Données manquantes et fusion de fichiers

Nicolas Fischer
LNE/Cedric
nicolas.fischer@lne.fr

Problématique générale

L’ensemble des données avec lequel on doit travailler n’est pas toujours complet

Variables

Observations

	1	2	3	…	P
1		NA			
2					
3					
.		NA			
.					
.			NA		
N					NA

Problématique générale

Données manquantes

- Variable à expliquer
- Variable(s) explicative(s)

Différents cadres sont possibles (enquête)

- Données manquantes partielles
  - Incompréhension de la question
  - Incohérence
  - Refus de répondre à la question…
- Données manquantes totales
  - Absence de l'individu
  - Refus de répondre à l'enquête

Problématique générale

Impacts

Perte d'information non pertinente et/ou informative

Perte d'information pertinente et/ou informative

- Impact fonction du taux de NA
- Biais possible dans l'estimation de la précision et de l'exactitude

Solutions

- Ne rien faire
- Utiliser une procédure adaptée de remplacement des NA. Mener l'analyse sur les données complétées

Impacts

Analyse univariée

Variables

Observations

	1	2	3	4	5
1		NA		NA	
2					
3	NA				
4		NA		NA	NA
5					
6			NA	NA	
%NA	16,7	33,3	16,7	50,0	16,7

NA

Observations exclues de l'analyse

Impacts

Analyse multivariée

Variables

Observations

	1	2	3	4	5
1		NA		NA	
2					
3	NA				
4		NA		NA	NA
5					
6			NA	NA	

NA Observations exclues de l'analyse

Nature des données manquantes (Rubin,1976)

MCAR: manque complètement au hasard

- La probabilité qu'une observation soit manquante est indépendante de toutes les valeurs que prend l'individu qui présente cette donnée manquante
- I.e. le fait de ne pas avoir la valeur pour une variable Y est indépendant des autres variables X

Exemple

- X = sexe, Y = activité professionnelle
- La probabilité que l'activité soit NA est indépendante de celle-ci ainsi que du sexe
- La probabilité est la même pour tous les individus

Nature des données manquantes

MAR : manquant au hasard

- La probabilité qu'une observation soit manquante ne dépend pas de la valeur qu'elle prend
- I.e le fait de ne pas avoir la valeur pour une variable Y est uniquement dépendant d'autres variables X observées

Exemple

- X = sexe, Y = activité professionnelle
- La probabilité que l'activité soit NA ne dépendant que du sexe de l'individu
- La probabilité n'est pas la même pour tous les individus

Nature des données manquantes

NMAR: ne manquant pas au hasard (informative)

- La probabilité qu'une donnée soit manquante dépend de la valeur qu'elle prend (non observée)
- I.e le fait de ne pas avoir la valeur pour une variable Y est dépendant de la valeur non observée de celle-ci

Exemple

- X = sexe, Y = activité professionnelle
- La probabilité que l'activité soit NA dépend de celle-ci et éventuellement aussi de du sexe
- La probabilité n'est pas la même pour tous les individus

Commentaires

- Différence entre MAR et NMAR
  - MAR et méthode d'analyse pertinente permettent inférences correctes
- MCAR et MAR
  - On parle de processus ignorable ou non-informatif
  - Si analyse correcte, ne nécessite pas de modéliser le processus d'observation
- NMAR
  - On parle de processus non-ignorable ou informatif
  - Inférence sur la population étudiée nécessite
    - Poser des hypothèses fortes
    - Ou obtenir des informations complémentaires
  - Nécessité de modéliser le processus d'observation

Méthodes de traitement des données manquantes totales

Relance des non répondants (Grobras 1987)

- Population divisée en deux strates: les répondants et les non répondants
- On tire un sous-échantillon parmi les derniers que l'on ré-interroge
- Combinaison des estimateurs des deux strates pondérés par leur taux de sondage

Méthodes de traitement des données manquantes partielles

Analyse des données complètes

Indicateur de données manquantes

Compléter la non-réponse par une valeur plausible. Imputation simple ou multiple

- Méthodes implicites
- modèles

Analyse de données complètes

Stratégie la plus courante

Généralement imposée par les logiciels

Proportion d'observations complètes peut être faible même si pour chaque variable la probabilité qu'une donnée soit observée est grande

Résultats non biaisés si les données sont MCAR

Mais diminution de la précision

Sinon biais importants

Indicateur de données manquantes

- Suppose des données MCAR ou MAR
- Peut améliorer la précision de certains estimateurs
- Permet d'apprécier le risque de biais
  - Une interaction significative entre l'indicatrice de données manquantes et une variable explicative signale l'existence d'un problème
- Mais ne protège pas contre le risque de biais

Imputation simple

Au lieu de travailler sur les données complètes, on remplace chaque NA par une donnée prédite ou simulée

L'analyse porte sur toutes les données (observées et prédites)

Deux familles de méthodes d'imputation:

- Méthodes explicites
- Méthodes implicites

Imputation simple

Hypothèse d'un processus d'observation MAR

Produit une valeur estimée/simulée pour remplacer la valeur manquante

Les informations disponibles sur les individus qui ne fournissent qu'une réponse partielle peuvent être utilisées comme variables auxiliaires pour améliorer la qualité des valeurs imputées

Estimation par modèle de régression

Remplacement d'une valeur manquante Y par une valeur prédite Y* obtenue par régression sur les variables X

Possibilité d'ajouter un aléa à la prédiction

Estimation ponctuelle correcte

Variance sous-estimée

Estimation basée sur des modèles de régression

Une donnée manquante sur une variable Y est modélisée à partir des variables X selon un modèle de régression

- **régression simple en prenant la variable la plus corrélée.**
- **régression multiple**
- **modèle linéaire général si X est nominale et la variable à expliquer est quantitative.**
- **Analyse discriminante, ou régression logistique si Y nominal**

Maximum de vraisemblance

Estimation dans un cadre paramétrique

Littel et Rubin (1987)

Algorithme EM (espérance, maximisation)

L'étape E: espérance conditionnelle de chaque donnée manquante sachant les données observées et l'estimation des paramètres.

L'étape M calcule les estimateurs du maximum de vraisemblance des paramètres, avec les lois conditionnelles des données manquantes.

Convergence vers la valeur la plus probable de chaque donnée manquante pour l'estimation obtenue des paramètres

Deux inconvénients majeurs pour toutes ces méthodes

Risque d'incohérence: si plusieurs données manquantes sont estimées une par une et non conjointement, sans prendre en compte les corrélations

Variabilité sous-estimée: deux unités ayant les mêmes valeurs de X auront la même estimation pour la valeur manquante de Y

Maximisation de la cohérence interne, ou de l'homogénéité

Présentation hollandaise de l'ACM de G=(G1|G2|…|Gm) **comme la minimisation d 'une fonction de perte:**

$$\sigma(X, Y) = \frac{1}{m}\sum_{j=1}^{m}(X - G Y_j)(X_j' - G' Y_j)$$

$$X = \frac{1}{m}\sum_{j=1}^{m} G_j Y_j$$

Maximisation de la cohérence interne, ou de l'homogénéité

Les données manquantes sont complétées pour avoir σ minimal: ACM avec valeurs propres maximales

MCA with missing data

Unit	Income	Age	Car
1	***x***	young	am
2	medium	medium	am
3	***y***	old	jap
4	low	young	jap
5	medium	young	am
6	high	old	am
7	low	young	jap
8	high	medium	am
9	high	***z***	am
10	low	young	am

Maximisation de la cohérence interne, ou de l'homogénéité

Results of the 27 MCA

x	*y*	*z*	λ_1	*x*	*y*	*z*	λ_1	*x*	*y*	*z*	λ_1
l	l	j	.70104	m	l	y	.63594	h	l	y	.61671
l	l	m	.77590	m	l	m	.72943	h	l	m	.66458
l	l	o	.76956	m	l	o	.72636	h	l	o	.65907
l	m	j	.78043	m	m	y	.70106	h	m	y	.70106
l	m	m	.84394	m	m	m	.77839	h	m	m	.74342
l	m	o	.84394	m	m	o	.84394	h	m	o	.74342
l	h	j	.78321	m	h	y	.73319	h	h	y	.68827
l	h	m	.84907	m	h	m	.80643	h	h	m	.74193
l	**h**	**o**	***.84964**	m	h	o	.80949	h	h	o	.74198

La différence de cette méthode avec les autres est que toutes les variables agissent simultanément dans la fonction de perte

Imputation par la moyenne

Remplacement d'une valeur manquante par la moyenne des mesures disponibles

La même pour toutes les NA d'une même variable

Estimations non biaisées si les données sont MCAR

Les méthodes implicites type Hot-Deck

- **Hot-Deck**
  - La valeur manquante est remplacée par une valeur observée chez un individu ayant les mêmes caractéristiques
- **Cold-Deck**
  - La valeur manquante est remplacée par une valeur observée chez un individu ayant les mêmes caractéristiques, mais provenant d'une autre source d'information
- **Mêmes caractéristiques veut dire « plus proche voisin »**
  - Distance basée sur une ou plusieurs variables auxiliaires

Les méthodes Hot-Deck

la valeur manquante est remplacée par la valeur observée chez un répondant “ proche ”, le “ donneur ”.

