Méthode multivariée de fusion statistique de fichiers de données appliquée au marché résidentiel de l'électricité

Christian Derquenne[1] . Nicolas Fischer[2] . Gilbert Saporta[3]

E-Mail : christian.derquenne@edf.fr, nicolas.fischer@edf.fr, saporta@cnam.fr

1. Contexte – Objectif

Dans le cadre d'une ouverture prochaine du marché électrique résidentiel à la concurrence, le Groupe EDF porte un effort accru à la connaissance de ses clients. Cet effort passe notamment par l'analyse de l'opinion à l'égard du chauffage électrique. Pour y parvenir l'entreprise mène un certain nombre d'enquêtes auprès de sa clientèle. Toutefois, il est bien évident que la taille de ces échantillons d'enquêtes reste sans commune mesure avec le nombre de clients de l'entreprise disponibles dans la base de facturation. Donc pour des raisons de temps, mais surtout de coût, le Groupe EDF ne peut disposer d'une grande base informatique de clients très bien renseignés. La solution consiste alors à rapprocher les données issues des différents fichiers externes (enquêtes, INSEE,..). Une solution raisonnable est de mettre en œuvre une approche par fusion statistique de fichiers de données. Celle-ci permet ainsi à partir de l'information commune aux différentes sources de données, d'estimer dans chacun des fichiers la distribution des variables spécifiques à l'autre fichier. Dans le cas présent, afin de tester les nouvelles méthodes développées sur de petits échantillons d'individus avant de les appliquer à l'ensemble de la base de données clientèle, une étude a été menée à partir de deux fichiers d'enquêtes (**SOFRES et CREDOC**) réalisées en 1990 sur deux échantillons distincts de quelques milliers d'individus. Chacun des deux fichiers comporte un certain nombre de variables spécifiques en plus de leurs variables communes, mais la présente étude concerne la satisfaction clientèle à l'égard du chauffage électrique. Nous n'avons donc retenu que les variables de satisfaction spécifiques au fichier SOFRES. Le but de l'étude sera donc d'estimer de nouvelles variables de satisfaction pour le **fichier CREDOC**. Ce dernier sera donc appelé le **fichier receveur** tandis que le **fichier SOFRES** sera le **fichier donneur.** Toutes ces variables sont ordinales, discrètes, codées sur 4 modalités.

2. Méthodes de fusion statistique mises en œuvre

2.1 Démarche du traitement

Avant même d'exposer brièvement la démarche mise en œuvre, il est important de rappeler l'exigence de qualité des données sans laquelle la modélisation n'a plus de sens.

Postulat : pour obtenir des résultats corrects et valides, il est absolument nécessaire qu'il y ait une structure explicative des variables à expliquer.

Toutefois, si la qualité des données constitue un pré requis pour obtenir des résultats valides, il n'existe pas de critère théorique de validation pour les méthodes de fusion de données. Seules des validations empiriques permettent d'évaluer la qualité des résultats. Par conséquent, pour tester les méthodes de fusion proposées, il va falloir les tester sur des échantillons complètement connus dont on aura préalablement masqué une partie de l'information. Ce qui revient à ne travailler, pour la phase de validation, que sur notre fichier donneur que l'on aura partagé en deux :

- Un fichier d'apprentissage qui servira de nouveau fichier donneur, sur lequel on construit la modélisation.

[1] EDF – Division Recherche et Développement, groupe SOAD, 92141 Clamart cedex
[2] EDF, groupe SOAD et CNAM, laboratoire CEDRIC
[3] CNAM, chaire de statistique appliquée, 75141 Paris cedex 03

- Un fichier test sur lequel sera appliqué les règles issues du fichier d'apprentissage fournissant ainsi des estimations pour les variables de satisfaction.

La validation consiste alors sur le fichier test, à vérifier la reconstitution des données observées à l'aide des estimations. Trois critères principaux de validation empiriques compte tenu des attentes liées aux méthodes de fusion ont été utilisés:

1. **Reconstitution des marginales** : comparaison des distributions marginales des variables observées et estimées (test d'indépendance du Chi2). Le nombre de degrés de liberté et la probabilité figurent dans le tableau de résultats à côté de la valeur du Chi2.
2. **Taux de bien classés** : mesure du nombre d'individus bien classés pour chaque variable, c'est à dire ayant la même modalité pour les variables observée et estimée (lecture des résultats: %bien-classés, signe "+", "-" ou "0" précisant si le Bcl est au dessus de la borne supérieure de l'intervalle, en dessous de la borne inférieure ou compris entre les deux).
3. **Reconstitution des corrélations** : comparaison des distributions croisées de deux variables estimées avec leurs deux variables observées (test d'indépendance du Chi2, même principe de lecture que pour les résultats des marginales).

