Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

CONSERVATOIRE NATIONAL DES ARTS ET METIERS
PARIS

THESE

Pour obtenir le grade de

DOCTEUR DU CONSERVATOIRE NATIONAL
DES ARTS ET METIERS

Spécialité : **INFORMATIQUE**

Présentée et soutenue publiquement par
Nicolas FISCHER
Le 1er Octobre 2004

Titre :

FUSION STATISTIQUE DE FICHIERS DE DONNEES

Composition du jury :

M. G. SAPORTA – Professeur du CNAM – Directeur de thèse

M.J.P. AIMETTI – Professeur du CNAM

M. T. ALUJA-BANET – Professeur Universitat Politecnica de Catalunya – Rapporteur

M. M. LEJEUNE – Professeur à l'Université de Grenoble 2 - Rapporteur

M. C. DERQUENNE – Chercheur EDF-R&D

M. G. HEBRAIL – Professeur à l'ENST

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Remerciements

A Monsieur le Professeur Gilbert Saporta, qui a dirigé ma thèse avec beaucoup de patience. Il m'a fait bénéficier de sa grande compétence scientifique, de son encadrement rigoureux, de sa disponibilité et de ses conseils toujours pertinents et prodigués avec beaucoup de modestie et de bienveillance.

A Monsieur Christian Derquenne qui a co-encadré cette thèse. Qu'il soit remercié pour sa grande disponibilité, son attention portée à ce travail, son extrême gentillesse et sa très grande patience. Apprendre avec lui au quotidien, profiter au jour le jour de ses compétences scientifiques, de son enthousiasme et de ses capacités de dialogue ont constitué pour moi un réel plaisir.

A Monsieur le Professeur Aluja-Banet, qui m'a accueilli avec beaucoup de gentillesse et avec qui j'ai pu avoir des discussions stimulantes et agréables. Il m'a fait l'honneur d'être rapporteur de ma thèse et m'a fait bénéficier de ses remarques pertinentes.

A Monsieur le Professeur Lejeune, qui m'a fait l'honneur d'accepter d'être rapporteur de cette thèse; il m'a fait ainsi bénéficier de ses compétences par ses commentaires attentifs qui ont beaucoup apporté à ce mémoire.

A Monsieur le Professeur Hebrail, qui m'a accueilli avec chaleur dans son groupe et m'a permis de bénéficier de l'environnement stimulant et agréable qu'il y fait régner. Tout au long de cette thèse j'ai pu bénéficier de son dynamisme, de ses remarques judicieuses et de ses commentaires avisés.

A Monsieur le Professeur Aimetti, qui a accepté d'être dans ce jury de thèse et de me faire bénéficier de ses compétences.

A Sabine pour sa gentillesse, sa disponibilité, son aide au quotidien; à Bertrand, en souvenir de la complicité amicale qui nous a liée pendant tout ce travail; à tous les membres passés et actuels du groupe SOAD avec lesquels le travail s'est déroulé dans une atmosphère chaleureuse et amicale, que je n'oublierai pas.

A ma famille, à Stéphanie, à mes amis, pour leur constant soutien et leur affection.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Table des matières

Chapitre I : Introduction 5

Chapitre II : Etat de l'art de la fusion statistique 9

II – 1 Les données manquantes 9

II - 1.1 Généralités 9
II - 1.2 Origine et raisons des non-réponses 9
II - 1.3 Nature des données manquantes 10
II – 1.4 Traitement des données manquantes totales 11
II - 1.4.1 La mise en évidence des facteurs explicatifs 12
II - 1.4.2 Les différents traitements des données manquantes totales 13
II – 1.5 Le traitement des données manquantes partielles 13
II – 1.5.1 Les méthodes d'imputation 14
II - 1.5.2 L'estimation des données manquantes partielles par des modèles explicites 14
II - 1.5.3 L'estimation des données manquantes partielles par des modèles implicites 15
II - 1.5.4 La méthode de repondération 17

II – 2 La fusion statistique de données 17

II – 2.1 Généralités sur la fusion de fichiers 17
II – 2.2 Historique du développement de la fusion de fichiers 18
II – 2.3 Principales méthodes de fusion statistique de fichiers 20
II - 2.3.1 La fusion par appariement aléatoire intra-cellulaires 20
II - 2.3.2 L'appariement statistique 21
a - Notation 22
b - Appariement statistique avec ou sans contrainte 22
c - Les bases théoriques de l'appariement statistique selon Rodgers 23
II - 2.3.3 La fusion sur référentiel factoriel 24
a - La fusion par « mariage » 26
b - La fusion par recherche de sosies 30
c - La greffe de fichiers 31
II - 2.3.4 La fusion par analyse homogène 33
II - 2.3.5 L'approche régression 33

II – 3 En conclusion 35

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Chapitre III : Nouvelles approches développées 37

III – 1 Matériel statistique nécessaire au développement des nouvelles méthodes 37

III – 1.1 Rappel sur la régression logistique 37
III – 1.2 Rappel sur la régression PLS 43
III - 1.2.1 L'algorithme de la régression PLS1 44
III - 1.2.2 L'algorithme de la régression PLS2 46
III - 1.2.3 La régression logistique PLS 48

III – 2 La fusion par reconstitution à l'aide de variables latentes 50

III – 3 Méthode univariée (Logit classique) 52

III – 4 La greffe séquentielle de variables 52

III – 5 La méthode dite pseudo-multivariée (pseudo-PLS2) 56

III – 6 La régression PLS2 avec recodage des données (0-1) 57

III – 7 Méthode recodage logit suivi de la régression PLS2 59

Chapitre IV : Validation 63

IV – 1 Critères globaux de validation des résultats 63

IV – 1.1 Reconstitution des distributions marginales 65
IV - 1.2 Reconstitution des distributions croisées de deux variables 66
IV – 1.3 Taux d'individus bien-classés 66

IV – 2 Critère individuel de validation 68

IV – 3 Validation opérationnelle de la fusion 69

IV – 3.1 Définition de la courbe lift 70
IV – 3.2 Mesure d'adéquation fondée sur le lift 71
IV – 3.3 La courbe lift du gain potentiel 71
IV – 3.4 Relation entre $lift_{gain}$ et les courbes lift de concentration 72
IV – 3.5 Calcul exact des courbes lift et de l'indice de Gini 73
IV – 3.6 Aire sous la courbe lft – lien avec la courbe ROC 75
IV – 3.7 Illustration de la courbe lift sur un exemple 77

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Chapitre V : Application des nouvelles méthodes de fusion 83

V – 1 Validation statistique de la fusion effectuée entre les deux fichiers d'enquêtes Credoc et Sofres sur quelques milliers d'individus 83

V – 2 Fusion de fichiers d'enquête avec une base clientèle 86

V – 2.1 Contexte et objectif 86
V – 2.2 Description des bases de données 88
V – 2.3 Démarche de fusion statistique des bases de données 90
V – 2.4 Résultats obtenus 101
V – 2.5 Validation opérationnelle et courbe de lift 106
V – 2.6 Ciblage direct vs fusion statistique de données 108

Chapitre VI : Conclusion et perspectives 111

Annexe 1 115
Annexe 2 117

Bibliographie 127

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Chapitre I :

Introduction

Dans le cadre de l'analyse de données, le statisticien est très régulièrement confronté à la problématique de données manquantes. Que ce soit dans le cadre d'une enquête d'opinion, de collecte de résultats expérimentaux, de regroupement de fichiers de clientèle, il est très fréquent que l'analyste doive se contenter d'un fichier tronqué, partiel, ne présentant pas toute l'information originalement attendue. Quelle que soit la raison de la présence de données manquantes ou de non réponses dans un fichier (panne d'appareil, refus de réponse d'un interviewé, perte de document, etc…), le fichier final s'en trouvera affecté, du moins la population retenue différera de celle initialement étudiée.
Dans certains cas le taux de données manquantes peut être si important, que la quantité de données retenues sera trop faible pour pouvoir mener quelque étude à bien.
Pour éviter d'en arriver à ce type d'extrémités, c'est à dire l'élimination d'une partie de ces données, le statisticien devra procéder à un traitement des données manquantes. Dans un premier temps il devra s'interroger sur le mécanisme, l'origine de ces absences afin de choisir le traitement approprié à son cas et répondre quant à l'ignorabilité des différentes données manquantes. Selon la nature de ces données et de son problème ainsi posé il optera alors au choix entre deux démarches : estimer les données manquantes de son fichier au moyen de différents modèles explicites ou implicites puis procéder à une analyse de données sur le fichier ainsi obtenu (réunion de données mesurées et estimées) ou alors, ignorer les données manquantes et ne travailler que sur un fichier incomplet.

Par ailleurs, il est fréquent que les données manquantes dans un fichier apparaissent en bloc et non sous forme de données dispersées dans le fichier. Autrement dit il s'agit d'un cas particulier de données manquantes de type « variables manquantes ». Dans la pratique, il n'est pas rare de rencontrer ce phénomène : par exemple dans le cas d'un questionnaire trop long ou d'une étude marketing insuffisante nécessitant l'apport d'autres études pour compléter un manque d'information.
D'une manière plus générale, pour optimiser la productivité et /ou améliorer la qualité des services auprès de la clientèle, anticiper sur la recherche, l'industrie comme les services ont souvent recours à l'analyse statistique de différents types de données (nouveaux clients, qualité de fourniture, enquêtes de satisfaction…). Pour cela, en général, une entreprise dispose de plusieurs bases de données de renseignements sur les différents clients. Ces sources peuvent être de nature différente: interne ou externe, statique ou dynamique…

Malheureusement, l'ensemble de ces informations n'est en général pas disponible dans une seule base de données regroupant les différents

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

renseignements sur le même client. Il arrive aussi que le fichier de données soit abîmé ou introuvable, que les questionnaires ne soient pas complets c'est à dire qu'il manque des réponses ou que les questions n'aient pas été posées parce que le questionnaire était trop long.
Cette situation se produit pour des raisons diverses.
Parfois il existe un manque de communication des informations au sein de l'entreprise, ayant pour principale conséquence une absence de fusion des résultats obtenus au cours de différentes enquêtes menées par l'entreprise.
Il en résulte une perte de renseignements essentiels sur le client, l'absence d'interprétation des attentes du client, la non anticipation de la demande du client, l'absence de réponse aux besoins du client.
Ce qui implique la nécessité de répertorier et de combiner ces informations de sources différentes dans une unique base de données et de simuler les informations manquantes dont l'entreprise a besoin grâce aux informations existantes.
Plus spécifiquement et en prenant l'exemple des enquêtes médias ou marketing, il est des cas où ne souhaitant pas interroger le même échantillon sur la totalité des questions d'intérêt, afin de ne pas engendrer un harassement susceptible de biaiser les réponses, on choisit de procéder par deux enquêtes distinctes auprès de deux échantillons indépendants appareillés sur les principales variables socio-démographiques. L'objectif pratique poursuivi est alors de mettre ensemble, "fusionner" les deux sources d'information. J. M. Gautier (Gautier 1997) rappelle justement l'intérêt et la problématique de la fusion dans le cadre d'enrichissement de larges bases de données.

Il existe quatre démarches possibles pour la combinaison de ces données :
la méta-analyse, l'appariement hiérarchique, l'appariement statistique et la fusion statistique de données. Néanmoins, ces quatre méthodes statistiques ont le même objectif : enrichir l'information détenue au sein de l'entreprise et estimer les données manquantes du fichier de renseignements.

La méta-analyse est une méthode statistique qui a pour objectif de faire apparaître les accords et les désaccords entre différentes études ayant la même vocation (par exemple, différents instituts de sondage ayant pour but d'étudier la fidélisation des clients " chauffage électrique " ne choisissent pas obligatoirement le même plan de sondage et n'ont pas forcément le même type d'analyses statistiques). L'appariement hiérarchique se fait quand on est en présence de deux fichiers ayant les mêmes individus ou encore quand on dispose de variables discriminantes qui identifient les paires d'individus. L'appariement statistique permet d'apparier deux fichiers avec des unités différentes : en supposant que l'on a établi une relation des variables avec la population entière, on peut effectuer de l'inférence sur la distribution jointe des variables. La fusion statistique de données est très proche de la précédente méthode, car on est également en présence de deux échantillons distincts (ou plus) n'ayant pas les mêmes unités sur le même univers. Les méthodes entrant dans ce domaine consistent à transférer des variables d'un échantillon appelé " donneur " vers un autre échantillon

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

nommé “ receveur ” en se fondant sur une passerelle de variables communes.

On peut considérer le problème de **la fusion de fichiers** comme un cas particulier **de traitement de données manquantes**. Néanmoins la fusion se distingue au sens où elle met en jeu plusieurs sources de données mais aussi, parce qu'elle traite uniquement des données manquantes apparaissant en bloc et non des données manquantes dispersées dans tout le tableau contenant les informations recueillies au cours des diverses enquêtes menées.

Présentons maintenant notre travail.
Le deuxième chapitre nous permet tout d’abord de revenir sur la problématique générale des données manquantes. Après avoir brièvement vu un certain nombre de notions générales concernant la nature, l’origine et les différents types de données manquantes, nous rappelons les principales méthodes existantes de traitement des données manquantes.
Nous nous intéressons ensuite plus spécifiquement aux données manquantes apparaissant en bloc et donc à la problématique de la fusion statistique de fichiers de données. Après avoir retracé succinctement l’historique de la fusion statistique, nous exposerons un certain nombre de techniques existantes de fusion statistique développées récemment.

Dans **le troisième chapitre** nous présentons successivement de nouvelles contributions à la fusion statistique de fichiers. Compte tenu du fait que la fusion statistique concerne souvent des données provenant d’enquêtes et que celles-ci sont à une écrasante majorité qualitatives, nous avons porté notre attention sur des démarches de fusion appropriées à ce type de données. Le problème posé à la fusion étant celui de la qualité de la reconstitution des données et leur utilisation ultérieure, nous nous sommes attachés à développer des méthodes de fusion ayant pour but de préserver au mieux la cohérence des données initiales. De là vient notre volonté de définir une méthode multivariée la plus appropriée pour conserver les corrélations entre variables d’intérêt.

Le quatrième chapitre est entièrement consacré à la validation des méthodes de fusion statistique. Ces méthodes ayant pour but de fournir des estimations en lieu et place des données manquantes, il est légitime de se demander dans quelle mesure ces estimations fournies sont éloignées des données réelles, ou encore s’il est possible de les utiliser pour faire de la statistique inférentielle au niveau collectif ou individuel. Si une fusion statistique ne peut se faire sans validation de son modèle, il n’existe toutefois pas à proprement parler actuellement dans la littérature de méthodologie de la validation de la fusion statistique. Nous présentons ainsi plusieurs critères de validation des modèles de fusion que nous illustrons par un exemple. Si ces critères fournissent une validation statistique des techniques de fusion développées, ils ne garantissent pas par la suite à l’utilisateur (spécialiste clientèle ou autre) un quelconque apport de la fusion en terme opérationnel. C’est pourquoi nous proposons une validation dite

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

opérationnelle de la fusion, basée sur un indicateur, "le lift" très utilisé dans le domaine marketing, permettant de caractériser l'apport de la fusion statistique dans un cadre opérationnel.

Le cinquième chapitre est entièrement consacré à l'étude pratique d'un exemple de fusion statistique effectué sur des données réelles. Dans le cadre d'un projet du groupe statistique de la division Recherche et Développement de EDF nous avons appliqué la méthodologie précédemment décrite : nouvelles approches de fusion statistique et validation opérationnelle à une problématique marketing. L'entreprise possédant de nombreuses sources de données sur sa clientèle mais non disponibles dans une seule et unique base, l'objectif via la fusion était de regrouper cette information en une seule source plus riche. Il est alors possible de définir des cibles plus fiables de clients dans un but marketing. Nous nous attachons tout particulièrement à illustrer au moyen de cibles prédéterminées l'apport que peut fournir la fusion statistique dans l'affinement des résultats.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Chapitre II :
Etat de l'art de la fusion statistique

II-1. Les données manquantes

II-1.1. *Généralités*

Dans un fichier, les données statistiques sont présentées sous forme d'un tableau ou matrice à n lignes et p colonnes, sur n unités-observations et pour chaque unité il y a p caractères-variables. Les éléments du tableau sont notés "$x_{i,j}$", i indique la ligne du tableau et j la colonne du tableau. "x_j" signifie la j-ième colonne du tableau.
Avec la présence de données manquantes, le tableau contenant les informations se présente différemment.
Deux situations se présentent :
On rencontre des cases vides ce qui caractérise **les données manquantes partielles**.
On rencontre un bloc de lignes vides ce qui caractérise **les données manquantes totales**.

II-1.2. *Origine et raisons des non-réponses*

Les **données manquantes partielles** se rencontrent par exemple au cours d'une enquête où une partie des réponses de l'individu interrogé manque soit par :

- l'incompréhension de la question (par exemple la question est mal formulée, un manque de vocabulaire chez l'interviewé car celui-ci est de nationalité étrangère…)
- l'impossibilité de répondre à la question
- l'incohérence de la réponse
- l'invalidation de la réponse (par exemple les dépenses du foyer dépassent le capital)
- le refus de répondre à une question (par exemple combien gagnez vous?)
- au cours d'un processus expérimental industriel où l'on constate l'absence d'information due par exemple à la panne d'un appareil.

Les **données manquantes totales** se manifestent par exemple par :

- le refus de répondre à l'enquête (par exemple une enquête qui concerne les moyens contraceptifs ou les intentions de vote)
- l'abandon de l'enquête après plusieurs visites, ou lors d'une enquête postale
- l'incapacité de répondre à l'enquête (par exemple une personne âgée)
- la négligence de l'interviewé ou de l'intervieweur.
- l'absence de personnes à interroger.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Bien évidemment, cette liste n'est pas exhaustive. Elle est juste ici pour se faire une opinion sur les origines des non-réponses.
Pour traiter les non-réponses, une méthode assez naturelle consiste à probabiliser l'univers des comportements que peut avoir l'interviewé. Ainsi, on considère que chaque individu de la population a deux attitudes possibles face à une question d'une enquête : répondre ou ne pas répondre, chacune de ces deux attitudes étant pondérée d'une certaine probabilité.

II-1.3. *Nature des données manquantes*

Suivant les relations des données manquantes d'une variable avec celles observées et avec des données d'autres variables, on distingue trois types de données manquantes partielles (Rubin, 1976):

- les Données Manquantes Aléatoires (D.M.A.)
- les Données Manquantes Complètement Aléatoires (D.M.C.A.)
- les Données Manquantes Non Aléatoires (D.M.N.A.)

- **D.M.A.** : On dit qu'une donnée manque au hasard si l'occurrence de son absence est indépendante de la valeur qu'elle prend.
  Autrement dit, les données manquantes pour une même variable peuvent être considérées comme un sous-échantillon aléatoire de l'échantillon initial.
  La prise en compte de ce type de données manquantes doit ignorer les raisons de leur absence et s'appuyer sur les relations entre les variables pour les reconstituer.
  Exemple : Dans un tableau de données, considérons parmi d'autres les deux variables, X = sexe de l'individu et Y = activité professionnelle de l'individu.
  Si la probabilité de réponse pour l'activité professionnelle est indépendante du salaire de celle-ci, mais dépend du sexe, alors les données manquantes sur l'activité professionnelle sont aléatoires.

- **D.M.C.A.** : On dit qu'une donnée manque complètement au hasard si l'occurrence de son absence est indépendante de toutes les autres valeurs que prend l'individu qui présente cette donnée manquante.
  Autrement dit, l'ensemble des données manquantes complètement aléatoires constitue un sous-échantillon aléatoire des valeurs prises par l'échantillon initial. Par contre, l'ensemble des données manquantes aléatoires ne constitue pas nécessairement un sous-échantillon aléatoire des valeurs initiales.
  Exemple : Dans un tableau de données, considérons parmi d'autres les deux variables X = sexe de l'individu et Y = activité professionnelle de l'individu.
  Si la probabilité de réponse pour l'activité professionnelle est indépendante du salaire de celle-ci et de celle du sexe, les

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

données manquantes sur l'activité professionnelle sont dites complètement aléatoires.

- **D.M.N.A**. : On dit qu'une donnée ne manque pas au hasard si l'occurrence de son absence dépend de la valeur qu'elle prend et (ou) de toutes les valeurs que prend l'individu ayant cette donnée manquante.
  La prise en compte de ce type de données manquantes doit s'appuyer sur les raisons de leur absence. En effet, les ignorer entraîne non seulement une perte de l'information, mais aussi cela peut générer des résultats aberrants statistiquement.
  Exemple : Dans un tableau de données, considérons parmi d'autres les deux variables, X = sexe de l'individu et Y = activité professionnelle de l'individu.
  Si la probabilité de réponse pour l'activité professionnelle dépend de celle-ci (et éventuellement du sexe de l'individu), les données manquantes sur l'activité professionnelle sont dites non aléatoires.

II-1.4. *Traitement des données manquantes totales*

Pour traiter ces données manquantes, on distinguera les données manquantes totales des données manquantes partielles car les méthodes statistiques de traitement des données manquantes diffèrent en fonction de ces deux types. (Comyn 1999)

Dans un fichier de données, les unités manquantes peuvent causer des problèmes de distorsion entre l'échantillon de réponses récoltées après les différentes enquêtes menées et la représentativité de la population. Cette distorsion peut alors conduire à une mauvaise interprétation des résultats qui découlent de l'interprétation des informations tirées des fichiers de données. D'autre part, ces non-réponses peuvent introduire un biais et provoquer une diminution de la précision dans les résultats ce qui fausse les résultats du statisticien. Les non-réponses méritent donc que le statisticien s'y attarde pour les étudier afin d'établir de façon précise un rapport proche de la situation.
Les questions que se pose alors le statisticien devant ces non-réponses sont les suivantes :
Peut-on distinguer au sein des non-réponses des groupes de non-répondants?
Quelle est la cause des non-réponses ?
Quels sont les motifs principaux ?
J.M. Grosbras (1987) propose des techniques pour trouver les facteurs explicatifs des non-réponses :

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

II-1.4.1 - La mise en évidence des facteurs explicatifs

1. L'analyse des données permet de décrire les non-répondants.
En effet, on peut procéder :
- soit à une analyse des correspondances multiples sur l'échantillon global. On affecte aux variables actives les variables signalétiques et à la variable supplémentaire la variable "réponse ou non-réponse".
- soit à une analyse des correspondances multiples sur le sous-échantillon des non-répondants en attribuant à la variable supplémentaire le critère "refus motivé ou pas".
- soit à une analyse discriminante en distinguant les répondants des non-répondants.

Les techniques d'analyse de données précédemment citées sont décrites dans l'ouvrage de G. Saporta (Saporta 1990).

2. Les techniques classiques d'économétrie des variables qualitatives (Gourieroux 1984) peuvent s'avérer utiles pour modéliser les deux variables "réponses" ou "non –réponses".
La modélisation est la suivante : soit Y la fonction indicatrice définie par :
- elle prend la valeur 1 si l'individu i ne répond pas
- elle prend la valeur 0 si l'individu répond

On note X_{1i} la valeur signalétique X_1 pour l'individu i.
On suppose que la probabilité de non-réponse est une certaine fonction F des variables $X_1,\ldots, X_r$.
On estimera les paramètres de la fonction F par la méthode la plus classique celle du maximum de vraisemblance. Selon le choix de la fonction F, plusieurs modèles apparaissent.
Deux modèles se distinguent des autres :
- Le modèle LOGIT: F est définie par une loi logistique où les β_j sont estimés par la méthode du maximum de vraisemblance.

$$F_i = \left\{1 + \exp-(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ji})\right\}^{-1}$$

- Le modèle PROBIT: Φ représentant la fonction de répartition de la loi normale centrée réduite, nous avons,

$$F_i = \Phi(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})$$

Les résultats obtenus par ces deux modèles sont généralement assez proches.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

II-1.4.2 -Les méthodes de traitement des données manquantes totales

Les principales méthodes de traitement des données manquantes totales sont :

- la relance des non répondants, méthode développée par J.M. Grosbras en 1987. Cette méthode suppose que l'on peut diviser la population en deux strates : la première représente les répondants, la seconde les non répondants. Pour obtenir une estimation des éléments de la seconde strate (celle des non-répondants) on tire un sous-échantillon parmi ces derniers que l'on ré-interroge. Ensuite on combine les estimateurs obtenus dans chacune des deux strates proportionnellement à leur taux de sondage.

- la re-pondération des individus répondants. Cette méthode consiste à traiter la non-réponse en augmentant le poids des répondants. On est donc amené à faire des hypothèses sur la distribution des non-répondants.

- la re-pondération fondée sur la post-stratification. Cette méthode revient à remplacer les valeurs manquantes de l'échantillon par les valeurs moyennes des sous-populations correspondantes. On est donc amené à faire l'hypothèse que les non-répondants des catégories ne se distinguent pas en moyenne de celle des répondants.

En conclusion, il faut insister sur le fait que le traitement de données manquantes totales est un cas spécial de la fusion de fichiers qui lui-même est un cas particulier du traitement des données manquantes.

II-1.5. *Le traitement des données manquantes partielles*

Il existe deux méthodologies pour traiter les données manquantes partielles. La première méthode consiste à remplacer la donnée absente par une valeur "plausible" qui est en général issue ou estimée à partir des répondants. Cette méthode fait partie des méthodes d'imputation puisqu'on impute à la valeur manquante une valeur calculée à partir des renseignements existants. La seconde méthode consiste à s'intéresser uniquement à l'échantillon des répondants (technique à utiliser avec la plus grande vigilance car on se contente de travailler sur un fichier incomplet sans chercher à estimer explicitement les données manquantes.), et à associer des poids de pondération à chaque individu pour compenser l'absence de réponses et faire par conséquent des extrapolations. Cette méthode s'inscrit dans les méthodes de repondération.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

II-1.5.1 - Les méthodes d'imputation

Toute méthode d'imputation s'appuie sur un modèle de prévision de la ou des variables manquantes à partir des variables existantes. Une méthode d'imputation se base donc sur une estimation de la loi de probabilité suivie par la ou les variables à imputer, en fonction des facteurs observés. Suivant les méthodes, l'imputation remplace la valeur manquante par l'espérance de la valeur prédite ou par une valeur aléatoire obtenue par la loi de probabilité estimée auparavant.
Les principales méthodes d'imputation sont :

- l'imputation par modèle probabiliste (régression, analyse de la variance, analyse discriminante…),
- l'imputation basée sur la vraisemblance,
- l'imputation par prédicteur,
- l'imputation par "hot-deck",
- l'imputation multiple,

L'ensemble des procédures qui traitent des données manquantes peuvent être regroupées au sein de deux grandes "familles" de techniques. On va donc distinguer deux types de modèles :

- les modèles explicites
- les modèles implicites

II-1.5.2 - L'estimation des données manquantes partielles par des modèles explicites

Parmi les modèles explicites, on rencontre :

<u>les méthodes de régression</u> dont le principe est de prédire les valeurs manquantes en utilisant un modèle adapté aux variables observées. Il y a plusieurs possibilités et nous allons en citer quelques unes :

- une régression simple en prenant la variable la plus corrélée.
- une régression multiple en prenant le meilleur sous-ensemble de variables explicatives, utilisant un modèle pas à pas ou plus communément appelée la méthode du stepwise ou la méthode de Furnival et Wilson. (La méthode du stepwise consiste à effectuer une régression simple sur une variable puis à ajouter les variables qui font augmenter le coefficient de corrélation multiple, avec remise en question éventuelle des choix antérieurs. La méthode de Furnival et Wilson permet de calculer les meilleures régressions pour 1 à p variables explicatives, par une exploration optimisée de toutes les possibilités.)
- une analyse de variance qui n'est rien d'autre qu'un cas particulier de la régression lorsque la variable explicative X est nominale et la variable à expliquer est quantitative.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

- Une régression logistique ou une analyse discriminante lorsque la variable expliquée est qualitative (booléen, nominale ou ordinale). Le principe est simple : il suffit d'estimer la probabilité d'obtenir une des deux valeurs et d'imputer par la catégorie la plus proche.

les méthodes basées sur la vraisemblance qui s'appliquent dans un cadre bien particulier. En effet, l 'attribution de valeurs 'plausibles' aux données manquantes nécessite de se placer dans un cadre paramétrique. L'estimation de paramètres d'une loi d'une variable présentant des données manquantes doit se référer à toute l'information existante sur cette variable et également sur les autres. De ce point de vue, l'approche la plus répandue vise à maximiser la vraisemblance de l'échantillon incomplet (Little et Rubin 1987). Son principe est brièvement résumé comme suit:
les paramètres de la loi multidimensionnelle sont estimés par l'algorithme EM (Dempster, Laird et Rubin 1977) dont le principe est le suivant :
partant d'une estimation des paramètres de la loi, l'algorithme utilise alternativement deux étapes. L'étape E (espérance) consiste à calculer l'espérance conditionnelle de chaque donnée manquante sachant les données observées et l'estimation des paramètres. L'étape M (maximisation) consiste à calculer les estimateurs du maximum de vraisemblance des paramètres. On attribue lors de la convergence de l'algorithme ainsi à chaque donnée manquante la valeur la plus probable pour l'estimation obtenue des paramètres de la loi multidimensionnelle.

les méthodes d'imputation par prédicteur qui s'énoncent ainsi :

on remplace la valeur manquante Y_k dans le fichier par une valeur $\hat{Y}(X_k; \hat{\beta})$ estimée par la connaissance d'un vecteur connu noté X_k et d'un paramètre estimé noté $\hat{\beta}$. On obtient ainsi que la loi de Y_k est estimée sachant X_k et $\hat{\beta}$.

La prédiction est alors : $\hat{Y} = E(Y_k / X_k; \hat{\beta})$. On notera par ailleurs qu'il existe plusieurs prédicteurs. Nous allons en citer quelques-uns :

- les prédicteurs par la moyenne : on impute la moyenne des répondants aux unités défaillantes.
- les prédicteurs par le ratio : $\hat{Y} = X_k \frac{\overline{Y_r}}{\overline{X_r}}$ où $\overline{X_r}$ et $\overline{Y_r}$ sont les moyennes des variables X et Y sur l'ensemble des répondants.
- les prédicteurs par un modèle de régression : $\hat{Y} = \hat{\beta}' X_k + \varepsilon$ (cette méthode vient d'être décrite précédemment).

II-1.5.3 - L'estimation des données manquantes partielles par des modèles implicites

On distingue deux méthodes statistiques qui utilisent les modèles implicites:

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

- la méthode d'imputation de type "hot-deck" dont le principe est le suivant :
  la valeur manquante est remplacée par la valeur observée chez un répondant "proche", le "donneur". Cette méthode se divise elle-même en plusieurs méthodes.
- le *hot-deck d'ensemble* : le donneur est choisi de façon aléatoire parmi les répondants.

- le *hot-deck par classe* : le donneur est choisi de façon aléatoire dans la classe à laquelle appartient le receveur.

- le *hot-deck séquentiel* : le fichier à compléter défile. Si une unité est manquante, alors on lui impute la valeur renseignée par l'individu le plus "récent" du tableau de données et appartenant à la même classe.

- le *hot-deck hiérarchisé* : une suite de critères C_1, C_2,..., C_k est utilisée. On remplace l'unité défaillante par une unité ayant les mêmes valeurs pour C_1, C_2,...,C_{k-1}. S'il n'en existe pas alors on la remplace par une unité ayant les mêmes valeurs pour C_1, C_2,..., C_{k-1}; etc. ...

- le *hot-deck métrique* ou méthode du plus proche voisin : on construit une distance d entre unités en fonction de variables clés bien renseignées qu'elles ont en commun. Si l'unité k est défaillante, on lui impute la valeur observée chez son plus proche voisin "donneur potentiel". S'il y a des ex aequo, alors on impute une de ces valeurs possibles aléatoirement.

- *le hot-deck basé sur l'homogénéité* (Buuren et al., 1991) : comme son nom l'indique cette méthode se base sur l'homogénéité-interne de toutes les variables simultanément. La méthode d'imputation se divise en deux étapes :
  - la définition d'une variable score d'individu pour mesurer la différence entre les individus.
  - la définition d'une règle d'imputation c'est à dire une reconstitution des données manquantes suivant la valeur du score individuel et la fonction de perte.

  Ces deux étapes sont étroitement liées lors du déroulement de la méthode.
  Une homogénéité simultanée sur toutes les variables s'exprime par une fonction score.
  Cette fonction score s'appuie sur un partitionnement des observations à l'intérieur de classes homogènes employées par les procédures hot-deck habituelles.
  La différence de cette méthode avec les autres est que toutes les variables agissent simultanément dans la variable score.

- la méthode d'imputation multiple :
  L'imputation simple peut présenter le désavantage de n'affecter qu'une seule valeur à la donnée manquante. D'ailleurs imputer une valeur unique équivaut à imputer une valeur certaine qui ne procure pas

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

forcément des variances correctes. C'est ainsi que Rubin a introduit la notion d'imputation multiple en 1987. Son principe repose sur le fait d'imputer chaque donnée par m (>1) valeurs obtenues par un ou plusieurs modèles d'estimation. Ensuite on procède à l'analyse des données sur chacun des m jeux de données ainsi complétés. L'estimateur final d'un paramètre quelconque peut être, par exemple, la moyenne des m estimations ainsi obtenues. L'imputation multiple sous un ou plusieurs modèles permet de simuler la distribution a posteriori des données manquantes sous ce ou ces modèles et d'obtenir des variances correctes. Malheureusement cette méthode présente des inconvénients majeurs qui sont un temps de calcul considérable, une quantité non négligeable de données à stocker et à gérer, et la complexité des calculs sur un ou plusieurs échantillons.

II-1.5.4 - La méthode de repondération

Cette méthode a été développée par Little (Little 1987). Son principe est le suivant :
les inférences à partir d'échantillons d'enquêtes de données sans réponses sont généralement basées sur des poids inversement proportionnels à la probabilité de sélection. La probabilité de sélection ou d'inclusion est la probabilité qu'un individu soit présent dans l'échantillon que l'on s'apprête à tirer.
Les procédures de pondération modifient les poids pour tenir compte des non-réponses.
La pondération est liée à la méthode d'imputation par la moyenne. Par exemple, si les poids sont constants dans des sous classes de l'échantillon, alors soit on impute la moyenne de la sous classe aux individus manquants de chaque sous classe, soit on impute la pondération des individus répondants par la proportion de répondants dans chaque sous classe. Ces deux estimateurs conduisent au même estimateur de la moyenne dans la population, mais à des estimateurs différents de la variance de l'échantillon (elle sera plus grande dans le premier cas que dans le second), à moins que des ajustements ne soient réalisés sur les données moyennes imputées.

II-2. La fusion statistique des données

II-2.1. *Généralités sur la fusion de fichiers*

On peut citer M. Lejeune (1995) comme première approche :
"la fusion de sources de données est un ensemble de techniques qui se sont développées à partir des années 80 dans le domaine des études média". En effet, les utilisateurs d'enquêtes des médias souhaitent combiner des données provenant de plusieurs fichiers sources dans le but de réduire les budgets consacrés aux enquêtes, d'avoir une connaissance du

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

comportement global du consommateur, d'avoir une estimation de la position des non-répondants dans une enquête par rapport à celle des répondants.

Le but de la fusion de fichiers est d'utiliser au mieux les informations existantes dans une base de données pour reconstituer les informations non-existantes dans une autre base de données. Le principe est d'estimer les valeurs des variables non-renseignées à partir d'un bloc de variables renseignées qui ont la propriété d'être corrélées avec le bloc de variables non-renseignées.

La différence avec la problématique des données manquantes traitée dans le partie II.1 est que, pour la fusion de fichiers, le bloc de données manquantes appartient à des groupes de variables non-renseignées, ce sont des variables inexistantes. Pour résumer la fusion de fichiers, on peut dire que son but est de rapprocher des fichiers de données issus de sources diverses.

Par rapport à d'autres méthodes qui s'appuient sur le transfert de variables d'un échantillon "donneur" vers l'autre échantillon "receveur", la fusion de fichiers comporte une étape intermédiaire qui caractérise cette méthode. Cette étape est celle de connaître les relations entre les variables d'un fichier "donneur" et d'un fichier "receveur". On dit que la meilleure fusion est celle qui respecte le mieux ces relations. D'ailleurs les principaux critères d'une fusion de fichiers sont de conserver les relations entre les variables du fichier "donneur" et du fichier "receveur" et de ne pas créer de relations inexistantes entre ces variables.

Deux conditions préalables sont requises pour apparier deux fichiers de données :

- Du point de vue de la population : le fichier donneur doit avoir suffisamment d'unités statistiques représentatives dans chaque sous-groupe de population contenu dans le fichier receveur. Il sera judicieux d'effectuer préalablement au processus de fusion un redressement sur le fichier donneur d'après le fichier receveur.

- Du point de vue des variables : il doit exister une assez forte corrélation entre les variables reconstituées et les variables prédictives (variables communes). Ceci est nécessaire à l'obtention d'une qualité satisfaisante des données reconstituées.

II-2.2. *Historique du développement de la fusion de fichiers (Co 1997)*

L'histoire de la fusion de fichiers est relativement récente comme M. Lejeune l'a déjà mentionné. Cette méthode est vieille de 20 ans et par conséquent la littérature sur cette technique se fait rare. Voici une liste des principaux auteurs ayant écrit sur cette technique:

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Wendt F. et Boucharenc L. sont considérés comme les premières personnes à présenter la méthode de fusion de fichiers. Cette dernière consiste à apparier un receveur dans le fichier receveur et un donneur dans le fichier donneur. Ensuite, on transfert les valeurs des variables correspondantes dans le fichier receveur.

Santini G. (1984) développe cette méthode en introduisant un type d'appariement entre le receveur et le donneur. Ce dernier permet de limiter les copies d'un même donneur et utilise ainsi le plus grand nombre possible de donneurs.

A nouveau, Wendt F. (1984) propose une autre méthode de fusion de fichiers. Cette dernière consiste à effectuer une typologie entre les variables communes des deux fichiers donneur et receveur. En fait, la fusion se fait à l'intérieur de chaque groupe.

Rubin D.B., célèbre dans le traitement des données manquantes, propose également une méthode de régression paramétrique dans le cas où l'on dispose de trois fichiers : un fichier receveur R sur (x, y), un fichier donneur D sur (y, z) et un fichier supplémentaire S sur (y, z) ou (x, y, z). Dans cette régression, on déduit d'abord une valeur intermédiaire z_int, une prédiction par la régression de la variable à transférer z par rapport aux variables communes (x, y). On détermine ensuite une valeur authentique z issue du fichier donneur D qui a la valeur la plus proche de la valeur intermédiaire z_int.

La société STATIRO (1995) utilise la technique de fusion dans les enquêtes du type 'auto-administré', plus précisément une technique d'injection, par la recherche de sosies en se basant sur des résultats d'analyse des correspondances.

L'INSEE à partir de 1994 utilise une technique de fusion par imputation aléatoire dans différentes cellules. Cette méthode consiste en un choix d'un critère de partition de cellule, la partition des individus des deux fichiers receveur et donneur en cellules et l'affectation à un individu d'un individu du fichier donneur pris de façon aléatoire sans remise dans la même cellule.

Jusqu'alors le problème de la validation des résultats de la fusion de fichiers n'est pas standardisé. En 1995, Lebart L. et Lejeune M. proposent des techniques de validation : la validation croisée et la validation basée sur une technique de bootstrap. La première est une méthode de comparaison de données réelles cachées, avec celles reconstituées par la fusion de fichiers. La seconde méthode est une procédure de ré-échantillonage afin d'évaluer les statistiques sur des variables fusionnées.