- le *hot-deck d'ensemble* : le donneur est choisi de façon aléatoire.
- le *hot-deck par classe* :
- le *hot-deck séquentiel* : l'individu le plus “ récent ” du tableau de données

Les méthodes Hot-Deck

le *hot-deck hiérarchisé* : On remplace l'unité défaillante par une unité ayant les mêmes valeurs pour C1, C2,…, Ck. S'il n'en existe pas alors on la remplace par une unité ayant les mêmes valeurs pour C1, C2,…, Ck-1; etc. …

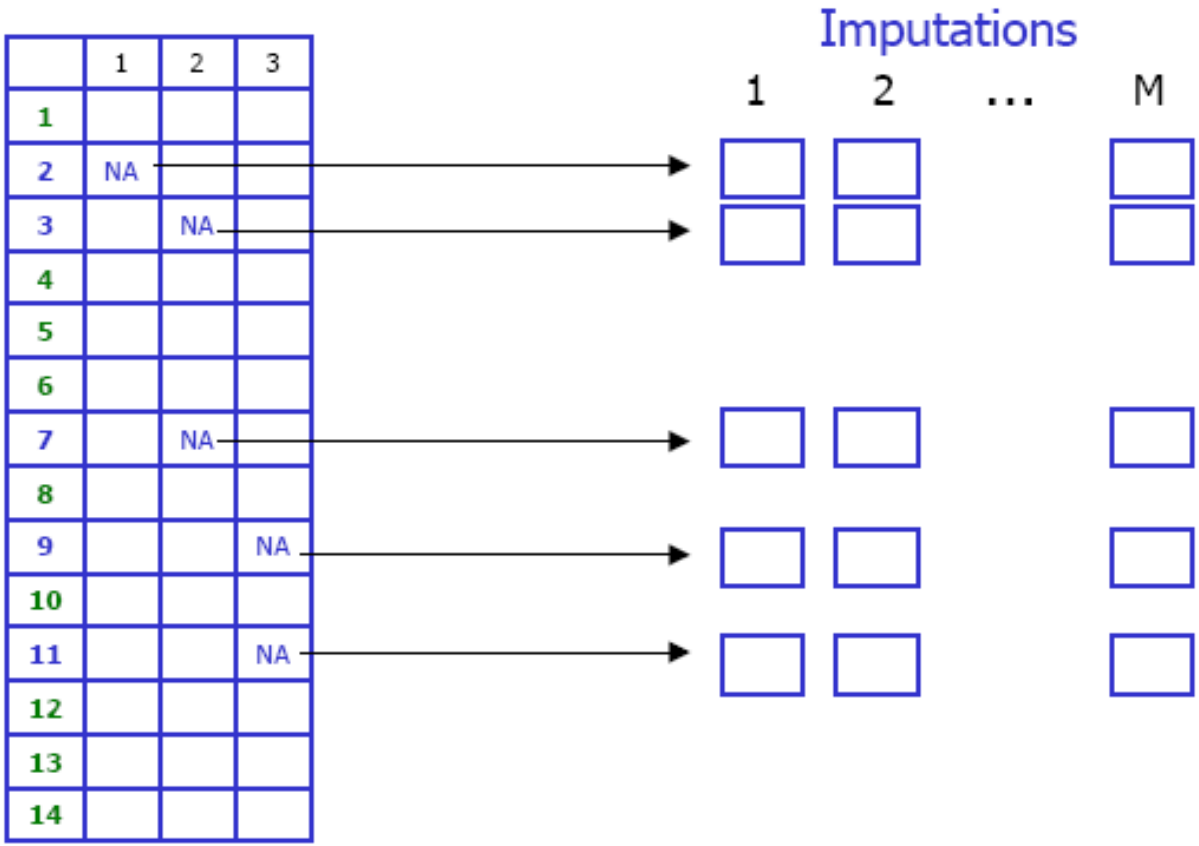
- le *hot-deck métrique* ou méthode du plus proche voisin avec une distance d(i,j)

Imputation Multiple

- Méthode consistant à créer plusieurs valeurs possibles d'une valeur manquante
- Les buts sont:
  - De refléter correctement l'incertitude des NA
  - De préserver les aspects importants des distributions
  - De préserver les relations importantes entre les variables
- Les buts ne sont pas:
  - De prédire les données manquantes avec la plus grande précision
  - De décrire les données de la meilleur façon possible

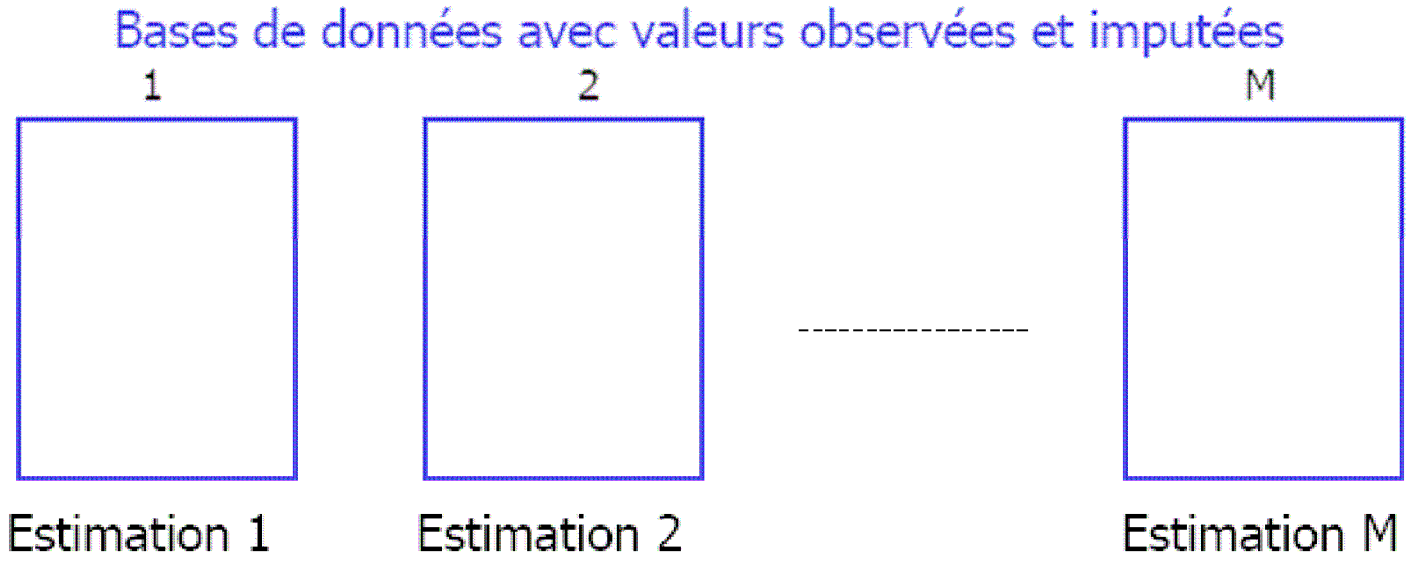
Imputation Multiple : les Étapes (1)

