pas été étendue au cas de plusieurs variables qualitatives. Cependant le cas d'une seule variable qualitative réponse dans le cadre de la régression logistique PLS a été développée [Tenenhaus, 2000]. Pour répondre à ce problème ouvert, nous avons proposé deux méthodes : **MUL** et **PML** permettant de modéliser un bloc de variables à expliquer qualitatives. La méthode **MUL** [Fischer & al., 2001] fournit une transformation (numérisation) adaptée à chaque type de variables catégorielles. Il y a alors  $R_j - 1$ , valeurs numériques possibles pour chaque  $Y_j$ , qui sont estimables par la régression PLS2 classique. Mais elle est limitée par la complexité croissante du nombre de variables à expliquer croisées. Pour pallier ces problèmes, la méthode **PML** (Partial Maximum Likelihood) fondée sur le maximum de vraisemblance partielle [Derquenne, 2003] et le principe de l'algorithme PLS2 a été mise en œuvre. Dans **PML**, l'itération initiale pour la recherche de la première composante PML, revient à choisir une variable de référence parmi les  $Y$ . Les itérations suivantes permettent de construire une composante (une variable latente) contenant les informations sur l'ensemble des variables à expliquer, en tenant compte de leur nature. Le processus itératif se déroule en un certain nombre de sous-étapes et s'arrête dès qu'il y a convergence de l'algorithme à un seuil fixé donné. Lorsque cette première composante PML est obtenue, les composantes suivantes sont simplement calculées à l'aide d'une régression PLS2 classique sur les résidus des  $X$  et des  $Y$ . L'avantage de cette méthode, par rapport à la précédente, est qu'elle ne dépend pas de la complexité croissante du nombre de variables « explicatives » croisées, car elle repose sur la prise en compte partielle et itérative des variables à greffer, et des variables candidates à l'explication pour construire une nouvelle variable (la première composante PML).

## 2.3 Appariement statistique par les plus proches voisins

Trois méthodes ont été mises en œuvre et reposent toutes sur le même principe : l'appariement à l'aide des plus proches voisins. Elles diffèrent selon la finesse du processus de recherche et de choix des individus donneurs. La méthode **PPV** consiste à rechercher un donneur le plus proche possible du receveur afin de

lui transférer entièrement ses variables spécifiques. Le rapprochement se fait par les variables communes. Dans ce cas, les deux types de variables sont de nature qualitative. Les individus donneurs et receveurs sont regroupés dans un même fichier. Une ACM (Analyse des Correspondances Multiples) est réalisée sur les variables communes ce qui permet d'obtenir les coordonnées de chaque individu du fichier sur le même espace (même référence factorielle). Puis, un tableau de distances euclidiennes entre chaque receveur et chaque donneur est construit. Le transfert des valeurs des donneurs des variables spécifiques aux receveurs repose sur un algorithme de mariage [Santini, 2002]. Mais cette méthode a tendance à sous-estimer la variance des valeurs transférées. La méthode **PPVDD** [Derquenne, 2004] permet de pallier ce problème. Elle débute aussi par une ACM, mais est suivie d'une classification automatique sur les composantes principales, puis utilise une pénalisation des donneurs ayant déjà donné, ainsi que la distribution des distances entre les donneurs et les receveurs. Les donneurs composant le voisinage d'un receveur sont ceux situés à une distance inférieure ou égale au 5<sup>ème</sup> percentile de la distribution des distances entre le receveur et les donneurs, par classe. Les receveurs sont choisis aléatoirement et pénalisés. Le voisinage est donc plus grand que dans le cas PPV. Cependant ces deux méthodes fournissent de très mauvais pourcentage de bien classés. Pour pallier cet inconvénient, nous avons mis en œuvre une approche dérivée de la précédente : **PPVDD bis** [Derquenne, 2004], en prédisant les valeurs respectives des variables spécifiques, par la modalité recueillant la plus forte proportion observée parmi celles de chaque variable spécifique.