Saporta G. et Co v. (1997) introduisent la fusion de fichiers basée sur l'analyse homogène. Cette nouvelle méthode réclame que le pouvoir prédictif des variables communes par rapport aux variables à transférer soit assez

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

fort : le niveau d'intercorrélation des données après quantification doit être assez élevé. C'est à dire que les variables sont suffisamment prédictives par rapport aux variables à transférer.

Aluja-Banet T. (1997) propose une méthode de greffe de fichiers s'inspirant du principe de la méthode de fusion sur référentiel factoriel défini par G. Santini, il définit à son tour un sous-espace commun aux données. Il utilise ensuite des méthodes de recherche des k-plus proches voisins pour imputer à un individu receveur l'information de ses donneurs les plus proches.

Santini (2000), poursuivant ses travaux sur les méthodes de fusion avec calcul de distance dans un référentiel factoriel, développe un exemple intéressant de fusion par "mariage" : la méthode de fusion procustéenne (décrite en fin de chapitre).

Enfin parmi les derniers travaux, Rassler (2002) présente une approche de la fusion statistique basée sur des modèles bayésiens.

II-2.3. *Principales méthodes de fusion statistique de fichiers*

Voici quelques méthodes que l'on peut trouver dans la littérature qui se rapporte à ce sujet. Il existe en fait deux principaux types de méthodes :

- la fusion par appariement.
- la fusion par prévision de variables.

II-2.3.1 - La fusion par appariements aléatoires intra-cellulaires

Souvent, on dispose de plusieurs fichiers de renseignements sur une population. Les domaines couverts sont spécifiques à chaque fichier mais relativement liés et une partie des informations est présente dans chaque fichier.
Un objectif naturel est alors d'apparier ces fichiers pour affiner et améliorer les liens entre ces domaines de renseignements.
La méthode statistique de l'appariement aléatoire par cellule repose sur le partitionnement des échantillons de chaque enquête en cellules regroupant des individus semblables. On affecte ensuite à un individu du fichier receveur un individu du fichier donneur pris de façon aléatoire sans remise dans la même cellule.
Pour résumer l'appariement aléatoire se décompose ainsi :

- le choix des critères de partitionnement.
- la partition des deux enquêtes selon les critères pour créer ces cellules.
- l'affectation d'un individu du fichier donneur, pris aléatoirement sans remise dans la cellule à un individu du fichier receveur de cette même cellule.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Une condition nécessaire d'utilisation de cette méthode est que les deux échantillons aient une distribution similaire suivant les variables communes. De plus, il faut que les tailles des deux échantillons ne diffèrent pas trop l'une de l'autre.

On rencontre souvent des échantillons pondérés aussi bien dans le fichier receveur que dans le fichier donneur. Utiliser l'appariement aléatoire par cellule sans précaution introduira des erreurs dans les estimations des variables à transférer. Au sein d'une même cellule, les poids peuvent être hétérogènes, c'est à dire que le fichier a été redressé par certaines variables qui ne sont pas incluses dans le critère de cellulage. A cet effet, l'INSEE répond par la démultiplication des individus du fichier donneur proportionnellement à leur poids. De même si l'échantillon receveur possède des pondérations hétérogènes, il est envisageable de démultiplier les individus du fichier receveur en fonction de leurs poids.

Les résultats de l'INSEE en 1994 donnent les conclusions suivantes :

- la démultiplication du fichier donneur dans le cas de poids hétérogènes donne de meilleurs résultats,
- la démultiplication du fichier receveur n'a pas de conséquence significative sur les résultats, elle réduit seulement le caractère aléatoire de la méthode.

Dans cette méthode, le choix du nombre de critères de cellulage et le rapport des tailles entre les deux fichiers sont essentiels.

Un avantage majeur de cette méthode est de respecter les relations entre les variables transférées. Malheureusement, elle a tendance à homogénéiser les comportements simulés pour les individus du fichier receveur entraînant ainsi une sous-estimation de la variance de l'échantillon.

II-2.3.2 - L'appariement statistique

"Statistical matching" est le terme anglais que l'on rencontre dans la littérature pour désigner des méthodes de fusion de fichiers. On traduit cette expression généralement par appariement statistique.

Souvent, on est amené à étudier des sujets se trouvant dans plusieurs enquêtes différentes. Mener une nouvelle enquête sur ces sujets prendrait du temps, ferait perdre de l'argent à ceux qui souhaitent obtenir les résultas et serait infaisable due à un questionnaire trop long. Une alternative est alors d'utiliser l'information contenue dans ces différentes enquêtes. Une solution à ces problèmes sont les procédures de "statistical matching".

Ces procédures recherchent pour un individu d'un fichier de renseignements un individu le plus similaire possible dans un autre fichier de

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

renseignements. Les procédures de "statistical matching" sont des méthodes d'appariement avec ou sans contraintes, ou par imputation multiple (cette méthode a été développée auparavant dans les traitements de données manquantes partielles).

II-2.3.2.a - Notation

Nous allons garder les notations de l'article de Rodgers (1984).

Soit S un vecteur de variables pour chaque enregistrement du fichier A de taille n, soit T un vecteur de variables pour chaque enregistrement du fichier supplémentaire B de taille m, tels que ces deux vecteurs doivent contenir un ensemble de caractéristiques communes. Ces variables communes sont notées $X = (X_1,\ldots, X_p)$. Le reste des variables est référencé dans chaque fichier par $Y = (Y_1,\ldots, Y_q)$ pour le fichier A et par $Z = (Z_1,\ldots, Z_r)$ pour le fichier B. De plus, un poids d'échantillonnage w est associé à chaque enregistrement dans les deux fichiers.

Le processus de "statistical matching" nécessite l'introduction d'une fonction de distance, laquelle permet d'estimer des similarités entre des paires d'individus des deux fichiers à partir d'une fonction D_{ij} sur les variables qui composent le vecteur X. De plus, certaines variables sont traitées comme des variables de cohorte. Les variables de cohorte sont identiques dans les deux fichiers et définissent des sous-classes d'individus dans chacun des fichiers. Les appariements sont alors possibles entre des paires d'individus uniquement à l'intérieur de la même sous-classe.

Une pondération est affectée à chaque individu du fichier fusionné notée w_{ij}. Elle est égale soit à w_i la pondération du fichier A, soit une pondération modifiée suivant la technique utilisée et le besoin d'ajuster les individus dans les deux fichiers.

II-2.3.2.b - Appariement statistique avec ou sans contraintes

Dans ces procédures de "statistical matching", deux types de méthodes sont reconnues.

La première méthode correspond au cas d'appariement sans contraintes, c'est à dire qu'il n'existe aucune restriction sur le nombre de fois qu'un individu du fichier A donnera son information Z au fichier B. L'appariement sans contrainte permet d'obtenir les mariages les plus proches pour chaque individu du fichier A, mais au prix d'une augmentation des estimateurs de variance des variables Z. Un appariement sans contraintes permet de prendre un échantillon aléatoire simple, avec remise, des enregistrements du fichier B. Les distributions des variables imputées Z ajoutées au fichier A, sont alors les distributions de l'échantillon sélectionné plutôt que les

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

distributions observées dans le fichier B. Malgré cet inconvénient, l'appariement sans contraintes est fréquemment utilisé.

La seconde méthode correspond au cas d'appariement avec contraintes. Les conditions posées pour ces contraintes se formulent par :

$$\sum_{j=1}^{m} w_{ij} = w_i \text{, pour } i = 1,\ldots, n,$$

et

$$\sum_{i=1}^{n} w_{ij} = w_j \text{, pour } j = 1,\ldots, m.$$

Maintenant, on pose $d_{ij} \geq 0$ la distance entre l'individu i du fichier A et l'individu j du fichier B, l'objectif de l'appariement sous contraintes est de minimiser la fonction suivante:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} * d_{ij}$$

sous réserve que les conditions ci-dessus soient admises et avec $w_{ij} \geq 0$ pour tout i et j.
L'avantage de l'appariement sous contraintes par rapport à l'appariement sans contrainte est que la distribution multivariée des variables Z du fichier B est précisément répliquée dans le fichier fusionné. Un inconvénient majeur est dû au fait que les paires d'individus fusionnées diffèrent plus sous contraintes en ce qui concerne les variables communes X. Un autre inconvénient est le coût considérable des calculs des opérations de minimisation de distance avec les données réelles.

II-2.3.2.c - Les bases théoriques de l'appariement statistique selon Rodgers (Rodgers 1984)

On pose X les variables communes aux deux fichiers et Y et Z les variables spécifiques des deux fichiers respectifs.

L'hypothèse inhérente aux méthodes d'appariement statistique est que le vecteur aléatoire Y sachant X est indépendant du vecteur aléatoire Z sachant X. Dans le cas particulier de la distribution normale multivariée, cette hypothèse est équivalente à la supposition que les corrélations partielles entre les variables Y et Z sachant X sont toutes nulles.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Cette hypothèse d'indépendance conditionnelle est forte par rapport aux faibles justifications qui sont généralement données. Cela implique que la relation de Y à Z peut être totalement inférée de la relation de Y à X et de celle de Z à X. Occasionnellement, l'information au sujet de la relation de la paire Y-Z est disponible dans une autre source, ce qui permet une amélioration dans l'hypothèse de l'indépendance conditionnelle. De toute évidence, cela ne change pas le fait que l'appariement statistique ne génère pas une nouvelle information sur la relation conditionnelle de la paire Y-Z. Il renvoie seulement les hypothèses émises lors de la création du nouveau fichier.

Pour illustrer l'importance de l'hypothèse de l'indépendance conditionnelle, on est souvent amené à considérer un intervalle de valeurs possibles du coefficient de corrélation entre la variable observée Y et la variable observée Z. Cet intervalle donne les contraintes qui proviennent des corrélations observées de ces variables avec la variable X. Dans le cas où il y a une corrélation multiple extrêmement élevée entre soit la variable Y, soit la variable Z et l'ensemble des variables X disponibles, l'intervalle des valeurs possibles pour la corrélation des variables Y et Z est alors petit. Dans ce cas, la corrélation multiple doit être si grande entre soit la variable Y, soit la variable Z et le jeu de variables X que l'une des deux variables observées Y ou Z est en fait proche d'une combinaison linéaire simple des variables X.

Pour remédier à cette importante hypothèse d'indépendance conditionnelle, Rubin en 1986 propose d'utiliser l'imputation multiple déjà évoquée au chapitre précédent.

On considère K méthodes d'imputation, chacune correspondant à différentes hypothèses. Par exemple, dans le cas simple de nos vecteurs Y et Z, les hypothèses variantes pourraient être la corrélation partielle r, entre Y et Z sachant X. On suppose qu'un ensemble imputé est créé sous chaque hypothèse. Chacun de ces fichiers sera analysé par une quelconque méthode utilisable sur un fichier ainsi au lieu d'un vecteur de statistiques, il y a K vecteurs de statistiques, T_1, ..., T_k pour chaque ensemble obtenu par l'imputation multiple. La variance des T_k représente le biais de l'estimation due à l'incertitude de chaque hypothèse testée (ici la corrélation partielle).

II-2.3.3 - La fusion sur référentiel factoriel

Une technique de fusion fréquemment utilisée en France par la société Statiro est la fusion basée sur "référentiel factoriel". Le principe de cette méthode est défini par G. Santini (Santini 1984), et repose sur les deux points suivants:

- les variables critiques : une partie des variables communes dans les deux fichiers receveur et donneur sert principalement à reconstituer les valeurs des variables manquantes. Ces variables sont prédictives par rapport aux

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

variables à reconstituer. Dans les méthodes classiques, ces variables critiques servent à déterminer pour l'individu du fichier receveur ses donneurs éligibles.

- les variables de rapprochement : une partie des variables communes, par un calcul de distance, permettant de choisir pour chaque receveur le donneur éligible le plus proche par rapport à ces variables dans les méthodes classiques.

D'autre part, cette méthode de fusion comprend la recherche de deux éléments essentiels :

- celle du référentiel factoriel. On commence par effectuer une analyse factorielle (par exemple une analyse des correspondances multiples) sur le tableau des variables critiques communes ou sur une sélection de variables communes les plus discriminantes des variables spécifiques, à l'ensemble des données disponibles c'est à dire les données relatives aux individus donneurs et receveurs. Ensuite on conserve les n premiers axes de l'analyse. Ainsi on positionne les observations dans un espace réel de dimension n, sur lequel on introduit une distance euclidienne calculée à partir des coordonnées factorielles qui mesure la similarité "donneur-receveur".

- celle de voisinage d'un receveur. Il nous faut introduire une nouvelle notion qui sera utilisée dans les différentes méthodes de fusion : celle de voisinage d'un point receveur dans un espace de dimension n. En effet, pour chaque individu receveur, on sélectionne un ensemble de donneurs dans un voisinage du receveur. Ce voisinage peut-être défini soit à partir d'une sphère centrée sur chaque receveur de rayon r, soit à partir de la recherche des k plus proches donneurs.

Après ces recherches, il faut choisir parmi les donneurs potentiels celui qui ressemble le plus au receveur sur les variables de rapprochement, qui sont des variables signalétiques comme l'age, le sexe, la classe socio-professionnelle, etc...

D'autre part, on évite d'utiliser trop souvent le même donneur en utilisant une fonction de pénalité introduite également par G. Santini. Le donneur le moins copié dans la méthode 'Statiro' est finalement retenu.

Les méthodes qui vont suivre utilisent un référentiel factoriel. Nous ne rappellerons pas ici l'ensemble des notations utilisées ainsi que les outils développés par l'analyse par axes principaux mais nous renvoyons à l'ouvrage de Lebart et Morineau (1995). L'intérêt d'une analyse en axes principaux est qu'elle filtre les informations, donne un meilleur pouvoir prédictif des distances. En effet, les méthodes factorielles sont aptes à

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

donner des représentations et des réductions de l'information contenue dans des tableaux où les données sont nombreuses. Les représentations simplifiées de ces tableaux, obtenues par ces méthodes sont des outils remarquables de synthèse. On extrait des données les tendances les plus marquantes et on élimine les effets marginaux en supprimant les données perturbatrices.

La possibilité de réduire la dimension de l'espace des données provient des corrélations entre les variables. Cette réduction provient de la construction de variables synthétiques, combinaisons linéaires des variables initiales.

D'autre part, dans une analyse par axes principaux, il est important d'éliminer les dimensions correspondant à de faibles valeurs propres qui sont sensibles à la perturbation du tableau, au bruit.

Ce rappel de l'analyse par axes principaux étant fait, on peut aborder les différentes méthodes de fusion utilisant un référentiel factoriel.

II-2.3.3.a - La fusion par "mariage"

Cette méthode de fusion s'appuie sur le calcul d'une distance dans le référentiel factoriel entre les individus donneurs et les individus receveurs.

Cette distance permet de définir ce que G. Santini appelle les "mariages" entre un individu receveur et un individu donneur (Santini 2000). Les individus sont mariés en fonction de leur proximité d calculée sur les coordonnées factorielles. Une méthode élémentaire de fusion serait une minimisation de la distance moyenne calculée sur les réponses. Cependant G. Santini veut éviter qu'un individu donneur soit marié plusieurs fois. En fait, il faut éviter qu'un même individu donneur transmette son information plusieurs fois c'est à dire à plusieurs individus receveurs. Pour pallier ce problème de mariages multiples, on peut introduire un système de pénalités pour les individus donneurs dans le cas où ces derniers donneraient leur information à plusieurs individus receveurs. On pénalise alors la distance, c'est à dire que si un individu donneur est déjà marié à n individus receveurs, cette distance d est pénalisée par la formule suivante :

$$d' = 1 - (1-d)^n$$

G. Santini a imaginé 6 types différents de relations de voisinage par "mariage" :
On note : A l'individu receveur, B l'individu donneur.

- le mariage par "coup de foudre" (voisins réciproques) : si A est le plus proche voisin de B et réciproquement B est le plus proche voisin de A et n'a jamais été marié, alors A et B sont immédiatement mariés.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

- le mariage avec "l'ami d'enfance" : si B est le plus proche voisin de A, mais que B est déjà marié à A' , alors A sera marié à B' qui est le plus proche voisin de A après B.

- le mariage par "adultère" : une variation du cas précédent est nécessaire quand la distance entre B' et A est plus grande que la distance pénalisée entre A et B (puisque B est déjà marié à A'). Donc nous marions ensemble A et B.

- le mariage par "assiduité" : un autre cas apparaît si l'on voulait unir A et B, mais B est le plus proche voisin de A', A'' et A''' avec lesquels il est déjà marié. Finalement nous marions A et B ensemble.

- le mariage de "raison" : après les cas simples illustrés ci-dessus, il existe d'autres mariages pour lesquels les décisions sont plus complexes. Ces mariages sont réalisés en utilisant des méthodes d'optimisation de distances à des niveaux globaux.

- le mariage des "irréductibles" : ce cas traite des individus restant n'ayant pas trouvé de
  mari; de tels cas sont dus principalement aux faiblesses des règles d'optimisation utilisées. Nous cherchons alors d'autres règles d'optimisation permettant d'obtenir un mariage avec cet irréductible.

La fusion sur référentiel factoriel permet d'imposer des contraintes à la fusion par mariage comme interdire des mariages entre des cellules disjointes définies sur des critères comme le sexe ou l'age ou encore la catégorie socio-professionnelle ou encore comme le nombre d'utilisations d'un même individu donneur.

Néanmoins la fusion par "mariage" n'est pas restreinte au seul cas de la fusion sur référentiel factoriel. En effet, Rassler (Rassler 1997) l'utilise dans un cadre plus théorique de fusion de fichiers de données continues ayant des distributions "classiques".

Explicitons un exemple intéressant de fusion par mariage : la **méthode de fusion procustéene**

(Santini 2000)

1. Contexte :

On dispose de deux échantillons indépendants qui proviennent de deux enquêtes distinctes. Notre objectif est de mettre ensemble , de compiler, de « fusionner » les deux sources d'information. Cet objectif nécessite que l'on se place dans un cadre de statistique bayésienne.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Lors de l'enquête, on distinguera les caractéristiques de l'individu de l'enquête (c'est à dire son âge, son sexe, sa profession, le lieu où il réside) supposés connus que l'on notera a_k avec k variant de 1 à K et les réponses de celui-ci aux autres questions de l'enquête (c'est à dire des questions d'attitude, d'usage) que l'on notera u_q avec q variant de 1 à Q.

D'autre part, on définit une distance qui sert à mesurer la similarité entre deux individus de la population de référence notée R. Cette distance est basée sur les caractéristiques seules. Cette distance étant construite, on peut associer à chaque individu noté j_0 de R, un ensemble d'individus notés j de R situés à l'intérieur d'une boule centrée sur l'individu j_0 et de diamètre δ. On a ainsi construit un voisinage de j_0 noté $V_\delta(j_0)$. Pour δ fixé, on construit le taux d'occurrence de la réponse u_q observée au sens du voisinage V_(j_0) et on suppose que ce taux tend vers une limite p_{uqj0} lorsque δ tend vers 0. Cette limite est appelée probabilité latente de u_q pour j_0.
Les réponses aux questions u_q de l'individu i sont considérées comme des réalisations de la probabilité latente $p_{uq(i)}$. Elles peuvent donc servir à estimer ces probabilités.

On définit une opération de fusion procustéene ainsi :
Soit i_1 un individu de la première enquête accompagné de ses variables u_q.
Soit i_2 un individu de la deuxième enquête accompagné de ses variables v_q.
i_2 est un donneur pour i_1 si on associe à i_1 les réponses de i_2.
Une fusion de i_1 et i_2 est dite possible si l'estimation des probabilités latentes $p_{vq(j)}$ est peu différente selon qu'elle est effectuée sur la base des informations recueillies dans la deuxième enquête ou celle des informations transférées dans la première.

2. Principes de la méthode :

Les principes de la méthode sont au nombre de quatre. Ils servent à créer les conditions d'un transfert sans distorsion des corrélations entre variables :

- 1. tout individu receveur est associé à un unique individu donneur.
- 2. toutes les informations recueillies sur un individu donneur sont transmises en bloc à l'individu receveur intéressé.
- 3. tout individu donneur déjà associé à un individu receveur sera pénalisé si il se lie à un autre individu receveur.
- 4. la similitude entre deux individus interrogés doit être prise dans son ensemble.

3. La fusion procustéene :

Pour prendre en considération le dernier principe, on construit à nouveau une distance qui a la particularité suivante :

Si la distance est faible, son rôle d'assurer la proximité entre les deux individus tant du point de vue des attributs que des variables spécifiques de la première enquête et celles de la seconde est maintenu.
On considère connues a_k, u_q, v_q.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Soit un ensemble de variables caractérisant des individus, on définit la distance entre deux individus comme étant la distance euclidienne des représentations de ces points dans l'espace associé à la correspondance multiple définie par ces variables. Cette distance peut s'identifier à celle du chi deux tronquée du fait qu'elle ne soit calculée que sur les premiers axes.

Le nom de fusion procustéenne provient lui du mode de calcul de la distance (Santini 2000).

4. Algorithme de mariage :

Si on assimile la distance entre deux informants à une mesure d'affinité au mariage, on lie prioritairement les individus qui présentent la plus forte affinité .
Pour résumer, ce processus se compose d'une chaîne qui construit alternativement un individu donneur et un individu receveur. Il faut donc initier le processus avec un individu donneur quelconque. La chaîne se ferme tôt ou tard. Cette chaîne va créer des mariages "coup de foudre", des mariages "ami d'enfance", ou encore des mariages "par assiduité".
Ils existent d'autres cas de figure que G. Santini appelle les "mariages de convenance" (cf mariage de "raison").

5. Optimisation :

L'ensemble des liens établis, on procède à une optimisation de ceux-ci en décidant de ne modifier que ceux qui sont associés à un mariage de convenance.
L'algorithme construit un tableau de contingence pour chaque couple de a_k et de v_q. On procède par itération jusqu'à l'impossibilité d'introduire une amélioration.

6. Mesure de validité :

L'étape optimisation étant réalisée on comptabilise le pourcentage de tableaux de contingence qui présente une différence au sens du chi deux.

Mais cette information restant imprécise , on examine les résultats variable par variable.
L'examen de tous les tableaux se révèle peu efficace compte tenu du nombre de tableaux.
On construit donc des tests qui permettent de conclure pour chaque variable v_q si elle est bien transférée ou mal transférée par la méthode de fusion procustéene. On dit qu'une variable v_q est bien transférée si en tenant compte des caractéristiques a_k les individus auxquels on affecte la variable v_q par fusion proviennent de la même population que ceux parmi les individus donneurs qui présentent la même valeur de variable.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

II-2.3.3.b - La fusion par recherche de sosies

Cette méthode de fusion a été créée par la société STATIRO et plus particulièrement par Sousselier en 1995. Son principe s'appuie sur la recherche du sosie d'un individu à l'aide d'une distance calculée dans le référentiel factoriel. Cette méthode et ses dérivées sont fréquemment utilisées dans le cas de fusion d'enquêtes de consommation.

Cette méthode se décompose en trois étapes :

- La première étape est la recherche du référentiel factoriel qui est obtenu par une analyse factorielle des variables critiques. On introduit ensuite une distance euclidienne qui mesure la similarité entre individus. Pour chaque individu receveur, on retient les m unités de la population des donneurs dont la distance D avec le receveur vérifie la condition suivante :

$$D(r, d) < S$$

où r est un individu du fichier receveur, d un individu du fichier donneur et S un seuil de distance.

- La seconde étape impose des contraintes pour choisir le donneur parmi les m unités de la population des donneurs (par exemple: l'impossibilité de réutiliser plus d'un certain nombre de fois un même donneur pour différents receveurs).

- La dernière étape est une comparaison de la ressemblance signalétique entre l'individu donneur et l'individu receveur. On utilise pour cela l'agrégation multicritère. On calcule un poids, suivant le degré de liaison entre les variables de signalétiques et les variables spécifiques à transférer du donneur vers le receveur, pour différencier l'importance des variables signalétiques. On a alors une note globale sur toutes les variables. Les individus les plus ressemblants sont retenus.

En cas d'échec c'est à dire dans le cas où on n'a pas trouvé de donneur pour le receveur, on recommence ces trois étapes en élargissant le rayon du voisinage S de la première étape.

Au niveau global, les résultats montrent que les distributions des variables sur les populations des répondants et des individus reconstitués sont très proches. En fait, il existe peu d'écarts significatifs dans les tableaux croisant les fichiers donneur et receveur.

Au niveau individuel, la signalétique des individus donneurs et celle des individus receveurs est sensiblement la même du fait que l'individu le plus ressemblant a été choisi comme l'individu donneur.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

II-2.3.3.c - La greffe de fichiers (Aluja-Banet 1997)

1 . Introduction :

La greffe de fichiers est une méthode qui permet de rassembler dans un même espace factoriel les nuages de points issus de deux tableaux de données ayant en commun un sous ensemble de variables. On exploite l'existence de variables communes pour trouver une relation éventuelle entre les variables qui ne sont pas communes aux deux enquêtes. Ce résultat n'est pas trivial car il n'existe pas à proprement parler de liaisons entre ces deux groupes de variables spécifiques. Mais la greffe de fichiers n'augmente pas l'information disponible dans les tableaux originaux. Elle sert juste à enrichir l'interprétation.
Les variables communes servent de béquilles à la greffe dans le sens où on suppose que les liaisons internes existantes entre les variables communes sont très semblables.

2. La greffe comme outil pour la fusion :

La première étape sert à définir l'espace de référence commun aux deux fichiers de données. On procède à une analyse factorielle de façon habituelle sur le premier fichier (donneur), le tableau X_0.

$$X_0 = V\Lambda\ U'$$

$$U'M_0U = I$$

$$V'N_0V = I$$

La représentation des éléments lignes et colonnes se fait respectivement sur les bases U et V par:

$$\Psi_0 = X_0M_0U \text{ et } \Phi_{X0} = f(X_0, \Psi_0)$$

$$\Phi_{Y0} = f(Y_0, \Psi_0)$$

sachant la relation de transition : $\Phi = X_0'N_0\ \Psi\Lambda^{-1/2}$

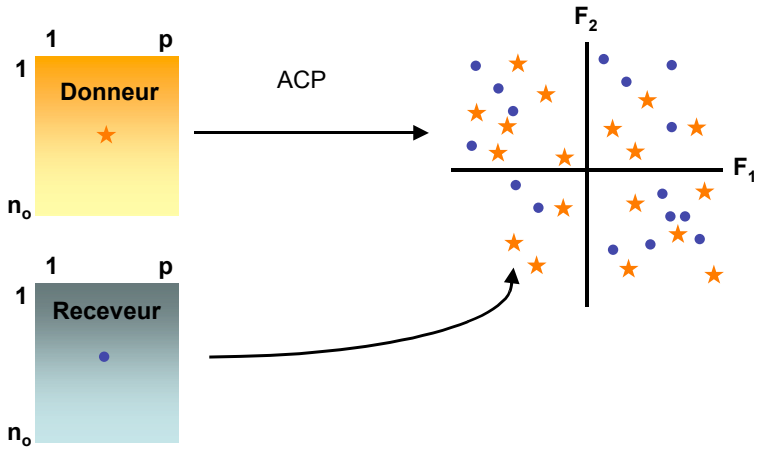
Les variables Y_0 (spécifiques au fichier donneur) sont positionnées en illustratives dans cette analyse.
Greffer le second fichier (receveur) signifie effectuer le positionnement de ses lignes et colonnes dans les bases issues de l'analyse factorielle précédente. Les individus de X_1 (fichier receveur) sont positionnés en illustratifs dans l'analyse de X_0 sans difficulté. Pour positionner les variables de Y_1 on s'appuie sur ces individus illustratifs en utilisant la relation de transition mentionnée auparavant. Pour cette raison on dit que les variables sont positionnées en double supplémentaire :

$$\Psi_1 = X_1M_0U \text{ et } \Phi_{X1} = f(X_1, \Psi_1)$$

$$\Phi_{Y1} = f(Y_1, \Psi_1)$$

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html



La position des variables communes aux deux fichiers indique la stabilité du sous espace commun et les directions stables. L'objectif est alors d'estimer les matrices de variables spécifiques à chacun des fichiers.

Une fois le référentiel factoriel défini, un algorithme de recherche des k-plus proches voisins est mis en œuvre. Pour les méthodes des k plus proches voisins on se référera au résumé de M.Comyn (Comyn 1999). En effet dans le sous-espace factoriel obtenu sur les variables communes la distance entre deux individus de chaque fichier a un sens. L'imputation des données spécifiques des receveurs vers leur donneur se fait alors par régression locale ou imputation stochastique locale.

Pour que tout ceci ait un sens, il est ainsi nécessaire que le sous espace commun soit prédictif des variables spécifiques que l'on veut transférer. C'est à dire que ces variables dans le sous-espace se positionnent dans des zones "intéressantes".

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

II-2.3.4 - La fusion par analyse homogène (Co, 1997)

L'analyse homogène permet d'estimer et d'imputer des données manquantes afin d'obtenir un ensemble de données les plus homogènes possibles. Cette méthode est basée sur un critère d'optimisation qui tend à maximiser l'homogénéité des données (Buuren 1991).

L'idée sous-jacente est similaire à celle de l'analyse factorielle ; si les variables mesurent plus ou moins la même propriété, il est alors possible de remplacer les différentes variables par une variable synthétique unique sans perdre trop d'information. Pour évaluer la qualité de cette substitution on définit un critère d'homogénéité et une fonction de perte. La maximisation de l'homogénéité des variables conduit à une analyse homogène semblable à l'analyse des correspondances multiples. (Saporta 2002)

Une condition pour l'application de l'analyse homogène dans la fusion est qu'il faut que les variables communes soient les plus prédictives possible des variables à transférer.

II-2.3.5 - L'approche régression

Mise à part la dernière méthode, toutes les méthodes entrevues précédemment consistent essentiellement à transférer de l'information d'un (ou plusieurs) individu(s) à un autre individu. D'autres méthodes récentes ont été développées résonnant à l'inverse sur les variables. Il n'est alors plus question d'apparier des individus mais de modéliser les variables spécifiques à partir des variables communes.

Derquenne (Derquenne 2000) présente une méthode pour construire un échantillon de données virtuelles à partir de plusieurs tableaux de données existants que nous résumons dans le cadre de la fusion par modélisation :

1. <u>Origine et contexte de cette méthode :</u>

Les grandes entreprises ne disposent pas en général d'une unique base de données qui regrouperait toutes les informations disponibles relevées au cours des différentes enquêtes ou sondages pour des raisons déjà évoquées. Pour renseigner leur base clientèle à partir d'enquêtes, un projet de simulateur a été lancé pour générer un échantillon d'individus virtuels. La démarche se décompose en deux étapes qui sont :

- la construction du premier échantillon virtuel fondé sur l'échantillon primaire (on développera la notion d'échantillon primaire par la suite). Cette construction est à base d'Analyse des Correspondances Multiples.
- une greffe statistique d'un échantillon secondaire sur le premier échantillon artificiel afin de construire le second échantillon

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

virtuel. La seconde étape utilise un outil statistique également : les modèles linéaires généralisés. Cette étape se répétera autant de fois voulues pour obtenir l'échantillon virtuel final.

2. Outils à notre disposition pour réaliser ce simulateur :

On a à notre disposition deux échantillons d'enquête qui sont respectivement :

- l'échantillon primaire X qui regroupe non seulement les variables du plan d'échantillonnage provenant de celui de l'enquête notées X_{mp} mais aussi des variables mesurées notées X_m.

- les échantillons secondaires qui regroupent non seulement des variables communes à X_{mp} que l'on notera $Y_{mp(k)}$ mais aussi d'autres variables qui peuvent être communes ou non à X_m notées $Y_{m(k)}$.

3. La méthode pour construire des échantillons virtuels:

On construit d'abord le *premier échantillon virtuel*.
Cette construction utilise l'échantillon primaire défini préalablement. On considère un vecteur noté X_{mp} de taille Q constitué de $X_{mp(1)},\ldots, X_{mp(Q)}$, c'est à dire des variables d'échantillonnage qui sont toutes qualitatives. On introduit un autre vecteur X_m de taille R constitué de R variables mesurées qui peuvent être nominales, ordinales ou discrétisées.
En fait, lors de cette étape, on applique sur les variables du plan d'échantillonnage notées X_{mp} une Analyse des Correspondances Multiples afin de constituer les composantes principales. Elles jouent le rôle de variables actives. Les variables supplémentaires sont les variables mesurées, X_m.
La raison pour laquelle on procède à une analyse des correspondances multiples est d'obtenir un espace réduit de nouvelles variables non corrélées tout en gardant le plus d'information possible.
Ensuite, des composantes principales obtenues répertoriées dans un vecteur Z de taille T, on sélectionne un sous–échantillon noté Z' de Z qui correspond aux valeurs propres significatives. Ce sous échantillon est de taille T' où T' est le nombre de valeurs propres supérieures à $1/mp(Q)$. Puis sont construits des groupes de variables les plus corrélées avec les composantes principales retenues à partir des rapports de corrélation entre les variables d'échantillonnage et les composantes significatives. Cette démarche est réitérée pour construire les groupes de variables supplémentaires notées $X_m(t)$.
Alors le tirage des individus virtuels se décompose en deux étapes :

La première étape est constituée de la discrétisation de chaque composante principale Z' en k_t intervalles afin de construire un espace pavé. C.Derquenne définit la distribution observée dans cet espace. On tire alors N individus de cette distribution. N est la taille de l'échantillon généré. On nomme ces individus virtuels les « individus muets » .

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

La seconde étape est la réplique de la première étape sur les $X_{mp(Q)}$ et sur les $X_{m(R)}$.
Donc on recueille N individus virtuels sachant les premiers, c'est à dire sachant les "individus muets". Ces nouveaux individus virtuels sont appelés " premières répliques ".

Cette construction achevée, on réalise la greffe statistique fondée sur les échantillons secondaires. Pour cela, on choisit le premier échantillon parmi l'ensemble des k échantillons secondaires . On fonde notre choix sur deux critères :
- le nombre de variables d'échantillonnage communes avec l'échantillon primaire.
- les variables de l'échantillon secondaire qui jouent un rôle important dans le cadre de l'enquête.

On introduit G_1 le nombre de variables à greffer sur le premier échantillon virtuel.
La construction du deuxième échantillon virtuel se fait en deux étapes.
La première étape est le redressement de l'échantillon secondaire par rapport au plan d'échantillonnage de l'échantillon primaire grâce à une méthode de calage sur marge.
La seconde étape est la greffe des G_1 variables une par une au premier échantillon virtuel.
Par les modèles linéaires généralisés est estimée la distribution de la variable à greffer. Des individus virtuels sont alors tirés selon la distribution estimée.

4. Conclusion

Cette méthode présente trois attraits qui sont :
- La capacité à construire un échantillon de taille variable
- Des résultats encourageants obtenus
- La génération d'individus artificiels à l'aide d'une connaissance a priori

Mais C. Derquenne avance des limites qui sont au nombre de deux :
- La complexité du processus de génération s'accroît avec le nombre de variables à greffer et d'échantillons secondaires.
- La taille des échantillons d'enquête est en général relativement basse.

II-3. En conclusion

Cette présente étude a permis tout d'abord de préciser le cadre de la fusion statistique de fichiers de données. Celle-ci apparaît comme un cas particulier des problèmes de traitement de données manquantes où les données sont manquantes par blocs. Ainsi ont été rappelées les principales méthodes de reconstitution des données, certaines techniques pouvant être transposées dans le cadre de la fusion. Après avoir défini ce que nous entendons par

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

fusion et appariement de données, nous pouvons distinguer deux grandes familles de méthodes : les fusions par appariement d'individus et celles par prévision de variables. Les premières techniques ne sont généralement pas optimales en terme d'estimation mais conservent les structures de covariance des variables transférées, les secondes sont probablement plus fiables en matière d'estimation de moyenne ou proportion mais moins bonnes pour l'estimation des covariances. Pour comparer et valider des méthodes de fusion, il n'existe pas de méthodologie à proprement parler mais on s'accorde pour dire qu'il est avant tout nécessaire que les distributions marginales et croisées des variables soient bien restituées. C. Derquenne propose quatre degrés de validation comme première approche méthodologique. Ils peuvent servir de critères de validation pour tout type de méthodes de fusion :

1. *Marginale univariée* : comparaison des distributions marginales de chaque variable entre une enquête existante et l'échantillon généré.
2. *Modalité univariée* : comparaison des proportions de modalités de chaque variable entre une enquête existante et l'échantillon généré.
3. *Corrélations* : comparaison des corrélations entre chaque couple de variables à estimer entre une enquête existante et l'échantillon généré.
4. *Modalités croisées* : comparaison des proportions de modalités croisées de chaque couple de variable entre une enquête existante et l'échantillon généré.

Nous pouvons cependant dire que la qualité d'une fusion dépendra de deux facteurs :

1. L'existence d'un ensemble important de variables communes entre les deux sources.
2. L'importance de la corrélation existant entre les variables communes et les variables de l'enquête que l'on souhaite transférer (estimer).

Certaines précautions sont également nécessaires préalablement à tout processus de fusion proprement dit, en particulier un redressement des fichiers donneurs conformément au fichier receveur.

En conclusion, les méthodes de fusion statistique de fichiers offrent aux grandes entreprises et aux instituts des perspectives nouvelles et bien moins coûteuses qui permettent de répondre aux difficultés de constituer de grandes bases de données. Néanmoins il est fondamental de rappeler que les résultats obtenus grâce aux diverses méthodes de fusion statistique et greffe de fichiers de données doivent être manipulées avec la plus grande précaution, l'utilisateur gardant toujours présent à l'esprit qu'elles n'augmentent pas l'information disponible dans les tableaux originaux mais servent à en enrichir l'interprétation.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Chapitre III :
Nouvelles approches développées

III-1. Matériel statistique nécessaire au développement des nouvelles méthodes

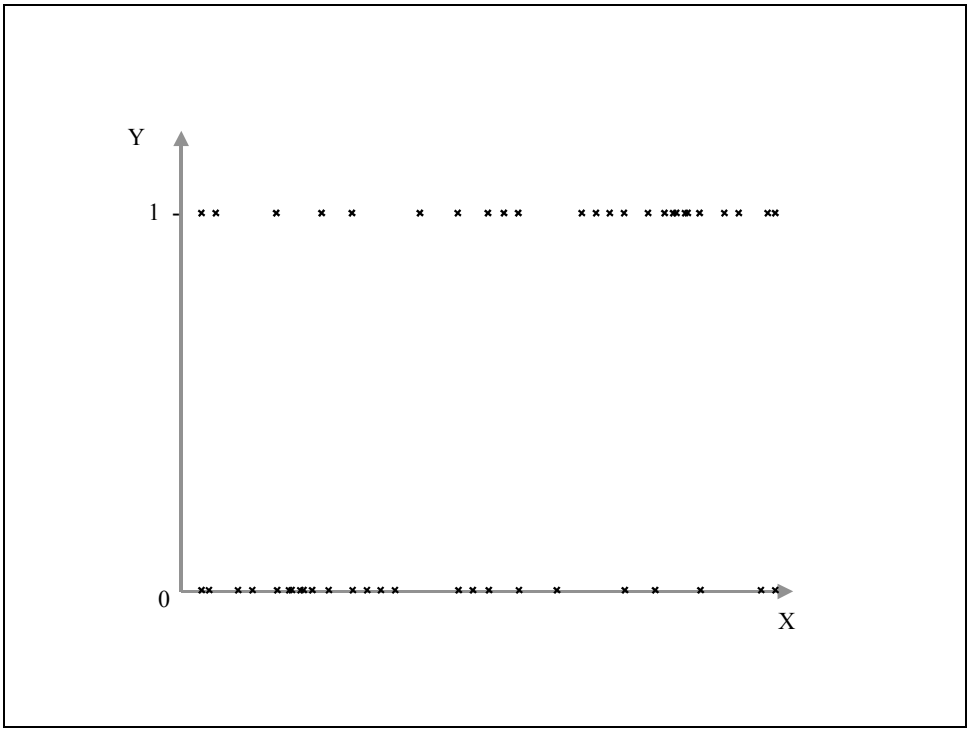
III-1.1. *Rappel sur la régression logistique*

Selon le type de variables d'intérêt que l'on cherche à étudier, on choisira le type de modélisation appropriée.
Dans le cadre de ce projet nous nous sommes intéressés à des variables d'intérêt qualitatives et principalement aux deux catégories suivantes :

- Variables ordinales (échelle de satisfaction,…)
- Variables booléennes (abandon ou non du chauffage électrique,…)

Nous allons développer la méthode de la régression logistique et montrer qu'elle est bien adaptée à la modélisation de variables qualitatives de types précédents par des variables candidates à l'explication, qui peuvent être quantitatives ou qualitatives.

