MISE EN ŒUVRE D'UNE DEMARCHE STATISTIQUE COMPLETE POUR LA PREDICTION DE VARIABLES DANS UNE BASE DE DONNEES CLIENTELE D'EDF

Christian Derquenne

EDF – Division Recherche et Développement
1 avenue du Général de Gaulle
92141 CLAMART Cedex

Cet article fait partie des Actes des 5èmes JOURNEES MODULAD qui ont eu lieu les 16 et 17 novembre 2000 à E.D.F. (Clamart).

1. Contexte, historique et objectif

Mieux connaître les attentes et les besoins de la clientèle est un des axes stratégiques majeurs d'EDF. Cette meilleure connaissance est notamment obtenue grâce à l'analyse statistique de différents types de données (consommation d'électricité, fidélisation, nouveaux clients, qualité de fourniture, enquêtes de satisfaction, ...). Pour cela, EDF possède plusieurs bases de données (fichiers de facturation, enquêtes d'opinion, ...) sur différents segments de clientèle (résidentielle, professionnelle, PME-PMI, grandes entreprises). L'objectif des études peut être par exemple de prédire l'énergie de chauffage ou d'identifier les clients mécontents.

La mise en œuvre de ce type d'analyse a fait l'objet d'un projet qui a débuté en 1998 à la Division Recherche et Développement d'EDF. A l'origine, ce projet avait pour but d'évaluer la faisabilité et l'intérêt de l'application des techniques de Data Mining sur les différentes données de la clientèle d'EDF. En résultat du projet, on attendait par conséquent, des

stratégies d'analyse et des propositions de mise en œuvre opérationnelle, pour l'utilisation de ces données dans un but d'une meilleure connaissance du client.

Pour atteindre ces objectifs, ce projet était constitué de quatre phases :

① un recensement des données disponibles sur la clientèle résidentielle d'EDF,

② une analyse des besoins,

③ des expérimentations sur des données clientèle d'EDF,

④ un dossier décisionnel sur la validité et l'éventuelle poursuite du projet.

Lors de l'année 1998, nous avons construit une démarche statistique que nous avons appliquée sur un « Centre EDF-GDF SERVICES » pilote. Différentes analyses ont été réalisées :

- l'enrichissement de certains champs de la base de facturation d'EDF (énergie de chauffage du logement, énergie de chauffage de l'eau sanitaire, ...),
- propension à demander des services concernant le code paiement (prélèvement automatique ou non) et la politique de facturation (mensualisation ou non),
- propension à modifier dans l'avenir le système de chauffage du logement, et si oui quel type de modification (modernisation, installation d'un autre moyen de chauffage en complément ou abandon de l'énergie de chauffage),
- propension à déménager dans les 10 ans à venir,
- propension à être satisfait ou non du chauffage électrique.

Fin 1998, il a été décidé de poursuivre le projet. L'année 1999 a alors été consacrée à la validation de l'ensemble de la démarche sur un autre Centre pilote, à l'amélioration de la démarche statistique et à l'application de celle-ci sur six nouveaux centres de clientèle résidentielle. Les résultats obtenus ont alors permis de décider de mettre en œuvre un

déploiement « industriel » sur 22 centres en 2000 et sur le reste des centres (70 centres) en 2001.

L'objectif de ce papier est de présenter la démarche statistique mise en œuvre dans le cadre de ce projet, et de fournir quelques résultats de l'application d'une telle démarche. Les apports, les critiques et les voies d'amélioration font l'objet de la conclusion.

2. Démarche statistique mise en oeuvre

Parmi, les nombreuses analyses statistiques réalisées pour ce projet, nous avons choisi de décrire la démarche statistique concernant l'enrichissement d'une donnée clientèle que l'on nommera *Y*, pour des raisons de confidentialité. En effet, la base de données de facturation d'EDF est constituée de plusieurs tables (CLIENT, LOCAL, CONTRAT, USAGE, ...). Certaines colonnes de ces tables sont renseignées à 100%, car elles correspondent à des données indispensables à la facturation (type de tarif, puissance souscrite, date de création du client, ...), alors que d'autres sont relatives à des informations supplémentaires (énergie de chauffage du logement, énergie de l'ECS, type de logement, ...). Ces dernières informations sont très utiles à EDF pour des actions marketing. L'enrichissement de telles variables doit se faire de la façon la plus rigoureuse possible, afin d'offrir une qualité et une validité statistique raisonnables.