### 3. COMPARAISON DES MÉTHODES ET VALIDATION

La validation est une étape essentielle de la fusion statistique. Nous avons utilisé une enquête EDF/SOFRES avec trois variables spécifiques sur l'appréciation du chauffage électrique (CE) : la satisfaction (3 modalités ordinales), le choix futur (CE ou non) et le conseil sur le CE auprès de l'entourage (3 modalités ordinales). Cet échantillon de 7114 interviewés a été découpé en deux parties : un échantillon d'apprentissage (77% des individus) sur lequel les règles d'attribution des variables spécifiques en fonction des variables communes sont construites ; un échantillon de validation sur lequel les règles précédemment élaborées sont appliquées. Chaque méthode de fusion est jugée selon trois critères de validation en terme : de préservation de distributions croisées entre les variables à expliquer, de reconstitution des distributions marginales, et de pourcentage de bien classés [Aluja-Banet, 2002, Santini, 2002]. Pour cela, nous avons construit quelques tests statistiques associés à chacune des trois formes de validation.

La grille de lecture du tableau 1 sur les pourcentages de bien classés est la suivante. Par exemple, la méthode **PML** appliquée à l'estimation de la variable « choix » obtient 64,4% de bien classés contre seulement 57,7% au hasard (sans modèle). Par ailleurs, le niveau de signification associé au test de comparaison des proportions issues du modèle et sans modèle est plus petit que  $10^{-4}$ . Ce qui signifie qu'il y a une différence significative entre ces deux proportions, et celle-ci est positive (un + est entre crochets) : le modèle fait donc mieux que le hasard. On peut constater qu'aucune méthode n'est performante sur la satisfaction, en particulier **PPV** et **PPVDD** sont très mauvaises. Cela est dû au fait que même les variables « explicatives » significatives de la satisfaction ne suffisent pas à bien discriminer les réponses. Les méthodes **MUL** et **PML** sur les deux autres variables sont les plus performantes, alors que **PPV** et **PPVDD** sont encore mauvaises.

<i>Méthodes</i>	<i>Satisfaction</i>	<i>Choix</i>	<i>Conseil</i>
Sans modèle	51,5%	57,7%	39,2%
LOGIT	52,3 (0,265) [0]	<b>63,9</b> ( $<10^{-4}$ ) [+]	<b>46,3</b> ( $<10^{-4}$ ) [+]
SEQ	53,5 (0,053) [0]	<b>64,3</b> ( $<10^{-4}$ ) [+]	40,7 (0,095) [0]
MUL	50,4 (0,823) [-]	<b>66,7</b> ( $<10^{-4}$ ) [+]	<b>45,7</b> ( $<10^{-4}$ ) [+]
PML	50,4 (0,831) [-]	<b>64,4</b> ( $<10^{-4}$ ) [+]	<b>44,3</b> ( $<10^{-4}$ ) [+]
PPV	42,4 (1,000) [-]	58,1 (0,358) [0]	39,2 (0,466) [0]
PPVDD	42,7 (1,000) [-]	57,7 (0,492) [-]	38,3 (0,769) [-]
PPVDD bis	49,9 (0,903) [-]	<b>62,9</b> ( $<10^{-4}$ ) [+]	<b>41,7</b> ( <b>0,016</b> ) [+]

Tableau 1 : Test statistique sur le pourcentage de bien classés

Le tableau 2 fournit les résultats sur la reconstitution des distributions marginales entre les variables observées et les variables estimées. Sur chaque ligne, sont données deux informations. La première est relative à la valeur de la statistique du  $\chi^2$ , alors que la seconde, entre parenthèses, fournit le niveau de

signification du test associé. Plus cette valeur est grande, moins les distributions marginales observées sont éloignées statistiquement des distributions estimées par le modèle. Par exemple, la méthode **PML** reconstitue assez bien les marginales de la variable choix, car son niveau de signification est relativement élevé : 0,1434. Ici les résultats obtenus sont à l'opposé de ceux concernant les pourcentages de bien classés. En effet, les méthodes précédemment médiocres permettent de bien reconstituer les distributions marginales, car comme on peut le constater les niveaux de signification sont très élevés pour **PPV** et **PPVDD**, pour les trois variables. Pour la méthode **MUL**, seule la variable « choix de l'énergie de chauffage » est bien reconstituée. Remarquons que cette méthode procure des statistiques du chi2 plus basses que **LOGIT** et **SEQ**. Enfin, la méthode **PPVDD bis** reconstitue bien les distributions marginales sur le choix et le conseil, et même si ce n'est pas le cas pour la satisfaction, son chi2 y est cependant relativement faible par rapport aux autres méthodes (sauf **PPV** et **PPVDD**).