Considérons tout d'abord le cas d'un échantillon sur lequel on a mesuré deux variables X et Y. On cherche à savoir si la variable quantitative X a un effet sur la variable booléenne Y. Dans un premier temps il est logique de regarder la distribution des individus dans le repère X-Y.



On observe deux traces de points suivant les demi-droites positives Y=0 et Y=1. On peut également remarquer que pour la valeur Y=0, les individus sont plus concentrés vers les petites valeurs de X tandis que pour Y=1 ils

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

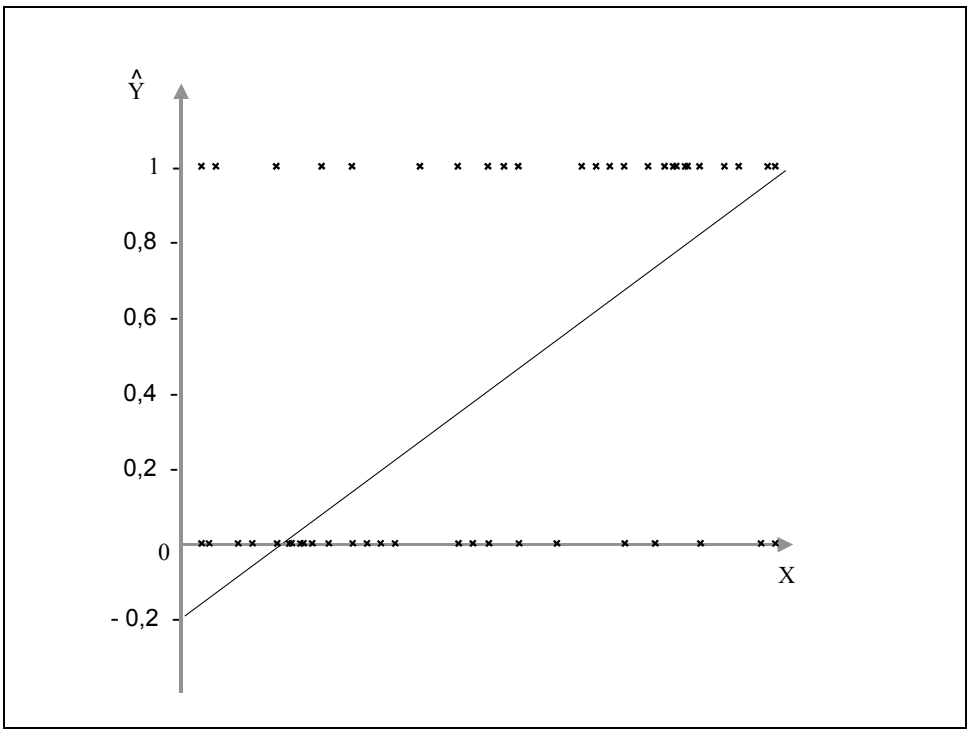
sont plus représentés vers les grandes valeurs de X. Ceci ne suffit toutefois pas pour affirmer l'existence d'une corrélation significative entre ces deux variables.
Le modèle usuel de régression linéaire simple peut s'écrire :

$$Y = E(Y/X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon \quad \text{avec} \quad \varepsilon \sim> N(0,\sigma)$$

alors le modèle estimé est :

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

où $\hat{\beta}_0$ et $\hat{\beta}_1$ sont estimés par la méthode des moindres carrés ordinaires. Le graphique suivant exprime le résultat de la régression :



Ce graphique montre bien que la prédisposition p de la modalité Y=1 croit linéairement avec la variable X d'un accroissement correspondant au coefficient de la pente de régression β_1. Cette prédisposition peut être considérée comme la probabilité p que Y soit égale à 1. On posera ainsi :

$$p(x) = \Pr(Y = 1/X = x)$$

On voit sur le graphique précédent les inconvénients d'un tel modèle linéaire en probabilité. Un accroissement constant de X produit également un accroissement constant de p. De plus p peut très bien avoir des valeurs hors de l'intervalle [0,1], ce qui est logiquement impossible pour une probabilité.

De plus pour une valeur de X donnée, les résidus ε ne peuvent prendre que deux valeurs possibles :

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

- Si Y=0, alors $\varepsilon = -p(x)$ avec une probabilité 1-*p(x)*
- Si Y=1, alors $\varepsilon = 1-p(x)$ avec une probabilité *p(x)*

Or ceci signifie que ε ne suit pas une loi continue ce qui est contraire à l'hypothèse de normalité des résidus introduite dans le cadre du modèle linéaire. De plus ε ayant une espérance nulle, sa variance est égale à [illegible]. Ce résultat est à nouveau en contradiction avec l'hypothèse de variance constante des résidus du modèle linéaire usuel car la variance dépend de *X*.
On en conclut donc que la régression linéaire simple usuelle n'est pas une bonne solution pour résoudre ce type de problèmes.

Enfin à la lumière de ces constatations, la distribution conditionnelle de Y suit une loi de Bernoulli avec une probabilité fournie par l'espérance conditionnelle :

$$p(x) = E[Y/X=x]$$

Donc si $Y/X=x$ ~> $\mathcal{B}(1,p(x))$ alors :

$$\Pr(Y=y/X=x) = p(x)^y(1-p(x))^{1-y}$$

Ceci peut encore s'écrire :

$$\Pr(Y=y/X=x) = \left(\frac{p(x)}{1-p(x)}\right)^y(1-p(x)) = \exp^{y\ln\left(\frac{p(x)}{1-p(x)}\right)+\ln(1-p(x))} = \exp^{yh(x)+\ln(1-p(x))}$$

où h(x) est la transformation logit suivante :

$$h(x) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0+\beta_1 x$$

d'où :

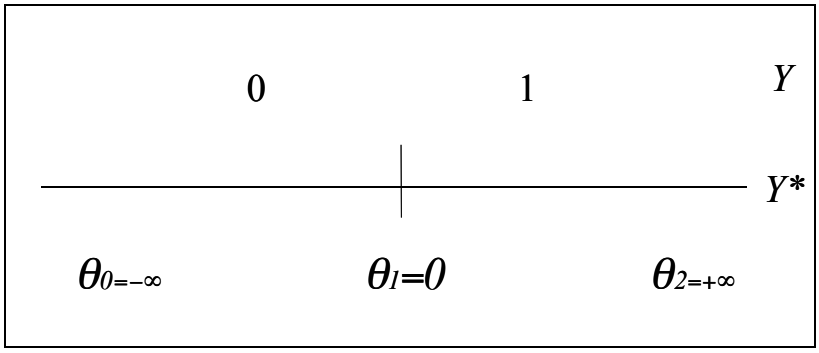
$$p(x) = \Pr(Y=1/X=x) = E[Y/X=x] = \frac{\exp^{\beta_0+\beta_1 x}}{1+\exp^{\beta_0+\beta_1 x}} = g(\beta_0+\beta_1 x)$$

Ayant introduit la fonction Logit nous allons désormais introduire le concept de variable latente qui sera essentielle lors de la description de la régression logistique ordinale.

Nous faisons l'hypothèse que la valeur de la variable booléenne Y précédente est le résultat d'une variable inobservable Y^*, dite latente, caractérisant le degré d'accord avec la modalité Y=1. Ceci peut se présenter sous la forme suivante :

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html



Où θ_1 représente la frontière entre le « 0 » et le « 1 », ou encore l'unique barreau d'une échelle. On peut poser $\theta_1 = 0$, sans perte de généralité, puisqu'il s'agit seulement d'une translation sur l'axe de Y^* qui appartient à la droite réelle. La variable inobservable Y^* est donc continue et nous permet d'avoir une équivalence avec Y la variable observée. En raisonnant en terme d'espérance conditionnelle, on peut écrire le modèle théorique précédemment vu :

$$E(Y^*/X) = X\beta \quad \text{avec} \quad Y^* = X\beta + \varepsilon$$

et

$$E(Y/X) = \frac{\exp^{X\beta}}{1+\exp^{X\beta}} = \Pr[Y=1/X]$$

alors :

$$\Pr[Y=1/X] = \Pr[Y^*>\theta_1/X] = \Pr[X\beta+\varepsilon>0/X] = \Pr[\varepsilon>-X\beta/X] = 1-F_\varepsilon(-X\beta) = F_\varepsilon(X\beta)$$

où F_ε est la fonction de répartition de ε, que l'on considère associée à une fonction logistique, telle que :

$$F_\varepsilon(X\beta) = \frac{\exp^{X\beta}}{1+\exp^{\beta}} = E(Y/X) = \Pr[Y=1/X]$$

On peut appliquer ce résultat sur un exemple simple où l'on dispose seulement de la proportion des modalités « 1 » et « 0 », alors le modèle théorique sera :

$$Y^* = \beta_0 + \varepsilon$$

d'où

$$E(Y/X) = \Pr[Y=1/X=cte]$$

alors on peut estimer cette probabilité à l'aide de la proportion de « 1 » observée, ce qui est la façon la plus naturelle, c'est à dire :

$$\hat{\Pr}[Y=1/X=cte] = \frac{\exp^{\hat{\beta}_0}}{1+\exp^{\hat{\beta}_0}} = \frac{n_1}{n}$$

alors :

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

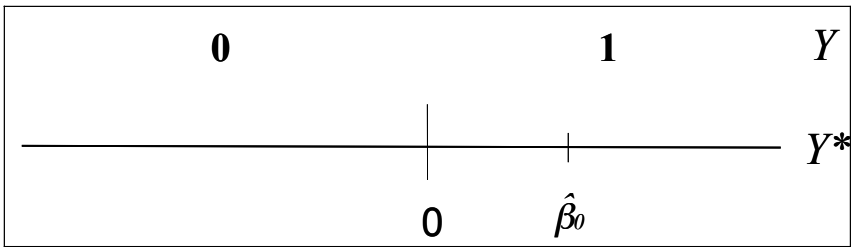
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

$$\hat{\beta}_0 = \ln\left(\frac{\hat{p}^1}{1-\hat{p}_1}\right)$$

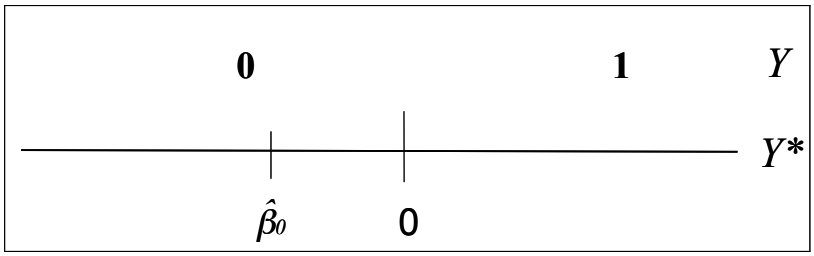
d'où :

$$\hat{y}^* = \hat{\beta}_0$$

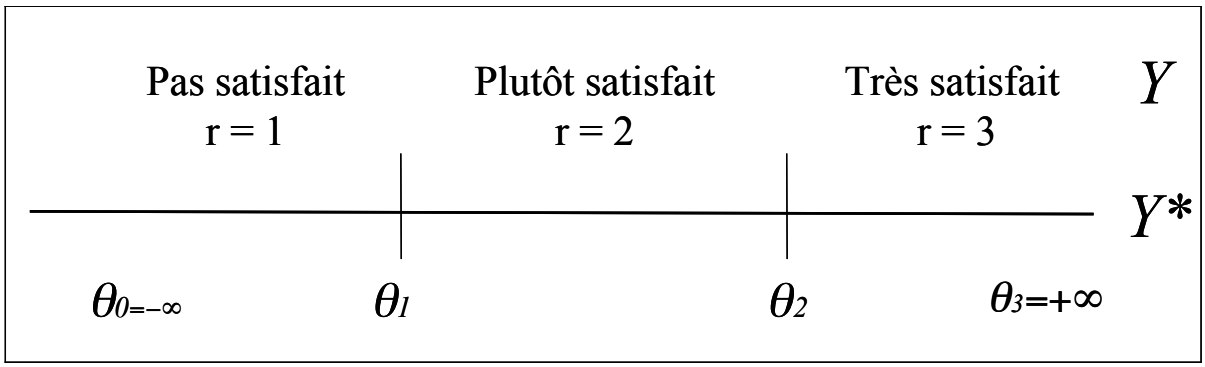
Ce qui peut se visualiser ainsi en reprenant la schématisation précédente :



$\hat{\beta}_0$ sera positif s'il y a plus d'individus ayant la modalité « 1 » et donc $\hat{y}^*$ le sera aussi.
Alors qu'il sera négatif si la proportion de 1 n'est pas majoritaire :



Dans le cadre de ce travail nous nous sommes plus particulièrement intéressés à des enquêtes de satisfaction clientèle. Les variables d'intérêt ne sont plus alors booléennes comme précédemment, mais des variables ordinales dont les modalités correspondant aux niveaux de la satisfaction, par exemple : « pas satisfait », « plutôt satisfait » et « très satisfait ».
On peut reprendre la notion de variable latente introduite précédemment et illustrée comme suit notre échelle ordinale :



où le nombre de catégories de réponse possible est bien R=3 (r=1 à 3).

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

De la même manière que dans le cadre développé en régression logistique binaire, nous pouvons raisonner en terme d'espérance conditionnelle :

$$E(Y^*/X) = \beta_0 + X\beta$$

avec :

$$Y^* = \beta_0 + X\beta + \varepsilon$$

A l'aide de l'échelle ordinale nous pouvons écrire :

(1) $\Pr[Y \le 1/X] = \Pr[Y^* \le \theta_1/X] = \Pr[\beta_0 + X\beta + \varepsilon \le \theta_1/X] = \Pr[\varepsilon \le \theta_1 - \beta_0 - X\beta/X] = F_\varepsilon(\theta_1 - \beta_0 - X\beta)$

(2) $\Pr[Y \le 2/X] = \Pr[Y^* \le \theta_2/X] = F_\varepsilon(\theta_2 - \beta_0 - X\beta)$

(3) $\Pr[Y \le 3/X] = \Pr[Y^* \le \theta_3/X] = F_\varepsilon(+\infty) = 1$

Par conséquent, chaque probabilité de réponse est égale à :

(1) $\Pr[Y = 1/X] = \Pr[Y^* \le \theta_1/X] - \Pr[Y^* \le \theta_0/X] = \Pr[Y^* \le \theta_1/:X] = F_\varepsilon(\theta_1 - \beta_0 - X\beta)$
car $\theta_0 = -\infty$, d'où $\Pr[Y^* \le \theta_0/X] = 0$.

(2) $\Pr[Y = 2/X] = \Pr[Y^* \le \theta_2/X] - \Pr[Y^* \le \theta_1/X] = F_\varepsilon(\theta_2 - \beta_0 - X\beta) - F_\varepsilon(\theta_1 - \beta_0 - X\beta)$

(3) $\Pr[Y = 3/X] = \Pr[Y^* \le \theta_3/X] - \Pr[Y^* \le \theta_2/X] = 1 - \Pr[Y^* \le \theta_2/X] = 1 - F_\varepsilon(\theta_2 - \beta_0 - X\beta)$

alors si F_ε est la fonction de répartition de la distribution logistique :

(1) $\Pr[Y = 1/X] = \dfrac{\exp^{\theta_1 - \beta_0 - X\beta}}{1 + \exp^{\theta_1 - \beta_0 - X\beta}}$

(2) $\Pr[Y = 2/X] = \dfrac{\exp^{\theta_2 - \beta_0 - X\beta}}{1 + \exp^{\theta_2 - \beta_0 - X\beta}} - \dfrac{\exp^{\theta_1 - \beta_0 - X\beta}}{1 + \exp^{\theta_1 - \beta_0 - X\beta}}$

(3) $\Pr[Y = 3/X] = 1 - \dfrac{\exp^{\theta_2 - \beta_0 - X\beta}}{1 + \exp^{\theta_2 - \beta_0 - X\beta}} = \dfrac{1}{1 + \exp^{\theta_2 - \beta_0 - X\beta}}$

Cependant, lors de l'estimation du modèle, on ne pourra pas séparer θ_1 et β_0 d'une part, et β_0 et θ_2 d'autre part. En d'autres termes, on obtiendra $\alpha_1 = \theta_1 - \beta_0$ et $\alpha_2 = \theta_2 - \beta_0$. Alors pour retrouver les deux θ, on posera $\beta_0 = -(\alpha_1 + \alpha_2)/2$, $\theta_1 = \alpha_1 + \beta_0$ et $\theta_2 = \alpha_2 + \beta_0$, sans aucune perte de généralité, car β_0 n'est qu'une translation. Par conséquent, $\theta_1 = (\alpha_1 - \alpha_2)/2$ et $\theta_2 = (\alpha_2 - \alpha_1)/2$, ce qui est appréciable puisque ces deux coefficients seront symétriquement opposés : $\theta_1 = -\theta_2$. On dit alors que le modèle est basé sur la probabilité de réponse cumulée $p_r = \Pr[Y \le r/X]$ et quand l'on utilise la fonction logistique, il s'agit d'un modèle logistique cumulé ou ordinale.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

En effet, on a :

1. $\Pr[Y \leq 1/X] = \dfrac{\exp^{\alpha_1 - X\beta}}{1+\exp^{\alpha_1 - X\beta}}$

2. $\Pr[Y \leq 2/X] = \dfrac{\exp^{\alpha_2 - X\beta}}{1+\exp^{\alpha_2 - X\beta}}$

Il est alors naturel de réécrire ces probabilités en fonction de la transformation logit :

$$\ln\left(\frac{\Pr[Y \leq r/X=x]}{1-\Pr[Y \leq r/X=x]}\right) = \alpha_r - X\beta \qquad \forall r=1,R=1$$

Ce modèle est encore appelé ”modèle à rapport de chances proportionnelles”, c’est à dire que pour deux valeurs de X, on aura :

$$\frac{\left[\dfrac{\Pr[Y \leq r/X=x_1]}{1-\Pr[Y \leq r/X=x_1]}\right]}{\left[\dfrac{\Pr[Y \leq r/X=x_2]}{1-\Pr[Y \leq r/X=x_2]}\right]} = \exp^{-(x_1 - x_2)\beta}$$

qui, comme on peut le constater, est indépendant de la réponse *r*. En d’autres termes, les chances de donner une réponse inférieure ou égale à *r* sont $exp^{-(x_1 - x_2)\beta}$ fois plus grandes en $X=x_1$, qu’en $X=x_2$.

De plus le signe négatif dans l’équation de la transformation logit assure que pour de grandes valeurs de $X\beta$, la probabilité associée aux catégories élevées de *Y* s’accroît.

III-1.2. *Rappel sur la régression PLS*

Dans le cas de données fortement corrélées (mais aussi en présence de données manquantes), la régression PLS apporte une alternative intéressante à la régression classique en permettant entre autres de tenir compte dans la modélisation de la structure de corrélation existante.

L'algorithme de régression PLS a été proposé par Wold, Albano *et al* (1983). C'est un algorithme itératif de décomposition orthogonale de l'espace engendré par *X* et *Y*, faite par une succession de régressions effectuées sur les résidus.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

On distingue néanmoins classiquement la régression PLS1 permettant de modéliser une seule variable Y de l'algorithme plus général de la régression PLS2 où l'on cherche à expliquer plusieurs variables Y simultanément.

Tenenhaus (1998) présente les propriétés mathématiques de la régression PLS2, et plus de développements sur la méthode.

III – 1.2.1 L'algorithme de la régression PLS1

La régression PLS1[1] consiste à réaliser la régression d'une variable à expliquer Y sur des variables explicatives $X_1,\ldots,X_p$ pouvant être hautement corrélées entre elles. On cherche de plus à fournir des coefficients de régression interprétables, c'est à dire que l'on puisse mesurer la contribution de la variable X_j à la variable Y à l'aide du coefficient de régression. En effet, si les signes du coefficient de régression et de la corrélation entre la variable explicative et la variable d'intérêt sont opposés, l'interprétation sera délicate pour le praticien.

L'algorithme peut se décomposer succinctement comme suit :

(1) On construit tout d'abord une composante :

$$t_1 = w_1 X_1 + \ldots + w_{1p} X_p$$

où

$$w_{1j} = \frac{\mathrm{cov}(X_j, Y)}{\sqrt{\sum_{j=1}^{p} \mathrm{cov}^2(X_j, Y)}}$$

Ces coefficients sont ensuite normalisés :

$$w_{1j} = \frac{w_{1j}}{\sqrt{\sum_{j=1}^{p} (w_{1j})^2}}$$

(2) Puis on effectue une régression simple de Y sur t_1 :

$$\hat{Y} = c_1 t_1 + Y_1$$

ou c_1 est le coefficient de régression et Y_1 le vecteur des résidus. D'où une première équation de régression :

$$\hat{Y} = c_1 w_{11} X_1 + \ldots + c_1 w_{1p} X_p + Y_1$$

dont les coefficients sont facilement interprétables.

[1] Le 1 signifie univariée

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

(3) Toutefois si l'on estime que le pouvoir explicatif de ce modèle est faible, on cherche une deuxième composante t_2, combinaison linéaire des X_j, non corrélée à t_1 et expliquant bien le résidu Y_1. Cette composante t_2 est combinaison linéaire des résidus X_{1j} des régressions des variables X_j sur la composante t_1 :

$$X_{1j} = X_j - p_{1j}t_1$$

où :

$$p_{1j} = \frac{\text{cov}(X_j, t_1)}{\text{var}(t_1)}$$

On obtient t_2 à l'aide de la formule suivante :

$$t_2 = w_{21}X_{11} + \ldots + w_{2p}X_{1p}$$

où :

$$w_{2j} = \frac{\text{cov}(X_{1j}, Y_1)}{\sqrt{\sum_{j=1}^{p} \text{cov}^2(X_{1j}, Y_1)}}$$

(4) On procède ensuite à une régression multiple de Y sur t_1, t_2 :

$$\hat{Y} = c_1t_1 + c_2t_2 + Y_2$$

Equation qui s'écrit en fonction des variables explicatives en remplaçant les composantes $t_1$et t_2 par leur expression en fonction des X_j :

$$\hat{Y} = c_1 \sum_{j=1}^{p} w_{1j}X_j + c_2 \sum_{j=1}^{p} w_{2j}\left(X_j - p_{1j} \sum_{j=1}^{p} w_{1j}X_j \right)$$

(5) On peut itérer ce processus en utilisant de la même manières les résidus Y_2, $X_{21},\ldots,X_{2p}$ des régressions de Y, $X_1,\ldots,$ X_p sur t_1, t_2.
Le nombre de composantes t_h à retenir est habituellement déterminé par validation croisée.
Pour chaque valeur h, on calcule des prédictions de la variable d'intérêt Y. Plus précisément on va calculer les prédictions $\hat{Y}_{hi}$ de Y_i en utilisant toutes les individus puis $\hat{Y}_{h(-i)}$ sans utiliser l'individu i. Puis on calcule les critères RSS_h (Residual Sum of Squares) et $PRESS_h$ (Prediction Error Sum of Squares) définis par :

$$RSS_h = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{hi} \right)^2$$

et

$$PRESS_h = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{h(-i)} \right)^2$$

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

La composante t_h sera retenu si :

$$\sqrt{PRESS_h} \leq 0.95 \leq \sqrt{RSS_{h-1}}$$

III – 1.2.2 L'algorithme de la régression PLS2

Dans toute cette partie, nous notons Y une matrice nxq dont les colonnes $y_1,\ldots,y_q$ sont des variables appartenant à R^n qui représentent les réponses, et X une matrice nxp formée de p variables prédicteurs $X_1,\ldots,X_p$. Soit $R_{11} = (k_{ij})$, $i, j = 1, p$ la matrice de covariance de X où les éléments $k_{i,j} = \mathrm{Cov}(X_i,X_j)$ et $R_{12} = Cov(X,Y)$ la matrice de pxq des covariances entre X et Y, et $R_{12} = R_{21}$.

Etape 0 : On considère X_0 et Y_0 les tableaux centrés réduits obtenus à partir de X et Y respectivement.

Etape 1 : On cherche $t_1 = X_0\ w_1$ et $u_1 = Y_0\ c_1$, les deux variables qui maximisent $cov(u_1, t_1)$ sous la contrainte $||w_1||=||c_1||=1$.
On régresse ensuite les deux variables X_0 et Y_0 sur la variable t_1 obtenant les résidus X_1 et Y_1 respectivement; d'où les équations : $X_0 = t_1\ p'_1+X_1$ et $Y_0= t_1\ c'_1+Y_1$. Par la méthode des multiplicateurs de Lagrange, on trouve que w_1 et b_1 sont respectivement, les vecteurs propres associés aux plus grandes valeurs propres de $X'_0Y_0Y'_0X_0$ et de $Y'_0X_0X'_0Y_0$. Et dans le cas particulier de la régression PLS1, où Y_0 est réduit à une seule variable réponse Y_1, on a c_1 et :

$$w_1 = \frac{[cor(X_1,Y_1),\ldots,cor(X_p,Y_1)]}{\sqrt{\sum_{j=1}^{p} cor^2(X_j;Y_1)}}$$

Etape 2 : On itère l'*étape 1* en remplaçant les tableaux de départ X_0 et Y_0 par les tableaux des résidus X_1 et Y_1.
On commence par initialiser u_2 à la première colonne du résidu Y_1, ce qui va nous permettre de calculer le coefficient w_2 :

$$w_2 = X'_1u_2/u'_2u_2$$

Une fois a_2 normalisé, on obtient t_2 par régression de X_1 sur w_2 :

$$t_2 = X_1\ w_2/\ w'_2\ w_2$$

On régresse ensuite les deux tableaux X_1 et Y_1 sur l'axe t_2 d'où les équations:

$$X_1 = t_2\ p'_2 + X_2 \quad et \quad Y_1= t_2\ c'_2 + Y_2.$$

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Ou encore :

$$X_0 = t_1 p'_1 + t_2 p'_2 + X_2 \quad et \quad Y_0 = t_1 c'_1 + t_2 c'_2 + Y_2.$$

On recalcule le nouveau vecteur u_2 :

$$u_2 = Y_1 c_2 / c'_2 c_2$$

De là on reprend *l'étape 2* au départ avec un nouveau w_2 et ce jusqu'à convergence de w_2.

Le schéma suivant illustre le rôle de composantes PLS que sont t et u pour les tableaux de données X et Y. Il rappelle aussi que la première composante principale de X (resp.de Y) « explique » mieux les données que la première composante PLS t_1 (resp. t_2).

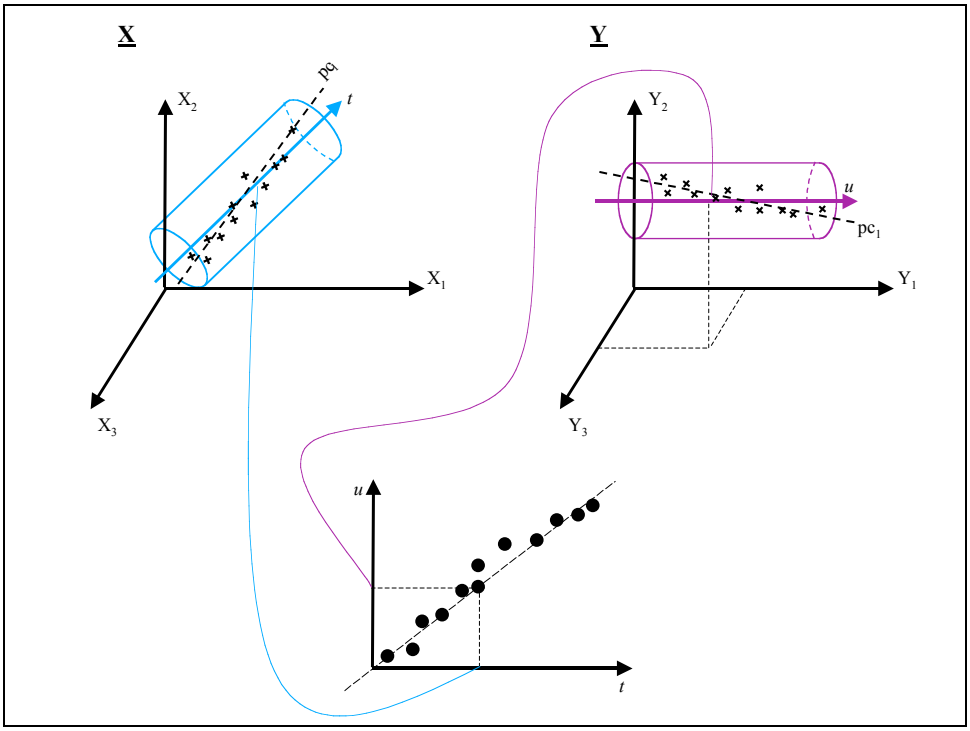


Figure 1: Représentation et lien des composantes PLS

Etape 3 : On réitère la procédure H fois, jusqu'à ce que les composantes $t_1,...,t_H$ expliquent suffisamment Y_0. Les $t_1,...,t_H$ sont des combinaisons linéaires des p colonnes de X_0 de variance 1 et non corrélées mutuellement, et H est un entier compris entre 1 et le rang de X_0. Enfin de l'expression de Y_0 on obtient les équations de régression PLS :

$$Y_k = m_{k0} + m_{k1} X_1 + ... + m_{kp} X_p + Y_{Hk} \text{ pour } k=1,q.$$

où : Y_{Ak} est la $k^{ième}$ colonne de Y_k résidu à la $H^{ième}$ étape.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Et $m_{kj} = wk_j (p'k_j wk_j)^{-1} r'k$

Pour déterminer le nombre de composantes PLS $t_1,\ldots,t_H$ à retenir, on procède généralement par validation croisée. Comme vu précédemment en régression PLS1, à chaque étape h, on calcule pour chaque variable Y_k les quantités $PRESS_{kh}$ et RSS_{kh}. Ce qui permet de calculer l'indice :

$$Q^2_{kh} = 1 - \frac{PRESS_{kh}}{RSS_{k(h-1)}}$$

Puis on calcule un critère global sur l'ensemble des variables Y :

$$Q^2_h = 1 - \frac{\sum_{k=1}^{q} PRESS_{kh}}{\sum_{k=1}^{q} RSS_{k(h-1)}}$$

Avec ces indices on peut ainsi mesurer l'apport marginal de chaque composante PLS t_h au pouvoir prédictif du modèle.
On se fixe ensuite deux critères possibles permettant de décider si l'apport d'une nouvelle composante t_h est significative :

1. $Q^2_h \geq (1-0.95^2) = 0.0975$

2. Il y a au moins un k tel que : $Q^2_{hk} \geq 0.0975$

III – 1.2.3 La régression logistique PLS

Cette extension a été proposée par Tenenhaus (2000). Nous venons de voir l'algorithme de la régression PLS, univariée et multivariée. Ces méthodes sont particulièrement utiles dans des conditions où la régression multiple fonctionne mal. Or on peut rencontrer des problèmes analogues en régression logistique. Cependant il est tout à fait possible de transposer les principes de la régression PLS à la régression logistique. Nous allons ainsi décrire l'algorithme de régression logistique PLS.

La régression logistique PLS consiste à adapter l'algorithme de régression PLS1 au cas d'une variable d'intérêt qualitative, binaire ou ordinale. Il est utile de rappeler que la régression logistique PLS est univariée et ne s'applique donc qu'à la modélisation d'une seule variable d'intérêt.
Considérons la variable d'intérêt Y que l'on cherche à modéliser à l'aide des variables $x_1,\ldots,x_p$ candidates à l'explication. Dans le cas d'une variable d'intérêt Y ordinale, la régression logistique est construite en supposant un modèle à rapport des chances proportionnel comme cela a été exposé auparavant.
Comme précédemment on recherche successivement les composantes PLS orthogonales t_h, combinaisons des variables explicatives X_0.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

- Recherche de la première composante PLS t_1 :

Etape1 : on calcule les coefficients de régression w_{1j} dans la régression logistique simple de Y sur chaque variable x_j.

Etape2 : la composante PLS s'écrit alors :

$$t_1 = Xw_1 / w'_1 w_1$$

- Recherche de la deuxième composante PLS t_2 :

Etape 1 : on calcule le résidu X_1 de la régression de X sur la première composante t_1. On notera x_{1j} la j-ème colonne de la matrice X_1 :

$$X = t_1 p'_1 + X_1$$

Etape 2 : pour chaque valeur de j, on calcule le coefficient de régression w_{2j} de x_{1j} dans la régression logistique multiple de Y sur les variables t_1 et x_{1j}.

Etape 3 : la composante PLS t_2 s'écrit alors :

$$t_2 = X_1 w_2 / w'_2 w_2$$

On peut aussi réécrire la composante t_2 en fonction de X_0 :

$$t_2 = \frac{w_2}{w_2 w_2}\left(\frac{1 - w_1 p_1}{w_1 w_1}\right) X_0 = X_0 w_2^*$$

- Recherche de la *h-ième* composante PLS t_h :

Etape 1 : on calcule le résidu X_{h-1} de la régression de X sur les premières composantes $t_1,\ldots,t_{h-1}$. On notera $x_{h-1,j}$ la j-ème colonne de la matrice X_{h-1} :

$$X = t_1 p'_1 + \ldots + t_{h-1} p'_{h-1} + X_{h-1}$$

Etape 2 : pour chaque valeur de j, on calcule le coefficient de régression w_{hj} de $x_{h-1,j}$ dans la régression logistique multiple de Y sur les variables $t_1,\ldots,t_{h-1}$ et $x_{h-1,j}$.

Etape 3 : la composante PLS t_2 s'écrit alors :

$$t_h = X_{h-1} w_h / w'_h w_h$$

Le choix du nombre de composantes PLS t_h est déterminé en régression PLS par validation croisée. On peut adapter ce critère vu précédemment au cas de la régression logistique PLS selon que la variable d'intérêt est

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

binaire ou ordinale. Tenenhaus [2000] présente en détails l'expression de ce critère.

Maintenant que nous avons vu comment calculer les composantes PLS, nous allons donner l'équation de la régression logistique PLS.
Pour chaque nouvelle composante t_h calculée, on procède à la régression logistique de Y sur les composantes $t_1,\ldots,t_h$. L'équation de régression logistique PLS est obtenue en exprimant cette équation en fonction des variables d'origine.
Considérons une variable booléenne Y, notant p la probabilité de l'événement Y=1, on obtient :

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = c_1 t_1 + \ldots + c_h t_h = c_1 X w_1^* + \ldots + c_h X w_h^*$$

Considérons une variable ordinale Y ayant trois modalités ("1","2","3") ainsi que la variable inobservable Y^*, dite latente, que nous avons étudiée précédemment. Nous avons alors vu que l'on peut écrire le modèle théorique suivant:

$$E(Y^*/X) = X\beta$$

D'après les résultats obtenus dans le cas d'un modèle logistique ordinal, les probabilités estimées de réponse s'écrivent alors :

$$\Pr[Y=1/X] = \frac{\exp^{\alpha_1 - X\beta}}{1+\exp^{\alpha_1 - X\beta}}$$

$$\Pr[Y\leq 2/X] = \frac{\exp^{\alpha_2 - X\beta}}{1+\exp^{\alpha_2 - X\beta}}$$

$$\Pr[Y=3/X] = \frac{1}{1+\exp^{\alpha_2 - X\beta}}$$

III-2. La fusion par reconstitution à l'aide de variables latentes

Considérons un ensemble de variables $Y_1,\ldots,Y_Q$, qualitatives ordinales ayant chacune $r_1,\ldots r_Q$ modalités, que l'on cherche à expliquer et un ensemble de variables communes aux deux fichiers $X_1,\ldots X_P$ de variables qualitatives (nominales, ordinales, booléennes) candidates à l'explication.
On commence par transformer les variables d'intérêt en variables-modalités binaires Y^*. Puis l'on procède à une analyse des correspondances multiples sur les nouvelles variables Y^* ainsi définies. On obtient ainsi un jeu $Z_1,\ldots,Z_r$ de composantes principales. Nous allons considérer ces composantes comme des variables latentes, étapes utiles dans l'estimation des variables d'intérêt.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Ainsi on va procéder à une régression PLS2 des composantes principales sur les variables candidates à l'explication initiales. Les nouvelles estimations des composantes principales permettront, en inversant la formule de l'analyse des correspondances multiples, de reconstituer les variables d'intérêts.

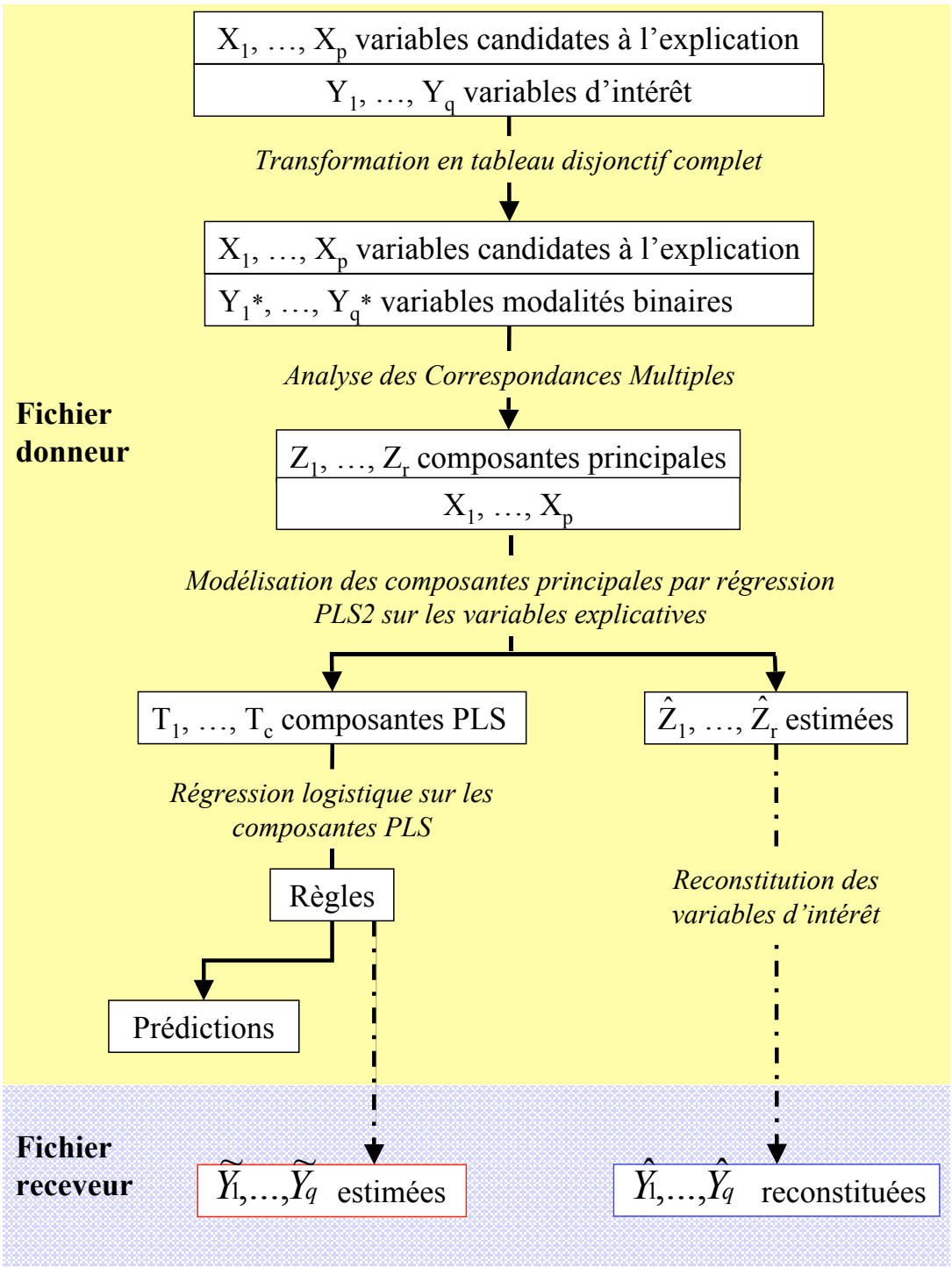


Figure2: Schéma de la méthode de reconstitution

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

III-3. Méthode univariée (Logit classique)

Considérons un ensemble de variables $Y_1,\ldots,Y_Q$, qualitatives ordinales ayant chacune $r_1,\ldots r_Q$ modalités, que l'on cherche à expliquer et un ensemble de variables communes aux deux fichiers $X_1,\ldots,X_P$ de variables qualitatives (nominales, ordinales, booléennes) candidates à l'explication.
Une première approche naturelle consiste à estimer chacune des Q variables Y_j au moyen de Q modèles de régression logistique ordinale sur les différents X_i. Cette méthode est univariée dans le sens où chaque variable Y_j est modélisée indépendamment des autres variables Y_j, mais elle permet de fournir un premier jeu de résultats indicatifs.