Remplacer chaque valeur manquante par M > 1 valeurs tirées d'une distribution appropriée



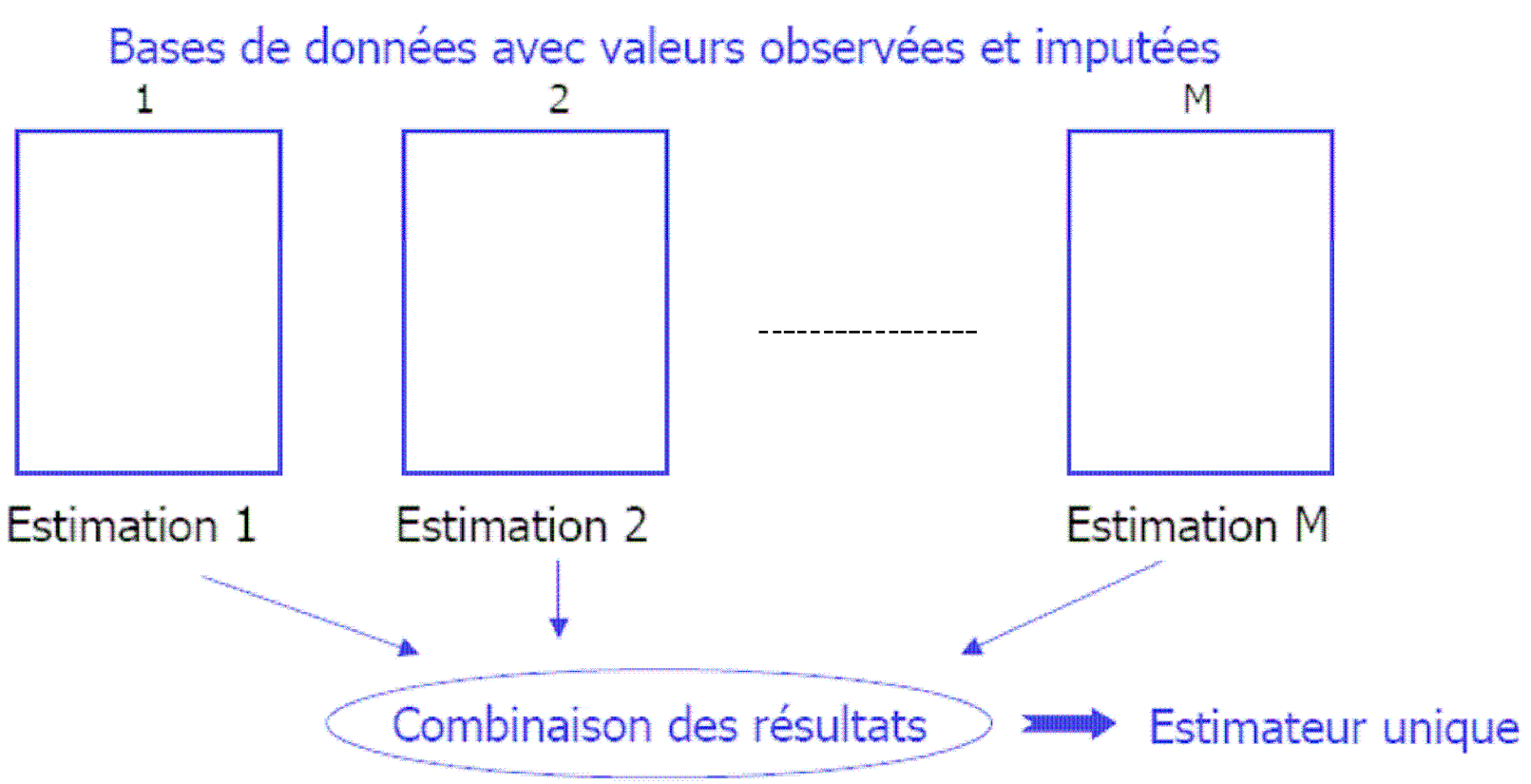
Imputation Multiple : les Étapes (2)

Analyses indépendantes et avec la même méthode standard des M > 1 bases de données complètes



Imputation Multiple : les Étapes (3)

Combiner les résultats des analyses afin de refléter la variabilité supplémentaire due aux données manquantes



IM: Commentaires

Approche générale
Un ensemble d'imputation peut servir pour plusieurs analyses
Les résultats définitifs incorporent l'incertitude des non réponses
Très efficace même pour des petites valeurs de M (>3)
Hypothèse d'un processus d'observation MAR
Les distributions prédictives des données manquantes peuvent être très compliquées
Le temps de calcul peut être long

Imputation Multiple : Logiciels

SOLAS
http://www.statsol.ie/html/solas/solas_home.html

SAS PROC MI et PROC MIANALYZE
http://support.sas.com/rnd/app/da/new/dami.html

S-Plus

http://www.insightful.com/

MICE*pour R et S-Plus
http://www.multiple-imputation.com

Références : Données Manquantes (1)

Allison PD. Missing Data. Sage Publications Inc.: Beverley Hills, CA, 2001.

Co V. (1997) Méthodes statistiques et informatiques pour le traitement des données manquantes. Thèse de doctorat, ENST, Paris.

Dempster A. P., Laird N. M., Rubin D. B. (1977) Maxium likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B39.

Grobras J. M. (1987) Méthodes statistiques des sondages, Economica, Paris.

Harel O, Zhou XH. Multiple imputation: Review of theory, implementation and software. Statistics in Medicine2007; 26(16): 3057-77.

Références : Données Manquantes (2)

Horton NJ, Stuart RL. Multiple imputation in practice: comparison of software packages for regressionmodels withmissingvariables. TheAmericanStatistician2001; 55: 244-254. http://www.biostat.harvard.edu/~horton/tasimpute.pdf

Horton NJ, KleinmanKP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. The American Statistician2007; 61(1): 79-90. http://maven.smith.edu/~nhorton/muchado.pdf

Kenward MG, Carpenter J. Multiple imputation: current perspectives. Statistical Methods in Medical Research2007; 16(3): 199-218.

Little RJA, Rubin DB. Statistical Analysis with Missing Data, second edition. Wiley: New York, 2002.

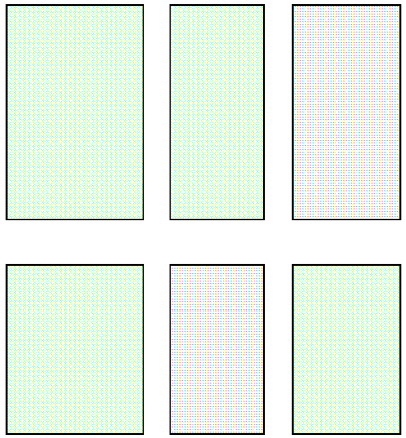
Multiple imputation online: http://www.multiple-imputation.com

Schafer JL. Analysis of incomplete multivariate data. Chapman & Hall, 2000.

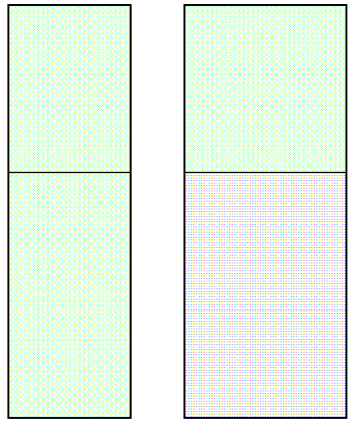
Fusion de données et thèmes similaires

- **Missing data problem**
  - Data comes from one source of iid random draws of a parametric distribution, with scattered missing values following some randomness pattern (mcar, mar, mnar).
  - Interest: to estimate global statistics (macrodata) taking into account the uncertainty of missing values.
  - The missing data techniques based on probabilistic framework.
- **Record linkage**
  - Matching the same individual in both sources.
  - Record matching techniques based on distances among individuals.
- **Data fusion**
  - Data fusion may be considered as a particular problem of missing data, where the missing data are *blocks of variables* missing by design.
  - Data usually comes from two independent sources, not necessarily representative.
  - Data Fusion is more ambitious. Interest is in the individual imputed data (microdata), and not only in the global statistics (macrodata).