2.2 De l'univarié au multivarié : 3 méthodes

Les méthodes de fusion statistique sont de deux types. Le premier regroupe des approches par appariement d'individus, alors que le second rassemble des méthodes par modélisation. C'est ce dernier type qui est développé dans le présent papier. La plupart des méthodes de fusion statistique existantes reposent sur une modélisation univariée de chacune des variables d'intérêt. Cette modélisation se faisant donc au détriment des liens pouvant exister entre ces différentes variables, l'idée sous-jacente de la nouvelle approche, développée ci-dessous, a été de mettre au point et d'expérimenter une modélisation multivariée.

2.2.1. Approche univariée

Considérons un ensemble de variables $Y_1,..Y_q..,Y_Q$, qualitatives ordinales ayant chacune $R_1,..Rq..,R_Q$ modalités, que l'on cherche à expliquer et un ensemble de variables communes aux deux fichiers $X_1,..X_j,..X_P$ de variables qualitatives (nominales, ordinales, booléennes) candidates à l'explication.

Une première approche naturelle consiste à estimer chacune des Q variables Y_q au moyen de Q modèles de régression logistique ordinale sur les différents X_j. Cette méthode est univariée dans le sens où chaque variable Y_q est modélisée indépendamment des autres variables *Y's*, mais elle permet de fournir un premier jeu de résultats indicatifs.

2.2.2. Approche hybride

Considérons toujours un ensemble de variables $Y_1,..Y_q,..Y_Q$, qualitatives ordinales ayant chacune $r_1,…r_Q$ modalités, que l'on cherche à expliquer et un ensemble $X_1,..X_j,..X_P$ de variables qualitatives (nominales, ordinales, booléennes) candidates à l'explication. En s'appuyant toujours sur la régression logistique ordinale, l'idée est toutefois de tenir compte dans la modélisation, des liens existant entre les différentes variables *Y*. La méthode est appelée "hybride" car elle n'est pas à proprement parlé multivariée. Dans un premier temps la démarche est la même que précédemment : chaque *Y* est estimée à l'aide d'un modèle de régression logistique ordinale sur les *X* candidates à l'explication. Mais au lieu de greffer chacune des *Y* indépendamment comme précédemment, C. Derquenne (1999) ne conserve que le modèle donnant le meilleur taux de bien classés pour la variable considérée (Y_k par exemple), 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 Y_k observée 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 seconde étape. On continue ainsi jusqu'à ce que toutes les variables *Y* soient greffées. Il y aura donc Q étapes.

Le net apport de cette méthode, par rapport au Logit ordinal univarié, est de tenir compte pour chaque futur *Y* à estimer, de l'éventuel pouvoir explicatif des autres *Y* déjà greffées.

2.2.3. Approche multivariée

Toutefois dans le cas de variables quantitatives, il est possible de modéliser conjointement le bloc des variables *Y* par régression PLS2 (Wold H. 1983) sur le bloc de variables communes. Dans le cas de variables qualitatives, un algorithme de régression logistique PLS1 a été développé par M. Tenenhaus (2000) mais n'a pour l'instant pas d'équivalent multivarié. En reprenant les notations précédentes: $Y_1,\ldots Y_q,\ldots,Y_Q$, variables ordinales à expliquer ayant respectivement $R_1,\ldots,R_q,\ldots,R_Q$ réponses possibles et $X_1,\ldots X_j,\ldots X_P$ variables qualitatives candidates à l'explication, l'objectif est d'estimer un modèle logit ordinal multivarié, c'est-à-dire fournissant l'estimation simultanée des *Y*'s. Les deux principales propriétés de ce modèle devant être la qualité de reconstitution de l'information (bon pourcentage de bien classés, respect des distributions marginales), et la préservation des corrélations entre variables à expliquer observées et estimées. Une solution raisonnable est d'utiliser la régression PLS2. Cependant, celle-ci est bien adaptée à des variables à expliquer numériques, alors que nous sommes en présence de variables qualitatives ordinales. Pour pallier cette différence, nous construisons tout d'abord des groupes d'individus à l'aide du croisement des modalités des variables candidates à l'explication *X*'s. On obtient *G* groupes notés : $v_1,\ldots, v_i,\ldots, v_G$, contenant respectivement n_i individus ayant les mêmes caractéristiques v_i, mais différentes sur les variables à expliquer. L'objectif est alors d'obtenir des *Y*'s adaptées et transformées en tenant compte de la nature de la variable ordinale. Pour cela, nous utilisons la fonction de lien naturelle d'une réponse ordinale qui est la fonction logit cumulée :