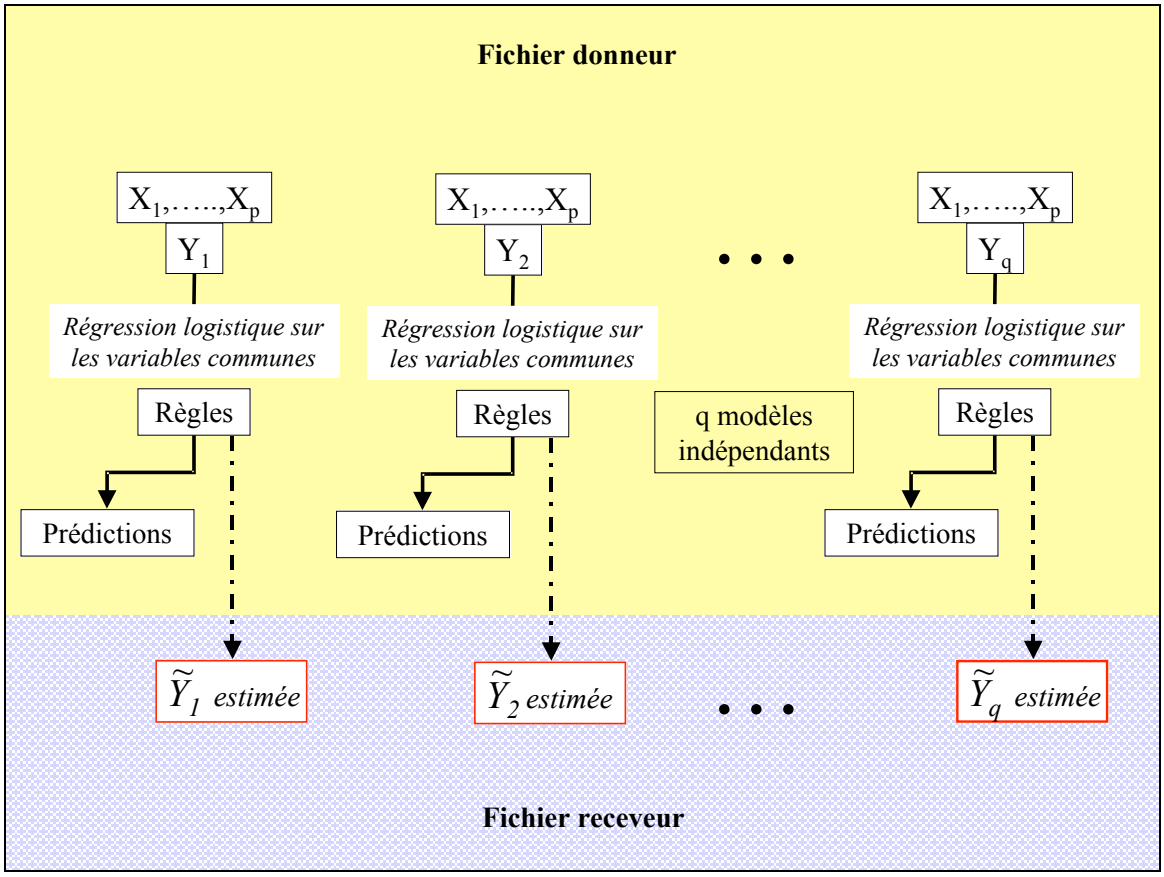


Figure 3: Schéma de la méthode univariée

III-4. Greffe séquentielle de variables

Cette méthode peut à proprement parler être considérée comme la véritable méthode de référence, fondée sur les travaux de C. Derquenne (Derquenne 98).
Considérons un ensemble de variables $Y_1,\ldots,Y_Q$ qualitatives ordinales ayant chacune $r_1,\ldots,r_Q$ modalités, que l'on cherche à expliquer et un ensemble $X_1,\ldots,X_P$ de variables qualitatives (nominales, ordinales, booléennes) candidates à l'explication. En s'appuyant toujours sur la régression

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

logistique ordinale l'idée est toutefois de tenir compte dans la modélisation des liens existant entre les différentes variables *Y*.

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 ne conserve que le modèle donnant le meilleur taux de bien classés pour la variable considérée (Y_k par exemple); et seule celle-ci sera greffée dans le fichier donneur.

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 dans les variables candidates à l'explication. 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 simple, est de tenir compte pour chaque futur *Y* à estimer de l'éventuel pouvoir explicatif des autres *Y* déjà greffées.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

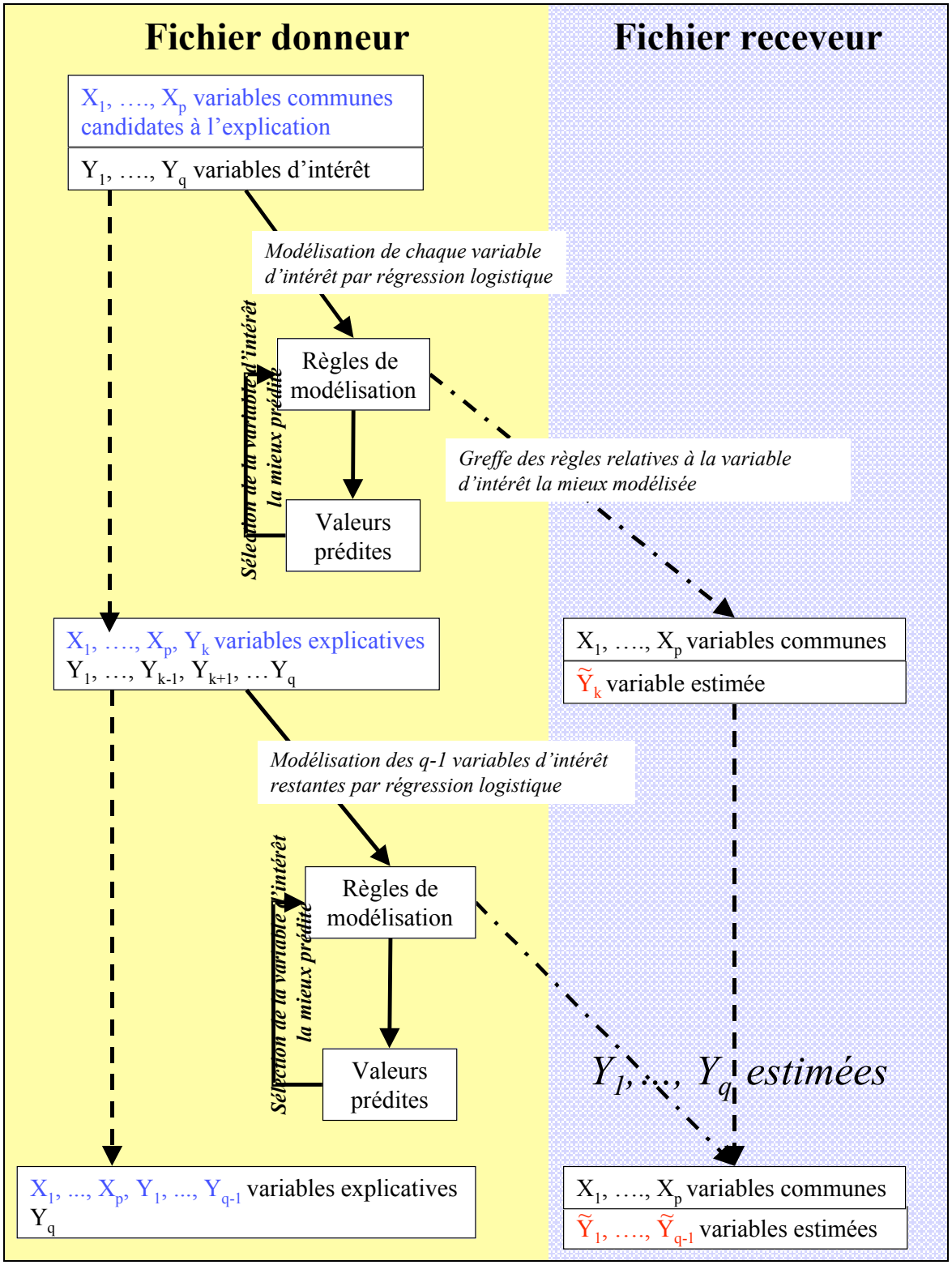


Figure 4: Schéma de la greffe séquentielle

Nous avons brièvement rappelé le principe de la régression PLS2 ; toutefois selon le nombre de variables à modéliser (une ou plusieurs) et le type de variables, nous ne pourrons pas appliquer le même type d'algorithme. En effet on peut différencier 4 situations :

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

1. *Une seule variable* Y_1 *quantitative* est modélisée par un ensemble de variables $X_1,...Xp$ quantitatives. C'est le cas de la **régression PLS1** (univariée).

2. *Plusieurs variables* $Y_1,..,Y_Q$ *quantitatives* sont modélisées par un ensemble de variables $X_1,...Xp$, quantitatives. C'est le cas classique de la **régression PLS2** (multivariée).

3. Une seule variable Y_1 qualitative est modélisée par un ensemble de variables $X_1,...Xp$, qualitatives ou quantitatives. C'est la cas de la **régression logistique PLS1**, méthode alternative développée par M. Tenenhaus (Tenenhaus 2000) à la régression PLS1 lorsque la variable à modéliser n'est pas quantitative.

4. Plusieurs variables $Y_1,...Y_Q$ qualitatives sont modélisées par un ensemble de variables $X_1,...Xp$, qualitatives ou quantitatives. Il n'existe pas à l'heure actuelle d'algorithme correspondant à ce cas de figure.

Par conséquent lors d'une analyse multivariée où les variables à expliquer sont qualitatives, deux solutions s'offrent à nous si nous souhaitons utiliser les algorithmes de la régression PLS :

1. Considérer les variables séparément et se ramener ainsi à des analyses univariées (méthode Logit PLS1), quitte à mettre en œuvre une seconde étape permettant de combiner par la suite les résultats issus de ces différentes modélisations (méthode Pseudo-PLS2).

2. Se ramener à des variables d'étude quantitatives de manière à pouvoir appliquer l'algorithme de régression PLS2 (Recodage 0-1 suivi de PLS2 et recodage Logit, suivi de PLS2).

Ce sont ces approches que nous allons présenter dans les paragraphes suivants.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

III-5. La méthode dite pseudo-multivariée (pseudo-PLS2)

Considérons un ensemble de variables $Y_1,\ldots,Y_q$ qualitatives ordinales ayant chacune $r_1,\ldots,r_Q$ modalités, que l'on cherche à expliquer et un ensemble $X_1,\ldots,X_P$ de variables qualitatives (nominales, ordinales, booléennes) candidates à l'explication.

Dans un premier temps, on va considérer les variables d'intérêt séparément et ainsi continuer à faire de l'analyse univariée. Mais contrairement aux méthodes vues précédemment, on va modéliser les variables d'intérêt, non plus par régression logistique ordinale, mais par régression logistique PLS 1 sur les variables candidates à l'explication. Cette première étape fournit ainsi pour chaque variable Y_j un jeu de composantes PLS T_{rj}.

Puis on va regrouper l'ensemble des composantes PLS précédentes issues des q modèles indépendants. L'idée sous-jacente ici est d'introduire une seconde étape dans la modélisation, combinant les résultats de la première étape de telle sorte que l'analyse des Y_j ne soit plus univariée mais multivariée.
En effet dans cette seconde étape on procède à nouveau à une régression logistique PLS 1 de chaque variable d'intérêt mais cette fois sur l'ensemble des composantes PLS précédemment regroupées. De ce fait chacune des variables Y_j aura été estimée au moyen d'informations, les composantes PLS de la première étape, issues de la modélisation de toutes les variables Y_j. Cependant on ne peut pas considérer cette méthode d'analyse comme une approche multivariée (car il s'agit seulement d'un regroupement d'analyses univariées). Par conséquent nous nommons cette méthode approche « pseudo-multivariée ».

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

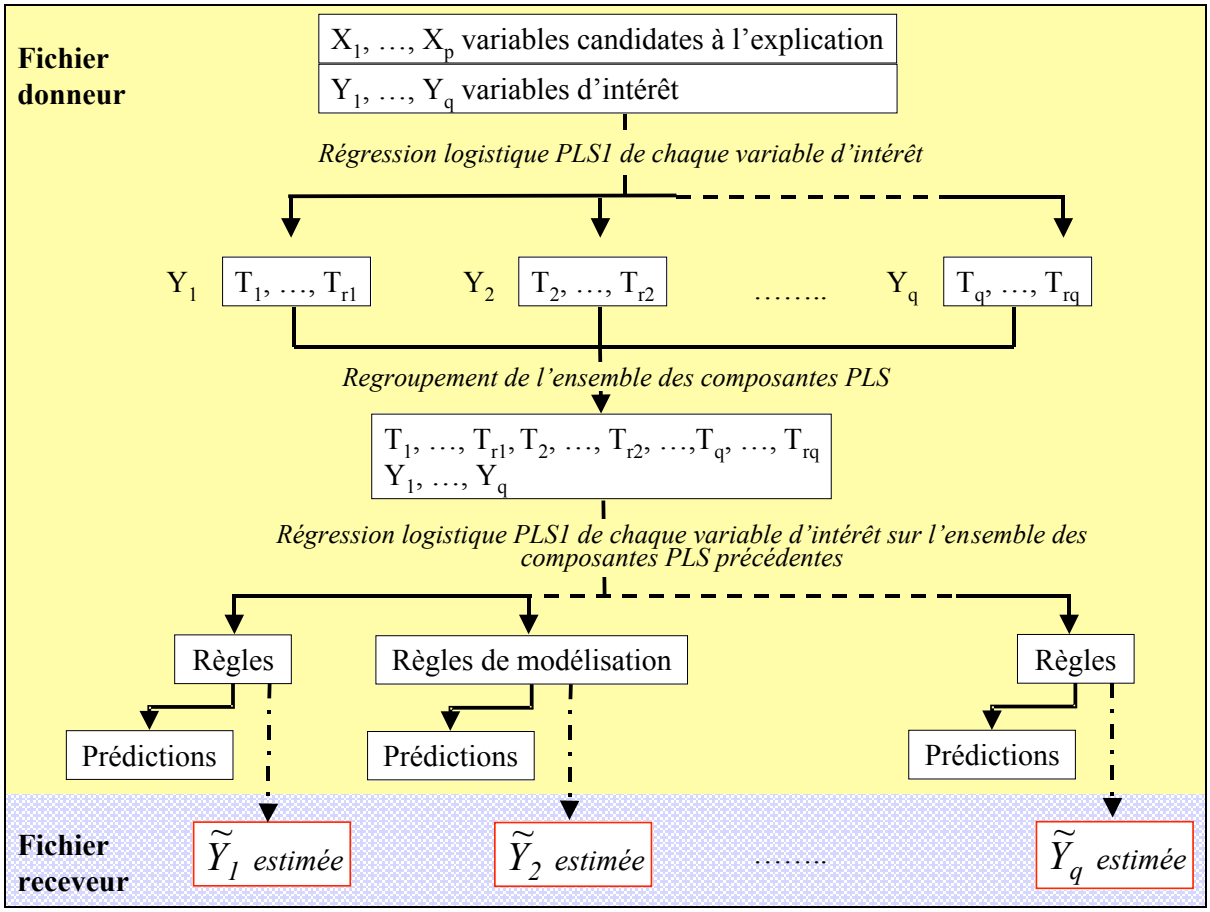


Figure 5: Schéma de la méthode pseudo-multivariée

III-6. La régression PLS2 avec recodage des données (0-1)

Considérons un ensemble de variables $Y_1,\ldots,Y_Q$, qualitatives ordinales ayant chacune $r_1,\ldots r_Q$ modalités, que l'on cherche à expliquer et un ensemble $X_1,\ldots X_P$ de variables qualitatives (nominales, ordinales, booléennes) candidates à l'explication. L'objectif de cette méthode est d'estimer simultanément les variables Y_i afin de préserver au mieux lors de la modélisation, la structure de corrélation existante. Dans pareil cas, une structure de corrélation forte entre variables explicatives ainsi qu'entre variables expliquées, l'utilisation de la régression PLS multivariée ou PLS2 peut s'avérer très efficace. Néanmoins cette méthode a été développée dans le cas où les variables à expliquer et les variables explicatives sont numériques ; c'est pourquoi il est nécessaire de transformer au préalable les variables de l'étude.

Une transformation relativement simple consiste à partir d'une tableau disjonctif complet formé par les jeux de variables X et Y de recoder comme suit les variables Y :

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

	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			.....
	$Y1_1$	$Y1_2$	$Y1_3$	$Y1_4$	$Y2_1$	$Y2_1$	$Y2_1$	.....
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	
.....								.....

On aurait théoriquement très bien pu utiliser le tableau disjonctif complet tel quel pour utiliser la régression PLS mais sans oublier que les variables-modalités sont considérées comme numériques et non booléennes dans cet algorithme. Ce recodage, s'il ne permet pas de se ramener à des variables continues, a le mérite de rendre le caractère ordinal des variables à expliquer grâce à une fonction de répartition à plusieurs bits (0-1). On peut ainsi espérer ne pas trop dégrader la reconstitution des variables marginales ainsi que le taux de bien classés obtenus tout en préservant les corrélations entre variables grâce à l'utilisation de l'algorithme PLS2.

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

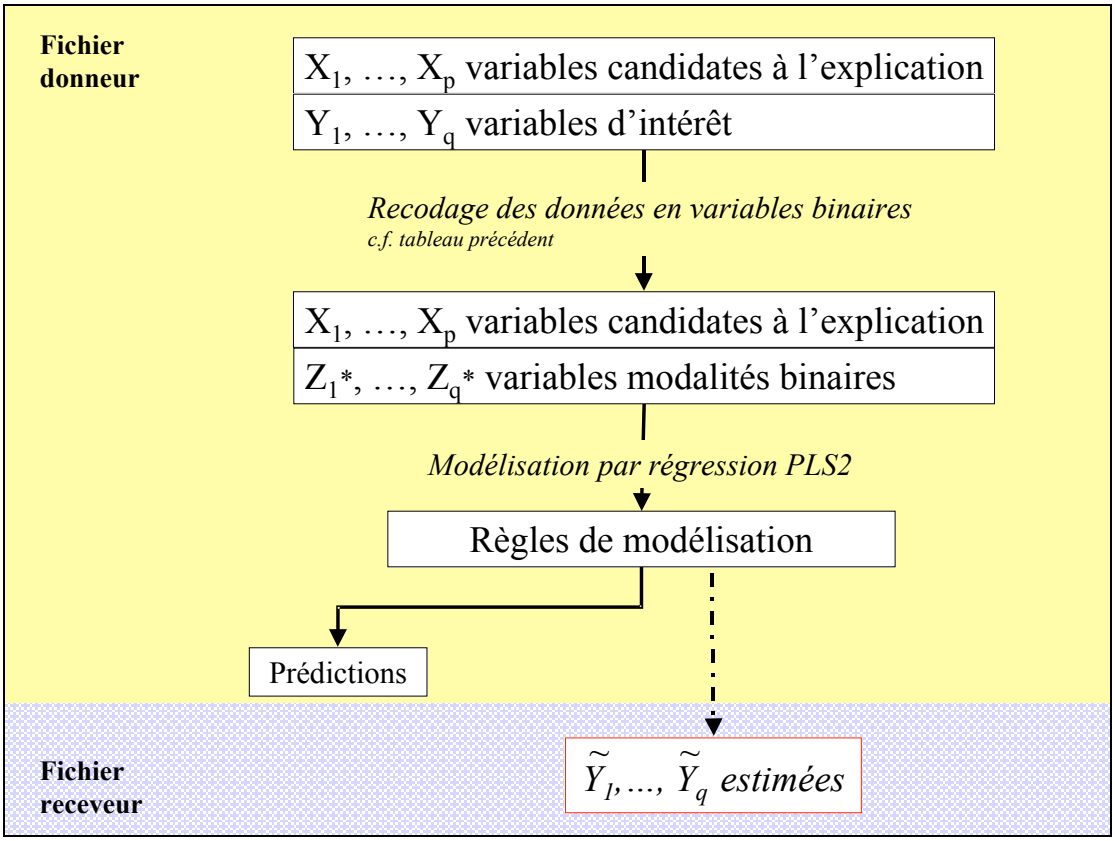


Figure 6: Schéma de la méthode PLS après recodage (0-1)

III-7. Méthode recodage logit suivi de la régression PLS2

Soient $Y_1,\ldots Y_q,\ldots,Y_Q$, Q variables ordinales à expliquer ayant respectivement $R_1,\ldots,R_q,\ldots,R_Q$ réponses possibles et $X_1,\ldots X_j,\ldots,X_P$, 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 :

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

$$g(\mu_{r(t)}^{(q)}) = \log\left(\frac{\Pr[Y_q \le r / t \in v_i]}{1-\Pr[Y_q \le r / t \in v_i]}\right) \qquad (1)$$

où r 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 :

$$g(\mu_{r(t)}^{(q)}) = \beta_0^{(q)} + \sum_{j=1}^{p}\sum_{l=1}^{m_j} a_{jl(t)} 1_{[t \in jl]}(jl) + \varepsilon_t \qquad (2)$$

où (2) est le logit cumulé théorique, alors que le logit observé est :

$$\log\left(\frac{\tilde{\Pr}[Y_g^{(r_q)} / v_i]}{1-\tilde{\Pr}[Y_q^{(r_q)} / v_i]}\right) = \log\left(\frac{\sum_{s=1}^{r} n_{qi}^{(s)}}{n_i - \sum_{s=1}^{r} n_{qi}^{(s)}}\right) \qquad (3)$$

Les probabilités cumulées théoriques prennent alors la forme suivante :

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

On crée alors de nouvelles variables Z_q à partir des logits observés.

$$Z_{qi}^{(r)} = \log\left(\frac{\sum_{s=1}^{r_q} n_{qi}^{(s)}}{n_i - \sum_{s=1}^{r_q} n_{qi}^{(s)}}\right) \quad (r_q = 1 \text{ à } R_q) \qquad (5)$$

Le nouveau tableau de données a toujours n individus, mais possède $H = \sum_{q=1}^{Q}(R_q - 1)$ 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 $\hat{Z}_{qi}^{(r)}$ 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 :

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

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

alors :

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

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 impossible de prendre en compte cette contrainte. Par conséquent, il faut fixer une valeur arbitraire à ces deux bornes qui fasse que les probabilités cumulées soient proches de zéro et de l'unité respectivement (par exemple -15, 3.10^{-7} et 15, $1\text{-}10^{-7}$).

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

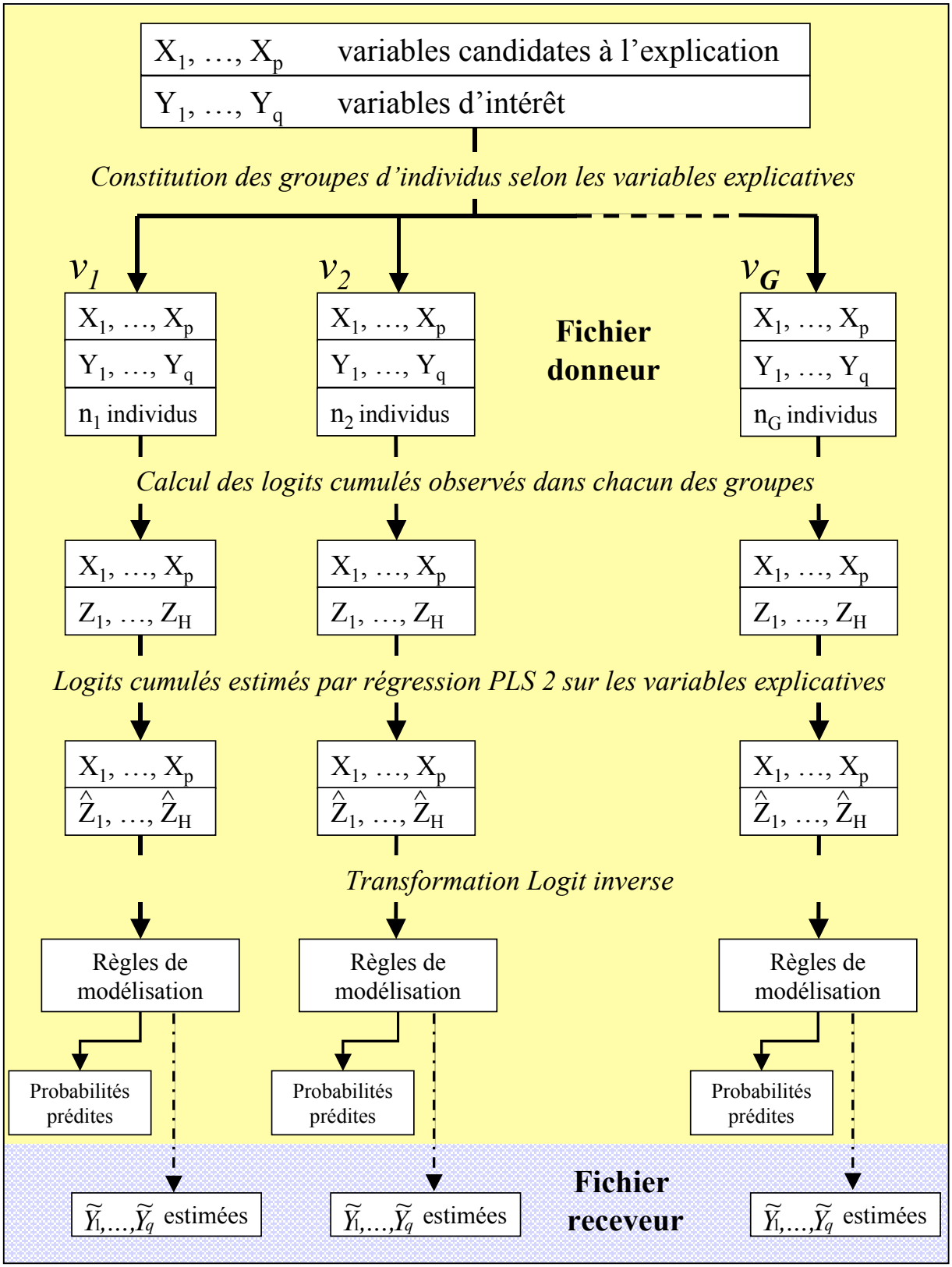


Figure 7: Schéma récapitulatif méthode Logit PLS2

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Chapitre IV :

Validation

IV-1. Critères globaux de validation des résultats

Afin d'illustrer les différents critères de validation statistique utilisés et de mieux les comprendre nous allons les définir à l'aide d'un exemple de modélisation, utilisant l'une des méthodes décrites dans le chapitre précédent appliqué à un petit échantillon de 200 individus. L'exemple que nous allons traiter a été mené à bien via le logiciel SAS version 8.2, tout comme l'ensemble des applications que nous avons développées.

Dans le cas présent le fichier donneur initial sera découpé aléatoirement en 2 échantillons :

- Un échantillon d'apprentissage constitué de 80% des individus.
- Un échantillon test comportant les 20% d'individus restant.

De manière plus générale, il est courant de découper l'échantillon initial en cinq échantillons de même taille, chacun servant à tour de rôle d'échantillon test, le reste du fichier constituant l'échantillon d'apprentissage. Ce principe de validation, consistant à faire permuter échantillon test et échantillon d'apprentissage, se nomme validation croisée et permet ainsi de fournir plusieurs jeux de résultats au lieu d'un seul. Chaque ligne des tableaux résultats correspond donc à une valeur de test effectué sur l'un des 5 échantillons tests (d'où les 5 lignes par case des différents tableaux de résultats présentés en annexe).

Nous avons ensuite défini trois critères principaux de validation empiriques compte tenu des attentes liées aux méthodes de fusion :

1. Reconstitution des marginales : comparaison des distributions marginales des variables observées et estimées (test d'indépendance du Chi2).
   Ex: pour la variable Thisyr qui a 2 modalités, on procède à un test d'indépendance du Chi2 à 1 degré de liberté et au calcul de la probabilité liée au rejet du test d'adéquation de la distribution estimée.

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.
   De plus pour chaque variable est calculé le taux de bien classés obtenu par la règle du maximum, autour duquel est construit un intervalle de confiance en considérant l'échantillon test comme un échantillon représentatif de la population (technique de sondage avec construction de l'intervalle de confiance pour la loi normale). Cet

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

intervalle est indiqué à titre de comparaison à côté du taux de bien classés (Bcl) de même qu'un 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 croisements : comparaison des distributions croisées de deux variables observées avec deux variables estimées (test du Chi2 sur le même modèle que celui utilisé pour les marginales mais avec par exemple, des degrés de liberté égaux à 8 pour deux variables ayant trois modalités chacune).

Nous présentons tout d'abord le fichier utilisé lors de la modélisation par régression logistique effectué avec le logiciel SAS.

Le fichier d'exemple (voir annexe1 TABLEV13) comporte un échantillon composé de 200 ménages caractérisés par un certain nombre de variables *X* pour lesquels on a mesuré deux variables d'intérêt *Y*. Les variables *X*, caractérisant les ménages, sont candidates à l'explication des variables *Y*.
Ces variables candidates à l'explication sont l'âge, le niveau d'études, la présence d'un enfant à charge, le revenu du ménage. Les variables d'intérêt, booléennes, sont Thisyr et Lastyr.

Sous SAS les caractéristiques du modèle sont les suivantes :

Model Information

Data Set	WORK.SAMPLE
Response Variable	THISYR
Number of Response Levels	2
Number of Observations	146
Model	binary logit
Optimization Technique	Fisher's scoring

Une fois la séparation faite entre échantillon d'apprentissage (servant d'échantillon donneur) et échantillon test (faisant office d'échantillon receveur), on définit un modèle de régression logistique des variables *Y* sur les variables *X*. L'objectif étant de mettre en évidence la validation statistique du modèle de fusion; on va alors fusionner les règles du modèle obtenues sur l'échantillon donneur à l'échantillon receveur et comparer les prédictions ainsi obtenues aux valeurs préalablement mesurées *Y* de cet échantillon receveur. Les critères de "comparaison" ou plutôt de validation étant ceux globaux et individuel définis en début de chapitre.

Suite à différents essais, on a retenu comme variables explicatives des variables d'intérêt les deux variables *age_s* (4 modalités) et *child2* (2 modalités).

Les sorties du logiciel nous permettent de valider le modèle retenu :

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	185.466	176.809
SC	188.449	191.727
-2 Log L	183.466	166.809

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.6561	4	0.0023
Score	16.8270	4	0.0021
Wald	15.0090	4	0.0047

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age_s	3	9.5066	0.0233
CHILD2	1	3.2292	0.0723

Ayant explicité le modèle de fusion statistique retenu sur le fichier d'apprentissage, nous allons illustrer à partir des résultats obtenus après fusion des règles de modélisation sur le fichier test, les différents critères de validation exposés précédemment.

IV-1.1. *Reconstitution des distributions marginales*

Nous comparons les distributions marginales des variables d'intérêt, Thisyr et Lastyr, observées et estimées après fusion statistique.
Pour cela il suffit tout simplement de sortir les tableaux croisés de chaque variable observée (Thisyr, Lastyr) par sa variable prédite (pred_thisyr et pred_lastyr).

Les résultats pour la variable Thisyr se présentent ainsi :

Thisyr	Frequency, Percent, Row Pct, Col Pct	pred_thisyr 0	pred_thisyr 1	Total
0		4	14	18
		7.41	25.93	33.33
		22.22	77.78	
		30.77	34.15	
1		9	27	36
		16.67	50.00	66.67
		25.00	75.00	
		69.23	65.85	
Total		13	41	54
		24.07	75.93	100.00

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Le test d'adéquation entre les deux variables a pour critère:

$$\hat{c}^2 = \sum_{i=1}^{k} \frac{(n_{iobs} - n_{iest})^2}{n_{iobs}}$$

où n_{iest} et n_{iobs} désignent respectivement, pour chacune des deux classes les effectifs estimés et observés ; c'est à dire les effectifs des variables thisyr et pred_thisyr.
Ce qui nous donne pour cet exemple : $\hat{c}^2 = 2{,}08$

Or, pour un risque α de 1ère espèce de 0,05 la variable de Pearson c^2 vaut 3,84. On vérifie donc : $\hat{c}^2 \leq c^2_{0.95}$

Ce qui conduit à accepter l'hypothèse d'adéquation entre les distributions des variables thisyr et pred_thisyr.

IV-1.2. *Reconstitution des distributions croisées de deux variables*

Après avoir testé l'adéquation entre distribution observée et prédite après fusion pour chacune des variables d'intérêt, il faut désormais vérifier si le modèle de fusion proposé respecte les corrélations initiales existantes entre les deux variables d'intérêt thisyr et lastyr.
Ceci revient à comparer les distributions jointes observée (thisyr,lastyr) et estimée après fusion statistique (pred_thisyr,pred_lastyr).
De la même manière que pour les distributions marginales, on procède à un test du c^2 d'adéquation des deux distributions.

Chacune des variables ayant deux modalités, la distribution croisée comporte quatre modalités ; par conséquent il s'agit d'un test du c^2 à trois degrés de liberté. Les résultats du test sont les suivants :

nomvar	chi2	dl	prob
thisyr*lastyr	20.4314	3	.00013815

Dans le cas présent nous devons rejeter l'hypothèse d'adéquation pour un risque a de première espèce de 5%. Avec ce modèle de fusion (« simple ») nous conservons donc mal le lien entre les deux variables d'intérêt. Dans ce cas cela vient tout simplement du fait que la seconde variable d'intérêt est déjà très mal reconstituée par le modèle.

IV-1.3. *Taux d'individus bien-classés (Tbcl)*

Enfin le troisième critère global concernant la validité statistique du modèle Est le taux d'individus bien classés (Tbcl) après fusion.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ss2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Table of THISYR by pred_thisyr

THISYR	pred_thisyr(Formatted Value of the Predicted Response)		
Frequency, Percent, Row Pct, Col Pct	0	1	Total
0	4 7.41 22.22 30.77	14 25.93 77.78 34.15	18 33.33
1	9 16.67 25.00 69.23	27 50.00 75.00 65.85	36 66.67
Total	13 24.07	41 75.93	54 100.00

Pour la variable thisyr en reprenant le même tableau de contingence on obtient le résultat suivant :

Bcl = (27+4) / 54 = 0.57.
On obtient donc 57% d'individus bien classés.

Ce Tbcl peut être comparé à la proportion de bien classés obtenue avec la règle du maximum et ainsi construire un intervalle de confiance autour de celle-ci pour voir si le Tbcl obtenu après fusion en est significativement différent (en mieux ou en moins bien).

Si l'on note P_{max} la proportion d'individus bien classés obtenus ave la règle du maximum ; c'est à dire que parmi les individus observés on regarde la modalité ayant la probabilité la plus forte et on attribue à tous la valeur de cette modalité. L'intervalle de confiance autour de cette valeur est alors :

$$\left[P_{\max} - 1.96\sqrt{\frac{P_{\max}(1-P_{\max})}{n_{test}}} ; P_{\max} + 1.96\sqrt{\frac{P_{\max}(1-P_{\max})}{n_{test}}} \right]$$

où n_{test} est le nombre d'individus de l 'échantillon test (ici 54). Si le Tbcl est supérieur à la borne sup. de l'intervalle on pourra dire qu'il est significativement différent du P_{max}.

Dans le cas présent où l'on dispose de peu de données et avec un modèle de fusion aussi simple, le Tbcl n'est pas significativement différent.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Table of LASTYR by pred_lastyr

LASTYR	pred_lastyr(Formatted Value of the Predicted Response)		
Frequency, Percent, Row Pct, Col Pct	0	1	Total
0	5 9.26 26.32 50.00	14 25.93 73.68 31.82	19 35.19
1	5 9.26 14.29 50.00	30 55.56 85.71 68.18	35 64.81
Total	10 18.52	44 81.48	54 100.00

On reprend de la même manière le tableau de contingence pour la variable lastyr, pour laquelle on obtient un Bcl de 0.65.
Ce qui signifie que bien qu'ayant une distribution moins bien reconstituée, le taux d'individus bien classés est meilleur pour la variable lastyr.

IV-2. Critère individuel de validation

Le critère individuel de validation repose sur le taux d'individus bien classés au sein du groupe auquel il appartient.
Par exemple pour le groupe d'individus dont l'âge est compris entre 21 et 27 ans et pour les quels Child2 vaut 1, on a le tableau de confusion suivant ; soit un taux d'individus bien classés relatif à ce groupe de : 9/13 = 0,69.

Table of _FROM_ by _INTO_

FROM_(Formatted Value of the Observed Response)	_INTO_(Formatted Value of the Predicted Response)	
Frequency, Percent, Row Pct, Col Pct	0	Total
0	9 69.23 100.00 69.23	9 69.23
1	4 30.77 100.00 30.77	4 30.77
Total	13 100.00	13 100.00

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

IV-3. Validation opérationnelle de la fusion

Validation de la fusion en terme marketing: efficacité et coût?

Si la validation statistique du modèle, au moyen des critères vus précédemment, est un outil qui permet à un statisticien de s'assurer de la qualité des traitements statistiques effectués, la validation opérationnelle sera elle un outil d'aide à la décision. Elle pourra, par exemple, servir à un expert clientèle dans le but d'atteindre telle ou telle cible de son marché de clients. Cette validation opérationnelle s'appuie sur un indicateur très connu des marketeurs qu'est la courbe lift. Toutefois selon le type d'étude que l 'on mène et l'exploitation que l'on cherchera à faire des résultas on ne s'en tiendra pas forcément à l'observation de la courbe lift pour elle même. D'autres indicateurs comme l'aire comprise sous la courbe lift ou encore des rapports d'aires sous les courbes offrent une très bonne appréciation de la qualité globale et donc de l'acuité d'un modèle. Par conséquent après avoir présenté la courbe lift et défini les indicateurs utiles à la validation, on s'attachera à calculer l'aire sous cette courbe. Un court exemple permettra d'illustrer l'intérêt de la validation opérationnelle basée sur le lift. Enfin nous reviendrons sur le calcul de l'aire sous la courbe lift en fin de chapitre, en exposant une autre méthode donnant une valeur exacte de cette aire.