Dans un premier temps nous décrivons brièvement les neuf étapes de la démarche statistique, alors que dans un second temps nous nous focalisons sur certaines d'entre elles, que nous avons jugées primordiales pour la bonne réussite d'une telle démarche. Afin d'avoir une approche la plus pédagogique possible, chaque étape sera discutée méthodologiquement et sera suivie immédiatement de l'application illustrée par des résultats. Signalons que l'ensemble des neuf étapes a été programmé avec le logiciel *SAS* produit par *SAS Institute*, sous forme de macros *SAS* (sous-programme *SAS* permettant d'automatiser certaines tâches à effectuer).

La démarche statistique adoptée est la suivante :

(E1) Importation des données de la base de facturation de la clientèle résidentielle d'EDF sous forme d'un certain nombre de fichiers dits de « transport *SAS* » en tables *SAS*. Celles-ci correspondent aux informations dont nous avons besoin pour mettre en œuvre le processus d'enrichissement de champs peu renseignés.

(E2) Fusion de ces tables *SAS* en une seule table *SAS*.

(E3) Construction du plan de sondage post-stratifié.

(E4) Redressement de l'échantillon renseigné.

(E5) Constitution de l'échantillon d'apprentissage en vue de la modélisation de la variable à enrichir à l'aide de la régression logistique.

(E6a) Sélection des variables « explicatives » les plus significatives.

(E6b) Modélisation sur l'ensemble réduit des variables sélectionnées précédemment, avec éventuellement simplification en regroupant les modalités ayant le même apport statistique pour chaque variable explicative.

(E7) Validation du modèle statistique sur l'échantillon test en appliquant les règles précédemment construites à partir de l'échantillon d'apprentissage.

(E8) Enrichissement des données non renseignées de la variable à prédire sur la population totale à l'aide des règles issues de l'étape **E6b** et fourniture d'un niveau de confiance précédemment construit.

(E9) Exportation de la nouvelle table de données ainsi obtenue dans un format permettant le reversement dans la base de facturation d'EDF.

Seules les étapes **E3** à **E8** sont décrites, les autres sont de la gestion informatique.

2.1. Etapes E3 et E4 : Structuration de la population à étudier

La variable Y possède deux modalités A et B. Cette variable dans la population des clients n'est renseignée qu'à 13,4% (22004 sur 1643548). Le tableau suivant montre comment se répartissent les catégories A et B.

	Echantillon renseigné		*Non renseigné*		*Population*	
Y	Nb de clients (n)	%	Nb de clients (N - n)	%	Nb de clients (N)	%
A	3810	17,3	?	?	?	?
B	18194	82,7	?	?	?	?
Total	22004	100,0	142354	100,0	164358	100,0

Tableau 1 : ***Répartition de A et B sur l'échantillon renseigné***

Dans un tel cas, un premier objectif peut être d'estimer la distribution sur la population entière de Y, c'est-à-dire les proportions $\hat{p}_A$ et $\hat{p}_B$. Rappelons que la proportion inconnue de A est :

$$p_A=\frac{1}{N}\sum_{i\in U} y_i \times \mathbb{1}_{[i\in A]} \qquad (1)$$

où U représente l'univers, c'est-à-dire la population des 164358 clients et y_i prend la valeur 1 si le client i possède la caractéristique A, 0 sinon (c'est-à-dire quand i a la caractéristique B).

La solution au problème d'estimation la plus naturelle est alors de considérer que l'on est en présence d'un plan de sondage simple sans remise. Dans ce cas, la proportion de A peut être estimée par :

$$\hat{p}_A=\frac{1}{n}\sum_{i\in S} y_i \times \mathbb{1}_{[i\in A]}=\frac{3810}{22004}=0,173 \qquad (2)$$

où S représente le sondage, c'est-à-dire l'échantillon renseignée des 22004 clients.

La variance associée vaut $5,636\times10^{-6}$.

Cependant, ce type de plan est seulement intéressant si aucune information sur la population n'est disponible. En effet, les clients pour lesquels un champ est renseigné dans la population des clients n'ont pas obligatoirement les mêmes comportements, les mêmes caractéristiques, etc., que ceux pour lesquels cette information est manquante. Il peut donc apparaître un déséquilibre entre certains champs (par exemple, sous-estimation ou surestimation du nombre de clients « chauffage électrique »). Par contre, si nous avons des informations supplémentaires (nommées habituellement informations auxiliaires), il est possible d'en tenir compte pour construire un plan de sondage plus judicieux [TIL97], tels que la stratification, les plans par grappe ou encore les plans à deux degrés. Les informations auxiliaires peuvent être, par exemple, des moyennes, des proportions, des totaux, des variances ou encore les valeurs d'un caractère sur toutes les unités du plan de sondage.