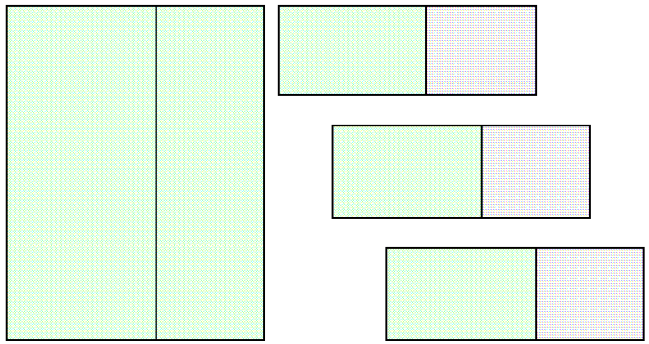
Types de Fusion



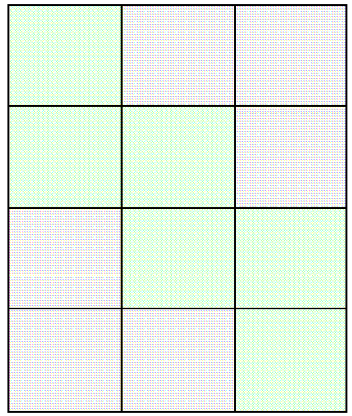
Parallel (or shared) questionnaire



Self-administered witth re-interview



Base survey with punctual enrichment



Panels

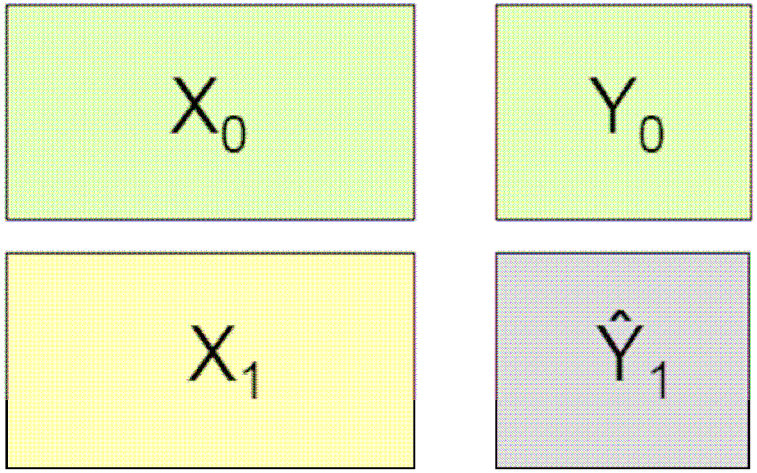
Principales applications: statistiques officielles, enquêtes d'opinion, de satisfaction, habitudes de consommation, …

Hypothèses de la fusion de données

Préservation des distributions conditionnelles

Hypothèse: mêmes distributions conditionnelles dans les deux fichiers même si les distributions jointes sont différentes (Lebart, Lejeune 1995)

$$f(Y_1/X_1)=f(Y_0/X_0)$$



$$f(X,Y/\theta_{f(X,Y)}) = f(Y/X,\theta_{Y/X})f(X/\theta_{f(X)})$$

$$g(X,Y/\theta_{g(X,Y)}) = f(Y/X,\theta_{Y/X})g(X/\theta_{g(X)})$$

Hypothèses de la fusion de données

Représentativité du fichier donneur

Hypothèse : Le fichier donneur est un échantillon représentatif de la population (Santini 1998, Van der Putten et al. 2002) , $f(\theta/X) = f(\theta/X_0)$

Nous ne faisons pas l'hypothèse que les deux fichiers sont des échantillons aléatoires d'une même population , $f(X_0, Y_0) \neq f(X_1, Y_1)$

Par contre, si les deux fichiers sont issus d'une même population, il est nécessaire de vérifier l'hypothèse suivante : $f(X_0) = g(X_1)$.

Questions clés préalable à une fusion de données

Comment et combien de variables communes doivent être retenues?

Quel algorithme utilisé?

Quelle validation mettre en œuvre?

Quels objectifs pour la fusion de données?

Minimiser l'erreur de prédiction : Min $E[y_1-\hat{y}_1]^2$

Les valeurs estimées doivent être aussi proches que possible des valeurs « vraies » inconnues

Préservation des distributions marginales du fichier donneur dans le fichier receveur, $f(\hat{Y}_1/X_1) \approx f(Y_0/X_0)$

Ces deux objectifs ne peuvent pas être optimisés en même temps

Si les deux fichiers sont issus d'une même population, préservation des lois marginales des variables communes et spécifiques : $f(X_1) \approx f(X_0)$, $f(Y_1) \approx f(Y_0)$

Objectifs complémentaires

Cohérence individuelle des données

Éviter les valeurs irréalistes pour l'ensemble des variables Y imputées

Reproduire la variabilité du fichier donneur

Ne pas réutiliser constamment les mêmes donneurs, introduire de la variabilité dans le processus d'imputation

Éviter le biais d'imputation : $E[Y_{@}] - E[\hat{Y}_{@}]$

Etc…

Fusions et greffes

Fusions de fichiers et greffes d'enquêtes: combiner des données provenant de sources différentes.

en amont du processus de « data mining » .

fusionner différentes bases: enquêtes, sources administratives, fichiers clients, données socio-économiques agrégées, etc.

Chaque base peut être constituée d'unités statistiques différentes ou d'agrégation de ces unités à différents niveaux.

Fusions et greffes

Fusion de fichiers. Cas élémentaire:
deux fichiers: F1 $p+q$ variables mesurées sur n_0 unités, F2 sous-ensemble de p variables pour n_1 unités. Souvent n_0 est faible par rapport à n_1 .

X_0	Y_0
X_1	?

Fusions et greffes

Greffes d’enquêtes
"coller" ou projeter les résultats d'une enquête S1 sur l'espace de référence défini par une enquête S0.

X_0	Y_0	
X_1		Y_1

Fusions et greffes

Étapes :

ACP de (X_0 Y_0), on retient k composantes C_0 que l 'on régresse sur les variables communes X_0. $\hat{C}_0 = X_0 b_O$

On positionne les individus de S1 dans le plan principal de S0: $C_1 = X_1 b_O$

- **On positionne les variables de Y_1 dans S0 en calculant les corrélations entre Y_1 et C_1**

Fusions et greffes

On utilise donc deux fois la méthode des points supplémentaires (les variables supplémentaires sont positionnées grâce aux individus supplémentaires) combinée avec une approximation des composantes principales.

Pour de bons résultats:

X_0 et Y_0 bien corrélées , pour pouvoir reconstituer les composantes principales de S0

X_1 et Y_1 aussi bien corrélées.

Modèles et méthodes pour la fusion de données

Appliquer industriellement une technique de traitement de données manquantes.