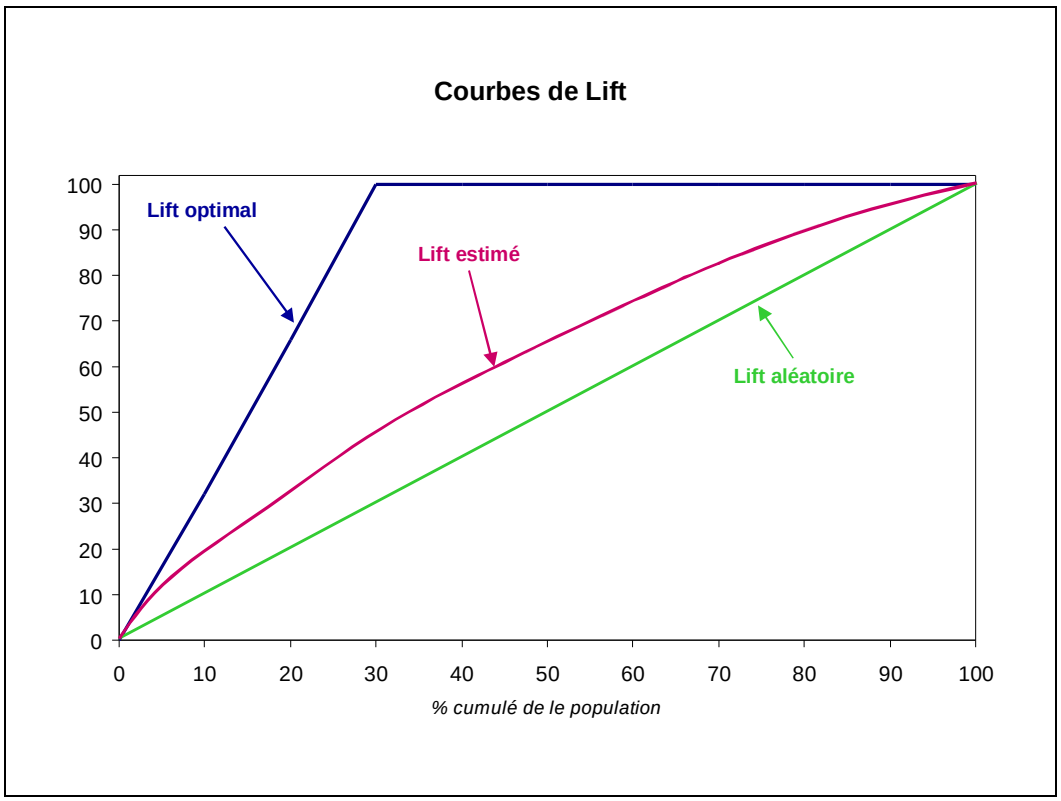
L'un des principaux outils d'aide à la décision dont dispose un expert clientèle est une courbe lift. Elle lui permet en effet pour une modélisation donnée de sélectionner une partie de son échantillon, à un seuil donné, « le plus probable d'être dans la cible » qu'il s'est fixé. Outre le choix d'un seuil qui entraîne un certain arbitraire dans la validation, il peut être intéressant d'avoir une mesure globale de la qualité permettant éventuellement de comparer les modèles entre eux et ce sur l'ensemble de la courbe lift. C'est pourquoi il est intéressant de s'intéresser au calcul de l'aire sous la courbe lift ainsi qu'au rapport entre les aires sous les courbes lift obtenues après modélisation et aléatoirement.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

IV-3.1. *Définition de la courbe lift*

Comment se calculent les courbes lift ?



On pose :

$$lift_{est}(x\%) = 100\left(\frac{\sum_{i=1}^{x} n_i^{y=cible}}{n^{y=cible}}\right)\%$$

où $n_i^{y=cible}$ désigne le nombre d'individus observés appartenant à la cible ordonnés de façon décroissante selon leurs probabilités estimées par le modèle, d'appartenir à la cible dans l'intervalle i : [x% ; x%+k[. En général, la largeur des intervalles est fixe (le pas k est constant). Enfin $n^{y=cible}$ correspond au nombre d'individus observé dans la cible.

$$Lift_{random}(x\%) = 100x\%$$

On considère que les nombres d'individus observés appartenant à la cible se répartissent aléatoirement dans l'échantillon (pas de modèle).

$$lift_{max}(x\%) = 100\left(\frac{\sum_{i=1}^{x} n_i}{n_{y=cible}}\right)\% \quad \text{si } n^{y=cible} \leq \sum_{i=0}^{x} n_i$$

$$= 100\% \quad \text{sinon}$$

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Le nombre d'individus observés dans la cible est cumulé jusqu'à ce que l'on ait atteint le nombre total d'individus dans la cible $n^{y=cible}$.

IV-3.2. *Mesures d'adéquation fondées sur le lift*

On peut calculer le rapport entre le lift au hasard et le lift estimé grâce au modèle par rapport à l'aire entre le lift maximum et le lift au hasard. Pour cela une façon simple est de calculer deux intégrales numériques classiques de la forme :

$$I_{est}=\left[0{,}5\left(lift_{est}(0\%)+lift_{est}(100\%)\right)+\sum_{i=1}^{99}lift_{est}(t\%)\right]-0{,}5\times100\times100$$

$$I_{\max}=\left[0{,}5\left(lift_{\max}(0\%)+lift_{\max}(100\%)\right)+\sum_{i=1}^{99}lift_{\max}(t\%)\right]-0{,}5\times100\times100$$

alors le rapport est égal à :

$$R_{\text{est}/\max}=I_{\text{est}}/I_{max}$$

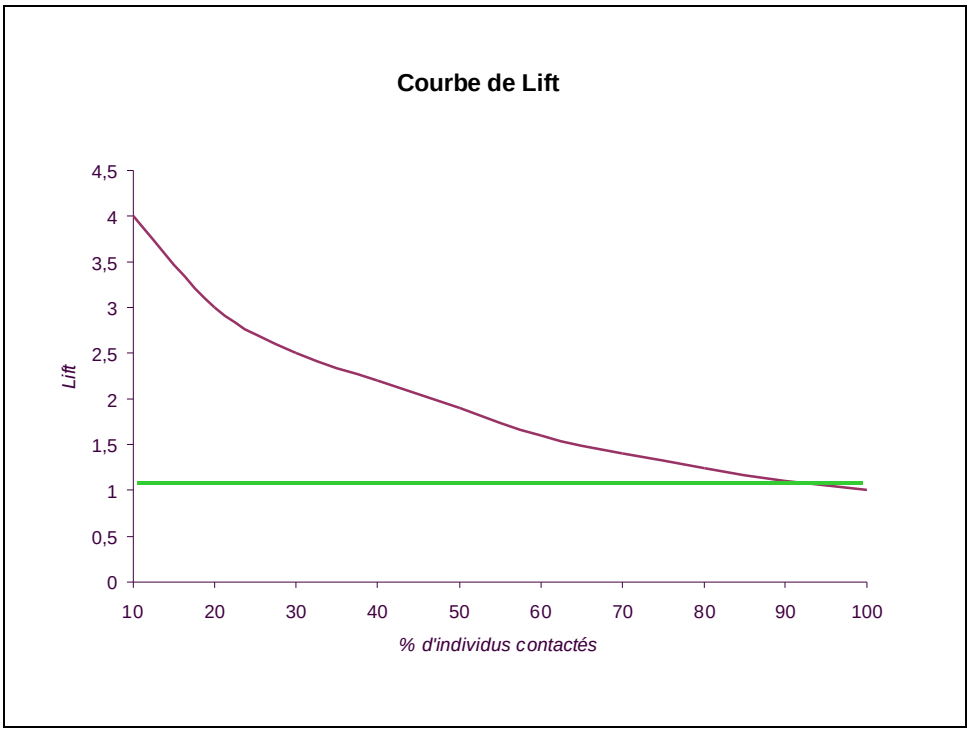
Ce rapport varie entre 0 et 1. Plus la valeur obtenue est proche de l'unité, plus le modèle estime bien la probabilité d'appartenir à la cible. Sur le graphique cela revient à ce que la courbe estimée du lift se rapproche très fortement de la courbe max du lift.

IV-3.3. *La courbe lift du gain potentiel*

La courbe lift du gain potentiel se présente comme suit. Cette courbe est également très utilisée pour fournir le gain potentiel associé aux probabilités estimées décroissantes, comme dans la courbe suivante. La courbe se calcule de la façon suivante.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html



$$lift_{gain}(x\%) = 100\left(\frac{\sum_{i=1}^{x} n_i^{y=cible}}{\frac{n}{100}x}\right)\%$$

Le nombre entre parenthèse représente le gain potentiel du lift par rapport au pourcentage observé de la cible dans l'échantillon : %cible :

$$gain(x\%) = \left(\frac{lift_{gain}(x\%)}{\%cible}\right)$$

IV-3.4. *Relation entre $lift_{gain}$ et les courbes lift de concentration*

Le $lift_{gain}$ peut s'exprimer en fonction du $lift_{est}$ et du $lift_{nul}$ à l'aide de la relation suivante :

$$lift_{gain}(x\%) = \left(\frac{lift_{est}(x\%)}{lift_{nul}(x\%)}\right) \times \%cible$$

De façon duale, nous avons la relation suivante :

$$lift_{est}(x\%) = \left(\frac{lift_{gain}(x\%)}{\%cible}\right) \times x\%$$

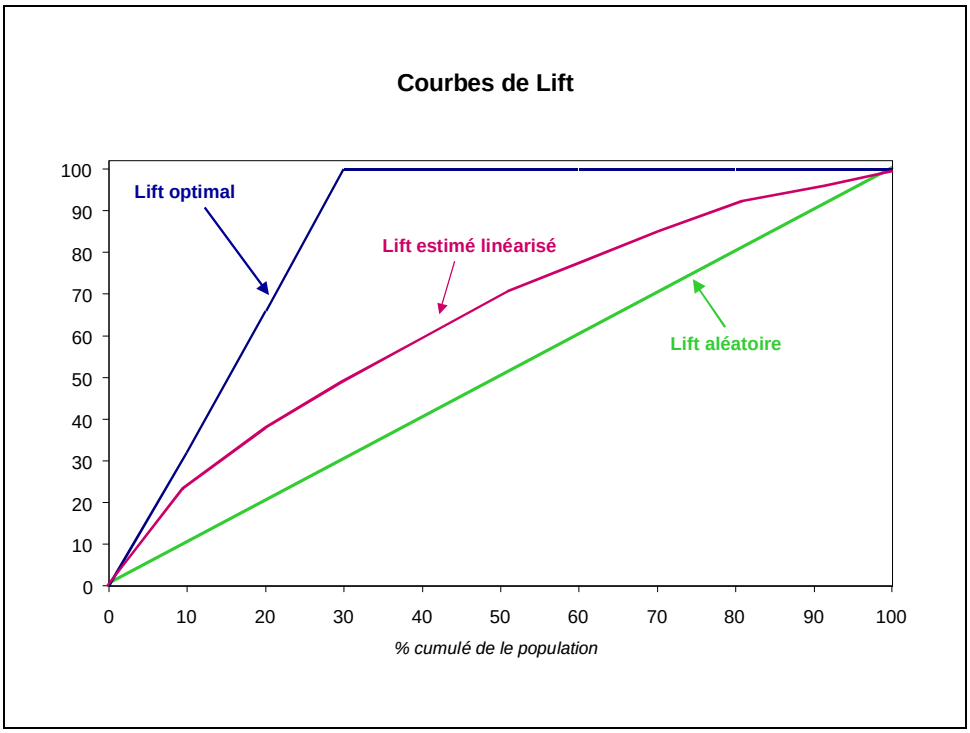
Il est alors intéressant de constater que lorsque l'ensemble de l'échantillon est cumulé, (les deux lift : $lift_{est}$ et $lift_{nul}$ se rejoignent, cf figure précédente),

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

alors $lift_{gain}$ est exactement égal au pourcentage observé dans la cible (%cible). Cela correspond bien au trait parallèle à l'axe des abscisses, ayant pour coordonnées %cible dans la figure précédente. Bien évidemment, plus $lift_{est}$ et $lift_{nul}$ sont éloignés, plus le $lift_{gain}$ est élevé. De façon duale, plus le gain potentiel : gain(x%) est grand, plus $lift_{est}$ sera élevé, et donc d'autant plus éloigné du $lift_{nul}$.

IV-3.5. *Calcul exact des courbes lift et de l'indice de Gini*



Ces différents indices individuels peuvent être rassemblés pour former des mesures de qualité globale du modèle.

Calcul de l'indice de Gini

La mesure I_{est} suivante, représente l'apport du modèle estimé et correspond à la somme des $lift_{esti}$ sur les *n* individus.

$$I_{est}=\sum_{i=1}^{n}\sum_{j=1}^{i}y_i=\sum_{i=1}^{n}(n-i+1)y_i=(n+1)\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}iy_i=(n+1)\times n_1-\sum_{i=1}^{n}iy_i$$

I_{max} est égal au cumul des $lift_{maxi}$ et fournit la mesure maximale obtenue par le modèle parfait.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.gs2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

$$I_{\max}=\sum_{i=1}^{n_1} i+n_0 n_1=\frac{n_1(n_1+1)}{2}+n_0 n_1$$

I_{nul} est égal à la somme des $lift_{nul}$ et est relatif à la mesure obtenue sans modèle.

$$I_{nul}=\frac{n_1}{n}\sum_{i=1}^{n} i=\frac{n_1}{n}\frac{n(n+1)}{2}=\frac{n_1(n+1)}{2}$$

La différence entre les deux quantités suivantes fournit l'apport réel du modèle estimé, par rapport à la notion « sans modèle » :

$$L_{est}=I_{est}-I_{nul}=\frac{(n+1)n_1}{2}-\sum_{i=1}^{n} i y_i$$

L_{max} représente l'apport maximum que l'on puisse obtenir avec un modèle parfait, par rapport à la notion « sans modèle » :

$$L_{\max}=I_{\max}-I_{nul}=\frac{n_0 n_1}{2}$$

Enfin, G est la mesure de qualité globale du modèle estimé :

$$G=\frac{L_{est}}{L_{\max}}=\frac{(n+1)n_1-2\sum_{i=1}^{n} i y_i}{n_0 n_1}$$

Cet indice varie entre –1 et +1. En effet, dans le pire des cas, les n_1 dernières probabilités etimées (les plus petites) correspondent aux yi's égaux à 1. En fait, cela représente un I_{min}. Dans ce cas, il vaut en partant de I_{est} :

$$I_{\min}=\frac{(n+1)\times n_1}{2}-\frac{n_0 n_1}{2}$$

alors G=-1

A l'opposé, si les n1 plus grandes probabilités correspondent à l'ensemble des yi's égaux à 1, alors I_{est} vaut I_{max}, par conséquent G=1.

Si l'on désire avoir un indice entre 0 et 1, il convient de faire la transformation suivante :

$$G'=(G+1)/2$$

Ce rapport ressemble en fait au R^2 de Hosmer et Lemeshow (cas de donnes groupées).

$$R^2_{HL}=\frac{l(\hat{y},y)-l(y_0,y)}{l(\hat{y},y)-l(y_0,y)}$$

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Les relations entre les différents lift sont les suivantes :

$$lift_{gain_i} = \frac{lift_{est_i}}{lift_{max_i}}$$

$$lift_{est_i} = i \times lift_{gain_i}$$

$$gain_i = \frac{lift_{gain_i}}{n_1/n} = \frac{lift_{est_i}}{lift_{nul_i}}$$

Par conséquent, de ces relations, on tire les quantités suivantes :

$$I_{est} = \sum_{i=1}^{n} lift_{est_i} = \sum_{i=1}^{n} i \times lift_{gain_i}$$

$$I_{gain} = \sum_{i=1}^{n} lift_{gain_i} = \sum_{i=1}^{n} \frac{lift_{est_i}}{lift_{max_i}}$$

IV-3.6. *Aire sous la courbe Lift – lien avec la courbe ROC*

Il est intéressant de remarquer que l'aire sous la courbe lift qui constitue un très bon indicateur opérationnel peut être obtenu au moyen d'une autre courbe, la courbe ROC. Nous présentons ici à titre indicatif une autre manière d'obtenir une valeur exacte de l'aire sous la courbe lift et ce au moyen de la courbe ROC. Dans l'étude expérimentale développée au chapitreV la validation opérationnelle est néanmoins effectuée à partir de la méthode du calcul de l'aire sous la courbe lift décrite dans le paragraphe précédent.

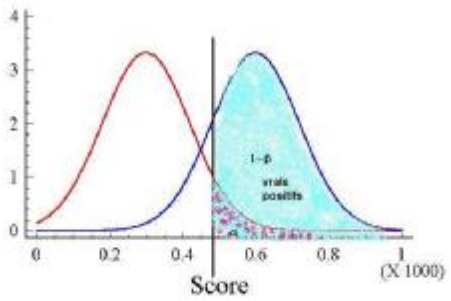
1. *Surface sous la courbe ROC.*

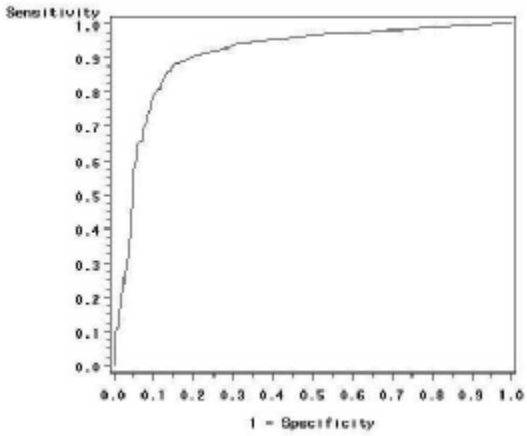
La courbe ROC donne les variations de la spécificité en fonction de la sensibilité, en d'autres termes de 1- β en fonction de α quand le seuil de sélection *s* varie. On supposera par la suite que le groupe à détecter G_1 est celui qui a les scores élevés et correspond à une loi de densité f_1.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

On rappelle que :





La surface théorique sous la courbe ROC varie de zéro à un et est égale $P(X_1>X_2)$ si on tire au hasard et indépendamment une observation de G_1 et une observation de G_2

On notera R la surface sous la courbe ROC dans ce qui suit. $R=\int_{s=+\infty}^{s=-\infty}(1-b(s))da(s)$

Cette probabilité, donc la surface sous la courbe observée, peut être estimée par le pourcentage de paires concordantes parmi les n_1n_2 paires possibles d'observations : l'une de G_1, l'autre de G_2 , noté *c* dans les sorties de la proc logistic de SAS

$$R=c=\frac{n_c}{n_1n_2}$$

Une paire est concordante si le score d'un individu du groupe 1 est supérieur au score d'un individu du groupe 2.

n_c se confond avec la statistique U de Mann-Whitney

La courbe ROC et donc la surface est invariante pour toute transformation monotone croissante du score. On peut d'ailleurs prendre pour score la probabilité d'appartenance au groupe 1 estimée par un modèle quelconque.

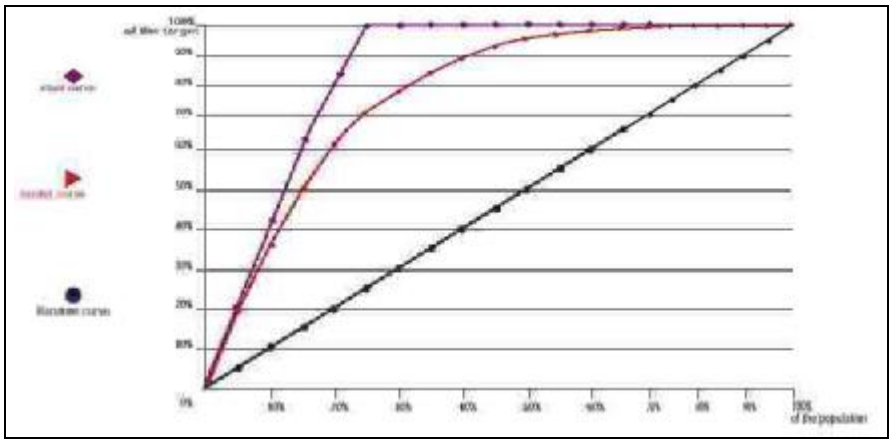
2. *Surface sous la courbe lift (lift chart ou gain chart selon les auteurs)*

La courbe lift donne la proportion du groupe 1 ayant un score supérieur à s , 1- β, en fonction du pourcentage d'individus ayant un score supérieur à s . Ce dernier pourcentage vaut $p_1(1-b)+(1-p_1)a$ où p_1 désigne la proportion du groupe 1.

Dans le cas d'un score parfait , la surface L sous la courbe de lift vaut 1- $p_1/2$

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ss2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html



Un calcul élémentaire donne :

$$L = \int_{s=-\infty}^{s=+\infty} (1-b(s))\, d\left(p_1(1-b(s)) + (1-p_1)a(s)\right)$$

$$L = p_1 \int_{s=-\infty}^{s=+\infty} (1-b(s))\, d(1-b(s)) + (1-p_1) \int_{s=-\infty}^{s=+\infty} (1-b(s))\, da(s)$$

Or :

$$R = \int_{s=-\infty}^{s=+\infty} (1-b(s))\, da(s)$$

$$\text{et} \quad \int_{s=-\infty}^{s=+\infty} (1-b(s))\, d(1-b(s)) = \frac{1}{2}$$

Soit :

$$\boxed{L = \frac{p_1}{2} + (1-p_1)\, R}$$

Si on considère les surfaces comprises entre les courbes et la diagonale, il suffit de soustraire 1/2 : la surface entre la courbe idéale et la diagonale vaut alors $(1- p_1)/2$.

Le rapport de ces surfaces $\dfrac{L-\frac{1}{2}}{\frac{1-p_1}{2}}$ vaut alors $\dfrac{p_1 + 2(1-p_1)R - 1}{1-p_1} = 2R-1$

IV-3.7. *Illustration de la courbe lift sur un exemple.*

L'exemple qui suit est extrait d'un échantillon de fichier clientèle d'EDF et correspond à une problématique réelle de l'entreprise. Les clients particuliers chauffant leur logement à l'électricité représentent une cible intéressante pour EDF, car il est possible de leur proposer des services et des tarifs adaptés à leurs comportements de consommations d'électricité. Cependant,

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

l'information sur la possession de l'énergie de chauffage n'est pas toujours disponible dans les bases de données de facturation d'EDF. La seule solution raisonnable est l'enrichissement de la base de données sur les seules informations disponibles (dans ce cas pas de fichier d'enquête complémentaire).

Dans un premier temps on ne prend en compte que deux informations dans la base, la consommation annuelle d'électricité et la possession ou non du chauffage électrique.

L'objectif de cette étude est de prévoir la possession du chauffage à l'aide de la consommation d'électricité. En d'autres termes, il s'agit de fournir une règle statistiquement fiable pour décider, avec un certain taux d'erreur tout de même, si un client chauffe son logement à l'électricité.

Les résultats obtenus sur l'échantillon test avec un modèle de régression logistique se présentent ainsi :

Matrice de confusion sur l'échantillon test

Table of chf_elec by pred_chf_elec

chf_elec / pred_chf_elec	0	1	Total
0	867	56	923
	65.93	4.26	70.19
	93.93	6.07	
	82.18	21.54	
1	188	204	392
	14.30	15.51	29.81
	47.96	52.04	
	17.82	78.46	
Total	1055	260	1315
	80.23	19.77	100.00

On obtient un bon pourcentage de bien classés de :
100((867 + 204)/1315)% =81,44%
Ce qui est d'ores et déjà mieux que si l'on avait appliqué dans ce cas la règle du maximum (70,19%).

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Les résultats se présentent ainsi :

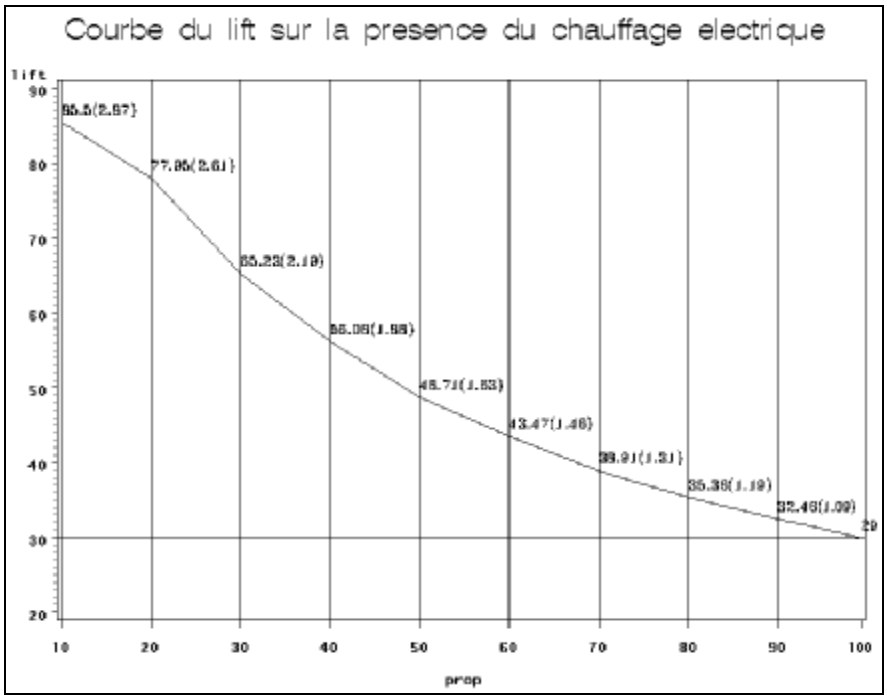
Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.040	1178	0	2822	0	29.5	100.0	0.0	70.6	.
0.060	1170	138	2684	8	32.7	99.3	4.9	69.6	5.5
0.080	1134	536	2286	44	41.8	96.3	19.0	66.8	7.6
0.100	1101	925	1897	77	50.7	93.5	32.8	63.3	7.7
0.120	1072	1252	1570	106	58.1	91.0	44.4	59.4	7.8
0.140	1050	1531	1291	128	64.5	89.1	54.3	55.1	7.7
0.160	1020	1712	1110	158	68.3	86.6	60.7	52.1	8.4
0.180	984	1870	952	194	71.4	83.5	66.3	49.2	9.4
0.200	955	1990	832	223	73.6	81.1	70.5	46.6	10.1
0.220	934	2123	699	244	76.4	79.3	75.2	42.8	10.3
0.240	900	2189	633	278	77.2	76.4	77.6	41.3	11.3
0.260	878	2250	572	300	78.2	74.5	79.7	39.4	11.8
0.280	856	2311	511	322	79.2	72.7	81.9	37.4	12.2
0.300	831	2369	453	347	80.0	70.5	83.9	35.3	12.8
0.320	806	2415	407	372	80.5	68.4	85.6	33.6	13.3
0.340	779	2459	363	399	81.0	66.1	87.1	31.8	14.0
0.360	761	2488	334	417	81.2	64.6	88.2	30.5	14.4
0.380	744	2514	308	434	81.5	63.2	89.1	29.3	14.7
0.400	720	2541	281	458	81.5	61.1	90.0	28.1	15.3
0.420	700	2563	259	478	81.6	59.4	90.8	27.0	15.7
0.440	681	2589	233	497	81.8	57.8	91.7	25.5	16.1
0.460	655	2597	225	523	81.3	55.6	92.0	25.6	16.8
0.480	632	2619	203	546	81.3	53.7	92.8	24.3	17.3
0.500	614	2635	187	564	81.2	52.1	93.4	23.3	17.6
0.520	596	2650	172	582	81.2	50.6	93.9	22.4	18.0
0.540	575	2658	164	603	80.8	48.8	94.2	22.2	18.5
0.560	566	2668	154	612	80.9	48.0	94.5	21.4	18.7
0.580	554	2678	144	624	80.8	47.0	94.9	20.6	18.9
0.600	539	2687	135	639	80.7	45.8	95.2	20.0	19.2
0.620	521	2698	124	657	80.5	44.2	95.6	19.2	19.6
0.640	500	2701	121	678	80.0	42.4	95.7	19.5	20.1
0.660	481	2711	111	697	79.8	40.8	96.1	18.8	20.5
0.680	456	2720	102	722	79.4	38.7	96.4	18.3	21.0
0.700	441	2725	97	737	79.2	37.4	96.6	18.0	21.3
0.720	423	2735	87	755	79.0	35.9	96.9	17.1	21.6
0.740	410	2746	76	768	78.9	34.8	97.3	15.6	21.9
0.760	378	2755	67	800	78.3	32.1	97.6	15.1	22.5
0.780	358	2767	55	820	78.1	30.4	98.1	13.3	22.9
0.800	340	2770	52	838	77.8	28.9	98.2	13.3	23.2
0.820	306	2777	45	872	77.1	26.0	98.4	12.8	23.9
0.840	290	2784	38	888	76.9	24.6	98.7	11.6	24.2
0.860	263	2792	30	915	76.4	22.3	98.9	10.2	24.7
0.880	233	2801	21	945	75.9	19.8	99.3	8.3	25.2
0.900	198	2805	17	980	75.1	16.8	99.4	7.9	25.9
0.920	174	2807	15	1004	74.5	14.8	99.5	7.9	26.3
0.940	111	2812	10	1067	73.1	9.4	99.6	8.3	27.5
0.960	78	2817	5	1100	72.4	6.6	99.8	6.0	28.1
0.980	1	2821	1	1177	70.6	0.1	100.0	50.0	29.4
1.000	0	2822	0	1178	70.6	0.0	100.0	.	29.5

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Le graphique du lift sur l'échantillon test est le suivant :

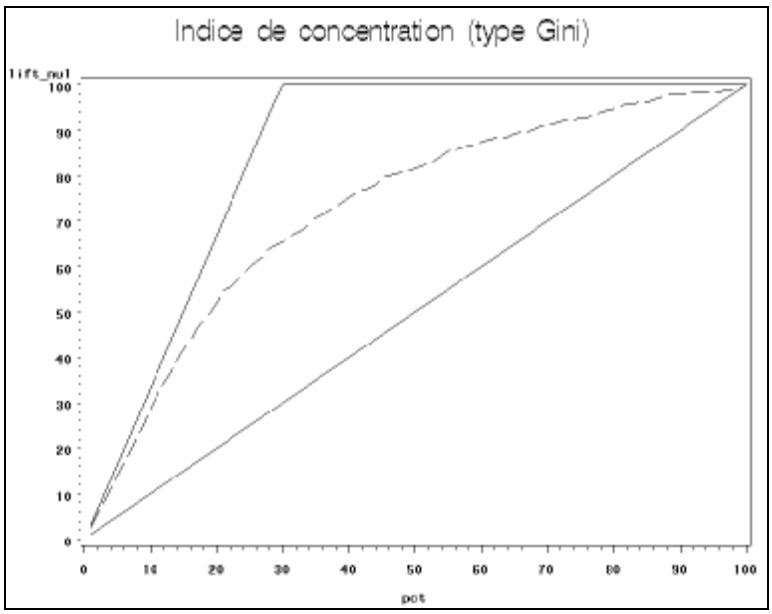


Il montre qu'en sélectionnant les 10 premiers % de plus forte prédiction de la présence du chauffage électrique, alors on récupérera 85,5% de clients se chauffant réellement à l'électricité, c'est-à-dire un de (85,5% / 29,81%) = 2,87. En d'autres termes, avec seulement un dixième des clients ayant la plus forte probabilité de se chauffer à l 'électricité, on aura environ trois plus de chance de détecter des clients se chauffant réellement à l'électricité que si nous n'avions pas utilisé de modèle.

La courbe de concentration confirme la qualité de bonne reconstitution de ce modèle, car l'aire entre la première bissectrice (pas de modèle) et la courbe en traits interrompus représente une large part.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.gs2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html



Nous allons maintenant reprendre la même étude avec davantage de variables candidates à l'explication et observer l'impact sur les résultats et en particulier sur la courbe lift obtenue.
Résultats obtenus avec le second modèle:

Table of chf_elec by I_chf_elec

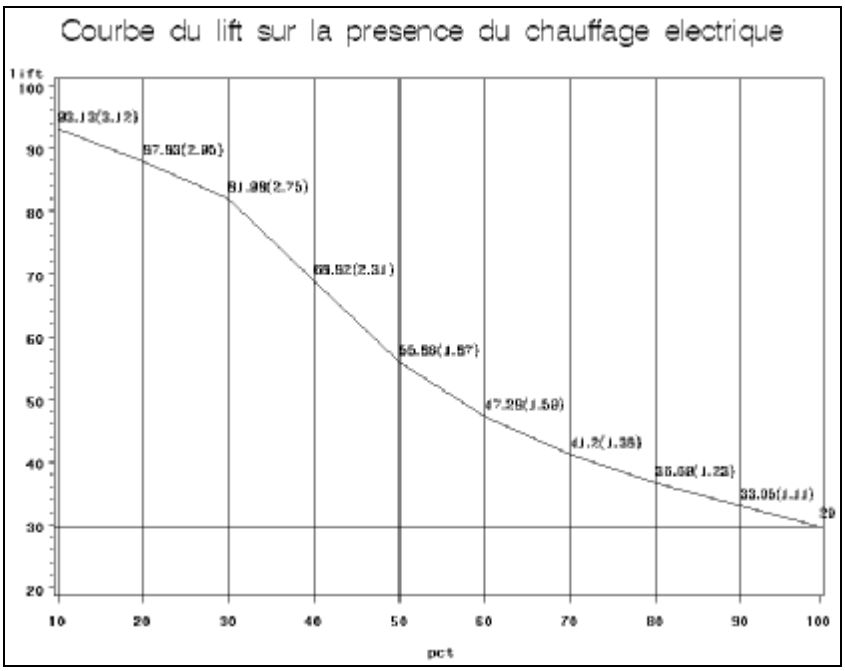
chf_elec	I_chf_elec(Into: chf_elec)		
Frequency , Percent , Row Pct , Col Pct	,CHF ELEC ,TRIC	,PAS CHF ELEC	, Total ,
Pas chf elec	68 5.17 7.37 17.66	855 65.02 92.63 91.94	923 70.19
Chf electric	317 24.11 80.87 82.34	75 5.70 19.13 8.06	392 29.81
Total	385 29.28	930 70.72	1315 100.00

Le pourcentage de bien classé est très élevé : (65,02% + 24,11% = 89,13%) avec le modèle 70,19% sans modèle. L'amélioration est très importante. Le modèle est donc tout à fait opérationnel.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

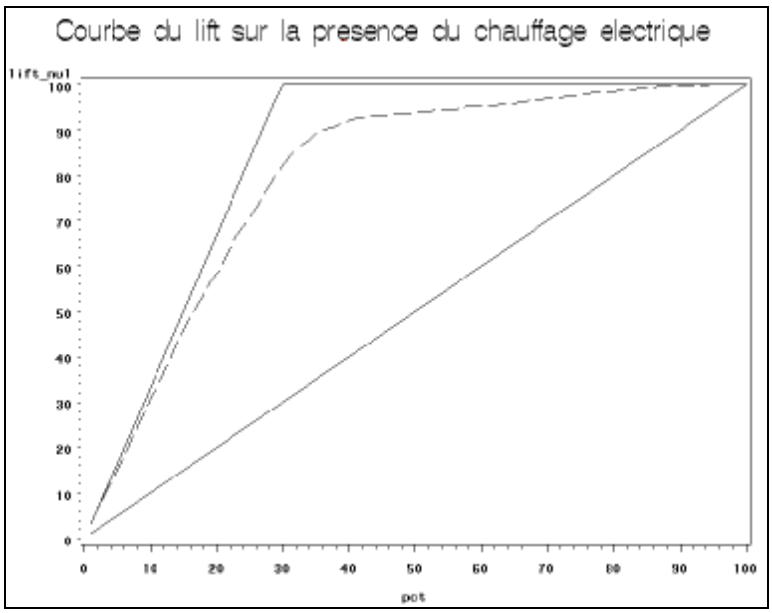
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

La courbe de lift sur l'échantillon test est la suivante :



On peut constater que pour les 10 premiers %, on obtient 93,13% contre 85,5% dans la première étude où l'on utilisé seulement la variable de consommation d'électricité. Le gain passe de 2,87 à 3,12. Mais le plus remarquable est qu'avec les 30 premiers %, on arrive à obtenir un gain de 2,75 contre 2,19 dans la première étude.

La courbe de concentration confirme la qualité de bonne reconstitution de ce modèle, car l'aire entre la première bissectrice (pas de modèle) et la courbe en traits interrompus représente une très large part.



PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Chapitre V :

Application des nouvelles méthodes de fusion

V-1. Validation statistique de la fusion effectuée entre les deux fichiers d'enquêtes Credoc et Sofres sur quelques milliers d'individus.

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 comprenant respectivement 8000 et 2000 individus.

Liste des variables communes aux deux fichiers utilisées lors de cette étude :

- Année de construction du logement
- Taille de l'agglomération
- CSP du chef de famille
- Type de logement
- Statut d'occupation
- Mode de chauffage principal
- Age du chef de famille
- Nombre de personnes par logement
- Toutes ces variables sont qualitatives.

Chacun des deux fichiers comporte un certain nombre de variables spécifiques mais, comme nous l'avons spécifié un peu plus haut, la présente étude concerne la satisfaction clientèle à l'égard du chauffage électrique. Nous n'avons donc retenu que les variables de satisfaction qui sont 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.

Liste des 9 variables de satisfaction du fichier SOFRES :

§	Satisfaction générale à l'égard du mode de chauffage,	Gene_sat
§	Satisfaction concernant le niveau de température,	Nivt_sat
§	Satisfaction à l'égard de l'esthétique,	Esth_sat
§	Satisfaction à l'égard de la régulation du chauffage,	Regl_sat
§	Satisfaction vis à vis de la rapidité de variation de la température,	Rapi_sat
§	Satisfaction concernant la facilité d'emploi,	Empl_sat
§	Satisfaction sur la sécurité du moyen de chauffage,	Sécu_sat
§	Satisfaction quant au taux d'humidité qui en résulte,	Humi_sat
§	Satisfaction vis à vis du coût du mode de chauffage	Coût_sat

Toutes ces variables sont ordinales, discrètes, codées sur 4 modalités, de 1 "TRES SATISFAIT" à 4 "PAS DU TOUT SATISFAIT". Toutefois compte tenu du faible nombre de réponses pour les modalités 3 et 4, ces deux modalités

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

ont été regroupées en une seule ("PEU ET PAS DU TOUT SATISFAIT") pour toutes les variables, mise à part Coût_sat qui ne présentait pas cette particularité.

A titre d'exemple nous présentons ci-dessous les résultats des critères globaux de la validation statistique obtenus par la méthode la plus « simple », la régression logistique.

La validation statistique a été effectuée sur le fichier donneur (SOFRES). Ce fichier a été divisé en cinq échantillons de même taille. Quatre de ces échantillons ont été regroupés pour jouer le rôle d'échantillon d'apprentissage tandis que le dernier constituait l'échantillon test. On a ensuite permuté les échantillons afin d'obtenir cinq jeux de résultats. Pour chacun les trois étapes de la validation statistique ont été effectuées.