Dans notre étude, nous avons choisi le plan de sondage stratifié car notre objectif était d'améliorer la précision de l'estimateur. En effet, les proportions estimées doivent être les plus fines possibles, car elles servent pour les agents commerciaux et peuvent influencer le plan stratégique des Centres EDF-GDF SERVICES. De plus, comme notre connaissance sur les informations auxiliaires est a posteriori et puisque les données renseignées n'ont pas été construites selon un plan de sondage stratifié comme dans le cas d'une enquête, on utilisera un plan de sondage post-stratifié. Les informations auxiliaires seront dans notre cas des variables présentes dans l'ensemble de la population. En fait, ce plan de sondage va nous permettre de structurer la population étudiée, afin d'estimer des proportions dans un premier temps, et de construire un modèle statistique dans un second temps, pour extrapoler les résultats obtenus à la population entière. La correction du déséquilibre discuté précédemment se fait alors grâce au redressement. Il consiste à calculer un « poids » pour chaque client en fonction de certaines caractéristiques personnelles (type de tarif, type de logement, etc.).

L'oubli d'un tel plan, et donc d'une absence de redressement, pourrait provoquer un biais sérieux sur les résultats et par voie de conséquence sur l'enrichissement de la base de données des clients.

La méthode consiste à choisir des variables auxiliaires les plus indépendantes, discriminantes

et significatives possibles pour stratifier la population en fonction du problème posé[1].

Ces variables doivent être également renseignées à 100% dans la table de données à analyser, afin d'avoir une information complète sur les distributions réelles de celles-ci. De plus, le découpage en strates marginales pour chaque variable auxiliaire est complètement dépendant de la population étudiée. Elles doivent répondre à deux critères souvent antagonistes : une répartition relativement équilibrée (en tous cas, éviter de basses fréquences) et des bornes interprétables et réalistes pour l'agent commercial dans le Centre EDF-GDF SERVICES. Il est par conséquent délicat de vouloir complètement automatiser cette étape.

Les variables choisies sont les suivantes :

- *la date de création du contrat* permet non seulement de faire apparaître l'ancienneté du lien entre le client et EDF, mais aussi d'avoir une idée, bien qu'approximative, de l'âge du client (plus la date de création est ancienne, plus l'intervalle des âges possibles est étroit),

- *le type de logement* (individuel ou collectif) dans lequel il habite, nous fournit des renseignements sur son environnement proche (caractéristiques sociales),

- *le nombre de clients dans la ville d'habitation*, nous donne des informations sur son environnement plus large (caractéristiques démographiques),

- *le tarif croisé avec la puissance souscrite* exhibe son lien avec EDF.

En fait, ces quatre variables permettent de faire apparaître à la fois les caractéristiques personnelles et relativement exhaustives du client, ainsi que des caractéristiques économiques entre lui et EDF. Les quatre tableaux suivants fournissent les distributions des variables auxiliaires introduites précédemment.

---

[1] Techniquement, cela revient à construire un hyperpallélépipède dont chaque dimension représente une variable auxiliaire. Chaque dimension est découpée en un certain nombre de segments correspondant aux modalités de la variable. Puis chaque pavé multidimensionnel de l'hyperpallélépipède symbolise une strate. Enfin, la densité de chaque pavé exhibe la proportion de clients associée à chaque strate par rapport à l'ensemble des clients.

Année de création	*Nombre de clients*	*%*
1970 et avant	35333	21,5
1971 à 1985	46233	28,1
1986 à 1996	49583	30,2
1997 à 1999	33209	20,2
Total	**164358**	**100,0**

Tableau 2 : ***Répartition de l'année de création sur la population***

Type de logement	*Nombre de clients*	*%*
Collectif	33191	20,2
Individuel	131167	79,8
Total	**164358**	**100,0**

Tableau 3 : ***Répartition du type de logement sur la population***

Nombre de clients dans la ville	Nombre de clients	*%*
Moins de 3000	46022	28,0
3000 à moins de 5000	34226	20,8
5000 à moins de 20000	54615	33,3
20000 et plus	29495	17,9
Total	**164358**	**100,0**

Tableau 4 : ***Répartition du nombre de clients dans la ville sur la population***

Tarif×Puissance en kVA	Nombre de clients	*%*
Simple et (1 à 5)	17124	10,4
Simple et (6 à 8)	58010	35,3
Simple et (9 et plus)	10895	6,6
Double et (8 et moins)	38002	23,1
Double et (9 et plus)	40327	24,5
Total	**164358**	**100,0**

Tableau 5 : ***Répartition tarif×puissance sur la population***

Le croisement des quatre variables auxiliaires permet de construire H=160 strates. Chacune a un poids dans la population (généralement différent) qui est égal à $w_h=N_h/N$ (h=1,H), où N

correspond à la taille de la population (= 164358) et N_h est relatif au nombre de clients dans la strate croisée S_h. Par exemple, la strate S_h contenant les clients récents (1997 et après), ayant souscrit un double tarif avec une puissance supérieure ou égale à 8 kVA, vivant en logement collectif situé dans une ville de 5000 à moins de 20000 habitants est de taille $N_h = 1911$, alors $w_h = 1911/164358 = 0{,}0116$ (1,16%). Enfin, la somme des w_h est égale à l'unité.