<i>Variables spécifiques</i> <i>Méthodes</i>	<i>Satisfaction</i>	<i>Choix</i>	<i>Conseil</i>
LOGIT	530,59 (<0,0001)	35,87 (<0,0001)	60,75 (<0,0001)
SEQ	615,61 (<0,0001)	45,24 (<0,0001)	833,30 (<0,0001)
MUL	304,06 (<0,0001)	<b>0,10 (0,7508)</b>	28,43 (<0,0001)
PML	264,52 (<0,0001)	<b>2,14 (0,1434)</b>	31,69 (<0,0001)
PPV	<b>0,75 (0,6882)</b>	<b>0,05 (0,8157)</b>	<b>0,03 (0,9833)</b>
PPVDD	<b>1,93 (0,3816)</b>	<b>0,45 (0,5028)</b>	<b>1,78 (0,4110)</b>
PPVDD bis	173,45 (<0,0001)	<b>1,14 (0,2866)</b>	<b>20,35 (0,4292)</b>

Tableau 2 : Test statistique sur les distributions marginales

Les résultats du tableau 3 sont relatifs à la préservation des distributions croisées entre les variables spécifiques observées et estimées. Ils confirment ceux sur la reconstitution des marginales pour les méthodes **PPV** et **PPVDD**. De plus, les approches **PPVDD bis**, **MUL** et **PML**, même si leurs niveaux de signification sont petits, offrent des statistiques du chi2 malgré tout nettement plus basses que les autres.

<i>Variables spécifiques</i> <i>Méthodes</i>	<i>Satisfaction×choix</i>	<i>Satisfaction×conseil</i>	<i>Choix×conseil</i>
LOGIT	576,36 (<0,0001)	632,76 (<0,0001)	98,60 (<0,0001)
SEQ	636,94 (<0,0001)	1079,52 (<0,0001)	859,27 (<0,0001)
MUL	325,95 (<0,0001)	352,52 (<0,0001)	51,10 (<0,0001)
PML	282,81 (<0,0001)	312,43 (<0,0001)	47,82 (<0,0001)
PPV	<b>1,45 (0,9190)</b>	<b>4,33 (0,8258)</b>	<b>1,16 (0,9484)</b>
PPVDD	9,05 (0,1070)	24,39 (0,0020)	<b>2,65 (0,7534)</b>
PPVDD bis	176,65 (<0,0001)	230,36 (<0,0001)	42,27 (<0,0001)

Tableau 3 : Test statistique sur les distributions croisées

Il est clair que ces résultats montrent que les méthodes peuvent se classer en deux groupes qui correspondent d'ailleurs aux types de méthodes exhibés dans les paragraphes 2.2 et 2.3 : modélisation et appariement. Le premier est caractéristique des approches possédant un bon pourcentage de bien classés, mais une reconstitution des marginales et des croisements très médiocre, alors que pour le second, c'est le contraire. Cependant, de façon globale et au vu des résultats, **MUL** et **PML** sont les seules méthodes à obtenir des résultats partagés entre les deux tendances. D'un point de vue plus général, l'approche univariée (**LOGIT**) a tendance à optimiser sur la prédiction, donc sur le pourcentage de bien classés. Mais les deux contraintes liées à l'adéquation des distributions marginales et croisées estimées aux distributions observées sont très nettement non prises en compte. Cependant, la régression logistique est très facile à mettre en oeuvre, car toute la chaîne d'inférence et, la validation statistique et opérationnelle sont bien balisées. L'approche séquentielle (**SEQ**) qui devrait pourtant être meilleure sur l'adéquation des distributions croisées ne l'est pas sur ce jeu de données. Cela peut s'expliquer par le fait que les résultats sont très dépendants de l'ordre dans lequel les variables spécifiques sont estimées. Mais cette méthode offre généralement des résultats corrects sur les prédictions. Les deux méthodes (**PPV** et **PPVDD**) sont quant à elles construites pour optimiser au mieux l'adéquation des distributions (marginales et croisées). Cependant, cette contrainte peut se révéler trop importante et pénaliser fortement la prédiction, ce qui entraîne mécaniquement des pourcentages de bien classés médiocres et donc non significatifs, voire en dessous du pourcentage d'un "modèle aléatoire". Une solution mixte a donc été mise en oeuvre avec