Tableau 1: Logit classique - Marginales

Ech	Gene_sat			Nivt_sat			Esth_sat			Regl_sat			Rapi_sat		
1	363.01	2	0	258.95	2	0	797.33	2	0	123.27	2	0	190.38	2	0
2	311.16	2	0	258.35	2	0	620.87	2	0	96.43	2	0	255.54	2	0
3	396.31	2	0	422.17	2	0	573.19	2	0	138.36	2	0	262.87	2	0
4	389.86	2	0	364.48	2	0	679.79	2	0	231.77	2	0	388.14	2	0
5	366.12	2	0	525.34	2	0	702.61	2	0	257.19	2	0	487.52	2	0
	Empl_sat			Sécu_sat			Humi_sat			Coût_sat					
1	168.80	2	0	126.87	2	0	580.79	2	0	472.32	3	0			
2	244.23	2	0	93.46	2	0	481.58	2	0	640.33	3	0			
3	73.14	2	1.1^{E}-16	81.91	2	0	532.73	2	0	512.02	3	0			
4	175.34	2	0	80.02	2	0	501.37	2	0	613.24	3	0			
5	174.23	2	0	101.96	2	0	636.70	2	0	494.69	3	0			

Dans chaque case de résultats la première colonne correspond à la valeur du Chi2, la deuxième donne les degré de liberté et la troisième la probabilité obtenue.
Les tests du Chi2 effectués sur les marginales conduisent à rejeter l'hypothèse de bonne adéquation des variables estimées aux variables observées. Toutefois compte tenu des résultats médiocres obtenus sur ce test pour toutes les méthodes, il est plus judicieux de comparer les Chi2 eux-mêmes par la suite.

Dans le deuxième tableau relatif aux taux de bien classés, le premier chiffre de résultat correspond au taux de bien classés pour chaque variable dans chacun des échantillons test. Il lui est associé l'intervalle de confiance construit autour de la règle du maximum. Les symboles ++, -- et 0 indiquent simplement si le taux de bien classés obtenus est significativement meilleur, moins bon ou peu différent de la valeur obtenue avec la règle du maximum.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Tableau 2: Logit classique - Taux de bien classés

Gene_sat	Nivt_sat	Esth_sat	Regl_sat	Rapi_sat
0.5977 ++	0.5463 0	0.5020 0	0.4458 0	0.5271 0
[0.5272;0.5804]	[0.5174;0.5708]	[0.4717;0.5252]	[0.4077;0.4608]	[0.5072;0.5607]
0.5403 --	0.4772 --	0.4963 0	0.4537 0	0.4529 --
[0.5530;0.6058]	[0.5208;0.5741]	[0.4876;0.5412]	[0.4226;0.4759]	[0.4675;0.5211]
0.5919 0	0.5079 0	0.5136 0	0.4327 0	0.4816 0
[0.5487;0.6015]	[0.4877;0.5411]	[0.5013;0.5547]	[0.3855;0.4382]	[0.4614;0.5149]
0.5661 0	0.5241 0	0.5118 0	0.4674 ++	0.4758 0
[0.5297;0.5826]	[0.4887;0.5420]	[0.4987;0.5520]	[0.3694;0.4216]	[0.4329;0.4860]
0.5491 0	0.4738 0	0.4898 0	0.4435 ++	0.4844 ++
[0.5157;0.5676]	[0.4381;0.4901]	[0.4643;0.5164]	[0.3598;0.4105]	[0.4070;0.4587]

Empl_sat	Sécu_sat	Humi_sat	Coût_sat
0.5443 0	0.4822 0	0.5337 0	0.4995 0
[0.5372;0.5903]	[0.4495;0.5030]	[0.4840;0.5376]	[0.4723;0.5259]
0.5682 0	0.5282 ++	0.5545 0	0.4827 0
[0.5150;0.5684]	[0.4526;0.5061]	[0.5210;0.5744]	[0.4375;0.4909]
0.6026 0	0.5326 0	0.5219 0	0.5310 0
[0.5809;0.6332]	[0.4974;0.5509]	[0.4944;0.5479]	[0.4911;0.5445]
0.5943 0	0.5257 0	0.5456 0	0.4995 ++
[0.5653;0.6177]	[0.4863;0.5397]	[0.5181;0.5713]	[0.4429;0.4962]
0.6061 ++	0.5375 ++	0.4968 0	0.5025 0
[0.5363;0.5880]	[0.4814;0.5335]	[0.4681;0.5202]	(0.4857:0.5378]

Les résultats obtenus sur les taux de bien classés vont de médiocre pour certaines variables (Regl, Nivt, Coût) à correct pour certaines autres (Empl, Gene).

Tableau 3: Logit classique - Corrélations

	Gene_sat	Nivt_sat	Esth_sat	Regl_sat	Rapi_sat	Empl_sat	Sécu_sat	Humi_sat	Coût_sat
Gene_sat		524.70	524.70	480.35	590.18	540.18	448.38	767.29	799.65
Nivt_sat	524.70		907.85	460.44	588.78	549.96	464.45	770.99	767.16
Esth_sat	899.71	907.85		765.64	765.64	838.26	762.23	922.96	996.02
Regl_sat	480.35	460.44	765.64		458.12	406.05	299.72	655.48	693.32
Rapi_sat	590.18	588.78	867.12	458.12		592.27	419.05	738.25	781.57
Empl_sat	540.73	549.96	838.26	406.05	592.27		233.05	702.97	699.98
Sécu_sat	448.38	464.45	762.23	299.72	419.05	233.05		636.03	636.89
Humi_sat	767.29	770.99	922.96	655.48	738.25	702.97	636.03		879.80
Coût_sat	799.65	767.16	996.02	693.32	781.57	699.98	636.89	879.80	

Pour les tests effectués sur les reconstitutions de corrélations, les Chi2 étant tous significatifs (quel que soit la méthode considérée), seule apparaît la valeur du Chi2 elle même (mauvaise adéquation des croisements entre 2 variables estimées et 2 variables observées). Toutefois les valeurs importantes des Chi2 s 'expliquent également par la taille importante des échantillons testés. Ceci met en évidence les limites d'utilisation de ce test

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

quant à l'interprétation des résultats. Il faut toutefois noter que chaque valeur de ce tableau est la moyenne des 5 Chi2 obtenus sur chacun des 5 échantillons tests de la validation croisée, ceci de manière à ne pas surcharger inutilement la quantité de résultats fournis.

L'ensemble des résultats obtenus par validation statistique est présentée en annexe2. A chaque variable d'intérêt est associée un tableau de résultats permettant de comparer les différentes méthodes exposées au chapitre3. On y trouvera également un tableau récapitulatif de ces résultats permettant de comparer l'efficacité de ces méthodes. En effet, pour chacune des méthodes développées trois colonnes (blanche, gris clair et gris foncé) donnent le nombre de fois où cette méthode arrive respectivement en première, deuxième et dernière position pour chacun des critères de validation statistique appliquée aux neuf variables d'intérêt de l'étude.
C'est à partir de ces résultats que nous avons sélectionné les méthodes de fusion statistique les plus pertinentes (greffe séquentielle ou univariée et recodage Logit avec PLS2 ou multivariée). C'est donc celles-ci qui ont ensuite été mises en œuvre lors de l'expérimentation sur la base clientèle EDF.

V-2. Fusion de fichiers d'enquête avec une base clientèle

V-2.1.*Contexte et objectif*

Cette étude menée au sein du groupe statistique SOAD de la EDF R&D a permis d'expérimenter la fusion statistique de fichiers de données sur des données clientèles enrichies par des fichiers d'enquêtes externes. Elle s'inscrit dans le cadre du projet TAVEL dirigé par C. Derquenne.
L'objectif du projet TAVEL est d'approfondir l'étude de la clientèle de masse dans le contexte actuel de développement du Groupe EDF. Il s'agit d'étendre le champ des problématiques traitées par l'utilisation de techniques avancées de Data Mining, c'est-à-dire d'étudier, de proposer et de tester de nouvelles méthodes statistiques pour analyser les bases de données clientèle dans une optique marketing. La démarche statistique du projet *DECLIC* avait déjà permis, dans le cadre des transactions commerciales sortantes (TCS), d'enrichir des informations très utiles telles que l'énergie de chauffage du logement, celle de l'eau chaude sanitaire et le statut d'occupation, présentes dans la base de données clientèle OPTIMIA. Cependant, cet apport d'information est interne dans le sens où les données enrichies le sont à l'aide de la base de données, elle-même, c'est-à-dire sans information externe.

Par contre, le projet *TAVEL* utilise non seulement des *données internes* (base de données clientèle OPTIMIA, IMPALA, et plus particulièrement dans cette expérimentation, des données de facturation du Centre EDF-GDF SERVICES * datant de 1998). Mais aussi, le projet TAVEL se sert de *fichiers externes* (INSEE, CREDOC, SOFRES) ou encore de systèmes d'informations géographiques (SIG). Dans cette expérimentation, nous avons utilisé l'enquête « Chauffage électrique » EDF/SOFRES 98.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Cependant, l'ensemble de ces informations n'est pas disponible dans une seule et même base de données regroupant les mêmes clients. Cela permettrait d'avoir une vision plus large de l'ensemble des données clientèle et de réaliser des projections sur le comportement de ces clients (impact de scénarios tarifaires, de création de services, concurrence, ciblage pour actions marketing, ...).

Une des solutions possibles pour résoudre ce problème, est la combinaison de données de sources différentes et plus particulièrement, la fusion statistique de données. En d'autres termes, elle permet de greffer des informations d'un ou plusieurs fichiers de données, (nommés fichiers donneurs, ici l'enquête EDF/SOFRES 98) sur une base de données (dite base receveuse, ici le fichier de facturation du Centre EDF-GDF * , rendu anonyme). L'opération de greffe ne peut être réalisée que s'il y a des informations communes (variables communes, par exemple le type de tarif, l'année de création du client) dans les fichiers donneur et receveur. Si ce n'est pas le cas, il est tout de même possible d'utiliser des variables analogues entre les deux fichiers. Mais en aucun cas, il ne s'agit d'une fusion informatique où les clés communes seraient les variables communes. Il s'agit d'une fusion statistique, dans le sens où ces variables vont permettre de modéliser des informations présentes dans le fichier donneur et de les prédire (donc de les greffer) sur le fichier receveur où elles sont absentes

Une première méthode nommée « méthode séquentielle » a déjà été développée et expérimentée sur deux enquêtes : CREDOC (fichier receveur) et SOFRES (fichier donneur). Elle a fourni de bons résultats (statistiquement corrects). Une deuxième méthode dite « méthode multivariée » a été construite afin d'éliminer un inconvénient de la première. En effet, celle-ci greffait les variables à prédire une à une sur le fichier receveur (séquentiellement), tout en tenant compte de celles déjà introduites. Cela avait pour effet de reconstituer partiellement les corrélations des variables greffées du fichier donneur sur le fichier receveur. La méthode multivariée permet de greffer les variables en bloc, ce qui a pour effet de mieux préserver, dans une certaine mesure, leurs corrélations mutuelles.

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. Cela signifie que le niveau de confiance que nous pouvons accorder dans ces prédictions est inégal. Les deux méthodes développées fournissent donc des niveaux de confiance individuels, ce qui permet de ne garder des individus (ici des clients) dont nous jugeons que la confiance de la prédiction est suffisante statistiquement.

Cette étude a pour objectif de présenter une expérimentation des deux méthodes sur les données de facturation du Centre EGS * , sur lesquelles

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

nous fusionnons des données issues de l'enquête de satisfaction EDF/SOFRES. Les deux bases de données datent de 1998.

Dans une première partie, nous décrivons les données, puis nous explicitons la démarche statistique employée, c'est-à-dire le déroulement de l'expérimentation, dans la deuxième. Dans la troisième partie nous fournissons les résultats et les comparons, et nous expliquons les différents indicateurs mis en place déterminer la qualité des prédictions et leur confiance associée, ainsi qu'une aide à la décision pour les experts clientèle, au moyen d'une validation opérationnelle. Enfin, une quatrième et dernière partie développe les apports et les critiques de ces deux méthodes.

V-2.2. *Description des bases de données :* Base EGS * et enquête "Chauffage électrique" EDF/SOFRES

La base de données OPTIMIA EGS * correspond au fichier receveur et l'enquête « Chauffage électrique » EDF/SOFRES représente le fichier donneur.

*La base de facturation OPTIMIA EGS** date de 1998 et avait déjà servi pour la première expérimentation du projet DECLIC. Nous avions notamment prédit l'énergie de chauffage du logement (électricité, gaz et autre). Le tableau suivant fournit les variables[1] que nous avons sélectionnées pour réaliser cette nouvelle expérimentation sur la fusion statistique de données.

Nom de la variable	*Description des variables d'un client*
CLI_C_LOCAB	Numéro de locabonné
CLI_C_PMT	Code paiement
CLI_D_CRE	Date de création
CLI_Q_PAY	Qualité payeur
CLI_VL_FACT	Politique de facturation
CNS_HC	Consommation annuelle heures creuses estimée
CNS_HP	Consommation annuelle heures pleines estimée
CNS_TOTA	Consommation annuelle totale estimée
CNS_TOTM	Consommation mensuelle totale estimée
CTE_VL_CNSUPN	Consommation mensuelle heures creuses estimée
CTE_VL_CNSUPJ	Consommation mensuelle heures pleines estimée
LOC_C_POSTAL	Code postal où habite le client
LOC_T_LOG	Type de logement
HPL_VL_WSSCR	Puissance souscrite
PUISS	Puissance souscrite (en classes de puissance)
TAR_C	Tarif
TARIF	Tarif (simple ou double)
USG_T_CHAUFP	Type d'énergie du chauffage principal
CHP	Type d'énergie du chauffage principal (recodé)

Tableau 1 : Liste des variables utilisées dans la base de facturation OPTIMIA

[1] celles notées en gras correspondent à des variables d'origine ; les autres sont des variables seulement construites à partir des précédentes.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

De plus, cette base contient 246279 clients mais nous verrons dans la troisième étape de la démarche statistique de fusion statistique que nous sélectionnerons à l'aide des variables USG_T_CHAUP et CHP, les clients renseignés comme se chauffant à l'électricité, car l'enquête EDF/SOFRES ne concerne que ces clients.

L'enquête « chauffage électrique », commanditée par EDF à la SOFRES, a été réalisée sur l'ensemble des 98 centres EGS. Le questionnaire associé contient huit thèmes, dont les quatre premiers n'ont pas servi dans l'expérimentation :

- « Votre chauffage actuel » (type d'appareil composant le chauffage électrique, autre moyen de chauffage, ...),
- « Les services rendus par EDF » (conseils d'utilisation du CE, du meilleur tarif, ...),
- « Votre logement et votre foyer » (année de construction, type de logement, statut d'occupation, âge du chef de famille, ...),
- « Vos remarques et suggestions » (deux questions en texte libre).
- « Ce que vous pensez de votre chauffage électrique » (satisfaction générale, choix d'un système de chauffage et conseil du système de CE dans l'entourage du client),
- « Les évolutions de votre chauffage » (année d'installation du chauffage électrique dans le logement, modification en vue d'améliorer son installation de CE, propension à modifier son système de chauffage, et notamment à son remplacement),
- « Ce que vous pensez d'EDF » (7 affirmations concernant EDF : informations et conseils, proximité d'EDF et de ses clients, ...),
- « Ce que vous attendez d'EDF » (9 propositions de la part d'EDF contacts plus fréquents, intérêt pour les besoins personnels du client, ...).

La base de données associée à ce questionnaire contient non seulement l'ensemble de ces questions, mais aussi des variables internes au fichier de facturation OPTIMIA. Sept variables sont communes aux deux fichiers de données : date de création du client, puissance souscrite, type de logement, qualité payeur, politique de facturation, code paiement, tranche de consommation d'électricité, ainsi que le numéro du centre EGS et une variable de poids de sondage.

Enfin, cette base de données d'enquête contient 59106 personnes interrogées correspondant à des clients chauffant leur logement à l'électricité (principal ou d'appoint).

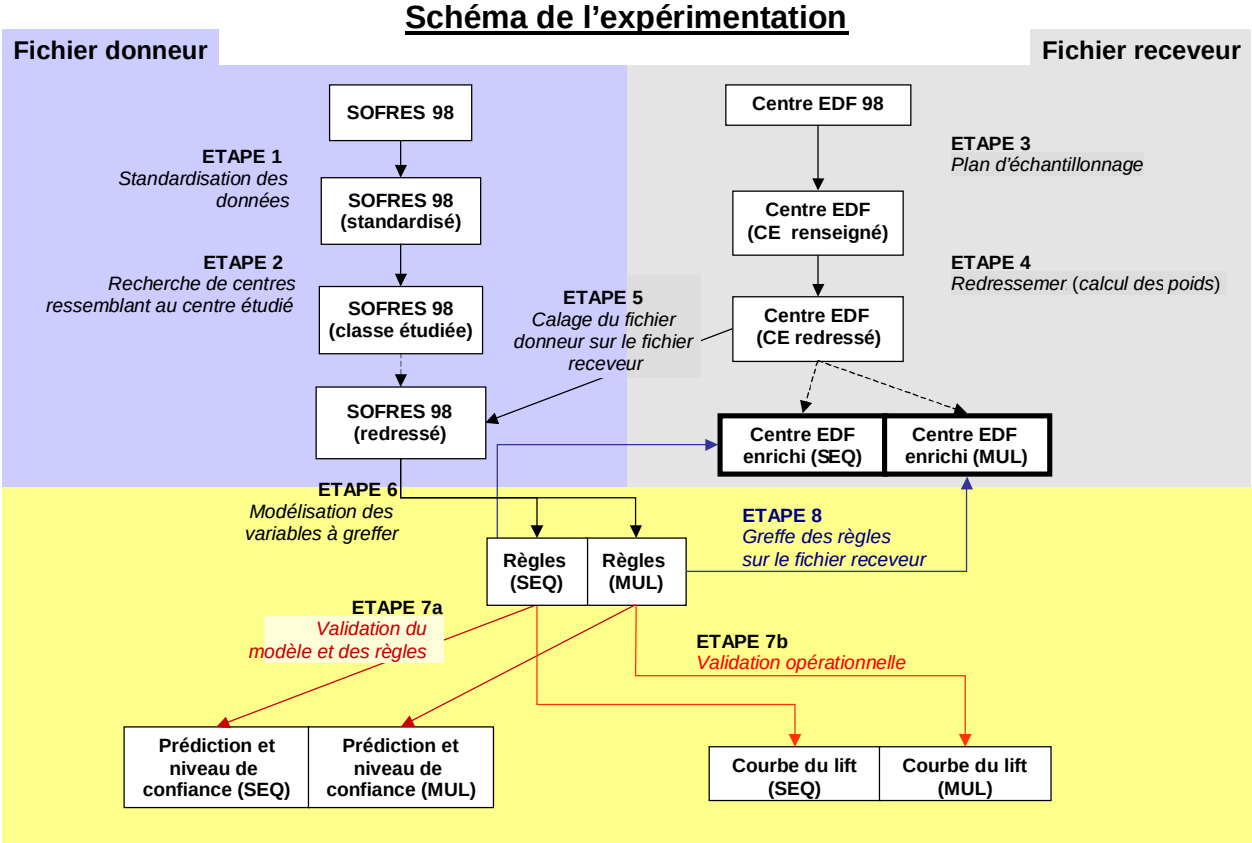
PDF créé avec la version d'essai FinePrint pdfFactory Pro www.gs2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

V-2.3.*Démarche de fusion statistique des bases de données*

Le schéma ci-dessous représente l'ensemble de la démarche de fusion statistique de données entre les deux bases de données EGS* et EDF/SOFRES datant de 1998.

Schéma de l'expérimentation



La démarche employée se décompose en huit étapes, comme nous avons pu le voir sur le schéma précédent.

· ETAPE 1 [*Standardisation des données*]
Dans le fichier EDF/SOFRES, à chaque enquêté est attribué un poids de sondage w_I, pour qu'il soit représentatif des clients auxquels il équivaut en terme de caractéristiques socio-démographiques (taille d'agglomération, âge et profession) et caractéristiques EDF (tarif, puissance souscrite et niveau de consommation). La somme des poids de sondage des 59105 clients est de 340.293.020 [(2)]. La standardisation revient simplement à effectuer une règle de trois afin d'obtenir une nouvelle somme de poids égale à 59105, c'est-à-dire $w_i^* = (w_i/\Sigma w_i)\times 59105$. Cette opération, purement technique, est utile dans le cadre de la modélisation effectuée dans l'étape 6.

[2]Cette valeur correspond a une estimation de la consommation totale d'électricité annuelle en kWh du chauffage électrique pour les 59105 enquêtés (la moyenne est de près de 5800 kWh).

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

· ETAPE 2 [*Recherche de centres ressemblant à EGS* *]
L'enquête EDF/SOFRES contient les 98 centres EGS de la métropole. Dans chaque centre, un certain nombre de personnes ont répondu à l'enquête. Il y a 591 clients concernés dans le Centre EGS *, ce qui est proche d'ailleurs du nombre moyen de répondants par centre (603 en moyenne). Cependant, ce nombre d'individus est trop peu élevé pour appliquer raisonnablement les modèles statistiques de l'étape 6. La solution la plus rapide serait de retenir l'ensemble des 59105 clients du fichier d'enquête. Cependant, cette solution risque de faire se mélanger des centres qui ne se ressemblent pas d'un point de vue des caractéristiques socio-démographiques (âge, enfants, …) de logement (type, statut, année de construction) et leur relation vis-à-vis d'EDF (puissance souscrite, qualité payeur, années de création du contrat, …). Par conséquent, il est préférable de rechercher et de sélectionner des centres proches du Centre EGS *, en fonction des caractéristiques citées ci-dessus. Nous pouvons alors espérer obtenir suffisamment de centres ressemblants à EGS * pour être en mesure d'effectuer les modélisations.

Brièvement, la démarche mise en œuvre consiste tout d'abord à résumer l'ensemble des informations (59105 clients croisés avec leurs caractéristiques), à l'aide de « l' analyse des Correspondances Multiples ». Puis, sur ce résumé d'information, une typologie des centres est construite permettant de regrouper les centres se ressemblant entre eux, au moyen d'une classification automatique. 6 classes sont obtenues. Le Centre EGS * est dans le groupe des centres suivants :

Nom du Centre	*Nombre d'enquêtés*
Hainaut-Cambrésis	505
Douai	453
Arras	526
Côte d'Opale	524
Pays de Somme	507
Pays de l'Aisne	561
Ardennes	520
Nancy Lorraine	477
Haute-Marne et Meuse	571
Vosges	569
Metz Lorraine	568
Lorraine 3 Frontières	513
Alsace	591
Franche Comté Nord	524
Franche Comté Sud	625
Total	8034

Tableau 2 : Liste des centres ressemblants à EGS Alsace

Nous pouvons constater d'une part, que ce regroupement de centres est au moins tout à fait naturel, en terme de proximité géographique, et d'autre part que le nombre de clients enquêtés passe de 591 à 8034, ce qui nous permet de réaliser l'étape 6 de modélisation.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

· Etape 3 [Construction du plan d'échantillonnage] et étape 4 [redressement]

La base de données EGS * 98 contient 246279 clients, mais seuls 130185 sont renseignés sur l'énergie de chauffage à l'aide de la variable USG_T_CHAUP. Cependant, 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 »). Pour pallier ce problème, nous avons construit un plan d'échantillonnage (post-stratifié). Ce plan de sondage va nous permettre de structurer la population étudiée, afin d'estimer des proportions dans un premier temps, et de construire des modèles statistiques 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.). Celles-ci sont nommées variables d'échantillonnage (ou auxiliaires) et doivent être les plus indépendantes, discriminantes et significatives possibles pour stratifier la population en fonction du problème posé[3].

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, à l'aide de la fusion statistique.

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.

[3] 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 alors une strate. Enfin, la densité de chaque pavé exhibe la proportion de clients associée à chaque strate par rapport à l'ensemble des clients.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Année de création	*Nombre de clients*	%
1970 et avant	48776	19,81
1971 à 1980	44313	17,99
1981 à 1990	60633	24,62
1991 à 1995	51085	20,74
1996 à 1998	41472	16,84
Total	246279	100,0

Tableau 3 : Répartition de l'année de création sur la population

Type de logement	*Nombre de clients*	%
Collectif	100723	40,90
Individuel	145556	59,10
Total	246279	100,0

Tableau 4 : Répartition du type de logement sur la population

Nombre de clients dans la ville	*Nombre de clients*	%
Moins de 2000	62688	25,45
2000 à moins de 5000	100424	40,78
5000 à moins de 10000	29718	12,07
10000 à moins de 20000	28621	11,62
20000 et plus	24828	10,08
Total	246279	100,0

Tableau 5 : Répartition du nombre de clients dans la ville sur la population

Tarif×Puissance souscrite	*Nombre de clients*	%
Simple et ≤ 3 kVA	20533	8,34
Simple et 4 à 8 kVA	97771	39,70
Simple et 9 à 14 kVA	14381	5,84
Simple et ≥ 15 kVA)	4702	1,91
Double et ≤ 6 kVA	50497	20,50
Double et 8 à 9 kVA	23138	9,40
Double et 10 à 15 kVA	24840	10,09
Double et ≥ 16 kVA)	10417	4,23
Total	246279	100,0

Tableau 6 : Répartition tarif´puissance sur la population

Enfin, comme le fichier d'enquête EDF/SOFRES ne concerne que les clients « chauffage électrique », nous avons seulement gardé dans le fichier EGS Alsace, ceux dont la variable USG_T_CHAUP prenaient les valeurs ("C1", "C2" et "C4")[4] et CHP = "E"[5]. Il reste donc 27785 clients munis de leurs poids

[4] C1 = Convecteurs électriques ; C2 = Chaudière. indiv. électrique ; C4 = Pompe à chaleur.

[5] lors de l'expérimentation de 1998, nous nous sommes aperçus que certains clients avaient un renseignement erroné sur le chauffage électrique. Par conséquent, nous les avons différenciés : CHP = E correspond à des clients dont le renseignement est jugé fiable (CHP = e, à des informations non corrects qui ont été éliminées).

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

d'échantillonnage calculés lors de l'étape de redressement dans le fichier EGS *.

Etape 5 [*Calage du fichier donneur sur le fichier receveur*]
Cette étape est primordiale dans la démarche de fusion statistique de données. En effet, même si les deux fichiers possèdent des variables communes, cela est seulement une condition nécessaire, mais pas suffisante. En effet, imaginons que ces variables communes correspondent à des rails sur un pont entre deux rives (les deux bases de données). Pour qu'un train puisse arriver sans encombre de l'autre coté, il faut que les rails aient le même écartement. En d'autres termes, les clients de la base de données EGS * doivent être comparables aux clients interrogés dans l'enquête EDF/SOFRES. Sans quoi, il y aurait un déséquilibre, en terme de poids, entre les données à greffer issues du fichier donneur et les données greffées sur le fichier receveur. Plus formellement, on désigne par F_s et f_s les proportions respectives de la strate s dans EGS * et dans EDF/SOFRES. Si $w_{i\in s}$* est le poids d'échantillonnage de l'enquêté i de la strate s du fichier EDF/SOFRES, alors ce poids est modifié par la formule suivante : $w_{i\in s}$**=$w_{i\in s}$*×(F_s/f_s).

Par exemple, la strate F_s des clients ayant établi un contrat entre 1971 et 1980, en double tarif, avec 8 ou 9 kVA et habitant en logement collectif vaut 0,0066 dans le fichier EGS *, alors que f_s = 0,0050 dans le fichier EDF/SOFRES. Dans ce cas, un enquêté qui a un poids de sondage dans EDF/SOFRES de 0,6018, aura un nouveau poids de : $w_{i\in s}$** = 0,6018×(0,0066/0,0050) = 0,7944.

Etape 6 : [*Modélisation des variables à greffer*]
L'objectif de cette étape est de construire des modèles statistiques permettant de prédire les variables à greffer (variables à expliquer) du fichier EDF/SOFRES sur le fichier EGS * à l'aide des variables communes (variables candidates à l'explication) aux deux bases de données. Ces modèles statistiques fournissent finalement des probabilités estimées de choisir telle ou telle réponse à la question. La plus grande probabilité estimée déterminera une règle de prédiction (pour l'attente numéro 5 de la question 25) du type :
SI l'enquêté *i* habite *en logement individuel*
ET a une *puissance souscrite inférieure ou égale à 6 kVA*
ET *consomme entre 0 et 5000 kWh*
ET est *mensualisé*
ET a souscrit un *contrat avec EDF entre 1981 et 1990*
ALORS il n'est « *pas intéressé par une visite à son domicile pour qu'EDF examine son installation de chauffage électrique* »

En effet, la probabilité estimée associée est de 0,47, alors que celle associée à « plutôt intéressé » est de 0,34 et celle correspondant à « très intéressé » vaut 0,19.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Comme nous l'avons vu, il y a quatre paquets de variables à greffer concernant :

- « Ce que vous pensez de votre chauffage électrique » (3 variables)
- « Les évolutions de votre chauffage » (1 variable)
- « Ce que vous pensez d'EDF » (7 variables)
- « Ce que vous attendez d'EDF » (9 variables)

et 7 variables communes :

- Tarif/puissance souscrite
- Année de création
- Type de logement
- Politique de facturation
- Qualité payeur
- Code paiement
- Tranche de consommation annuelle d'électricité

Les deux approches : méthode séquentielle et méthode multivariée (notées respectivement SEQ et MUL sont utilisées pour modéliser chaque paquet de variables à expliquer. La première approche consiste à modéliser séquentiellement chaque variable à expliquer en fonction des variables communes. De plus, afin de préserver une certaine structure de corrélation entre variables à expliquer observées et variables prédites, une variable déjà expliquée dans la séquence, devient elle-même une variable « explicative » , etc. Cependant, même si cette méthode offre de nombreux avantages, comme on le verra dans la partie sur la validation (pourcentage de bien classés, par exemple), la mauvaise reconstitution des corrélations discutée plus haut est tout de même un des inconvénients majeurs.

Par conséquent, nous avons développé une nouvelle approche (méthode multivariée), offrant la possibilité de prédire plusieurs variables à la fois et donc tenter de mieux préserver les corrélations entre variables à expliquer. Cela nous permet donc de réaliser cette opération sur chacun des quatre paquets de variables à greffer. Signalons cependant que si cette méthode améliore grandement la cohérence des corrélations entre variables à expliquer, elle fournit des taux de bien classés plus faibles.

De plus, nous utilisons une approche plus simple dite « méthode univariée » qui représente la méthode de référence naturelle. En effet, les variables à expliquer sont modélisées une par une, sans tenir compte de celles modélisées comme dans le cas de la méthode séquentielle, ou en bloc comme pour la méthode multivariée. Par conséquent, la méthode univariée ne nous servira dans la suite que comme point de comparaison dans les étapes 7a et 7b de validation.

Etape 7 : [*Validation*]
Cette étape est primordiale car d'elle dépend la décision de remettre en cause le ou les modèles mis en œuvre. En effet, elle consiste d'une part à valider statistiquement le modèle et les règles de prédiction associées, et d'autre part à effectuer une validation opérationnelle des résultats obtenus. En

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

d'autres termes, le premier type de validation correspond à un outil permettant au statisticien de s'assurer de la qualité des traitements statistiques effectués. Par contre, la validation opérationnelle fournit, par exemple, à l'expert clientèle une aide à la décision pour atteindre telle ou telle cible de son marché de clients.

Etape 7a : [*Validation du modèle et des règles*]
Cette étape contient d'une part la validation du modèle, c'est-à-dire sa qualité de reproduction de l'information à prédire et la fourniture d'un niveau de confiance individuel associé à chaque enquêté, puis, comme on le verra dans l'étape 8, aussi à certains clients d'EGS *.

La validation du modèle passe par 3 phases :
- la reconstitution des pourcentages de réponses marginales associées à chaque question,
- le pourcentage de bien classés, c'est-à-dire le pourcentage d'enquêtés dont la réponse prédite grâce à la règle correspond à ce qu'ils avaient effectivement répondu,
- l'adéquation des corrélations estimées aux corrélations observées entre les variables à expliquer (variable à prédire).

Cette phase de validation est effectué sur un échantillon-test. En effet, la table de données est découpée en deux, un échantillon d'apprentissage sur lequel on construit les modèles et on établit des règles de prédictions, et un échantillon-test sur lequel on applique les règles de prédiction.

Adéquation des pourcentages de réponses estimés aux observés

Par exemple, pour l'attente numéro 5, les répartitions marginales observée et estimée obtenues avec la méthode séquentielle sont les suivantes :

Réponses	*% observés*	*% estimés*
Pas intéressé	30,54	19,13
Plutôt intéressé	35,62	41,62
Très intéressé	33,84	39,25

Tableau 7 : Répartitions marginales de l'attente 5

Un test statistique d'adéquation des proportions (test du Chi^2) montre que ces deux répartitions ne sont pas en adéquation.
Par contre, la variable à expliquer sur l'attente numéro 7 (« attente vis-à-vis du détail de la consommation d'électricité par usage ») obtient une bonne adéquation statistique de la répartition marginale estimée à la répartition marginale observées.

Réponses	*% observés*	*% estimés*
Pas intéressé	20,58	22,77
Plutôt intéressé	38,65	38,00
Très intéressé	40,77	39,23

Tableau 8 : Répartitions marginales de l'attente 7

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Pourcentage de bien classés

Ce pourcentage est calculé grâce à la matrice de confusion qui croise les valeurs observées et les valeurs estimées. Sur le tableau suivant, l'attente 5 observée est représentée par les lignes, alors que son estimation est en colonne. Le pourcentage de bien classés est égal à 8,56 + 14,55 + 15,27 = 38,39% qui est à comparer à 35,62% (plus grand pourcentage observé). En d'autres termes, si l'on n'avait pas fait de modèle, le pourcentage de bien classés n'aurait été que de 35,62%. Cependant, ces deux pourcentages sont relativement proches. Par conséquent, de la même façon que pour les répartitions marginales, il est possible de tester statistiquement si un pourcentage de bien classés est significativement supérieur au taux sans modèle. Dans le cas de l'attente 5, le test est significatif.

Estimé Observé	*Pas intéressé*	*Plutôt intéressé*	*Très intéressé*	Total
Pas intéressé	8,56	12,52	9,45	30,54
Plutôt intéressé	6,54	14,55	14,52	35,62
Très intéressé	4,03	14,54	15,27	33,84
Total	19,13	41,62	39,25	100,00

Tableau 9 : Matrice de confusion sur attente 5

Pour l'attente 7, le pourcentage de bien classés est égal à 13,39 + 21,33 + 27,40 = 62,12%, comparé à 40,77% est très significatif.

Estimé Observé	*Pas intéressé*	*Plutôt intéressé*	*Très intéressé*	Total
Pas intéressé	13,39	5,78	1,41	20,58
Plutôt intéressé	6,90	21,33	10,42	38,65
Très intéressé	2,48	10,89	27,40	40,77
Total	22,77	38,00	39,23	100,00

Tableau 10 : Matrice de confusion sur attente 7

Adéquation des corrélations estimées aux corrélations observées

Dans cette phase, nous calculons le pourcentage croisé entre deux variables observées et celui entre deux variables estimées. Prenons les attentes 5 et 7.

Croisement	*Observé*	*Estimé*
Pas intéressé×pas intéressé	16,90	8,56
Pas intéressé×plutôt intéressé	9,82	6,54
Pas intéressé×très intéressé	3,82	4,03
Plutôt intéressé×pas intéressé	2,38	9,06
Plutôt intéressé×plutôt intéressé	21,45	16,68
Plutôt intéressé×très intéressé	11,79	15,87
Très intéressé×pas intéressé	1,30	5,14

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.gs2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Très intéressé×plutôt intéressé	7,38	14,78
Très intéressé×Très intéressé	25,16	19,33

Tableau 11 : Croisement des attentes 5 et 7 (estimé vs observé)

Les répartitions ne se ressemblent pas, d'ailleurs le test statistique associé montre qu'il n'y a pas d'adéquation des croisements estimés aux croisements observés.

Calcul du niveau de confiance individuel

Cette information permet de garantir une qualité de reconstitution des individus estimés d'un groupe par rapport à ce qui a été observé sur ce même groupe.

Par exemple, reprenons le groupe d'enquêtés introduit dans l'étape 6 :

SI l'enquêté i habite en logement individuel
ET a une puissance souscrite inférieure ou égale à 6 kVA
ET consomme entre 0 et 5000 kWh
ET est mensualisé
ET a souscrit un contrat avec EDF entre 1981 et 1990
ALORS il n'est « pas intéressé par une visite à son domicile pour qu'EDF examine son installation de chauffage électrique »

Dans ce cas, seule la colonne « Pas intéressé » est considérée puisque la probabilité estimée était la plus élevée : 0,47 (contre 0,34 et 0,19 pour les deux colonnes suivantes). Le nombre d'individus pondérés bien classés est de 8,43, alors que le nombre total d'individus pondérés vaut 11,02. Par conséquent, le niveau de confiance individuel est de : 8,43/11,02 = 0,7650. Cela correspond bien à 76,50 dans la case à l'intersection de la deuxième ligne et la deuxième colonne.

Le niveau individuel de confiance peut donc prendre ses valeurs entre 0 et 1. Plus il est proche de 1, plus l'adéquation de la prédiction à l'observation est élevée. Il n'y a pas de règle explicite pour fixer une valeur limite supérieure pour laquelle, on peut considérer que le niveau de confiance est statistiquement valide. Nous avions fixé une valeur arbitraire dans le projet DECLIC (dans le cadre des TCS) qui était de 0,75, ce que nous reprenons également ici. Par conséquent, la prédiction sur le groupe précédent est acceptée, car son niveau de confiance associé est inférieur à 0,75. Une autre valeur aurait pu être fixée en fonction du nombre de réponses possible à une question, soit R. Si le niveau de confiance est supérieur à $1{,}5\times(1/R)$, alors on accepte la règle. Dans notre cas, elle aurait été a fortiori acceptée puisque $0{,}7650 > 1{,}5\times(1/3) = 0{,}5$.

Enfin, le niveau de confiance individuel a un lien direct avec le pourcentage de bien classés. En effet, la moyenne pondérée des niveaux de confiance individuels (multiplié par 100) est égale au pourcentage de bien classés.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Etape 7b : [*Validation opérationnelle*]
Comme nous l'avons déjà expliqué, cette étape a pour objectif de fournir une aide à la décision pour l'expert du terrain.