Le redressement s'effectue sur les données renseignées de la variable Y, à l'aide d'un poids calculé de la façon suivante : $w_h=(N_h/N)/(n_h/n)$ où n et n_h sont respectivement le nombre total de données renseignées[2] et le nombre de données renseignées dans la strate h, alors $\sum_{h=1}^{H}\sum_{i\in S_h} w_{hi}=n$.

La proportion inconnue de A dans la population est bien évidemment la même qu'en (1), mais elle peut se décomposer également sous forme de strates :

$$p_A=\frac{1}{N}\sum_{i\in U} y_i\times 1_{[i\in A]}=\frac{1}{N}\sum_{h=1}^{H}\sum_{i\in S_h} y_i\times 1_{[i\in A]}=\frac{1}{N}\sum_{h=1}^{H} N_h p_h^A \qquad (3)$$

alors que la proportion estimée à l'aide de l'estimateur post-stratifié est égale à :

$$\hat{p}_{post}^A=\frac{1}{n}\sum_{h=1}^{H}\sum_{i\in S_h} w_{hi}\, y_i\times 1_{[i\in A]}=\frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i\in S_h} y_i\times 1_{[i\in A]}=\frac{1}{N}\sum_{h=1}^{H} N_h \hat{p}_h^A=\sum_{h=1}^{H} w_h \hat{p}_h^A \qquad (4)$$

Les répartitions respectives de Y selon les deux plans sont :

Y	A		B		
Plan	n	%	n	%	Variance
Simple	3810	17,3	18194	82,7	$5{,}636\times10^{-6}$
Post-stratifié	3134,5	14,3	18869,5	85,7	$2{,}792\times10^{-6}$

Tableau 6 : *Effet du plan post-stratifié*

On peut constater tout d'abord que la variance post-stratifiée est deux fois plus petite que la variance du plan simple (gain de précision escompté). De plus, bien que les pourcentages estimés de A (resp. de B) dans les deux plans soient assez proches, les deux intervalles de confiance à 95% associés ne se recouvrent pas.

[2] Il représente en quelque sorte l'échantillon au sens de la théorie des sondages.

Ce plan de sondage va également servir pour un second objectif relatif à la modélisation de Y, à l'aide de la régression logistique dans les étapes **E6a** et **E6b**.

2.2. Etape E5 : Constitution de l'échantillon d'apprentissage et de l'échantillon test

Quand on applique une méthode telle que la régression logistique, l'usage est de découper la population en deux échantillons distincts. Le premier correspond à l'« échantillon d'apprentissage », sur lequel on modélise les données et on construit des règles d'affectation d'un individu en fonction de ses caractéristiques à une des réponses de la variable à expliquer, c'est-à-dire la variable Y dans notre cas. Le second groupe est nommé « échantillon test » ; il a pour objectif de vérifier si le modèle construit sur l'échantillon d'apprentissage est valide statistiquement ou pas, en appliquant les règles d'affectation construites précédemment.

La constitution de ces échantillons ne pose pas de problème particulier. Le tirage s'effectue aléatoirement et sans remise dans chaque strate parmi les clients renseignés, avec un certain taux (73%, ici). Le reste des clients est mis de coté pour l'échantillon test.

2.3. Etape E6a et E6b : Construction du modèle statistique

La régression logistique[3] [HOS89] a pour objectif d'estimer, par exemple, la probabilité de posséder le caractére A, sachant que le client appartient à un groupe[4] de caractéristiques « significatives » (type de logement, puissance souscrite, code paiement, ...). Cette méthode permet d'éliminer les effets de structure entre les variables candidates à l'explication et de raisonner « toutes choses égales par ailleurs » sur ces mêmes effets.

Choix a priori des variables candidates à l'explication

Dans notre démarche, nous avons privilégié les variables dites **opérationnelles**, en ce sens qu'elles sont complètement renseignées dans les bases de données clientèle et à un degré

---

[3] Nous ne développons pas ici la méthode de la régression logistique, car elle est très bien expliquée dans de nombreux ouvrages de référence.