Deux approches:

Méthodes d'imputation:
compléter la non-réponse par une valeur plausible

Repondération: affecter aux répondants des pondérations pour compenser les non-réponses

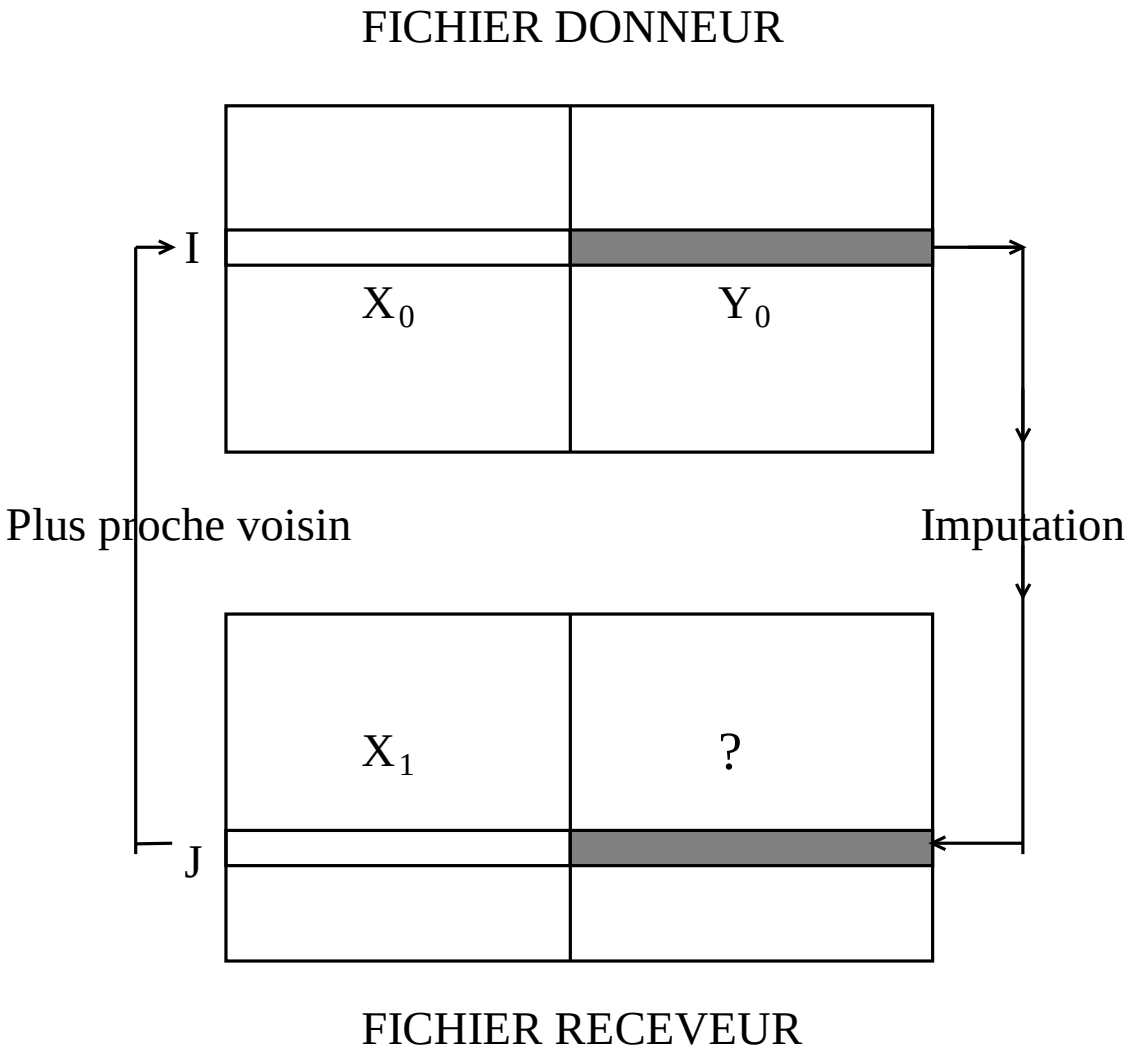
Conditions

- Vérifier préalablement que la taille de la population du fichier donneur est suffisamment importante par rapport au fichier receveur
- Les variables communes et les variables spécifiques doivent posséder des liaisons relativement fortes entre elles.

Les méthodes implicites:

- fusion par appariements intra-cellulaires,
- imputation par Hot-Deck,
- méthode des plus proches voisins etc….
- donner simultanément aux variables du fichier receveur toute l'information et les renseignements détenus par les variables du fichier donneur.

Schéma Hot-Deck



Processus d'imputation Hot deck

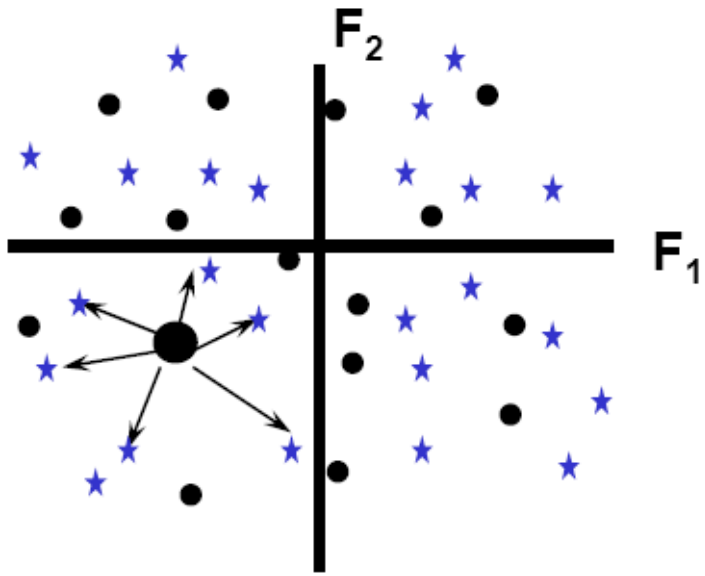
1. Definition of the common variables
   - Selection of minimum subset of common variables with maximum explicative power of the specific ones.
   - Positioning of all individuals in the same reference subspace defined by the common variables
2. Analysis of the conditions of application
   - Predictive relevance assumption
   - Equivalence of both samples
3. Determination of the imputation method and its parameters according the objectives of the operation
4. Imputation of the *Y* variables
5. Validation of the imputation

Fusion de données par K plus proches voisins

1. K-nn step

To find per each receptor the list of the most similar donors, according the common variables

A table with, say 40 neighbours per each receptor is produced and stored.



2. Imputation step

To transfer the specific vars. of neighbours to the receptor.

A complete file with imputed and observed data is produced

La fusion sur référentiel factoriel

fréquemment utilisée en France. Son principe (Santini 1984) repose sur :

- **- les 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**

La fusion sur référentiel factoriel

- Référentiel factoriel: ACM sur l'ensemble des variables critiques ou communes

- Détermination d'un voisinage du receveur

- Choix final parmi les donneurs éligibles selon les variables de rapprochement (sexe, age, …)

- Pénalisation pour éviter de prendre trop souvent les mêmes donneurs (voir fusion par mariage)

Fusion par mariages

éviter qu'un même donneur transmette son information à plusieurs receveurs (mariages multiples)

si un donneur est déjà marié à n receveurs, d est pénalisée par :

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

La fusion par mariage

G. Santini a imaginé 6 types différents de relations de voisinage par “ mariage ”: A receveur, B donneur.

- **le mariage par “ coup de foudre ” (voisins réciproques) : si A est le plus proche voisin de B et si B est le plus proche voisin de A et n'a jamais été marié, alors A et B sont immédiatement mariés.**
- **le mariage avec “ l'ami d'enfance ” : si B est le plus proche voisin de A, mais 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 ” : variante du cas précédent quand d(B ’,A) est plus grand que la distance pénalisée entre A et B (puisque B est déjà marié a A'). On marie alors A et B.**

Validation

Procédures empiriques où on estime des données connues mais cachées que l'on compare ensuite aux vraies valeurs : validation croisées, bootstrap …

Indicateurs:

- **reconstitutions de données individuelles**
- **prévisions au niveau de groupes**
- **reconstitutions de marges, de croisements**

Validation

Fusion avec collage du vecteur entier du donneur

moins bon pour la reconstitution de données individuelles,mais garde la structure de corrélation et évite les incohérences

Régression variable par variable.

C ’est l’inverse

Dans tous les cas il est nécessaire d'avoir:

- Un nombre suffisant de variables communes
- Des corrélations élevées entre variables communes et variables à imputer.
- Une structure commune entre fichier donneur et fichier receveur: distributions comparables des variables communes ou critiques, sinon résultats biaisés. Redressements souvent nécessaires.

Conclusions

Prudence quand on utilise des “ données ” qui sont en réalité des estimations et non des valeurs observées: ne jamais utiliser à un niveau individuel sans indicateur de donnée simulée, mais uniquement agrégé.