Prenons comme exemple la question 7 :
« Si vous aviez la possibilité de choisir vous-même le système de chauffage de votre logement choisiriez-vous … : Le chauffage électrique ; un système utilisant une autre énergie »
On comprend bien l'intérêt de cette question et surtout des réponses. Un client qui répond « un système utilisant une autre énergie » est un client potentiellement infidèle. Par conséquent, si l'on désire retenir ces clients, une action de ciblage marketing pourrait être mis en œuvre. Pour cela, les deux approches proposées (séquentielle et multivariée) permettent de prédire la réponse des clients quand les règles de prédictions sont greffées dans la base de données EGS *. Voyons quelle est la matrice de confusion de cette question prédite avec la méthode séquentielle :

Estimé Observé	*Le chauffage électrique*	*Un système utilisant une autre énergie*	Total
Le chauffage électrique	72,03 (33,78)	19,02 (10,10)	43,88
Un système utilisant une autre énergie	27,97 (13,12)	80,98 (43,00)	56,12
Total	100,00 (46,10)	100,00 (53,10)	100,00

Tableau 12 : Matrice de confusion sur le choix de l'énergie de chauffage pour l'ensemble des enquêtés

La colonne « *Un système utilisant une autre énergie* » correspond aux enquêtés prédits avec cette réponse dans 53,10% des cas. En décomposant cette colonne, on peut constater que 10,10% parmi l'ensemble des enquêtés ont été prédits avec cette réponse, alors qu'en réalité, ils avaient répondu « *Le chauffage électrique* », alors que le reste, c'est-à-dire 43% ont bien été prédits. Cela permet de calculer le pourcentage de bien classés dans cette colonne, c'est-à-dire : 100×(43,00/53,10) = 80,98%. Cela signifie que si l'on faisait un mailing ciblé sur les clients prédits avec cette réponse, il y en aurait en moyenne huit sur dix qui auraient probablement répondu cela. Ce pourcentage est à comparer avec 56,12% (dernière colonne) qui correspond au pourcentage observé et qui aurait naturellement obtenu sans modélisation. Ce résultat est donc très satisfaisant car le gain du modèle est de 80,98/56,12 = 1,44. En d'autres termes, si l'on faisait un mailing sur 2000 clients sans modèle, on atteindrait en moyenne : 2000×0,5612 = 1122 clients comme choisissant une autre énergie, alors qu'avec le modèle, on en obtiendrait : 2000×0,8098 = 1620 clients potentiels. Une autre façon de raisonner serait de se mettre dans la situation suivante : *combien le modèle me permet d'économiser sur un mailing ?* Par exemple, si l'on veut atteindre 1000 clients susceptible d'être dans cible des « infidèles », il faudra envoyer : 1000/0,8098 = 1235 courriers avec le modèle, alors qu'il faudra en envoyer 1000/0,5612 = 1782 sans le modèle. Ce dernier nombre peut être aussi obtenu en multipliant 1235 par 1,44. Cela fait donc un gain de 1782-1235 = 547 envois, ce qui n'est pas négligeable.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Cependant, il encore possible de faire augmenter ce gain. Cela a deux avantages : une plus grande fiabilité dans l'atteinte de la cible et une réduction de coût plus importante. Pour cela, nous mettons en ordre décroissant les probabilités prédites par le modèle d'avoir répondu « *Un système utilisant une autre énergie* » , puis nous découpons, par exemple par tranche de 10%. Les 10% qui ont obtenu les plus fortes probabilités, puis les 20% (les 10% précédents et 10% suivants), ..., jusqu'à 100%, c'est-à-dire tout l'échantillon, car tout est cumulé. Il est alors possible de faire une matrice de confusion, comme précédemment, pour les premiers 10%, puis les 20%, ...

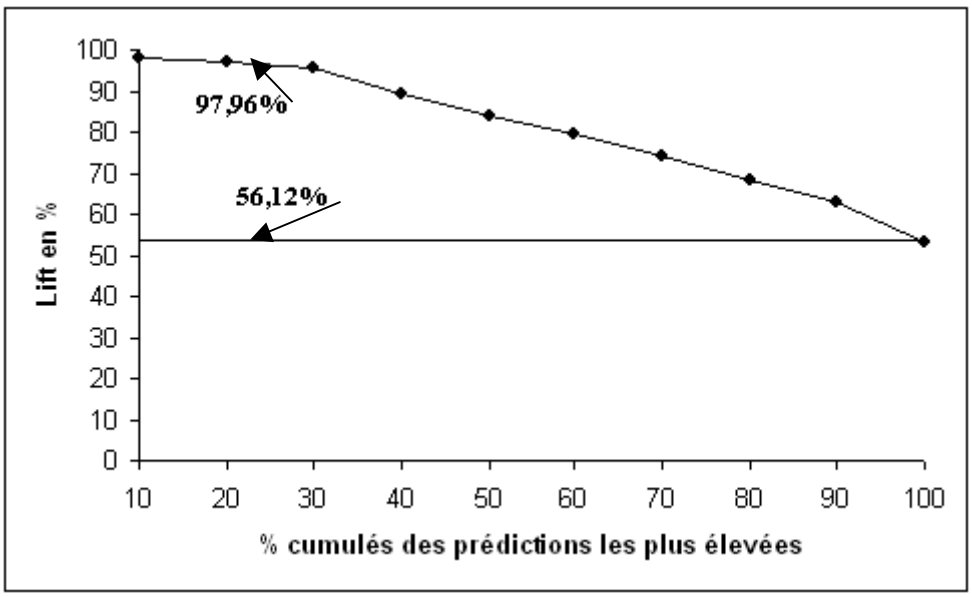
La matrice de confusion suivante concerne seulement les premiers 10% :

Estimé Observé	*Le chauffage électrique*	*Un système utilisant une autre énergie*	Total
Le chauffage électrique	0,00 (0,00)	2,04 (2,04)	2,04
Un système utilisant une autre énergie	0,00 (0,00)	97,96 (97,96)	97,96
Total	0,00 (0,00)	100,00 (53,10)	100,00

Tableau 13 : Matrice de confusion sur la question 7 pour les premiers 10%

On peut constater que si l'on prend seulement les 10% avec la probabilité la plus élevée d'être potentiellement infidèle, alors il y en a 97,96% qui ont réellement répondu « *Un système utilisant une autre énergie* ». Ceci est à comparer à 56,12% pour l'ensemble des enquêtés qui ont donné cette réponse (cf. dernière colonne tableau 12). Nous pourrions alors continuer cet exercice, jusqu'à ce que l'on arrive à 100%.

Le graphique suivant fournit la courbe de lift sur les tranches cumulées de 10%.



Graphique 1 : Courbe de lift pour le choix de l'énergie de chauffage

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.gs2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Le gain de performance calculé précédemment donne maintenant : 97,96/56,12 = 1,75. Cela permet cette fois-ci d'envoyer seulement 1000/0,9796 = 1021 courriers, à la place de 1285, en prenant les prédits « *Un système utilisant une autre énergie* » avec le modèle et encore plus sans modèle : 1782. cela fait donc un gain de 761 envois, ce qui est loin d'être négligeable.

On voit bien l'intérêt pratique et immédiatement utilisable d'un tel type de validation (opérationnelle), par rapport à la validation statistique.

Etape 8 : [*Greffe des règles sur le fichier receveur*]
Cette étape permet d'adapter les règles de prédiction établies précédemment grâce au fichier donneur sur le fichier receveur. En effet, l'ensemble des clients dans le fichier EGS * qui possèdent les caractéristiques prendront par exemple tous la valeur prédite sur l'attente 5 de la question 25. Dans notre exemple, 16 clients sont concernés et l'on considérera qu'ils ne sont « *pas intéressés par une visite à leur domicile pour qu'EDF examine leur installation de chauffage électrique* ».

Pour chaque groupe de clients et chaque variable à greffer, les prédictions sont fournies au fichier EGS *. L'extrait suivant du fichier montre ce qui est disponible dans le fichier receveur. La première colonne correspond à la référence client (CLI_C_LOCAB), la deuxième est relative à la prédiction de l'attente 5, le troisième colonne a trait au niveau de confiance individuel et la dernière fournit la probabilité estimée de prendre la réponse 1.

CLI_C_LOCAB	pred_attenteb5_SEQ	niv_cnf_ind_attenteb5_SEQ	prob_ind_attenteb5_SEQ
510112135197	1	0.76483	0.47092
593110020689	1	0.76483	0.47092
528905825337	1	0.76483	0.47092
504020150174	1	0.76483	0.47092
508916050730	1	0.76483	0.47092
511610295197	1	0.76483	0.47092
505321165288	1	0.76483	0.47092
509441855707	1	0.76483	0.47092
517520345210	1	0.76483	0.47092
590135330807	1	0.76483	0.47092
505816300890	1	0.76483	0.47092
510011301128	1	0.76483	0.47092
528110150120	1	0.76483	0.47092
545630250191	1	0.76483	0.47092
572420295146	1	0.76483	0.47092
575010010799	1	0.76483	0.47092

Tableau 14 : Extrait du fichier livré sur attente 5 pour un groupe donné d'enquêtés

L'ensemble de ce fichier permet aussi de sélectionner parmi les niveaux individuels de confiance supérieurs à 0,75, les premiers 10% de probabilités estimées les plus élevées et donc d'obtenir les clients les plus intéressants selon un ciblage marketing.

V-2.4.*Résultats obtenus*

Nous avons appliqué les deux méthodes de fusion statistique (séquentielle et multivariée) sur les quatre paquets de variables à greffer concernant :

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

- « Ce que vous pensez de votre chauffage électrique » (3 variables)
- « Les évolutions de votre chauffage » (1 variable)
- « Ce que vous pensez d'EDF » (7 variables)
- « Ce que vous attendez d'EDF » (9 variables)

Premièrement pour chacune d'elle, nous fournissons 6 nouvelles informations dans le fichier receveur EGS* :
- la prédiction
- la probabilité estimée
- le niveau de confiance individuel

et ceci pour chacune des deux méthodes.

Deuxièmement pour chaque paquet de variables, nous fournissons quatre tableaux résumant des résultats sur :
- le classement des deux méthodes[6] vis-à-vis de la reconstitution des pourcentages marginaux de réponses aux questions,
- le pourcentage de bien classés et le test statistique associé,
- le classement des deux méthodes sur l'adéquation des corrélations estimées aux corrélations observées entre les variables à expliquer,
- la proportion de clients ayant un niveau de confiance individuel supérieur à 0,75, sur le fichier receveur.

Enfin, pour le paquet concernant « Ce que vous pensez de votre chauffage électrique » , nous fournissons les courbes pour chaque variable et les méthodes univariée, séquentielle et multivariée.

Classement des deux méthodes sur les pourcentages marginaux

Le tableau suivant montre que sur le thème « Ce que vous pensez de votre chauffage électrique », l'adéquation des pourcentages estimés aux pourcentages observés est meilleure pour la méthode multivariée (MUL) (deux fois classée première) que pour la méthode séquentielle (SEQ), sauf pour la question sur le choix du système.

Question	*Classement sur* SEQ	*Classement sur* MUL
Satisfaction (Q6)	2	1
Choix du système (Q7)	1	2
Conseil du CE (Q8)	2	1

Tableau 15 : Test statistique sur les pourcentages marginaux

Pour le thème « Ce que vous pensez d'EDF », seule l'opinion 7 est mieux reconstituée par la nouvelle méthode (MUL), que par l'ancienne (SEQ).

Image	*Classement sur* SEQ	*Classement sur* MUL
1	1	2
2	1	2

[6] La méthode multivariée (**MUL**) n'a pas été appliquée sur le thème « Les évolutions de votre chauffage », car il n'y a qu'une seule question.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

3	1	2
4	1	2
5	1	2
6	1	2
7	2	1

Tableau 16 : Test statistique sur les pourcentages marginaux

Pour le thème « Ce que vous attendez d'EDF », seules les attentes 5 et 9 sont mieux reconstituées par la méthode multivariée, que par la méthode séquentielle.

Attente	*Classement sur* SEQ	*Classement sur* MUL
1	1	2
2	1	2
3	1	2
4	1	2
5	2	1
6	1	2
7	1	2
8	1	2
9	2	1

Tableau 17 : Test statistique sur les pourcentages marginaux

Pourcentages de bien classés

Pour la question sur la satisfaction (première ligne), la deuxième colonne contient le pourcentage de bien classés (%bcl) contre le pourcentage maximum observé (%max) pour la méthode séquentielle (SEQ), alors que la quatrième colonne a trait au test statistique associé. On peut d'ailleurs constater que celui-ci indique que %bcl (54,1%) et %max (52,6%) ne sont pas significativement différents. En d'autres termes, cela signifie que le modèle n'apporte pas plus d'information que si l'on avait pris la réponse relative au pourcentage maximum. La troisième et la cinquième colonnes concernent la nouvelle méthode (MUL). La dernière colonne montre notamment que le pourcentage de bien classés (49,7%) est significativement inférieur au pourcentage maximum (52,6%). Par contre, les deux questions sur le choix du système et le conseil du CE ont un taux de bien classés significativement supérieur au pourcentage maximum. Cependant, l'ancienne méthode est plus performante que la nouvelle.

Question	*%bcl/%max sur* SEQ	*%bcl/%max sur* MUL	*Test sur* SEQ	*Test sur* MUL
Satisfaction (Q6)	54,1/52,6	49,7/52,6	0	-
Choix du système (Q7)	76,6/56,0	65,4/56,0	+	+
Conseil du CE (Q8)	69,5/38,0	43,9/38,0	+	+

Tableau 18 : Test statistique sur le pourcentage de bien classés

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Ceci se confirme sur les résultats concernant le thème « Ce que vous pensez d'EDF ».

Image	*%bcl/%max sur* SEQ	*%bcl/%max sur* MUL	*Test sur* SEQ	*Test sur* MUL
1	84,9/64,9	65,1/64,9	+	0
2	82,2/69,0	69,1/69,0	+	0
3	85,1/77,6	77,7/77,6	+	0
4	80,1/56,7	59,0/56,7	+	+
5	72,6/56,1	57,1/56,1	+	0
6	83,2/62,2	61,4/62,2	+	0
7	59,9/56,6	59,1/56,6	+	+

Tableau 19 : Test statistique sur le pourcentage de bien classés

Le tableau relatif aux résultats du thème « Ce que vous attendez d'EDF » montre que la nouvelle méthode est encore moins bonne dans ce cas.

Attente	*%bcl/%max sur* SEQ	*%bcl/%max sur* MUL	*Test sur* SEQ	*Test sur* MUL
1	68,2/56,5	55,0/56,5	+	-
2	71,9/56,7	56,0/56,7	+	0
3	63,7/46,7	45,2/46,7	+	-
4	62,7/40,9	40,4/40,9	+	0
5	38,4/35,6	35,5/35,6	+	0
6	67,0/58,9	58,2/58,9	+	0
7	62,1/40,8	43,1/40,8	+	+
8	59,7/44,5	45,9/44,5	+	+
9	61,4/46,5	46,1/46,5	+	0

Tableau 20 : Test statistique sur le pourcentage de bien classés

Classement des deux méthodes sur les croisements.

Prenons le cas du thème « Ce que vous pensez de votre chauffage électrique », la nouvelle méthode (MUL) arrive toujours en premier par rapport à l'ancienne méthode (SEQ). Ceci est tout à fait logique puisque la nouvelle méthode a été spécialement conçue pour préserver les croisements entre les variables à expliquer.

Question	chi2 sur SEQ	chi2 sur MUL
Satisfaction x choix	636,94 (<0,0001)	630,92 (<0,0001)
Satisfaction x conseil	1079,52 (<0,0001)	352,52 (<0,0001)
Choix x conseil	859,27 (<0,0001)	51,10 (<0,0001)

Tableau 21: Test statistique sur les croisements de distributions

Proportion de clients avec une confiance supérieure à 0,75

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ss2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Comme on l'a précisé dans l'étape 7a, le niveau de confiance individuel est le garant d'une bonne prédiction et donc d'une bonne qualité de l'évaluation de la greffe effectuée du fichier donneur sur le fichier receveur. Nous fournissons deux indicateurs. Le premier correspond à la moyenne des niveaux de confiance individuels supérieurs à 0,75. Le second est relatif à la proportion de clients répondant à ce critère. Plus cette proportion est élevée, plus le nombre de clients sélectionnés sera élevé.
Dans le cas du thème « Ce que vous pensez de votre chauffage électrique », les résultats suivants montrent que si les deux méthodes font à peu près jeu égal pour l'indicateur de la moyenne, les proportions de clients que l'on pourrait retenir pour des actions marketing sont plus élevées pour la méthode séquentielle.

Question	*Moyenne sur* SEQ	*Moyenne sur* MUL	*Proportion sur* SEQ	*Proportion sur* MUL
Satisfaction (Q6)	0,93	0,90	0,14	0,05
Choix du système (Q7)	0,95	0,87	0,36	0,26
Conseil du CE (Q8)	0,94	0,91	0,16	0,04

Tableau 22 : Proportion de clients avec un niveau de confiance > 0,75

Le tableau suivant sur le thème « Ce que vous pensez d'EDF », accentue encore plus ce résultat.

Image	*Moyenne sur* SEQ	*Moyenne sur* MUL	*Proportion sur* SEQ	*Proportion sur* MUL
1	0,96	0,85	0,97	0,21
2	0,96	0,85	0,85	0,28
3	0,98	0,86	0,87	0,64
4	0,88	0,84	0,78	0,11
5	0,92	0,86	0,35	0,07
6	0,93	0,81	0,97	0,11
7	0,87	0,84	0,35	0,14

Tableau 23 : Proportion de clients avec un niveau de confiance > 0,75

Enfin, les résultats obtenus par la nouvelle méthode sont encore moins bons pour le thème « Ce que vous attendez d'EDF », que pour le précédent. Seule la proportion associée à l'attente 1 est plus élevé pour la nouvelle méthode.

Attente	*Moyenne sur* SEQ	*Moyenne sur* MUL	*Proportion sur* SEQ	*Proportion sur* MUL
1	0,74	0,02	0,75	0,24
2	0,86	0,83	0,66	0,06
3	0,91	0,81	0,71	0,03
4	0,81	0,87	0,40	0,01
5	0,93	0,88	0,03	0,01
6	0,89	0,81	0,62	0,11
7	0,87	0,92	0,35	0,01
8	0,83	0,88	0,35	0,01
9	0,97	0,91	0,58	0,01

Tableau 24 : Proportion de clients avec un niveau de confiance > 0,75

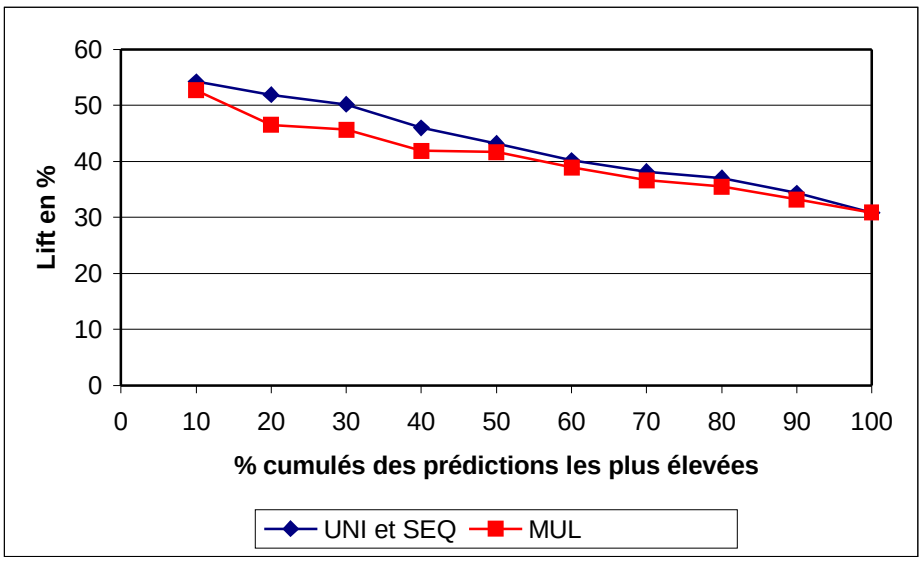
PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

V-2.5. *Validation opérationnelle et courbe de lift*

La satisfaction vis-à-vis du chauffage électrique

Les deux courbes de lift[7] concernent la question 6 sur la satisfaction, les prédictions cumulées correspondent aux insatisfaits. Avec les premiers 10% des prédictions les plus élevés permettent d'obtenir, 54% des enquêtés qui sont réellement insatisfaits, pour les méthodes univariée et séquentielle, alors que pour la méthode multivariée, on atteint 53%. Ceci est à comparer avec 31% correspondant aux insatisfaits sur l'ensemble des clients (ordonnée correspondante à 100% en abscisse). En d'autres termes, le modèle statistique permet de faire 54%/31% = 1,74 fois mieux que si l'on avait tiré au hasard 10% parmi l'ensemble des clients. Cela représente un gain important en terme ciblage marketing (efficacité et moindre coût d'envoi) :



Graphique 2 : Courbes de lift sur l'insatisfaction pour les trois méthodes

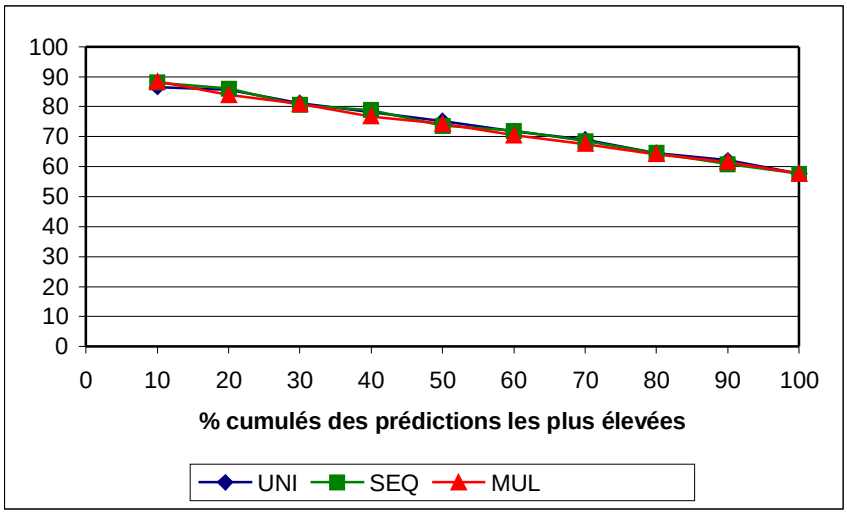
La possibilité de choisir le système de chauffage du logement

On constate que les courbes de lift associées aux méthodes séquentielle et multivariée font à peu près jeu égal concernant le choix du système de chauffage. En effet, les premiers 10% des enquêtés prédits comme désirant remplacer leur chauffage électrique au profit d'une autre énergie ont de bonnes courbes de lift : 88% respectivement. Cela permet de dégager un gain du modèle de : 88%/57,5% = 1,53. Par contre, la méthode univariée fait légèrement moins bien puisqu' elle obtient 86,5%.

[7] Les courbes de lift des méthodes univariée (UNI) et séquentielle (SEQ) sont confondues car, c'est la satisfaction qui a été modélisée en premier dans la méthode séquentielle.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

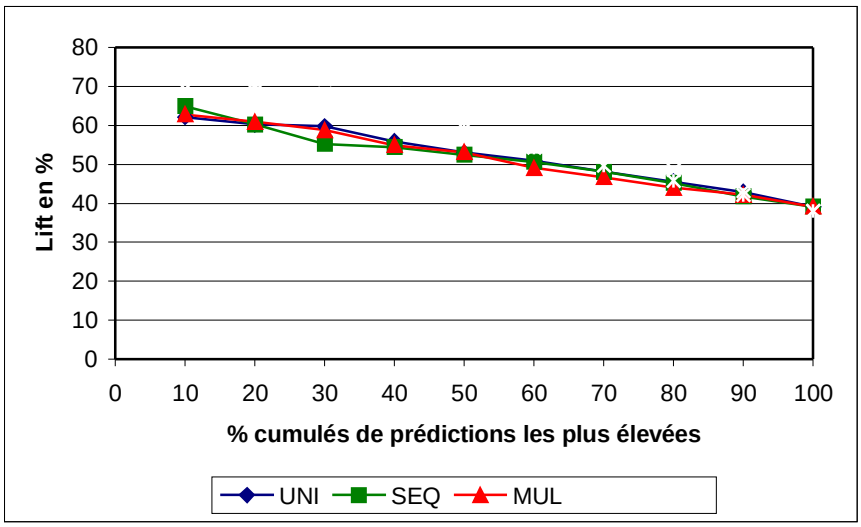
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html



Graphique 3 : Courbes de lift sur le choix d'un autre système de chauffage que le CE

Le conseil du chauffage électrique à l'entourage

Là aussi les courbes de lift des méthodes univariée, séquentielle et multivariée sont à peu près de même niveau pour les premiers 10% : 62%, 63% et 65% pour les premiers 10%. Le gain apporté par le modèle construit à l'aide de la méthode séquentielle est de 65% / 39% = 1,67.



Graphique 4 : Courbes de lift sur le conseil à l'entourage de prendre d'un autre système que le CE

En résumé, cette validation opérationnelle, à l'aide de l'indicateur du lift permet de constater d'une part, que les deux méthodes (séquentielle et multivariée) sont bien meilleures qu'une approche classique telle que la méthode univariée. D'autre part, la méthode multivariée offre de bien meilleurs résultats sur l'ensemble des variables car elle ne dépend pas, comme la méthode séquentielle, de l'ordre d'entrée des variables à prédire, puisque l'ensemble de celle-ci est modélisé en une seule fois.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

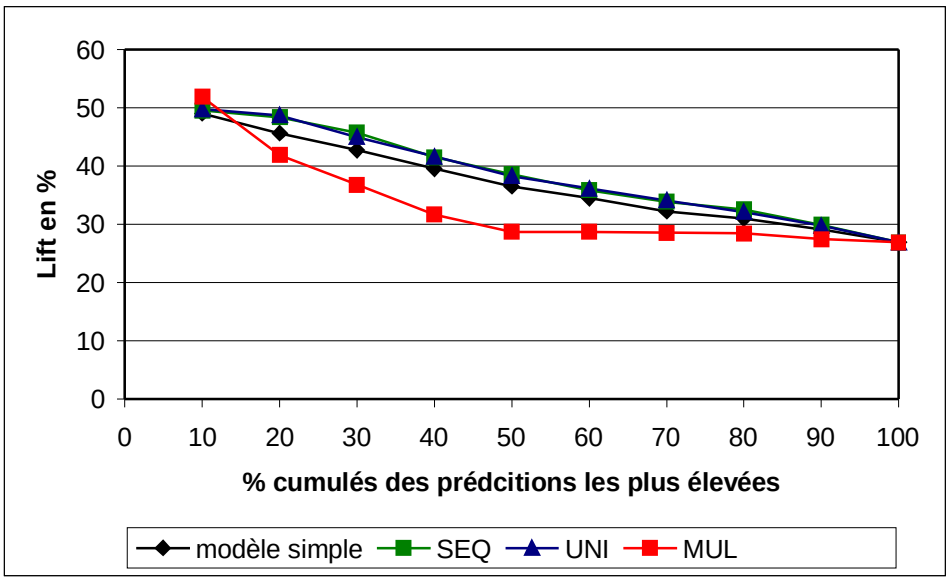
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

V-2.6.*Ciblage direct* vs *fusion statistique de données*

Considérons maintenant qu'un expert clientèle ait défini lui-même sa cible, par exemple, des clients potentiellement infidèles. Sa définition est la suivante : clients insatisfaits par le chauffage électrique, qui ne choisiraient pas ce système de chauffage s'ils le pouvaient et qui ne le conseilleraient pas à leur entourage (25,3% dans l'enquête EDF/SOFRES). Par conséquent, les clients présentant simultanément ces trois propriétés appartiendront à la cible, et les autres clients ne seront pas dans la cible. Une solution pour déterminer les caractéristiques de ces clients à l'aide des variables communes à l'enquête et au fichier EGS * est de construire un modèle prédictif très simple, dans lequel la variable à expliquer possédera seulement deux modalités : « appartenir à la cible » ; « ne pas appartenir à la cible ». Puis, un mailing ciblé pourra être réalisé en prenant pas exemple les premiers 10% ayant les plus fortes probabilités d'appartenir à la cible grâce à la courbe du lift.

Pour mettre en œuvre cette simulation, nous avons, à l'aide des trois méthodes : univariée, séquentielle et multivariée, déterminé les probabilités respectives d'appartenir à la cible et nous les avons comparées à la probabilité estimée par le modèle simple (être ou ne pas être dans la cible).

Le graphique suivant montre que le modèle simple et la méthode univariée font jeu égal, alors que les deux méthodes de fusion statistique de données : séquentielle et multivariée sont très nettement meilleures. En effet, les premiers 10% fournissent seulement que 47% contre 25% si on faisait un mailing à l'ensemble des clients, alors que la méthode multivariée obtient 81% et la méthode séquentielle permet d'atteindre 91% des clients.



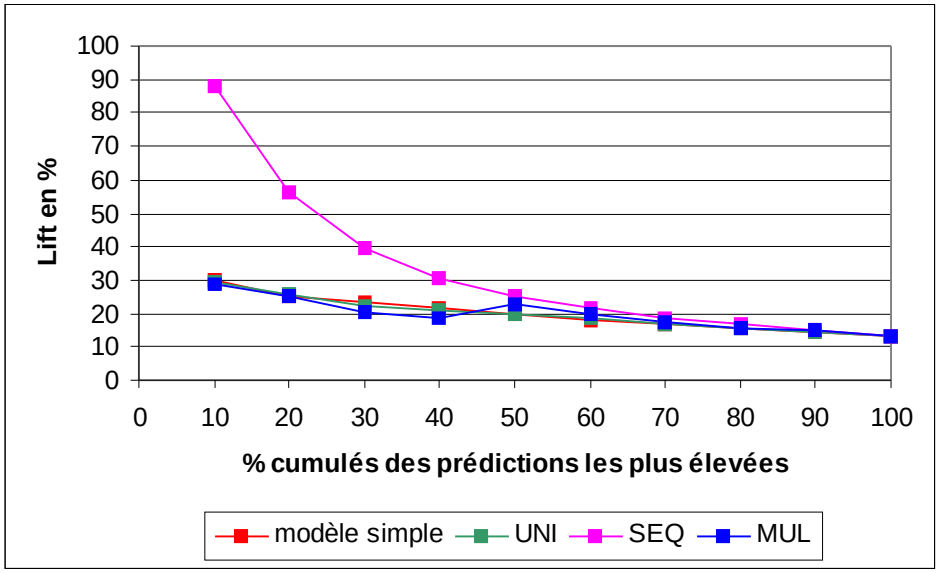
Graphique 5 : Courbes de lift sur la cible des clients infidèles

Prenons une seconde cible : les clients « hyper-fidèles » qui dans notre cas correspondent aux très satisfaits du chauffage électrique qui choisiraient à

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ss2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

nouveau ce type de système et le conseilleraient certainement à leur entourage. Ceux-ci représentent 13,4% dans l'enquête EDF/SOFRES. Le graphique suivant montre aussi que le modèle simple par ciblage direct ne serait pas très performant. En effet, seulement 30% des clients seraient atteints parmi les premiers 10% les mieux prédits, alors que la méthode séquentielle en fourniraient 89%, par contre les deux autres méthodes sont médiocres (30% elles aussi).



Graphique 6 : Courbes de lift sur la cible des clients fidèles

Si l'on continue ces essais sur d'autres cibles[8] à l'aide de ces trois questions, le ciblage direct (modèle simple) fait toujours jeu égal avec la méthode univariée, alors que la méthode séquentielle est constamment la meilleure. La qualité des résultats de la méthode multivariée est intermédiaire.

On peut donc tirer deux avantages non négligeables de la fusion statistique de données par rapport au ciblage direct. Le premier est purement technique, alors que le second est financier. Tout d'abord, les deux méthodes (séquentielle et multivariée) font nettement mieux que le ciblage direct d'un point de vue statistique et opérationnel, en effet les courbes de lift sont meilleures. Coté financier, on comprend bien que l'approche par ciblage direct peut se révéler très coûteuse si l'expert clientèle désire travailler sur plusieurs cibles. En effet, il faut refaire un modèle à chaque fois, ce qui n'est pas du tout performant. De plus, cela ne lui permet pas de disposer d'une panoplie de cibles dans laquelle il pourrait piocher. On voit alors bien l'utilité de la fusion statistique de données qui permet, quant à elle, de prédire un ensemble de variables et dont le croisement de celles-ci permet d'obtenir toutes les cibles possibles. Par conséquent, un expert clientèle peut construire une cible à sa guise à l'aide des variables greffées à partir d'une enquête au moyen de l'approche par fusion statistique de données, dans sa base de facturation EDF. Et surtout, il peut évaluer la qualité d'un mailing à

[8] A condition qu'elles représentent au moins 10% des enquêtés, il y en 4 autres.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

l'aide des niveaux de confiance individuels et sélectionner le nombre de clients qu'il désire au moyen de la courbe de lift.

De plus, un expert clientèle dispose, à partir de la courbe de lift, d'un autre outil d'aide à la décision l'indice de Gini, que nous avons développé au chapitre IV. On vient de voir qu'il peut à sa guise se donner un seuil « a priori » pour un client d'être dans sa cible et sélectionner en conséquence les individus sur la courbe de lift. Mais il peut également utiliser comme critère global de qualité de son modèle, l'ensemble de la courbe de lift. Ainsi l'indice de Gini, noté G est une mesure globale de la qualité permettant de comparer des modèles entre eux sur l'ensemble de leur courbe de lift.

Rappel des notations vues au chapitre IV :

I_{est} représente l'apport du modèle estimé (aire sous la courbe de lift).
I_{max} fournit la mesure maximale obtenue par le modèle parfait.
I_{nul} est égale à la mesure obtenue sans modèle.

Alors :

$$G = \frac{I_{est} - I_{nul}}{I_{\max} - I_{nul}}$$

G varie entre –1 et +1, avec G=1 dans le cas idéal.

Le calcul de cet indice de Gini a été effectué pour les questions du thème « Ce que vous pensez de votre chauffage électrique ». Chaque question a, tour à tour, été choisie comme variable cible et son indice de Gini a été calculé à partir de sa courbe de lift correspondante. Les résultats sont présentés dans le tableau suivant.

Question	Indice de Gini SEQ	Indice de Gini MUL
Satisfaction (Q6)	0,3487	0,2678
Choix du système (Q7)	0,3914	0,3739
Conseil du CE (Q8)	0,2934	0,2954

Tableau 25 : Indices de Gini

Les indices de Gini, obtenus sur ce thème de questions, sont supérieurs dans le cas de la méthode séquentielle. L'expert clientèle pourra donc en conclure que s'il choisit ces questions comme variable cible sur sa clientèle, la méthode séquentielle est le modèle lui offrant la meilleure qualité globale de résultats selon cet indice.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.gs2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Chapitre VI :

Conclusion et perspectives

Nous avons répertorié dans un premier temps les principales techniques statistiques permettant de traiter du problème d'apparition de données manquantes dans un fichier de données. Ces méthodes se révèlent intéressantes par rapport à la problématique de la fusion statistique dans le sens où cette dernière peut être considérée comme un cas particulier de traitement de données manquantes apparaissant par bloc. En effet qu'il s'agisse de modèles implicites ou explicites, les techniques statistiques utilisées pour l'estimation de données manquantes peuvent être réutilisables pour un problème approprié de fusion de données.

Nous avons ensuite vu qu'il existe déjà un certain nombre de techniques de fusion statistique ou d'appariement statistique que l'on peut plus généralement regrouper en deux grandes catégories : la fusion par imputation à un individu de l'ensemble de l'information relative à un autre individu et la fusion par estimation puis fusion des variables d'intérêt d'un fichier à un autre à l'aide de variables communes candidates à l'explication. Quel que soit le problème rencontré, le statisticien aura dans tous les cas, tout intérêt à privilégier une technique de fusion statistique parmi d'autres selon le type de données dont il dispose (variables d'intérêt et variables candidates à l'explication quantitatives ou qualitatives, numériques, ordinales ou nominales) et selon l'objectif préalablement défini de la fusion statistique effectuée. En effet selon le domaine d'études considéré, ciblage marketing, enquête d'opinion ou autre, la finalité et donc le critère de qualité d'une fusion statistique peut varier. Certains s'attacheront à obtenir des résultats dans le fichier « receveur » ayant un bon taux de confiance associé au niveau individuel tandis que d'autres privilégieront des indices plus globaux comme par exemple la préservation des distributions marginales, la moyenne ou la variance des variables d'intérêt de l'étude. Dans le cas de l'étude de plusieurs variables d'intérêt fortement corrélées, l'un des objectifs pourrait également être la préservation après fusion de la structure de corrélation existante. Dans tous les cas, les limites actuelles des techniques de fusion statistique imposent au statisticien de choisir une méthode offrant le meilleur compromis possible entre les différents objectifs de son étude (conservation de la variance versus taux d'individus bien classés, etc...).

Dans un second temps, nous avons concentré notre étude sur la problématique de la fusion statistique pour des données qualitatives. Nous avons ainsi proposé plusieurs nouvelles techniques dans le cadre d'une fusion par modélisation des variables d'intérêt sur un fichier dit « donneur ». Ces méthodes ont ensuite été testées sur un exemple de fusion statistique entre deux fichiers d'enquêtes. Deux méthodes (la méthode dite séquentielle et la méthode dite multivariée basée sur le codage Logit et la régression

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

PLS2) ont ainsi été retenues puis expérimentées sur des bases de données clients en vue d'un ciblage de clientèle.
De plus, les données obtenues après fusion sont des données estimées et non observées ; ainsi elles ne peuvent être considérées par la suite comme des observations réelles. C'est pourquoi il est indispensable, en vue de leur utilisation, de leur adjoindre un ou plusieurs indicateurs de qualité proposant une validité statistique du modèle de fusion utilisée. La validation ne pouvant s'effectuer que de manière empirique sur le fichier donneur, ce dernier est découpée en k parties, comme k fichiers receveurs, avant de procéder à une validation croisée. Nous avons ainsi proposé de résumer cette validation statistique de la qualité d'une fusion à trois critères globaux :

- la préservation des distributions marginales des variables d'intérêt
- la préservation des distributions croisées des variables d'intérêt
- le taux d'individus bien-classés après fusion

et un critère individuel :

- le taux d'individus bien classés dans le groupe considéré.

Les résultats obtenus via la démarche de validation statistique proposée dans le cadre de l'expérimentation, avec les méthodes séquentielle et multivariée sont tout à fait encourageants et permettent d'exploiter les résultats dans le cadre d'éventuelles études de ciblage. En effet, afin de souligner l'intérêt et l'apport de la fusion statistique nous avons développé puis mis en œuvre, sur une étude, une démarche de validation opérationnelle de celle-ci. En s'appuyant sur un indicateur de qualité qu'est la courbe lift, nous montrons que la fusion statistique de données présente bien des intérêts comparée à un ciblage direct. D'un point de vue statistique et opérationnel les courbes de lift obtenues sont en effet meilleures. Egalement d'un point de vue financier le coût de la fusion statistique peut s'avérer bien moindre ; si un expert clientèle désire travailler sur plusieurs cibles, il lui faut refaire dans le cas d'un ciblage direct un modèle à chaque fois, ce qui n'est pas du tout performant.

Le domaine de la fusion statistique de données est un sujet de recherche encore relativement jeune, une trentaine d'années au plus et offre de nombreuses perspectives de développement. Dans la continuité du travail que nous venons de présenter on peut envisager d'approfondir les pistes suivantes :

1. Mener des comparaisons avec d'autres méthodes de fusion statistique, notamment une méthode du type imputation par plus proches voisins adaptée à des données qualitatives.

2. Poursuivre le développement d'une méthode de fusion statistique multivariée prenant en compte la nature qualitative des données traitées en lieu et place du recodage logit suivi de la régression PLS2. C' est ce qu'a entrepris C. Derquenne en proposant une méthode dite « PML » (Partiel Maximum Likelyhood). Il propose ainsi un algorithme alternatif à la régression PLS2 (celle-ci n'étant utilisable qu'en présence de variables quantitatives) basée sur le maximum de vraisemblance dans le cas de

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

données qualitatives. Les résultats obtenus avec la même démarche de validation statistique sont très encourageants.