[4] On distinguera strate et groupe. Une strate est construite à partir du plan de sondage sans remise post-stratifié, alors qu'un groupe est déterminé par le croisement des modalités de variables « explicatives ».

moindre les variables **semi-opérationnelles** qui sont partiellement disponibles. Les exemples type de variables opérationnelles sont la puissance souscrite, le type de contrat ou la qualité payeur du client, alors que le type de logement entre plutôt dans la seconde catégorie de variables.

Dans notre cas, nous avons gardé le type de logement comme variable auxiliaire et comme variable candidate à l'explication, car elle était renseignée à un peu plus de 99%.

A l'inverse, d'autres variables absentes dans les bases de données clientèle peuvent néanmoins avoir une influence sur les variables à prédire. On parlera a contrario ici de variables **externes**.

- Dans certains cas, ces données peuvent être disponibles dans le Centre, dans des fichiers complémentaires (comme la description des communes par les fichiers INSEE). La fréquence de mise à jour de ces fichiers peut être plus rare, et la clé de fusion imprécise, de sorte que leur exploitation est parfois plus périlleuse.

- Au dernier niveau de l'échelle, on trouve des variables qui, sauf dans certaines enquêtes ponctuelles, n'appartiennent pas au système global d'information du Centre (la profession ou l'âge du chef de famille), mais que les Services commerciaux pourraient envisager de se procurer. Dans ce cas là, la pertinence statistique de la variable doit être comparée à son coût d'acquisition.

Choix du « bon » modèle

La difficulté quand on construit un modèle statistique est notamment de déterminer le « meilleur » modèle. L'analyse statistique nous fournit néanmoins des critères quantitatifs pour identifier des variables explicatives. En clair, des tests statistiques à un risque fixé, examinent si le pouvoir explicatif d'une variable est significatif. D'autre part, ce pouvoir explicatif est quantifiable ce qui nous permet de classer les variables les plus explicatives. Dans la pratique, la taille de l'échantillon test limite le nombre de variables de sorte qu'on utilise leur pouvoir explicatif pour ne conserver que les variables les plus pertinentes.

Nous avons par conséquent, procédé en deux temps. Tout d'abord, nous avons sélectionné les variables les plus « explicatives » (étape **E6a**), puis nous avons simplifié ce modèle réduit, le cas échéant en regroupant des modalités des variables « explicatives » restantes (étape **E6b**).

Dans l'étape **E6a**, nous avons étudié tout d'abord l'adéquation globale du modèle, puis la comparaison de modèles et enfin l'apport marginal des variables candidates à l'explication.

- Pour l'adéquation du modèle, nous nous sommes servis de différents coefficients de détermination plus particulièrement développés pour le modèle logit : (cf. [DER97], [HOS89] et [MCF74]) ainsi que de tests sur des statistiques classiques (Wald, Score et logarithme du rapport des vraisemblances).

- Pour la comparaison de modèle, nous avons utilisé des critères d'information standards (R^2 ajusté, *AIC* et *BIC)* [HOS89].

- Enfin, concernant l'apport marginal des variables « explicatives », nous avons employé des tests classiques [HOS89], ainsi qu'un test et trois coefficients de détermination développés par l'auteur [DER97]. Les tests de cette dernière partie permettent en première approche de hiérarchiser l'apport des variables candidates à l'explication, et éventuellement de sélectionner les plus significatives. Cependant, ce n'est pas parce que l'on hiérarchise les variables par niveau de signification croissant (risque associé au test statistique, sous forme d'une probabilité), que l'on obtient les meilleurs résultats. En effet, le nombre de modalités associées à chacune d'elle peut être élevé, ce qui mécaniquement produit beaucoup de groupes. Par conséquent, il est quelque fois préférable de choisir une variable un peu moins « explicative », mais avec moins de modalités. Pour pallier ce problème, nous avons construit un test statistique[5] pour savoir si une variable placée avant une autre explique significativement plus qu'elle. Si ce n'est pas le cas, l'une des deux variables peut être choisie.

Les variables candidates à l'explication choisies avant la sélection sont : *l'année de création, le type de logement, la politique de facturation, le tarif×puissance souscrite, la qualité du*

[5] la forme de ce test n'est pas donnée ici.

payeur, le code paiement et le nombre de clients dans la ville.

Nous avons effectué une première sélection permettant de ne garder que les cinq premières de la liste. Puis, nous avons réalisé une seconde sélection à l'aide de certains tests précédemment discutés.

Tout d'abord, les cinq modèles sont significativement acceptables selon le test du rapport des vraisemblances *RV*. De plus, le deuxième modèle « **Complet – {qualité payeur}** » a les plus petits *AIC* et *BIC*, et le plus grand R^2 ajusté, ce qui peut nous inciter à choisir ce sous-modèle.

