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

Nicolas FISCHER (CNAM-EDF)
Christian DERQUENNE (EDF)
Gilbert SAPORTA (CNAM)

Plan de la présentation

- Contexte de l’étude
- Définition de la fusion statistique
- Méthodes de références
- Présentation de nouvelles approches
- Validation
- Résultats et commentaires
- Perspectives

Contexte EDF

- Objectif: développer l'usage du chauffage électrique chez les clients résidentiels.
- Comment: par une meilleure compréhension du comportement de la clientèle ainsi que des divers acteurs (fournisseurs, constructeurs,…) engagés sur le marché résidentiel de l'électricité.
- Problème principal:l'ensemble de l'information concernant chaque client n'est pas disponible dans une unique base de données.

Information disponible

- Fichier de facturation, fichier receveur regroupant l'ensemble des clients, caractérisés par variables:
  - socio-démo (taille d'agglomération)
  - relatives au logement (type de logement,…)
  - relatives au client EDF (tarif,puissance,…)
- Fichiers d'enquêtes, fichier donneur (quelques milliers d'individus) comprenant:
  - Variables du fichier facturation.
  - Variables spécifiques, notamment d'opinions.

Données de l'étude

- Deux fichiers d'enquêtes, Sofres et Credoc 1990 comportant:
  - 8 variables communes qualitatives (booléennes, nominales, discrètes): âge, CSP, année du logement, type de logement, statut d'occupation, type de chauffage, taille d'agglomération.
  - Un certain nombre de variables spécifiques dont 9 variables de satisfaction.

Définition: fusion

- combinaison de données, provenant de sources différentes comprenant des données manquantes, pour obtenir un seul jeu de données dans lequel toutes les variables sont renseignées.
- cas particulier du traitement de données manquantes faisant intervenir diverses sources de données et dont les données manquantes apparaissent en bloc.

Deux grandes familles de méthodes

- la fusion par appariement d'individus
  - ➔Imputation de l'ensemble de l'information d'un individu donneur à un individu receveur après recherche de son plus proche voisin.
- le fusion par prévision de variables
  - ➔Estimation de chaque valeur manquante par modèles de régression, régression logistique,…

Première approche: Logit classique

- Toutes les variables de satisfaction sont estimées indépendamment les unes des autres.
- Régression logistique ordinale des Y sur les variables X candidates à l'explication.
- Greffe de chacune des variables au fichier receveur.

Greffe séquentielle de variables

- Chaque variable d'intérêt est estimée par régression logistique sur les variables candidates à l'explication.
- La variable la mieux expliquée est greffée au fichier donneur.
- 8 variables d'intérêt restantes estimées à nouveau par régression logistique avec la variable précédente comme variable explicative.
- La variable la mieux expliquée parmi les 8 dernières est greffée au fichier donneur.
- Le processus est itéré autant de fois qu'il y a de variables à greffer.

➔ Problème: préservation de la structure de corrélation des variables d'intérêt.

Rappel sur la régression PLS

- régression PLS [S. Wold]
  - Une alternative à la régression MCO et l'analyse canonique des corrélations (uni et multivarié Y) Résout des problèmes avec forte multicolinéarité.
  - **Composantes PLS**: combinaisons linéaires T des variables X maximisant simultanément la variance expliquée des Y et des X:

$$\max \left[V(T).\rho^2(Y,T) \right]$$

PCA et Régression

Algorithme: une **séquence de régressions simples**

- régression logistique PLS [M. Tenenhaus 2000]

  extension pour Y qualitative, utilisant la régression logistique à la place des régressions simples de l'algorithme.

Méthode Pseudo-PLS2

- Variables Y estimées indépendamment par régression logistique PLS sur les variables communes.
- Sélection de toutes les composantes PLS significatives obtenues à l'étape précédente.
- Nouvelles estimations des variables Y par régression logistique PLS sur les composantes PLS précédemment conservées.

Régression PLS2
recodage des données (Fdr 0/1)

- Pré-traitement des données

  But:transformer les variables Y afin de les considérer comme quantitatives pour appliquer l'algorithme de régression PLS2.

	Y1				Y2		
	$Y1_1$	$Y1_2$	$Y1_3$	$Y1_4$	$Y2_1$	$Y2_2$	$Y2_3$
Y1=1	1	0	0	0			
Y1=2	**0**	**1**	**0**	**0**			
Y1=3	**0**	**0**	**1**	**0**			
Y1=4	**0**	**0**	**0**	**1**			
Y2=1					1	0	0
Y2=2					0	1	0
Y2=3					0	0	1

	Y1				Y2		
	$Y1_1$	$Y1_2$	$Y1_3$	$Y1_4$	$Y2_1$	$Y2_2$	$Y2_3$
Y1=1	1	0	0	0			
Y1=2	**1**	**1**	**0**	**0**			
Y1=3	**1**	**1**	**1**	**0**			
Y1=4	**1**	**1**	**1**	**1**			
Y2=1					1	0	0
Y2=2					1	1	0
Y2=3					1	1	1

Régression PLS2 recodage des données (Logit ordinal)

- On note $Y_1, \ldots Y_q, \ldots, Y_Q$, variables ordinales à expliquer ayant respectivement $R_1, \ldots, R_q, \ldots, R_Q$ réponses possibles
- Construction de groupes d'individus à l'aide du croisement des modalités des variables candidates à l'explication.