$$g(\hat{t}_{r(t)}^{(q)}) = \log\left(\Pr\left[Y_q \le r_q / t \in v_i\right] \Big/ \left(1 - \Pr\left[Y_q \le r_q / t \in v_i\right]\right)\right) \quad (1)$$

où r_q est la réponse à la variable Y_q. Cette fonction permet de revenir à une fonction linéaire des variables candidates à l'explication, c'est-à-dire :

$$g(m_{r(t)}^{(q)}) = a_{r_q} + \sum_{j=1}^{p}\sum_{l=1}^{m_j} a_{jl(t)} 1_{[t\in jl]}(jl) \quad (2)$$

où (1) est le logit cumulé théorique, alors que le logit observé est :

$$\log\left(\tilde{\Pr}\left[Y_q \le r_q / v_i\right] \Big/ \left(1 - \tilde{\Pr}\left[Y_q \le r_q / v_i\right]\right)\right) = \log\left(\sum_{s=1}^{r_q} n_{qi}^{(s)} \Big/ n_i - \sum_{s=1}^{r_q} n_{qi}^{(s)}\right) \quad (3)$$

Les probabilités cumulées théoriques prennent alors la forme suivante :

$$\Pr\left[Y_q \le r_q / v_i\right] = \exp\left(a_{r_q} + \sum_{j=1}^{p}\sum_{l=1}^{m_j} a_{jl}\right) \Big/ \left[1 + \exp\left(a_{r_q} + \sum_{j=1}^{p}\sum_{l=1}^{m_j} a_{jl}\right)\right] \quad (4)$$

On crée alors de nouvelles variables $\tilde{Y}_q$ à partir des logits observés.

$$\tilde{y}_{qi}^{(r)} = \log\left(\sum_{s=1}^{r_q} n_{qi}^{(s)} \Big/ \left(n_i - \sum_{s=1}^{r_q} n_{qi}^{(s)}\right)\right) \quad (r_q = 1 \text{ à } R_q) \quad (5)$$

Le nouveau tableau de données a toujours *n* individus, mais possède $\sum_{q=1}^{Q}(R_q - 1)$ variables, à la place des *Q* variables initialement. Ces nouvelles variables sont quantitatives. Elles peuvent être traitées avec la régression PLS2 classique dans laquelle les variables candidates à l'explication sont qualitatives, donc un modèle du type ANOVA. Les résultats obtenus sur chaque paquet R_q, représente les logits cumulés estimés $\hat{y}_{qi}^{(r)}$ associés à variable Y_q et au groupe v_i. Pour retourner aux probabilités de chaque réponse, il suffit de prendre la fonction logit inverse et de faire la différence entre deux quantités calculées successives. De façon formelle, nous avons :

$$\hat{\Pr}\left[Y_q \le r_q / v_i\right] = \exp\left(\hat{a}_r + \sum_{j=1}^{p}\sum_{l=1}^{m_j} \hat{a}_{jl}\right) \Big/ \left[1 + \exp\left(\hat{a}_r + \sum_{j=1}^{p}\sum_{l=1}^{m_j} \hat{a}_{jl}\right)\right] = \frac{\exp\left(\hat{y}_{qi}^{(r)}\right)}{1 + \exp\left(\hat{y}_{qi}^{(r)}\right)} \quad (6)$$

alors :

$$\hat{\Pr}\left[Y_q = r_q / v_i\right] = \hat{\Pr}\left[Y_q \le r_q / v_i\right] - \hat{\Pr}\left[Y_q \le r_q - 1 / v_i\right] \quad (7)$$