Modèle	*RV*	*AIC*	*BIC*	R^2 ajusté
Complet	5355,6 [21] (≅0)	1476,9	1667,8	0,7812
Complet – {qualité payeur}	5226,0 [12] (≅0)	378,3	491,2	0,9339
Complet – {tarif×puissance}	5175,6 [17] (≅0)	605,6	761,9	0,8962
Complet – {politique de facturation}	5130,4 [18] (≅0)	642,5	807,4	0,8896
Complet – {type de logement}	4510,3 [20] (≅0)	1065,6	1274,9	0,8080
Complet – {année de création}	3727,3 [18] (≅0)	607,5	772,4	0,8611

Tableau 7 : *Adéquation globale du modèle et comparaison de modèles*

Ce choix est d'ailleurs confirmé par les résultats du tableau 8, où l'on peut constater que la qualité payeur est la variable la moins explicative, bien qu'elle le soit encore d'après le test du rapport des vraisemblances *D(X)*.

Variable « explicative »	*D(X)*	R^2 partiel
Année de création	1628,4 [3] (≅0)	0,7812
Type de logement	845,2 [1] (≅0)	0,9339
Politique de facturation	225,1 [3] (≅0)	0,8962
Tarif×puissance	180,0 [4] (≅0)	0,8896
Qualité payeur	129,5 [9] (≅0)	0,8080

Tableau 8 : *Apport de chaque variable explicative*

Enfin, le tableau 9 fournit les niveaux de signification du test de comparaison des apports de chaque variable « explicative » les unes par rapport aux autres.

Variable « explicative »	**An. Créa.**	**Type log.**	**Pol. Fact.**	**Tar×pui**	**Q. pay.**
Année de création	1	(≅0)	(≅0)	(≅0)	(≅0)
Type de logement		1	(≅0)	(≅0)	(≅0)
Politique de facturation			1	0,3309	0,1577
Tarif×puissance				1	0,2857
Qualité payeur					1

Tableau 9 : *Comparaison mutuelle de l'apport des variables explicatives*

Les résultats montrent que l'on peut distinguer trois classes de variables, en prenant comme critère de regroupement un niveau de signification supérieur à 5%.

C_1 = {*Année de création*}

C_2 = {*Type de logement*}

C_3 = {*Politique de facturation*, *Tarif×puissance*, *Qualité payeur*}

En d'autres termes, autant l'année de création et le type de logement ont des apports significativement différents, autant les trois variables de la classe C_3, ont le même apport significatif sur la construction du modèle.

Par conséquent, le modèle que nous avons sélectionné contient seulement les variables « explicatives » suivantes : *année de création, type de logement, politique de facturation et tarif×puissance souscrite*.

Nous pouvons maintenant passer à l'étape **E6b** pour construire un modèle sur ces quatre variables précédemment retenues.

Les résultats de la régression logistique sont les suivants :

Variable	Paramètre	Wald	pr>Wald
Constante	-3.7094	20.7423	0.0001
Log. collec.	1.8362	28.6534	0.0001
Log. indivi.	0	.	.
1971 à 1985	-0.2423	1.9151	0.0556
1986 à 1996	0.2684	2.4031	0.0164
1997 à 1999	2.5275	23.9969	0.0001
Avant 1970	0	.	.
DT <=8 kVA	0.7934	4.9012	0.0001
DT >=9 kVA	-0.3830	2.2292	0.0259
ST 1-5 kVA	0.8569	4.8450	0.0001
ST 6-8 kVA	0.3619	2.2442	0.0249
ST >=9 kVA	0	.	.
Rel. normal	-1.4740	12.2362	0.0001
Index mensu.	0.2426	3.1111	0.0020
Index déter.	0.0297	0.4009	0.6886
Mensualisé	0	.	.

Tableau 10 : ***Modèle avant regroupement des modalités***

Les résultats des tests croisés de Wald [DER98] permettent d'obtenir les modalités par variables « explicatives » qui ont la même contribution significative à la variable à expliquer

Y. Dans notre cas, nous avons seulement regroupé les clients ayant les services « Index estimé déterminé » et « Mensualisé » dans le cadre de la politique de facturation. Nous voyons sur la figure suivante, que les *A* sont plutôt des clients récents, qui habitent dans des logements collectifs, qui ont adopté plutôt la mensualisation et qui ont souscrit des puissances assez basses, toutes choses égales par ailleurs.