Conséquence perverse: un moindre effort de collecte, puisque l'on peut reconstituer des données…

Nécessité de valider

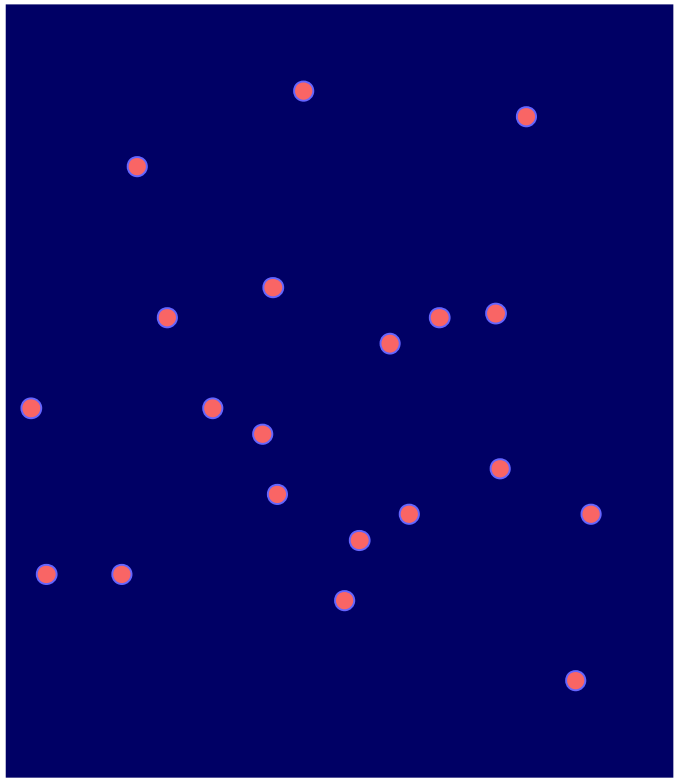
Déontologie, confidentialité et protection de la vie privée

Des données qui n'ont pas été recueillies mais estimées, peuvent être ajoutées dans des fichiers à l'insu des individus concernés. Quid de La loi “ Informatique et Liberté ” ?

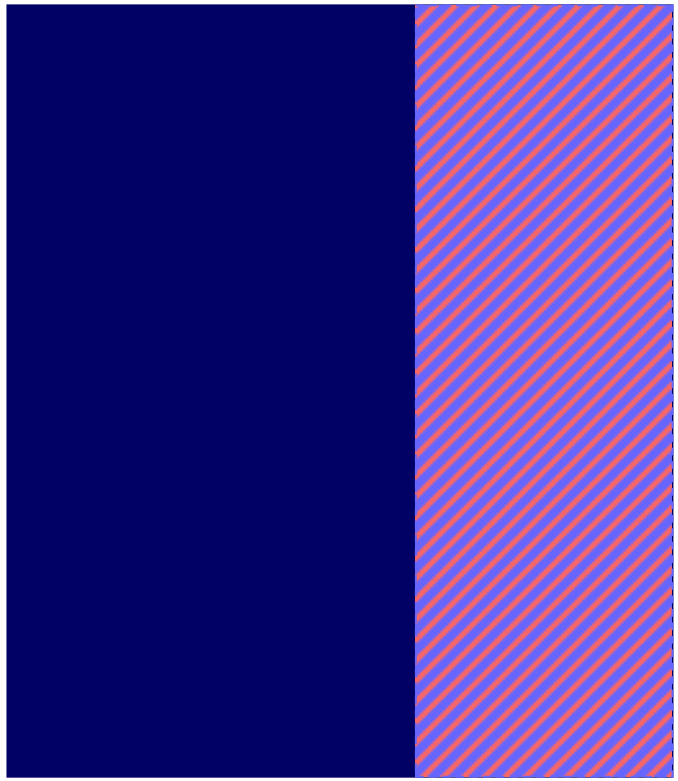
Paradoxe alors que les INS développent des techniques pour assurer la confidentialité

Les données manquantes peuvent apparaître :

Dispersées



Par bloc



Données manquantes par bloc

Besoin d'information complémentaire

Information dispersée entre différentes sources de données

Nécessité de réunir et d'exploiter toute cette information

Une solution raisonnable :

La fusion statistique de fichiers

Travaux précédents (liste non exhaustive)

Wendt F. (1980.) première personne à présenter une méthode de fusion de fichiers de données

Santini G. (1984) méthode d'appariement entre receveur et donneur

Rubin (1987) méthode d'imputation multiple

Lebart L. et Lejeune M. (1995) techniques de validation

Saporta G. et Co V. (1997) fusion par analyse homogène

Aluja-Banet T. (1997) fusion sur référentiel factoriel

Rässler (2002) fusion statistique par modèles bayésiens

Plan de la présentation

Contexte

Définition et Notations

Nouvelles méthodes de fusion statistique

Validation du modèle

Application à deux enquêtes

Application dans le domaine de la connaissance clientèle

Validation opérationnelle

Conclusion

Définition de la fusion statistique

Combinaison de données, provenant de sources différentes, pour obtenir un seul jeu de données dans lequel toutes les variables sont renseignées (présence obligatoire de variables communes)

Cas particulier du traitement de données manquantes faisant intervenir diverses sources de données et dont les données manquantes apparaissent en bloc

Deux grandes familles de méthode

La fusion par appariement d'individus

La fusion par prédiction de variables

La fusion par prédiction de variables

Estimation de chaque valeur manquante par des modèles de régression, régression logistique, …

Variables spécifiques **= variables à expliquer**

Variables communes **= variables candidates à l'explication**

Approche développée

Différents objectifs envisageables :

- Critères globaux : reconstitution des distributions marginales et des corrélations
- Reconstitution des données individuelles

Nombreuses méthodes de fusion statistique pour variables d'intérêt numériques
Développement des nouvelles méthodes pour données catégorielles (booléenne et ordinale)