3. Améliorer et étendre le processus de validation statistique à tous les types de données.

4. Etudier précisément et standardiser les conditions nécessaires pour réaliser une fusion à savoir :
   - Du point de vue de la population : le fichier dit « donneur » doit avoir suffisamment d'unités statistiques représentatives dans chaque sous-groupe de population contenu dans le fichier receveur. Si nécessaire, effectuer un redressement sur le fichier donneur.
   - Du point de vue des variables : la nécessité d'une forte liaison entre les variables d'intérêt et les variables prédictives. Ceci est absolument nécessaire pour obtenir une qualité satisfaisante des données reconstituées.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Annexe 1
Table V13

	THISYR	LASTYR	AGE	EDUC	CHILD1	CHILD2
1	0	1	27	10	0	0
2	1	0	35	12	0	0
3	1	1	40	12	0	0
4	1	1	35	12	0	0
5	1	0	28	10	0	1
6	1	1	30	10	0	1
7	1	1	35	10	0	1
8	1	1	46	7	0	0
9	1	0	32	10	0	0
10	1	1	30	10	0	0
11	1	1	45	12	0	1
12	1	1	33	16	0	0
13	1	1	39	12	0	0
14	1	1	22	12	0	1
15	1	0	27	12	0	0
16	0	0	23	10	0	1
17	0	1	25	12	0	1
18	0	0	43	12	0	1
19	1	1	38	12	0	0
20	0	0	40	12	0	1
21	1	0	44	12	0	0
22	0	0	31	12	0	1
23	1	1	42	12	0	0
24	0	0	24	12	0	0
25	1	1	42	12	0	0
26	1	1	31	12	0	0
27	1	1	38	12	0	0
28	0	0	43	12	0	0
29	0	1	24	12	0	1
30	0	0	29	12	0	0
31	0	0	31	12	0	0
32	1	0	37	12	0	0
33	1	1	42	12	0	0
34	1	1	34	12	0	0
35	0	1	25	10	0	1
36	0	0	25	10	1	0
37	1	1	31	12	0	0
38	0	0	35	12	1	0
39	0	0	40	12	0	1
40	0	0	26	12	0	1
41	1	1	34	12	0	0
42	1	1	39	12	0	0
43	1	1	43	12	0	0
44	0	0	23	12	0	1
45	1	1	40	12	0	0
46	0	0	36	12	0	0
47	1	0	25	12	0	1
48	1	1	30	12	0	1
49	1	1	27	12	0	0
50	1	1	32	12	0	1
51	1	1	37	12	0	0
52	1	1	37	10	0	0
53	0	0	27	12	0	1
54	0	1	35	10	0	1
55	0	0	39	12	0	1
56	1	1	33	16	0	0
57	1	1	46	12	0	0
58	0	0	32	7	0	0
59	0	0	28	12	0	1
60	0	0	33	12	0	1
61	1	1	38	12	0	0
62	1	1	40	12	0	0
63	1	1	40	12	0	1
64	1	1	46	12	0	1
65	0	0	46	12	0	0
66	1	1	40	12	0	0
67	0	0	26	12	0	1
68	1	1	21	12	0	0
69	0	0	26	12	0	1
70	1	1	27	16	0	0
71	1	1	44	12	0	0
72	0	0	36	12	0	0
73	1	1	39	10	0	0
74	1	1	45	10	0	0
75	0	0	31	12	0	0
76	1	1	35	12	0	1
77	1	1	40	12	0	0
78	0	0	43	12	0	0
79	0	1	24	12	1	0
80	0	0	34	12	0	1
81	1	0	25	12	0	1
82	1	1	30	12	0	1
83	1	1	43	7	0	1
84	0	0	30	12	0	1
85	0	0	45	12	0	0
86	1	1	39	12	0	0
87	0	0	41	12	0	0
88	1	1	27	12	0	1
89	1	1	32	12	0	0
90	1	1	33	18	0	0
91	1	1	31	12	0	0
92	1	1	35	12	0	1
93	1	1	23	12	1	0
94	1	1	29	12	0	1
95	1	1	32	12	0	1
96	0	1	44	12	0	0
97	1	0	34	12	0	0
98	1	0	40	12	0	0
99	0	0	46	12	0	0
100	1	1	36	12	0	0
101	0	0	42	12	0	0
102	0	0	31	12	0	1
103	1	1	38	10	0	1
104	1	0	46	12	0	0
105	1	0	46	18	0	0
106	1	0	29	10	0	1
107	1	1	25	12	0	1
108	0	0	32	12	0	1
109	0	1	26	12	0	0
110	1	1	29	12	0	0
111	1	1	34	12	0	0
112	1	0	38	16	0	0
113	1	1	32	12	0	0
114	1	1	36	12	0	1
115	1	1	28	18	1	0

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

116	1	1	37	12	0	0
117	1	1	38	7	0	0
118	1	1	43	7	0	0
119	0	0	23	12	0	1
120	0	1	24	12	0	0
121	1	1	40	12	0	0
122	1	1	46	12	0	0
123	1	0	42	12	0	0
124	1	1	46	12	0	0
125	0	0	42	12	0	0
126	1	1	31	12	0	0
127	0	0	39	12	0	1
128	1	0	46	12	0	1
129	0	0	28	16	1	0
130	0	0	23	12	1	0
131	1	1	32	12	0	0
132	1	1	40	12	0	0
133	1	1	45	12	0	0
134	0	1	31	12	0	1
135	1	1	46	7	0	0
136	1	1	38	12	0	0
137	1	1	29	10	0	1
138	0	0	44	12	0	0
139	1	1	31	18	0	0
140	1	1	35	18	0	0
141	1	1	24	12	0	0
142	0	0	34	12	0	0
143	1	0	28	16	0	1
144	0	0	27	12	0	1
145	1	1	46	12	0	0
146	1	1	27	12	0	1
147	1	1	32	12	0	0
148	1	1	38	16	0	0
149	1	0	26	16	0	1
150	0	0	31	16	1	0
151	0	0	30	12	0	1
152	1	1	42	12	0	0
153	1	1	46	18	0	0
154	1	1	44	12	0	0
155	1	1	31	12	0	0
156	1	1	36	12	0	1
157	1	0	45	12	0	0
158	1	1	42	16	0	0
159	0	0	29	12	1	0
160	1	1	43	10	0	0
161	1	1	27	16	0	0
162	1	1	40	12	0	0
163	1	1	30	12	0	0
164	0	1	29	10	0	1
165	1	1	39	12	0	0
166	1	1	30	12	0	1
167	1	1	35	12	0	0
168	0	0	40	12	0	0
169	0	0	46	12	0	0
170	1	1	36	12	0	1
171	1	1	28	12	0	1
172	1	1	24	12	0	0
173	0	1	34	12	0	0
174	0	1	33	12	0	0
175	1	1	26	10	0	0
176	1	1	35	12	0	0
177	1	1	40	12	0	0
178	1	1	42	12	0	0
179	1	0	34	12	0	0
180	1	0	45	12	0	0
181	1	1	40	10	0	0
182	1	1	45	10	0	0
183	1	1	41	12	0	0
184	1	0	44	12	0	0
185	1	1	24	12	0	0
186	1	1	42	12	0	0
187	1	1	46	12	0	0
188	1	1	46	10	0	0
189	1	0	32	12	0	0
190	0	0	45	12	0	0
191	1	1	31	12	0	1
192	0	0	32	12	1	0
193	1	0	42	12	0	0
194	0	0	32	12	0	1
195	1	0	40	12	0	0
196	1	1	45	12	0	0
197	0	0	35	12	0	1
198	0	1	31	10	0	0
199	0	1	31	12	0	1
200	1	1	25	10	0	0

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.gs2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Annexe 2

Tableau 1 : Validation des résultats sur la variable Gene_sat

	Greffe univariée	Rec-Logit PLS2	Recodage0-1PLS2	Logit PLS1	Pseudo-PLS2	Logit Classique
Reconstitution des marginales	19.642 2 5.4E-5 5.534 2 0.0628 0.017 2 0.9914 2.598 2 0.2728 0.972 2 0.6152	169.47 2 0 113.29 2 0 249.55 2 0 162.71 2 0 179.40 2 0	291.621 2 0 457.424 2 0 588.613 2 0 467.238 2 0 577.37 2 0	1344.46 2 0 1451.16 2 0 1511.01 2 0 1539.67 2 0 1322.30 2 0	1534.58 2 0 1465.56 2 0 1600.71 2 0 1574.11 2 0 1334.67 2 0	363.01 2 0 311.16 2 0 396.31 2 0 389.86 2 0 366.12 2 0
Corrélations	116.56 156.52 230.71 235.46 136.12 102.57 180.58 315.49	395.88 404.77 333.52 454.47 323.22 319.30 379.39 1525.40	769.16 1187.26 518.98 799.44 813.05 606.18 866.17 818.38	2124.91 3226.56 1795.90 2418.60 2163.56 1776.83 2810.13 2707.98	2133.50 3076.07 1883.85 2525.09 2175.03 1826.41 2162.99 2778.49	524.70 899.71 480.35 590.18 540.73 448.38 767.29 799.65
Taux de bien-classés	0.74657 ++ [0.52722;0.58039] 0.78254 ++ [0.55300;0.60583] 0.76763 ++ [0.54870;0.60150] 0.79076 ++ [0.52966;0.58258] 0.71923 ++ [0.51577;0.56762]	0.53331 -- [0.56138;0.61211] O.54763 -- [0.5606;0.6096] 0.52013 -- [0.5489;0.5950] 0.52667 -- [0.5379;0.5881] 0.54022 0 [0.4926;0.5417]	0.61097 0 [0.56238 ;0.61112] 0.55605 -- [0.56152 ;0.60866] 0.58737 0 [0.54923 ;0.59464] 0.57363 0 [0.53829 ;0.58769] 0.53585 0 [0.49427 ;0.54001]	0.5526 0 [0.5442;0.5789] 0.5602 0 [0.5436;0.5741] 0.5708 0 [0.5486;0.5753] 0.5737 0 [0.5490;0.5754] 0.5853 ++ [0.5528;0.5794]	0.5678 0 [0.5442;0.5789] 0.5721 0 [0.5436;0.5741] 0.5647 0 [0.5486;0.5753] 0.5716 0 [0.5490;0.5754] 0.5844 ++ [0.5528;0.5794]	0.5977 ++ [0.5272;0.5804] 0.5403 -- [0.5530;0.6058] 0.5919 0 [0.5487;0.6015] 0.5661 0 [0.5297;0.5826] 0.5491 0 [0.5157;0.5676]

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.os2.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Tableau 2 : Validation des résultats sur la variable Nivt_sat

	Greffe univariée	Rec-Logit PLS2	Recodage0-1PLS2	Logit PLS1	Pseudo-PLS2	Logit Classique
Reconstitution des marginales	27.109 2 1.3E-6 91.057 2 0 47.453 2 0 86.715 2 0 72.807 2 0	482.50 2 0 160.72 2 0 357.04 2 0 82.27 2 0 345.03 2 0	438.130 2 0 487.516 2 0 782.919 2 0 628.759 2 0 729.50 2 0	1325.89 2 0 1222.73 2 0 1533.33 2 0 1576.78 2 0 1549.40 2 0	1498.46 2 0 1567.86 2 0 1565.03 2 0 1581.23 2 0 1563.60 2 0	258.95 2 0 258.35 2 0 422.17 2 0 364.48 2 0 525.34 2 0
Corrélations	116.56 234.61 262.50 355.68 245.55 167.53 248.57 386.91	395.88 459.16 399.42 556.80 423.45 394.47 456.75 1572.22	769.16 1279.41 641.12 881.11 927.89 716.574 949.08 940.47	2124.91 3085.59 1834.21 2498.45 2256.83 1912.67 2789.82 2902.33	2133.50 3119.85 1869.25 2524.51 2269.63 1914.84 2187.24 2898.08	524.70 907.85 460.44 588.78 549.96 464.45 770.99 767.16
Taux de bien-classés	0.63058 ++ [0.51747;0.57075] 0.66453 ++ [0.52087;0.57413] 0.66055 ++ [0.48773;0.54111] 0.65020 ++ [0.48878;0.54201] 0.61495 ++ [0.43816;0.49005]	0.48325 -- [0.5129;0.5644] 0.46375 -- [0.5375;0.5868] 0.49644 -- (0.4945;0.5411] 0.45850 -- (0.4995;0.5450] 0.47687 0 [0.4336;0.4825]	0.55183 0 [0.51397 ;0.56331] 0.52702 -- [0.53844 ;0.58591] 0.53260 0 [0.49486 ;0.54072] 0.53766 0 [0.49990 ;0.54963] 0.48283 ++ [0.43528 ;0.48087]	0.5245 0 [0.5076;0.5295] 0.5204 0 [0.5063;0.5374] 0.5277 0 [0.5062;0.5330] 0.5146 0 [0.5035;0.5301] 0.5309 0 [0.5169;0.5437]	0.5234 0 [0.5076;0.5295] 0.5327 0 [0.5063;0.5374] 0.5313 0 [0.5062;0.5330] 0.5154 0 [0.5035;0.5301] 0.5349 0 [0.5169;0.5437]	0.5463 0 [0.5174;0.5708] 0.4772 -- [0.5208;0.5741] 0.5079 0 [0.4877;0.5411] 0.5241 0 [0.4887;0.5420] 0.4738 0 [0.4381;0.4901]

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.cs2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Tableau 3 : Validation des résultats sur la variable Esth_sat

	Greffe univariée	Rec-Logit PLS2	Recodage0-1PLS2	Logit PLS1	Pseudo-PLS2	Logit classique
Reconstitution des marginales	119.94 2 0 124.416 2 0 121.151 2 0 149.009 2 0 87.841 2 0	337.43 2 0 249.29 2 0 114.09 2 0 46.22 2 9.2^{E}-11 255.42 2 0	881.220 2 0 957.382 2 0 912.432 2 0 725.525 2 0 1035.94 2 0	1678.57 2 0 1475.39 2 0 1517.22 2 0 2572.87 2 0 2816.48 2 0	1579.23 2 0 1467.89 2 0 1501.10 2 0 2446.57 2 0 2910.58 2 0	797.33 2 0 620.87 2 0 573.19 2 0 679.79 2 0 702.61 2 0
Corrélations	156.52 234.61 288.18 347.55 241.04 232.57 222.31 412.45	404.77 459.16 363.02 388.53 329.96 349.69 353.60 1555.71	1187.26 1279.41 923.16 1144.38 1007.49 981.92 1131.77 1211.05	3226.56 3085.59 2567.21 2872.73 2765.38 2674.09 2988.68 3276.03	2133.50 3119.85 2489.03 2960.60 2672.03 2487.60 3011.13 3309.29	524.70 907.85 765.64 867.12 838.26 762.23 922.96 996.02
Taux de bien-classés	0.59356 ++ [0.47173;0.52521] 0.61735 ++ [0.48766;0.54114] 0.59313 ++ [0.50131;0.55463] 0.57664 ++ [0.49876;0.55196] 0.60177 ++ [0.46432;0.51634]	0.48302 0 [0.4482;0.4997] 0.41793 -- [0.4511;0.5008] O.46473 -- [0.4952;0.5412] 0.47869 -- [0.5269;0.5772] 0.46318 0 [0.4467;0.4957]	0.46966 0 [0.44920 ;0.49863] 0.47959 0 [0.45206 ;0.49985] 0.51880 0 [0.49557 ;0.54143] 0.55147 0 [0.52731 ;0.57683] 0.47114 0 [0.44834 ;0.49403]	0.5127 0 [0.4965;0.5272] 0.5353 ++ [0.4878;0.5184] 0.5202 ++ [0.4787;0.5055] 0.5064 0 [0.4942;0.5209] 0.5105 0 [0.5032;0.5299]	0.5245 0 [0.4965;0.5272] 0.5198 ++ [0.4878;0.5184] 0.5313 0 [0.4787;0.5055] 0.5154 0 [0.4942;0.5209] 0.5141 0 [0.5032;0.5300]	0.5020 0 [0.4717;0.5252] 0.4963 0 [0.4876;0.5412] 0.5136 0 [0.5013;0.5547] 0.5118 0 [0.4987;0.5520] 0.4898 0 [0.4643;0.5164]

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.cs2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Tableau 4 : Validation des résultats sur la variable Regl_sat

	Greffe univariée	Rec-Logit PLS2	Recodage0-1PLS2	Logit PLS1	Pseudo-PLS2	Logit classique
Reconstitution des marginales	123.266 2 0 96.426 2 0 138.335 2 0 231.765 2 0 257.185 2 0	43.10 2 4.4^{E}-10 151.09 2 0 55.27 2 0 61.94 2 3.5^{E}-14 131.73 2 0	138.306 2 0 58.173 2 2.3^{E}-13 29.837 2 3.2^{E}-7 166.458 2 0 52.38 2 4.2^{E}-12	567.83 2 0 461.74 2 0 654.42 2 0 648.84 2 0 600.12 2 0	567.99 2 0 493.17 2 0 468.81 2 0 701.32 2 0 596.05 2 0	123.27 2 0 96.43 2 0 138.36 2 0 231.77 2 0 257.19 2 0
Corrélations	230.71 262.50 288.18 369.60 322.49 285.54 257.39 440.58	333.52 399.42 363.02 272.28 202.24 208.28 333.67 1542.80	518.98 641.12 923.16 533.76 622.66 358.94 668.44 669.26	1795.90 1834.21 2567.21 1756.93 1452.78 1098.61 2144.17 2449.49	1883.85 1869.25 2489.03 1797.91 1461.06 1140.31 1977.22 2421.11	480.35 460.44 765.64 458.12 406.05 299.72 655.48 693.32
Taux de bien-classés	0.44578 0 [0.40770;0.46072] 0.45371 0 [0.42266;0.47589] 0.43268 0 [0.38558;0.43815] 0.46742 ++ [0.36945;0.42154] 0.44351 ++ [0.35982;0.41046]	0.45813 ++ [0.3908;0.4417] 0.42443 -- [0.4303;0.4798] 0.48132 0 [0.4675;0.5141] 0.37062 0 [0.3621;0.4113] 0.42167 0 [0.3793;0.4275]	0.51628 ++ [0.39187 ;0.44066] 0.46883 0 [0.43122 ;0.47887] 0.40324 -- [0.46787 ;0.51375] 0.41263 ++ [0.36246 ;0.41095] 0.47312 ++ [0.38095 ;0.42585]	0.4856 ++ [0.4049;0.4358] 0.4918 ++ [0.4125;0.4389] 0.4922 ++ [0.4024;0.4289] 0.4654 ++ [0.4065;0.4328] 0.4624 ++ [0.4094;0.4359]	0.4612 ++ [0.4049;0.4358] 0.4735 ++ [0.4125;0.4389] 0.4756 ++ [0.4024;0.4289] 0.4548 ++ [0.4065;0.4328] 0.4605 ++ [0.4094;0.4359]	0.4458 0 [0.4077;0.4608] 0.4537 0 [0.4226;0.4759] 0.4327 0 [0.3855;0.4382] 0.4674 ++ [0.3694;0.4216] 0.4435 ++ [0.3598;0.4105]

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.os2.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Tableau 5 : Validation des résultats sur la variable Rapi_sat

	Greffe univariée	Rec-Logit PLS2	Recodage0-1PLS2	Logit PLS1	Pseudo-PLS2	Logit classique
Reconstitution des marginales	165.908 2 0 214.173 2 0 89.181 2 0 231.801 2 0 272.468 2 0	178.47 2 0 263.53 2 0 166.58 2 0 90.63 2 0 286.83 2 0	454.406 2 0 424.892 2 0 418.907 2 0 432.476 2 0 624.54 2 0	1467.79 2 0 1238.58 2 0 1505.67 2 0 1337.65 2 0 1449.65 2 0	1335.89 2 0 1412.87 2 0 1372.99 2 0 1348.71 2 0 1437.82 2 0	190.38 2 0 255.54 2 0 262.87 2 0 388.14 2 0 487.52 2 0
Corrélations	235.46 355.68 347.55 369.60 302.50 272.67 362.77 504.89	454.47 556.80 388.53 272.28 314.19 320.37 462.32 1584.32	799.44 881.11 1144.38 533.76 817.61 610.06 856.13 875.75	2418.60 2498.45 2872.73 1756.93 2178.24 1805.19 2804.95 2936.37	2525.09 2524.51 2960.60 1797.91 2125.84 1819.17 2817.13 2928.98	590.18 588.78 765.64 458.12 592.27 419.05 738.25 781.57
Taux de bien-classés	0.64883 ++ [0.50729;0.56065] 0.64221 ++ [0.46754;0.52104] 0.67771 ++ [0.46144;0.51483] 0.64950 ++ [0.43292;0.48601] 0.61024 ++ [0.40707;0.45863]	0.49255 -- [0.5124;0.5638] 0.44460 -- (0.4568;0.5064] 0.53451 0 [0.4996;0.5462] 0.43973 -- [0.4735;0.5241] 0.42047 -- [0.4318;0.4808]	0.54615 0 [0.51343 ;0.56278] 0.44133 -- [0.45771 ;0.50552] 0.49598 -- [0.49998 ;0.54582] 0.49687 0 [0.47388 ;0.52367] 0.50961 ++ [0.43351 ;0.47910]	0.5132 ++ [0.4835;0.5129] 0.5231 ++ [0.4879;0.5186] 0.5219 ++ (0.4791;0.5059] 0.5009 ++ [0.4732;0.4999] 0.4992 0 [0.4805;0.5073]	0.5145 ++ [0.4835;0.5129] 0.5096 0 [0.4879;0.5186] 0.5198 ++ (0.4791;0.5059] 0.5020 ++ [0.4732;0.4999] 0.5002 0 [0.4805;0.5073]	0.5271 0 [0.5072;0.5607] 0.4529 -- [0.4675;0.5211] 0.4816 0 [0.4614;0.5149] 0.4758 0 [0.4329;0.4860] 0.4844 ++ [0.4070;0.4587]

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.os2.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Tableau 6 : Résultats sur la variable Empl_sat

	Greffe univariée	Rec-Logit PLS2	Recodage0-1PLS2	Logit PLS1	Pseudo-PLS2	Logit classique
Reconstitution des marginales	84.025 2 0 50.537 2 0 43.958 2 0 117.117 2 0 80.822 2 0	30.69 2 2.2E-7 90.63 2 0 4.68 2 0.0962 128.70 2 0 66.60 2 3.4E-15	216.803 2 0 312.031 2 0 259.276 2 0 201.561 2 0 279.75 2 0	595.74 2 0 623.45 2 0 508.22 2 0 643.54 2 0 666.29 2 0	576.89 2 0 546.92 2 0 505.96 2 0 633.04 2 0 672.35 2 0	168.80 2 0 244.23 2 0 73.14 2 1.1E-16 175.34 2 0 174.23 2 0
Corrélations	136.12 245.55 241.04 322.49 302.50 134.57 267.50 396.27	323.22 423.45 329.96 202.24 314.19 206.71 309.39 1560.99	813.05 927.89 1007.49 622.66 817.61 382.91 854.17 796.68	2163.56 2256.83 2765.38 1452.78 2178.24 829.78 2134.91 2522.07	2175.03 2269.63 2672.03 1461.06 2125.84 843.39 1898.69 2586.26	540.18 549.96 838.26 406.05 592.27 233.05 702.97 699.98
Taux de bien-classés	0.72772 ++ [0.53725;0.59030] 0.77944 ++ [0.51500;0.56832] 0.75514 ++ [0.58095;0.63312] 0.73355 ++ [0.56530;0.61767] 0.72708 ++ [0.53638;0.58801]	0.54448 0 [0.5395;0.5907] 0.54231 ++ [0.4685;0.5183] 0.52758 0 [0.5124;0.5589] 0.53749 0 [0.5223;0.5727] 0.56470 0 [0.5174;0.5663]	0.59477 ++ [0.54054 ;0.58961] 0.51734 ++ [0.46948 ;0.51732] 0.51570 0 [0.51277 ;0.55854] 0.58081 ++ [0.52274 ;0.57231] 0.55311 0 [0.51902 ;0.56462]	0.5845 ++ [0.5372;0.5712] 0.5783 ++ [0.5338;0.5603] 0.5716 ++ [0.5232;0.5499] 0.5971 ++ [0.5554;0.5818] 0.5881 0 [0.5628;0.5893]	0.5856 ++ [0.5372;0.5712] 0.5738 ++ [0.5338;0.5603] 0.5714 ++ [0.5232;0.5490] 0.5898 ++ [0.5554;0.5818] 0.5872 ++ [0.5628;0.5893]	0.5443 0 [0.5372;0.5903] 0.5682 0 [0.5150;0.5684] 0.6026 0 [0.5809;0.6332] 0.5943 0 [0.5653;0.6177] 0.6061 ++ [0.5363;0.5880]

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.os2.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Tableau 7 : Résultats sur la variable Sécu_sat

	Greffe univariée	Rec-Logit PLS2	Recodage0-1PLS2	Logit PLS1	Pseudo-PLS2	Logit classique
Reconstitution des marginales	78.107 2 0 44.318 2 0 70.217 2 0 63.848 2 0 79.209 2 0	99.47 2 0 109.76 2 0 100.94 2 0 113.13 2 0 190.85 2 0	77.774 2 0 104.041 2 0 110.023 2 0 85.269 2 0 166.95 2 0	402.27 2 0 431.29 2 0 420.26 2 0 354.63 2 0 339.34 2 0	425.87 2 0 384.18 2 0 417.80 2 0 359.01 2 0 337.62 2 0	126.87 2 0 93.46 2 0 81.91 2 0 80.02 2 0 101.96 2 0
Corrélations	102.57 167.53 232.57 285.54 272.67 134.57 218.73 361.47	319.30 394.47 349.69 208.28 320.37 206.71 352.34 1570.50	606.18 716.57 981.92 358.94 610.06 382.91 720.49 738.15	1776.83 1912.67 2674.09 1098.61 1805.19 829.78 2151.37 2304.87	1826.41 1914.84 2487.60 1140.31 1819.17 843.39 2213.23 2527.04	448.38 464.45 762.23 299.72 419.05 233.05 636.03 636.89
Taux de bien-classés	0.75943 ++ [0.44956;0.50299] 0.77168 ++ [0.45262;0.50608] 0.73522 ++ [0.49748;0.55082] 0.75498 ++ [0.48636;0.53961] 0.75699 ++ [0.48141;0.53343]	0.53386 ++ [0.4601;0.5117] 0.53944 ++ [0.4861;0.5358] 0.49966 ++ [0.4433;0.4898] 0.48342 0 [0.4743;0.5249] 0.49225 ++ [0.4429;0.4919]	0.54093 ++ [0.46116 ;0.51063] 0.50954 0 [0.48702 ;0.53485] 0.49482 ++ [0.44369 ;0.48947] 0.54827 ++ [0.47468 ;0.52447] 0.53103 ++ [0.44461 ;0.49028]	0.5384 ++ [0.4687;0.4939] 0.5314 ++ [0.4641;0.4907] 0.5278 ++ [0.4531;0.4799] 0.5448 ++ [0.4794;0.5060] 0.5465 ++ [0.4805;0.5073]	0.5317 ++ [0.4687;0.4939] 0.5296 ++ [0.4641;0.4907] 0.5282 ++ (0.4531;0.4799] 0.5482 ++ (0.4794;0.5060] 0.5467 ++ [0.4805;0.5073]	0.4822 0 [0.4495;0.5030] 0.5282 ++ [0.4526;0.5061] 0.5326 0 [0.4974;0.5509] 0.5257 0 [0.4863;0.5397] 0.5375 ++ [0.4814;0.5335]

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.cs2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Tableau 8 : Résultats sur la variable Humi_sat

	Greffe univariée	Rec-Logit PLS2	Recodage0-1PLS2	Logit-PLS1	Pseudo-PLS2	Logit classique
Reconstitution des marginales	106.75 2 0 79.773 2 0 122.837 2 0 36.114 2 0 133.842 2 0	338.30 2 0 178.32 2 0 206.08 2 0 137.24 2 0 166.54 2 0	552.002 2 0 576.024 2 0 547.405 2 0 505.566 2 0 757.84 2 0	1678.39 2 0 1429.02 2 0 1546.81 2 0 2070.05 2 0 2010.56 2 0	1879.56 2 0 1987.45 2 0 1611.46 2 0 2138.69 2 0 2038.90 2 0	580.79 2 0 481.58 2 0 532.73 2 0 501.37 2 0 636.70 2 0
Corrélations	180.58 248.57 222.31 257.39 362.77 267.50 218.73 396.13	379.39 456.75 353.60 333.67 462.32 309.39 352.34 1564.15	866.17 949.08 1131.77 668.44 856.13 854.17 720.49 1008.36	2810.13 2789.82 2988.68 2144.17 2804.95 2134.91 2151.37 3051.38	2162.99 2187.24 3011.13 1977.22 2817.13 1898.69 2213.23 3073.03	767.29 770.99 922.96 655.48 738.25 702.97 636.03 879.80
Taux de bien-classés	0.60608 ++ [0.48405;0.53752] 0.61375 ++ [0.52107;0.57433] 0.57022 ++ [0.49445;0.54781] 0.57707 ++ [0.51817;0.57122] 0.56931 ++ [0.46818;0.52020]	0.49118 -- [0.4923;0.5439] 0.50236 -- (0.5182;0.5677] 0.50573 -- [0.5166;0.5631] 0.49466 0 [0.4922;0.5427] 0.47485 0 [0.4514;0.5005]	0.52412 0 [0.49338 ;0.54284] 0.53899 0 [0.51914 ;0.56680] 0.54152 0 [0.51698 ;0.56272] 0.53105 0 [0.49257 ;0.54234] 0.47614 0 [0.45311 ;0.49881]	0.5463 ++ (0.4963;0.5258] 0.5506 ++ [0.4971;0.5279] 0.5540 ++ [0.4995;0.5263] 0.5330 ++ (0.5048;0.5314] 0.5465 ++ [0.5177;0.5445]	0.5347 ++ [0.4963;0.5258] 0.5478 ++ [0.4971;0.5279] 0.5579 ++ (0.4995;0.5263] 0.5496 ++ [0.5048;0.5314] 0.5322 ++ [0.5177;0.5445]	0.5337 0 [0.4840;0.5376] 0.5545 0 [0.5210;0.5744] 0.5219 0 [0.4944;0.5479] 0.5456 0 [0.5181;0.5713] 0.4968 0 [0.4681;0.5202]

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.cs2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Tableau 9 : Résultats sur la variable Coût_sat

	Greffe univariée	Rec-Logit PLS2	Recodage0-1PLS2	Logit-PLS1	Pseudo-PLS2	Logit classique
Reconstitution des marginales	252.111 3 0 293.636 3 0 282.930 3 0 276.629 3 0 283.388 3 0	1309.88 3 0 1400.12 3 0 1691.53 3 0 1142.42 3 0 1572.12 3 0	468.595 3 0 644.147 3 0 497.428 3 0 506.760 3 0 568.43 3 0	1678.29 3 0 1734.97 3 0 2090.73 3 0 1919.07 3 0 1587.97 3 0	1978.78 3 0 1839.82 3 0 2052.59 3 0 1931.83 3 0 1540.84 3 0	472.32 3 0 640.33 3 0 512.02 3 0 613.24 3 0 494.69 3 0
Corrélations	315.49 386.91 412.45 440.58 504.89 396.27 361.47 396.13	1525.40 1572.22 1555.71 1542.80 1584.32 1560.99 1570.50 1564.15	818.38 940.47 1211.05 669.26 875.75 796.68 738.15 1008.36	2707.98 2902.33 3276.03 2449.49 2936.37 2522.07 2304.87 3051.38	2778.49 2898.08 3309.29 2421.11 2928.98 2586.26 2527.04 3073.03	799.65 767.16 996.02 693.32 781.57 699.98 636.89 879.80
Taux de bien-classés	0.56896 ++ [0.47239;0.52588] 0.54495 ++ [0.43752;0.49089] 0.58943 ++ [0.49111;0.54448] 0.54993 ++ [0.44395;0.49612] 0.56283 ++ [0.48577;0.53779]	0.22749 -- [0.5036;0.5551] 0.24454 -- [0.4532;0.5029] 0.21692 -- [0.5246;0.5709] 0.28884 -- [0.4454;0.4959] 0.19687 -- [0.5062;0.5552]	0.51466 0 [0.50465 ;0.55406] 0.46930 0 [0.45416 ;0.50196] O.57804 ++ [0.52491 ;0.57059] 0.48859 0 [0.44579 ;0.49550] 0.51068 0 [0.50790 ;0.55357]	0.5109 0 [0.5029;0.5321] 0.5144 0 [0.5056;0.5363] 0.5087 0 [0.4895;0.5164] 0.5127 ++ (0.4851;0.5118) 0.5186 ++ [0.4743;0.5011]	0.5109 0 [0.5029;0.5321] 0.5213 0 [0.5056;0.5363] 0.5157 0 (0.4895;0.5164] 0.5129 ++ (0.4851;0.5118] 0.5153 ++ (0.4743;0.5011]	0.4995 0 [0.4723;0.5259] 0.4827 0 [0.4375;0.4909] 0.5310 0 [0.4911;0.5445] 0.4995 ++ [0.4429;0.4962] 0.5025 0 (0.4857:0.5378]

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.os2.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Tableau récapitulatif

		Greffe univariée			RecodageLogitPLS2			Recodage 0-1 PLS2			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**

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.os2.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

VI. Bibliographie

Aluja-Banet T., Rius R., Juarez C. (2002) Data fusion by PLS regression. XXXIV Journées de Statistique JSBL-2002. Bruxelles, Louvain-la-Neuve.

Aluja-Banet T., Morineau A., Rius R. (1999) La greffe de fichiers et ses conditions d'application. Méthode et exemple. Enquêtes et Sondages éd. Scientifiques Brossier-Dussaix, Dunod, p.94-102.

Aluja-Banet T., Martinez-Abarca M. J. (1999) Fusion de datos de audiencia. II Seminario sobre nuevas tecnologias en la investigacion el marketing y la comunicacion, 1999, Séville.

Ardilly P. (1994) Techniques de sondage, Editions Technip.

Buuren S., Rijckevorsel J.L.A. (1991) Fast least squares imputation of missing data. Leiden psychological reports.

Comyn M. (1999) Modélisation et validation des rapprochements et fusions de fichiers d'enquêtes. Thèse de doctorat, ENST, Paris.

Co V. (1997) Méthodes statistiques et informatiques pour le traitement des données manquantes. Thèse de doctorat, CNAM, Paris.

Deming W.-E., Stephan F.-F. (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, Annals of Mathematical Statistics, 11, pp327-444

Dempster A. P., Laird N. M., Rubin D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B39.

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

Derquenne C., Fischer N. (2001) A multivariate approach to combine data from different sources: application to customers knowledge of Electricité de France, 53rd ISI Conference 2001, Seoul.

Derquenne C. (2002) Projet TAVEL : Expérimentation de la fusion statistique de données sur un Centre EGS et l'enquête de satisfaction EDF/SOFRES, Note interne EDF R&D, HE-72/2002/018/A.

Derquenne C. (1999) Une méthode pour construire un échantillon de données virtuelles à partir de plusieurs tableaux de données existants.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Combiner des données de sources différentes, recueil du Symposium 99 de Statistique Canada.

Deville J.-C. (1999) Calage sur des totaux issus de sources différentes. Combiner des données de sources différentes, recueil du Symposium 99 de Statistique Canada.

Fischer N., Derquenne C., Saporta G. (2001) A method to match data set applied to electric market, ETK-NTTS 2001, Crête.

Fischer N., Derquenne C. (2002) Projet TAVEL : Nouvelle méthode de fusion statistique de fichiers : Application à une meilleure connaissance de la clientèle EDF, Note interne EDF R&D, HE-72/2002/017/A.

Fischer N. (2002) Projet TAVEL : Etat de l'art sur la fusion statistique de fichiers de données: *pour une meilleure connaissance de la connaissance clientèle*, Note interne EDF r&D, HE-72/2002/016/A.

Furnival G.M., Wilson R.W. (1974) Regressions by leaps and bounds. Technometrics, 16, p499-511.

Gautier J.-M. (1999) Mégabases de consommateurs, sondages et statistique. Enquêtes et sondages, ed. scientifiques Brossier-Dussaix, Dunod, p.73-81.

Gourieroux Ch. (1984) Econométrie des variables qualitatives, Economica, Paris.

Grosbras J.M. (1987) Méthodes statistiques des sondages, Economica, Paris.

INSEE (1994) Appariements aléatoires de deux fichiers : budgets de famille et revenus fiscaux, Conseil Economique et Social, CES/AC, 70/6, Genève.

Jong S. (1999) Geometrical view on multivariate regression methods and models, PLS'99 Symposium, Jouy-en-Josas.

Lebart L., Morineau A., Piron M. (1995) Statistique exploratoire multidimensionnelle, Dunod, Paris.

Lejeune M. (2001) Traitements des fichiers d'enquêtes : Redressements, injections, fusions. Presses Universitaires de Grenoble.

Lejeune M. (1995) De l'usage des fusions de données dans les études de marché. In Bulletin de l'Institut international de statistique, Actes, 50ème session, Tome LVI, livraison 3 p923-935.

Little R.J.A., Rubin D.B. (1987) Statistical analysis with missing data, Wiley, New York.

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Piatetsky-Shapiro G., Steingold S. (2000) Measuring lift quality in database marketing, SIGKDD Explorations, December 2000, Vol. 2, Issue 2, pp81-85.

Rassler S., Fleischer K. (1997) Aspects concerning data fusion techniques. Discussion paper 16/1997, Nuremberg.

Rassler S. (2002) Statistical Matching, collection Lecture Notes in Statistics, Springer.

Renssen R. H. (1998) Utilisation de méthodes d'appariement statistique dans l'estimation de calage. Techniques d'enquêtes, décembre 1998. Vol. 24, n. 2, pp185-199. Statistique Canada.

Rodgers W.L. (1984) An Evaluation of Statistical Matching. Journal of business and economic statistics, January 1984 p91-102.

Rosset S., Neumann E., Eick U., Vatnik N., Idan I. (2002) Evaluation of prediction models for marketing campaigns.

Rubin D.B. (1976), Inference and missing data, Biometrika 63, p581-592.

Rubin D.B. (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of business and economic statistics, January 1986, p87-94.

Rubin D.B. (1987) Multiple imputation for nonresponse in Surveys, Wiley, New York.

Santini G. (2002) Méthode de fusion procustéenne. Traitement des données d'enquêtes. XXXIV Journées de la Statistique 2002, Bruxelles-Louvain.

Santini G. (1984) La méthode de fusion sur référentiel factoriel. Proceedings IREP, mars 1984.

Saporta G. (2002) Data fusion and data grafting, Computanional statistics and data analysis, 38, p.465-473.

Saporta G., Co V. (1997) Fusion de fichier: une nouvelle méthode basée sur l'analyse homogène. Enquêtes et sondages, ed. scientifiques Brossier-Dussaix, Dunod, p.81-94.

Saporta G. (1990) Probabilités, Statistique et Analyse des Données, Editions Technip.

Sousselier J. (1995), Un programme de fusion de fichiers, STATIRO, Paris.

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

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html

Tenenhaus M. (1998) La régression PLS, théorie et pratique, Editions Technip.

Tenenhaus M. (1994) Méthodes statistiques en gestion, Editions Dunod.

Tufféry S. (2002) Data Mining et Scoring : Bases de données et gestion de la relation client, Editions Dunod, Paris.

Wells G. A., O'Rourke K. (1999) Combiner des données dans le domaine de la recherche en santé. Combiner des données de sources différentes. Recueil du symposium de Statistique Canada, mai 1999.

Wendt F. (1984) The AG-MA model, Prodeedings of the 2nd International Symposium Media Research, Montréal 1983, Editions H. Henry, North-Holland, pp393-403.

Wold S., Albano *et al* (1983) Pattern recognition : Finding and using regularities in multivariate data; Proc IUFOST conf. «Food research and data analysis», Martens J. Ed, Applied Sciences publications. London.

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

PDF créé avec la version d'essai FinePrint pdfFactory Pro www.ns2i.fr/fineprint/pdffactory.htm
Trial version of ABC Amber PDF Merger, http://www.processtext.com/abcpdfmg.html