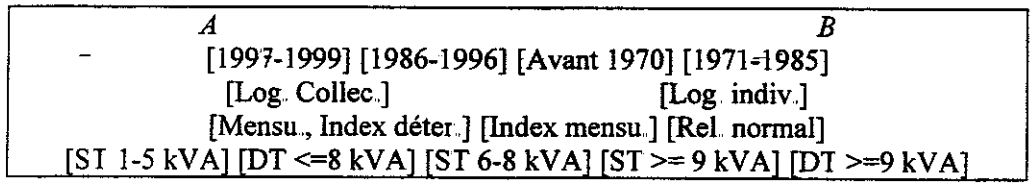

Figure 1 : *Synoptique des caractéristiques A et B*

2.4. Etape E7 : Validation du modèle sur l'échantillon test

Pour valider le modèle statistique précédemment construit nous avons utilisé la règle d'affectation du maximum. De façon pratique, pour chaque client, si la probabilité estimée d'être *A* est supérieure à 0,5, alors *Y* « estimée » prendra la valeur : *A*, sinon elle vaudra *B*[6]. Par exemple, on pourra obtenir une règle du type : *si le client habite en appartement et qu'il a souscrit un contrat avec EDF, il y a moins de cinq ans, alors il a 8 chances sur 10 d'être dans la catégorie A*. Dans ce cas, tous les clients ayant ces caractéristiques seront désignés comme *A*, car ils ont plus de chances d'être *A* que *B*. Ici cette règle d'affectation permettra d'attribuer à un client l'une des deux catégories de *Y* en fonction de son appartenance à un sous-ensemble de la population[7].

3 niveaux de validation sur l'échantillon test sont utilisés : Individuelle, marginale et globale.

Validation individuelle : Pour chaque groupe de clients ayant les mêmes caractéristiques, en terme de variables explicatives, nous fournissons un niveau de confiance individuel (entre 0 et 1 : plus la valeur est proche de l'unité, plus la confiance peut être élevée). Cela permet par

[6] Cette règle n'a pas été appliquée quand l'un des deux pourcentages de la variable à enrichir est inférieur à 10%, car l'autre est mécaniquement toujours choisi. Dans ce cas, on a utilisé la règle du pur hasard qui consiste à attribuer au hasard à chaque modalité *A* et *B*, sa probabilité prédite.

[7] correspondant aux groupes (cf. note de bas de page, numéro 4).

exemple, lors d'un mailing après enrichissement, d'avoir une cible plus fiable statistiquement, si on a choisit un groupe où le niveau de confiance est élevé. En d'autres termes, les clients inclus dans cette cible sont statistiquement plus susceptibles d'avoir été enrichis de façon « juste ».

La *validation marginale* consiste à calculer pour chaque catégorie (*A* et *B*, ici) le rapport du nombre de clients attribués à une catégorie par le modèle et possédant effectivement celle-ci sur le nombre total estimé de clients entrant dans cette catégorie. Le tableau 11 montre que ***pour les A*, 84%** sont bien classés grâce au modèle contre 0% sans le modèle, alors que ***pour les B*, 92,3%** sont bien classés contre 85,7% (comme ce dernier taux marginal est supérieur à 50%, on prend 85,7% car on utilise la règle d'affectation du maximum).

La *validation globale* revient à calculer le quotient du nombre de clients estimés et observés dans chaque catégorie sur le nombre total de clients. Dans notre cas, le tableau 11 fait apparaître **91,6%** de clients bien classés grâce au modèle contre 85,7% sans le modèle. Ce taux très correct permet de passer à l'étape E8, sans remettre en question la forme du modèle.

Y / *Statistique*	*A (%)*	*B (%)*	*sur l'ensemble (%)*
Pourcentage estimé	14,3	85,7	100,0
Sans modèle	0,0	100,0	85,7
Avec le modèle	**84,0**	**92,3**	**91,6**

Tableau 11 : *Validation du modèle*

2.5. Etape E8 : Enrichissement des données non renseignées

Cette étape consiste à prédire une des deux catégories de la variable *Y* en fonction des caractéristiques significatives communes à la table des valeurs renseignées pour une variable à prédire dans la base de données des 164358 clients. Pour ce faire, on utilise la règle d'affectation du maximum, pour enrichir les valeurs non renseignées[8].

Enfin, à chaque client dont le champ *Y* est enrichi, est attribué le niveau de confiance individuel (variant entre 0 et 1) associé à son groupe de modalités caractéristiques. Celui-ci

[8] Un champ de la variable *Y* déjà renseigné, n'est pas changé par la prédiction.

est calculé de la même façon que dans l'étape E7 de validation du modèle sur l'échantillon test pour le niveau individuel. Alors, les prédictions sont reversées dans la base de données clientèle à condition que leur niveau de confiance individuel soit supérieur à 0,8 ou 0,9.