Notations

Variables communes : X_1, …, X_p

- **Candidates à l'explication**
- **Qualitatives ou quantitatives**

Variables spécifiques : Y_1, …, Y_Q

- **Variables catégorielles**
- **Avec chacune, r_1, …, r_Q modalités**

Plan de la présentation

Contexte

Définition et Notations

Nouvelles méthodes de fusion statistique

Validation du modèle

Application à deux enquêtes

Application dans le domaine de la connaissance clientèle

Validation opérationnelle

Conclusion

Trois approches

Univariée

une par une, problème de corrélation entre les variables spécifiques

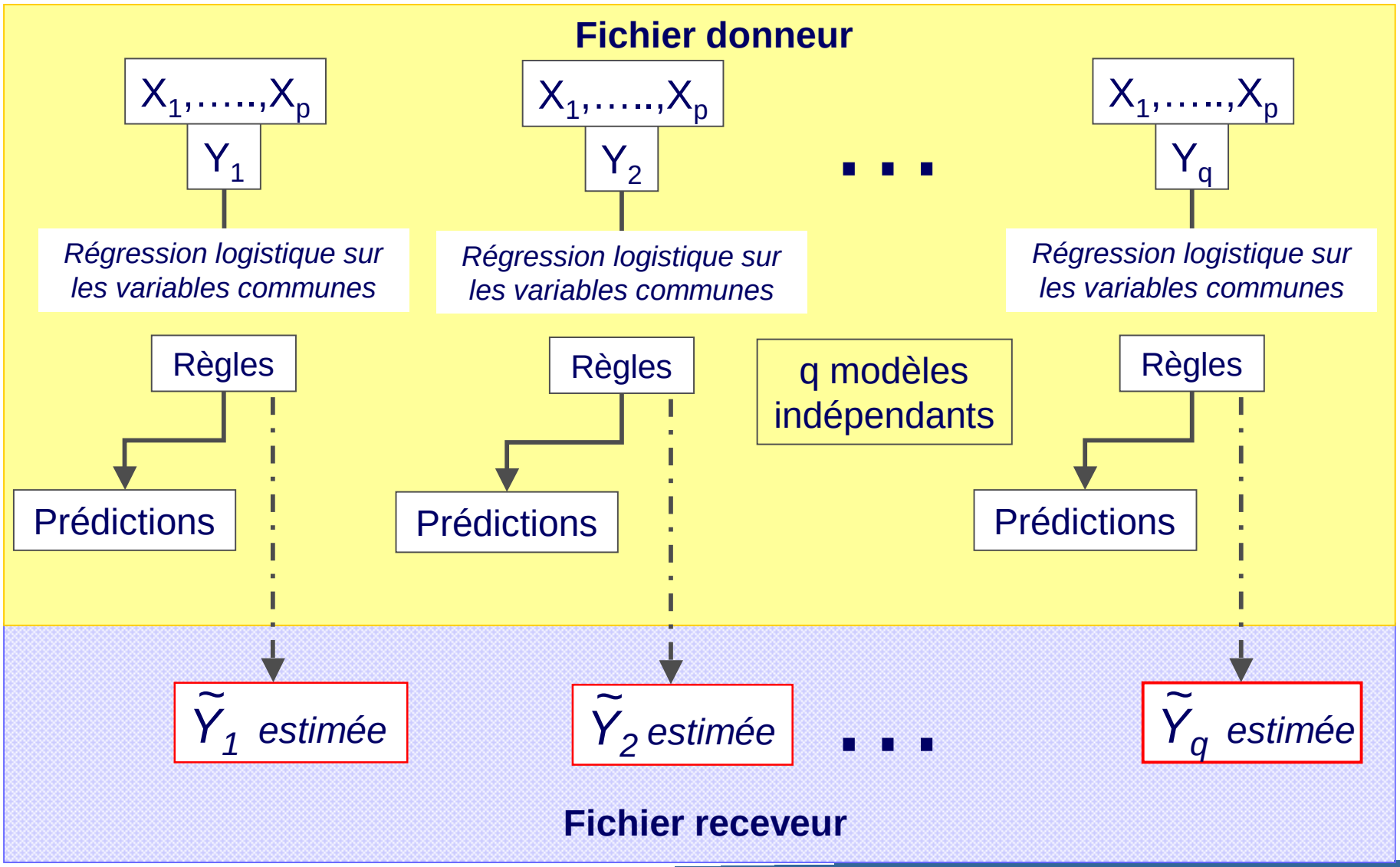
Séquentielle

conditionnel, début de prise en compte des corrélations

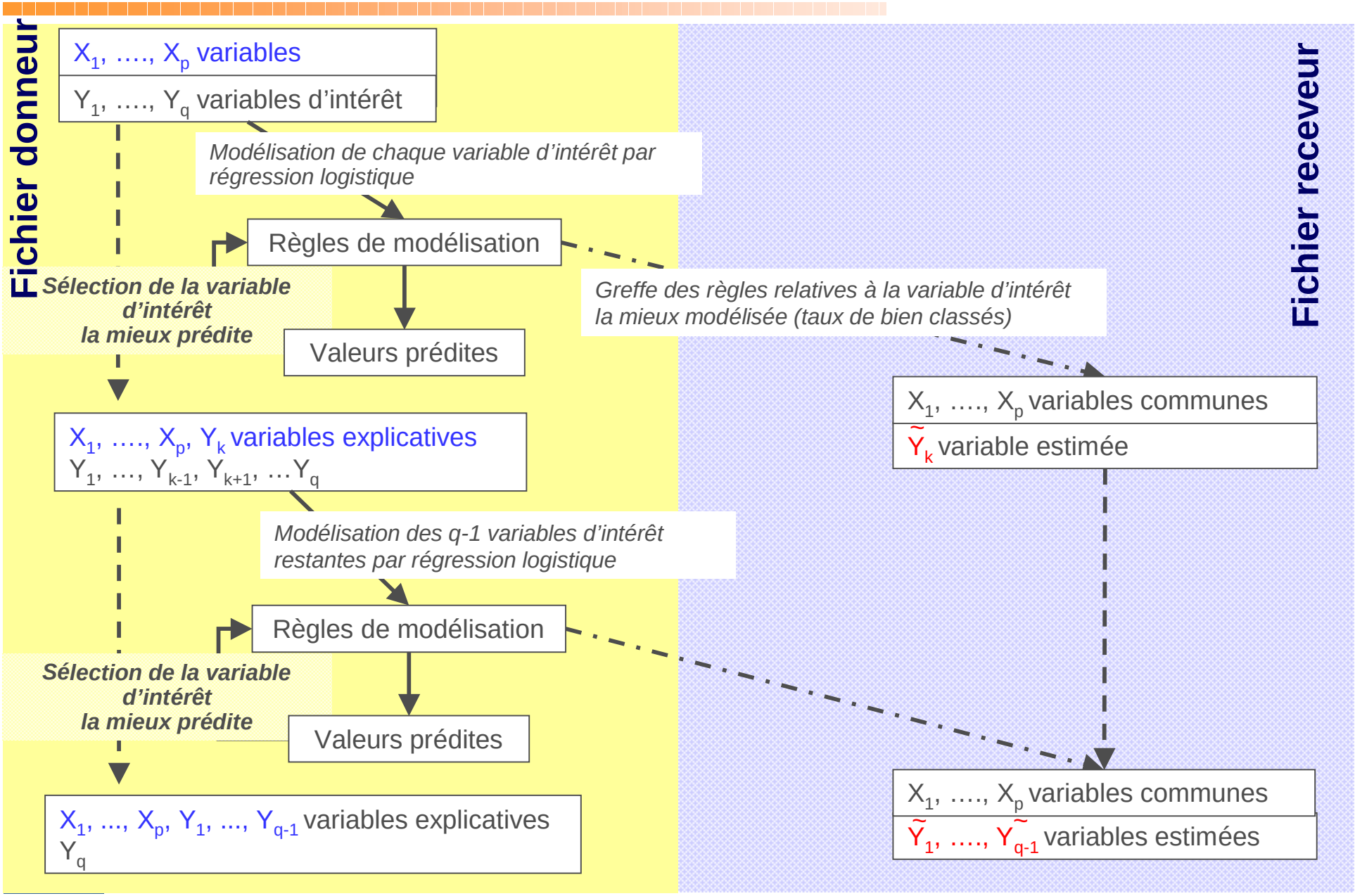
Multivariée

toutes en même temps, meilleure prise en compte des corrélations

Approche univariée : Logit classique



Approche séquentielle



Approches multivariées

Régression PLS [Wold, 1983]

Une alternative à la régression MCO et l'analyse canonique des corrélations (uni et multivarié Y) résout des problèmes avec forte multicolinéarité

Composantes PLS : combinaisons linéaires T des variables X maximisant simultanément la variance expliquée des Y et des X :

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

Algorithme : une séquence de régressions simples

Régression logistique PLS [Tenenhaus, 2000]

Extension pour un Y catégoriel, utilisant la régression logistique à la place des régressions simples de l'algorithme

Méthode Pseudo-PLS2

Variables Y estimées indépendamment par régression logistique PLS sur les variables communes

Sélection de toutes les composantes PLS significatives obtenues à l'étape précédente par validation croisée

Nouvelles estimations des variables Y par régression logistique PLS sur les composantes PLS précédemment conservées

Régression PLS2
recodage des données (0/1)

Pré-traitement des données

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

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

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

Régression PLS2
recodage des données (Logit ordinal) (1/3)

Y_1, …Y_q, …, Y_Q, variables ordinales à expliquer ayant respectivement R_1, …, R_q, …, R_Q réponses possibles
Construction de groupes d'individus à l'aide du croisement des modalités des variables candidates à l'explication

G groupes notés : v_1,…, v_i,…, v_G, contenant respectivement ni individus ayant les mêmes caractéristiques v_i, mais différentes sur les variables à expliquer

But : obtenir un nouveau jeu de variables Y's adaptées tenant compte du caractère de variable ordinale

Utilisation de la fonction de lien logit cumulé :

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

où r est la réponse à la variable Yq

Régression PLS2
recodage des données (Logit ordinal) (2/3)

Création de nouvelles variables quantitatives issues des logits cumulés observés

$$\tilde{y}_{qi}^{(r)} = \log\left(\frac{\left(\sum_{s=1}^{r_q} n_q^{(s)}\right)_i / n_i}{\left(n_{i-} \sum_{s=1}^{r_q} n^{(s)}\right)_i / n_i}\right)$$

avec r_q=1 à R_q

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

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

Régression PLS2
recodage des données (Logit ordinal) (3/3)