  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.
- But: obtenir un nouveau jeu de variables Y' adaptées tenant compte du caractère de variable ordinale.
- Utilisation de la fonction de lien Logit cumulé:

$$g(\mu_{r(t)}^{q)}) = \log\left(\frac{\Pr[Y_q \le r / t \in v_i]}{1 - \Pr[Y_q \le r / t \in v_i]}\right)$$ où r est la réponse à la variable Y_q

Régression PLS2
recodage des données (Logit ordinal)

- Création de nouvelles variables quantitatives issues des logits cumulés observés.

$$\tilde{y}_{qi}^{(r)} = \log\left(\frac{\sum_{s=1}^{r_q} n_{qi}^{(s)}}{n_{i-} - \sum_{s=1}^{r_q} n_{qi}^{(s)}}\right) \qquad \text{avec } r_q = 1 \text{ à } R_q$$

- Modélisation de ces nouvelles variables (*on en a désormais* $\sum_{q=1}^{Q}(R_q - 1)$ *au lieu de* Q) par régression PLS2 sur les variables candidates à l'explication.

- Obtention sur chaque ensemble R_q des logits cumulés estimés $\hat{y}_{qi}^{(r)}$ associés à la variable Y_q et au groupe v_i.

Régression PLS2
recodage des données (Logit ordinal)

- On retrouve les probabilités de chaque réponse en utilisant la fonction logit inverse et en faisant la différence entre deux quantités calculées successives:

$$\hat{\Pr}[Y_q \leq r_q / v_i] = \frac{\exp(\hat{y}_{qi}^{(r)})}{1+\exp(\hat{y}_{qi}^{(r)})}$$

alors $\hat{\Pr}[Y_q = r_q / v_i] = \hat{\Pr}[Y_q \leq r_q / v_i] - \hat{\Pr}[Y_q \leq r_q - 1 / v_i]$

Validation

- Validation empirique effectuée sur le fichier donneur, séparé en fichier d'apprentissage et fichier test.
- 3 critères de validation globaux:
  1. Reconstitution des distributions marginales.
  2. Taux de bien classés.
  3. Reconstitution des distributions croisées (corrélations) de deux variables à estimer.
- Critère individuel:niveau de confiance associé à l'individu, résultant du taux de bien classés.

		Greffe univariée			Logit PLS2			Tenenhaus			Logit PLS1			Pseudo-PLS2			Logit Classique		
Gene	marg	5				5										5			
	corr	8				7							2			6		1	
	bcl	5					3		1	1		2			1			1	1
Nivt	marg	4	1		1	3										5		1	
	corr	8				7							2			6		1	
	bcl	5					4		2						3				1
Esth	marg	3	2		2	3							4			1			
	corr	8				7							4			4		1	
	bcl	5					4		1	1		1			3				
Regl	marg		2ex		2	2		3	1				2			3		2ex	
	corr	5	3		3	4			1				3			5			
	bcl	1ex		1ex		1	3	2		1	2	3			1		1ex		1ex
Rapi	marg	4	1		1	3							3			2		1	
	corr	7	1		1	6							2			6		1	
	bcl	5				1	3		2	1		1			1				1
Empl	marg	2	3		3	2							4			1			
	corr	7	1		1	6							3			5		1	
	bcl	5					1		1	3		2						2	1
Sécu	marg	4	1					1					3			2		4	
	corr	7	1		1	6							1			7		1	
	bcl	5				1	2		2	2					1			1	1
Humi	marg	5				5										5			
	corr	8				7							4			4		1	
	bcl	5					5					2			2			1	
Coût	marg	5							3				2			3		2	
	corr	8							1				3			5		7	
	bcl	5					5		2			1			2				
Total	**marg**	**32**	**10**		**9**	**23**		**4**	**4**				**18**			**27**		**10**	
	corr	**66**	**6**		**6**	**50**			**2**				**24**			**48**		**14**	
	bcl	**41**		**1**		**3**	**30**	**2**	**11**	**9**	**2**	**12**			**14**		**1**	**5**	**6**
		84.13	10.32	0.79	9.52	40.48	23.81	4.76	12.70	7.14	1.59	9.52	23.81	0.00	11.11	40.48	0.79	17.46	4.76
Tot		**106**	**13**	**1**	**12**	**51**	**30**	**6**	**16**	**9**	**2**	**12**	**30**	**0**	**14**	**51**	**1**	**22**	**6**

Tableau récapitulatif

		Greffe univariée			**Logit PLS2**			Recodage PLS2			Logit PLS1			Pseudo-PLS2			Logit Classique		
Gen	mar	5				5										5			
	corr	8				7							2			6		1	
	bcl	5					3		1	1		2			1			1	1
Reg	mar		2ex		2	2		3	1				2			3		2ex	
	corr	5	3		3	4			1				3			5			
	bcl	1ex		1e		1	3	2		1	2	3			1		1e		1e
Coû	mar	5							3				2			3		2	
	corr	8							1				3			5		7	
	bcl	5					5		2			1			2				
Tot	**mar**	**32**	**10**		**9**	**23**		**4**	**4**				**18**			**27**		**10**	
	cor	**66**	**6**		**6**	**50**			**2**				**24**			**48**		**14**	
	bcl	**41**		**1**		**3**	**30**	**2**	**11**	**9**	**2**	**12**			**14**		**1**	**5**	**6**
Tot		**106**	**13**	**1**	**12**	**51**	**30**	**6**	**16**	**9**	**2**	**12**	**30**	**0**	**14**	**51**	**1**	**22**	**6**

Résultats et commentaires

- Greffe séquentielle donne de loin les meilleurs résultats.
- Seule la méthode du logit PLS2 donne des résultats comparables pour certaines variables sur les critères (1) et (3).
- Limites imposées par la taille des fichiers de données et le pouvoir explicatif des variables communes.

Perspectives

- Mise en oeuvre de ces méthodes sur le fichier de facturation EDF (expérimentation)
  - Effort de pédagogie pour l'utilisation des résultats au niveau des centres régionaux EDF.
- Comparaison aux méthodes d'imputation basées sur les plus proches voisins.