3. Apports, critiques et voies d'amélioration

Les apports de la démarche statistique mise œuvre pour prédire des variables dans la base de données clientèle d'EDF dans le cadre de ce projet sont au moins au nombre de quatre.

La première_contribution significative concerne la construction du plan de sondage post-stratifié (étape **E3**) qui permet d'améliorer la précision de l'estimateur des proportions et qui joue par voie de conséquence sur la précision du modèle.

Le deuxième apport a trait à la démarche de construction du modèle en deux étapes. La première étape (**E6a**) concerne la sélection d'un modèle possédant la qualité de parcimonie, du fait de différents choix possibles à l'aide de l'approche par hiérarchisation des variables candidates à l'explication et du regroupement par apport de contribution de ces mêmes variables. Cette approche est relativement novatrice car elle met en œuvre des tests statistiques nouveaux, voire non encore publiés, en plus de certains tests usuels. La seconde étape (**E6b**), plus classique, permet de simplifier le modèle, en regroupant le cas échéant des modalités de variables « explicatives » déjà sélectionnées dans l'étape précédente. Cependant ces deux étapes permettent d'obtenir des modèles robustes en terme de validité statistique.

La troisième contribution, tout à fait modeste, mais très utile pour l'enrichissement, est relative à la fourniture d'un niveau de confiance individuel attribué à chaque client dont la donnée n'est pas renseignée. Un critère jugé pertinent pour la phase d'enrichissement (E8), bien qu'empirique, est de ne reverser dans la base de données clientèle que les prédictions ayant un niveau de confiance individuel supérieur à 0,8 ou 0,9. La qualité de l'enrichissement a été déjà confirmée par un retour d'expérience dans un Centre EDF-GDF SERVICES.

Enfin, ***le quatrième*** et dernier ***apport*** a trait au fait que ce type d'expérimentations à grande échelle (traitement statistique sur de grosses bases de données non prévues à cet effet) a représenté une première pour EDF, et en particulier pour la Division Recherche et Développement dans le cadre d'une meilleure connaissance de la clientèle, axe stratégique majeur pour EDF, rappelons-le.

Développement dans le cadre d'une meilleure connaissance de la clientèle, axe stratégique majeur pour EDF, rappelons-le.

Cette démarche statistique, qui paraît simple de prime abord, a été assez lourde à mettre en œuvre au niveau informatique, car elle fait appel à l'écriture de programmes *SAS* (macro *SAS*) permettant d'automatiser au maximum le traitement des données. Mais, comme on l'a précisé dans ce papier, certaines étapes sont délicates à automatiser (**E3**, **E6a** et **E6b**), car elles font appel à l'expérience du statisticien.

Cependant, nous travaillons actuellement pour faciliter une semi-automatisation de l'étape **E3**, notamment dans le choix des bornes de strates marginales. Enfin, sur le déroulement des deux autres étapes, nous pouvons donner quelques voies raisonnables. L'étape **E6a** pourrait être automatisée puisqu'elle permet de hiérarchiser l'importance de contribution de chaque variable candidate à l'explication de la variable à enrichir. En effet, il suffirait de choisir les quatre ou cinq premières, mais seulement à condition d'utiliser les tests de hiérarchisation et de comparaison d'apport statistique des variables « explicatives ». L'étape **E6b** pourrait être également automatisée en se servant des résultats des tests statistiques de regroupement de modalités. Mais cette approche serait assez délicate, car d'une part les regroupements doivent avoir une certaine signification (démographique, sociologique, ...) et d'autre part, il peut se produire des effets de chaînage ou des problèmes de transitivité. Nous avons quelques idées à ce sujet, mais nous ne les avons pas encore mises en œuvre dans ce projet.

BIBLIOGRAPHIE

[DER97] Derquenne C., (1997), Goodness of Fit Measures based on Likelihood in Generalized Linear Models, *51° Session de l'Institut International de Statistique*, Istanbul, Turquie.

[DER98] Derquenne C., (1998), Analyse de la variance sur variables ordinales : Application aux enquêtes d'opinions, *Club SAS/STAT*, Paris, France.

[HOS89] Hosmer D. W. & Lemeshow S., (1989), *Applied Logistic Regression*, New-York, John Wiley & Sons.

[MCF74] Mc Fadden D., (1974), The Measurement of Urban Travel Demand, *Journal of Public Economics*, 3 302-328.

[TIL97] Tillé Y., (1997), *Théorie des Sondages*, support de cours ENSAI, Bruz.