**PPVDD bis** et a permis d'améliorer très significativement les pourcentages de bien classés, tout en restant relativement correct sur les deux autres critères. La méthode **MUL** obtient les résultats les meilleurs sur les prédictions, ainsi que sur la reconstitution des marginales, mais moyen sur la préservation des distributions croisées. Enfin, la méthode **PML** est comparable en terme de résultats à **MUL**. Ils sont assez variables sur les pourcentages de bien classés, mais **PML** obtient de bons résultats sur la reconstitution des distributions, en particulier sur les distributions croisées. Elle offre aussi le grand avantage par rapport à **MUL** de ne pas être limitée par le nombre de variables communes. De plus, elle n'est pas limitée en nombre de variables spécifiques. Enfin, elle a le grand avantage de traiter ces variables avec la méthode d'estimation adéquate quand on sort du cadre linéaire gaussien habituel, et de tenir compte dans l'estimation de la distribution de probabilité des variables spécifiques.

#### 4. APPORTS, UTILISATIONS ET PERSPECTIVES

Comme on a pu le constater il existe de nombreuses méthodes de fusion statistique de données ayant des objectifs différents, principalement bien prédire vs bien reconstituer les distributions. On pourrait dire sous forme de boutade « on ne peut avoir le beurre et l'argent du beurre ». Cette constatation est notamment apparue grâce à la mise en œuvre de la validation statistique et de la validation opérationnelle. Et seules les deux approches multivariées (**MUL** et **PML**) fournissent un compromis, avec un sérieux avantage pour la seconde méthode qui n'est pas limitée en terme de variables candidates à l'explication, et permet, avec le maximum de vraisemblance partielle, de tenir compte de la nature de chaque variable à greffer, grâce à la possibilité de spécifier complètement la loi de distribution associée. L'utilisation potentielle de ces méthodes pour répondre à des problématiques d'enrichissement de bases de données des clients EDF ou autres, dépend principalement du nombre de variables à greffer, de leurs corrélations mutuelles et du pouvoir d'explication des variables communes (explicatives). Signalons enfin que cette étude a été réalisée sur un seul jeu de données, même si celui-ci est complètement opérationnel pour la problématique d'enrichissement de base de données de facturation clientèle, il faudrait des expériences supplémentaires sur d'autres jeux de données en vraie grandeur.

#### RÉFÉRENCES

- Aluja-Banet T., Ruis R. & Juarez C., (2002), Data Fusion by PLS Regression, XXXIV<sup>èmes</sup> Journées de Statistique, Bruxelles, Belgique.
- Derquenne Ch. (1999), A Method of Generating a Sample of Artificial Data from Several Existing Data Tables : Application Based on the Residential Electric Power Market, *Proceeding of Statistics Canada Symposium 99, Combining Data from Different Sources*.
- Derquenne Ch. (2003), A Multivariate Modelling Method based on the Partial Maximum Likelihood, *PLS'03 : 3<sup>rd</sup> International Symposium on PLS and Related Methods*.
- Derquenne Ch. (2004), Méthode des plus proches voisins améliorée - document de travail, EDF R&D.
- Fischer N., Derquenne Ch., Saporta G. (2001), A method to match data set applied to electric market, *ETK-NTTS 2001*, Creta.
- Santini G. (2002), La fusion des données et des fichiers, XXXIV<sup>èmes</sup> Journées de Statistique, Bruxelles, Belgique.
- Tenenhaus M. (2000), La Régression Logistique PLS, *Journées d'Etudes en Statistique, Modèles Statistiques pour données Qualitatives*, 261-273.
- Wold S., Martens H. & Wold H. (1983), The Multivariate Calibration Problem in Chemistry Solved by PLS Method, In *Proc. Conf. Matrix Pencils*, Ruhe A. & Kågström B. (Eds), March 1982, Lecture notes in mathematics, Springer Verlag, Heidelberg, 286-293.