Le calcul des logits observés peut poser quelques problèmes. En effet, quand dans la formule (5) les $n_{qi}^{(1)} = 0$ et/ou $n_{qi}^{(R_q-1)} = n_i$, il y a des problèmes de bord car elle vaut $-\infty$ en r_q=1, alors qu'elle vaut $+\infty$ en r_q =R_q-1. Cependant, il est possible prendre en compte cette contrainte, en fixant une valeur arbitraire à ces deux bornes qui fasse que les probabilités cumulées soient proches de zéro et de l'unité respectivement (par exemple, $\tilde{y}_{q_i}^{(1)} = -15$ et $\tilde{p}_{q_i}^{(1)} = 3.10^{-7}$; $\tilde{y}_{q_i}^{(R_q-1)} = +15$et $\tilde{p}_{q_i}^{(R_q-1)} = 1-10^{-7}$).

3. Validation et Résultats

	Méthode **Univariée**		Méthode **Hybride**		Méthode **Multivariée**	
	Regl_sat	Gene_sat	Regl_sat	Gen_sat	Regl_sat	Gen_sat
Marg	231.7 2 0	389.86 2 0	231.77 2 0	2.60 2 0.2728	61.95 2 3.5E-14	162.71 2 0
Corr	480.35 8 0 460.44 8 0 765.64 8 0 458.12 8 0 406.05 8 0 299.72 8 0 655.48 8 0 693.32 11 0	524.70 8 0 899.71 8 0 480.35 8 0 590.18 8 0 540.73 8 0 448.38 8 0 767.29 8 0 799.65 11 0	230.71 8 0 262.50 8 0 288.18 8 0 369.60 8 0 322.49 8 0 285.54 8 0 257.39 8 0 440.58 11 0	116.56 8 0 156.52 8 0 230.71 8 0 235.46 8 0 136.12 8 0 102.57 8 0 180.58 8 0 315.49 11 0	333.52 8 0 399.42 8 0 363.02 8 0 272.28 8 0 202.24 8 0 208.28 8 0 333.67 8 0 1542.80 11 0	395.88 8 0 404.77 8 0 333.52 8 0 454.47 8 0 323.22 8 0 319.30 8 0 379.39 8 0 1525.40 11 0
Bcl	0.4674 ++ [0.3694;0.4216]	0.5661 0 [0.5297;0.5826]	0.4674 ++ [0.3694;0.4216]	0.7908 ++ [0.5296;0.5826]	0.3706 0 [0.3621;0.4113]	0.5267 -- [0.5379;0.5881]

Le tableau de résultats fait apparaître en premier lieu que la méthode dite hybride donne de meilleurs résultats sur l'ensemble des variables et ce sur les trois critères d'évaluation : la reconstitution des marginales des variables à transférer(1), le taux d'individus bien classés pour chaque variable fusionnée(2) et la reconstitution des corrélations entre variables transférées(3). Toutefois ces résultas par variable sont d'autant plus performants que les variables sont introduites au plus tard dans le modèle. Ainsi Regl_sat étant la première variable estimée, est finalement mieux reconstituée par la méthode multivariée tandis que les résultats sur Gen_sat, qui est modélisée après un certain nombre d'autres variables de satisfaction, sont bien meilleurs avec la méthode hybride qu'avec les deux autres méthodes. Deux autres méthodes sont actuellement en développement : la première est basée sur la recherche de plus proches voisins et l'autre sur la régression PLS2, adaptée au maximum de vraisemblance.

Références :

Derquenne C. (1999) *Une méthode pour construire un échantillon de données virtuelles à partir de plusieurs tableaux de données existants. Combiner des données de sources différentes*, recueil du Symposium 99 de Statistique Canada.

Fischer N., Derquenne C., Saporta G. (2002) "Méthode multivariée de fusion statistique de fichiers appliquée au marché résidentiel de l'électricité". *Journées de la Statistique 2002, Bruxelles.*

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

Tenenhaus M. (1998) *La régression PLS, théorie et pratique*; éditions TECHNIP.

Wold H. (1983) *Partial Least Square*, Encyclopedia of Statistical Sciences, vol. 6, Kotz S. & Johnson N.L. (Eds), John Wiley & Sons, New-York, pp. 581-591.