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

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

Plan de la présentation

Contexte

Définition et Notations

Nouvelles méthodes de fusion statistique

Validation du modèle

Application à deux enquêtes

Application dans le domaine de la connaissance clientèle

Validation opérationnelle

Conclusion et perspectives

Validation du modèle

La validation du modèle est une étape indispensable du processus de fusion statistique
Elle s'effectue sur le fichier donneur découpé en deux parties

- Fichier d'apprentissage (4/5)
  - construction des règles de modélisation
- Fichier test (1/5)
  - comparaison des données observées (masquées) aux données estimées à l'aide des règles

Test classique sur le modèle global
Test sur chaque variable explicative

- Choix des variables X candidates à l'explication, significatives sur tous les Y

Validation du modèle

Critère individuel de validation :

Taux d'individus bien classées au sein du groupe auquel il appartient

Trois critères globaux de reconstitution des données :

Reconstitution des distributions marginales

(test du $?^2$ entre les distributions marginales estimées et observées)

Préservation des croisements entre les Y

(test du $?^2$ entre les distributions croiseés, estimées et observées d'une paire de Y)

Pourcentage significatif de bien classés

(test de comparaison de deux proportions : avec et sans le modèle)

Plan de la présentation

Contexte

Définition et Notations

Nouvelles Méthodes de fusion statistique

Validation du modèle

Application à deux enquêtes

Application dans le domaine de la connaissance clientèle

Validation opérationnelle

Conclusion et perspectives

Deux fichiers d'enquêtes

Fichier donneur :

- enquête Sofres 1990 "le chauffage de votre logement"
- 8 000 individus

Fichier receveur :

- enquête CREDOC 1990 "Conditions de vie et aspiration des français"
- 2 000 individus

8 variables communes qualitatives (booléennes, nominales, discrètes) :

- âge, CSP, année du logement, type de logement, statut d'occupation, type de chauffage, taille d'agglomération…

Un certain nombre de variables spécifiques dont 9 variables ordinales de satisfaction

Tableau récapitulatif des résultats

		Approche séquentielle			Logit PLS2			Recodage PLS2			Logit PLS1			Pseudo – PLS2			Approche univarié		
Gen	mar	5				5										5			
	corr	8				7							2			6		1	
	bcl	5					3		1	1		2			1			1	1
Req	mar		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		1e		1e
Cot	mar	5							3				2			3		2	
	corr	8							1				3			5		7	
	bcl	5					5		2			1			2				
Tot	mar	32	10		9	23		4	4				18			27		10	
	corr	66	6		6	50			2				24			48		14	
	bcl	41		1		3	30	2	11	9	2	12			14			5	6
Tot		**106**	13	1	**12**	**51**	30	6	16	9	2	12	30	0	14	51	1	22	6

Plan de la présentation

Contexte

Définition et Notations

Nouvelles méthodes de fusion statistique

Validation du modèle

Application à deux enquêtes

Application dans le domaine de la connaissance clientèle

Validation opérationnelle

Conclusion et perspectives

Application à une enquête et à un fichier de facturation

Objectif de l'enquête :

Obtenir une meilleure connaissance de la clientèle afin de lui proposer des services et des produits adaptés à ses besoins

Fichiers utilisés :

Fichier receveur :

La base de facturation d'un centre EDF (1998)

27 785 individus

Fichier donneur :

L'enquête SOFRES "Chauffage électrique"

7 114 individus

Application à une enquête et à un fichier de facturation

Variables de l'enquête

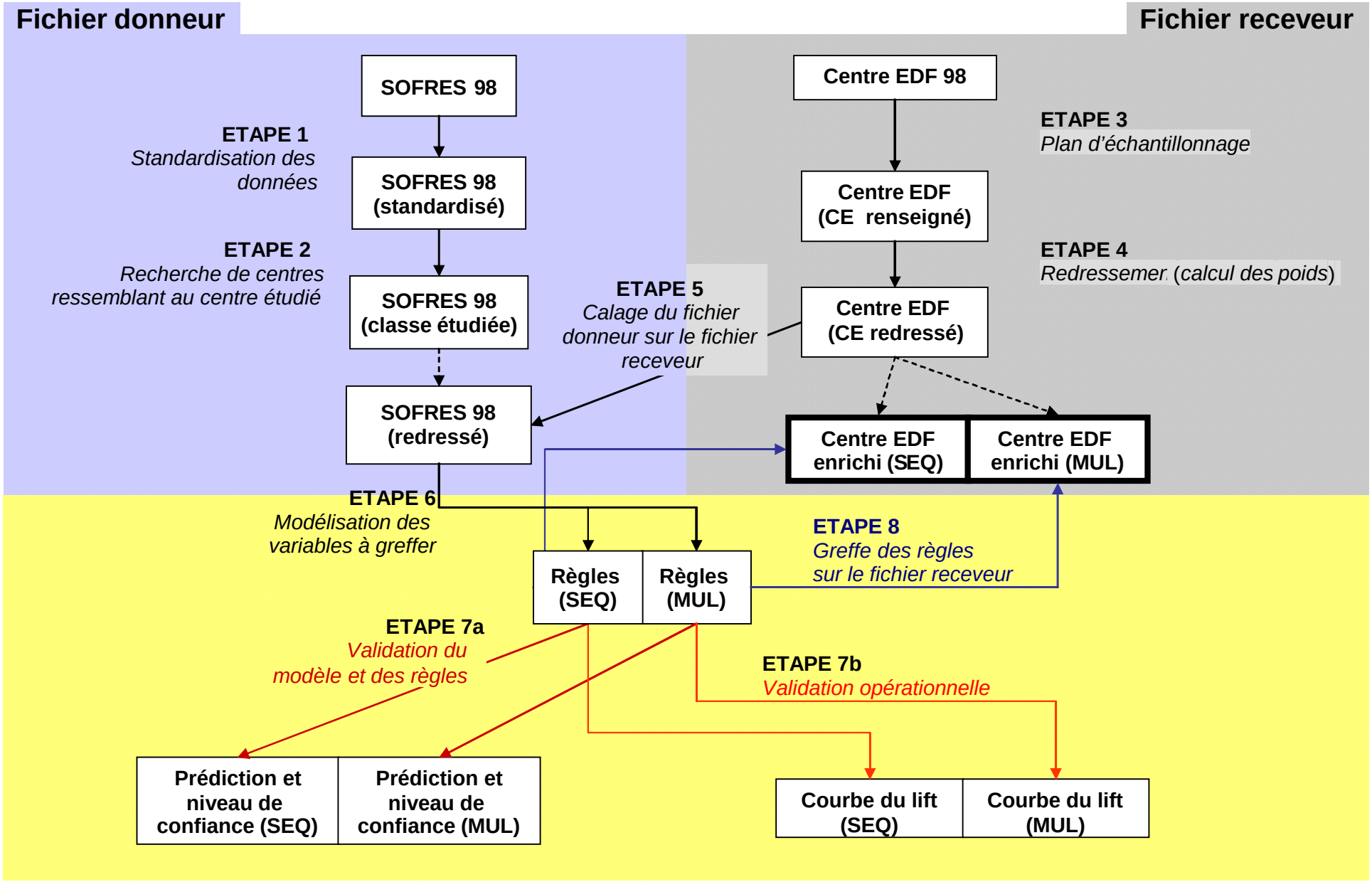
Quatre paquets de variables spécifiques :

- « ce que vous pensez de votre chauffage électrique »
- « les évolutions de votre chauffage »
- « ce que vous pensez d'EDF »
- « ce que vous attendez d'EDF »

Sept variables communes

- Tarif / puissance souscrite
- année de création
- Type de logement
- Politique de facturation
- Qualité payeur
- Code paiement
- Tranche de consommation annuelle d'électricité

Schéma de l'expérimentation



Résultats de la validation

Exemple pour la septième variable du thème « ce que vous attendez d'EDF » le pourcentage de bien classés égal à : 13,39 + 21,33 + 27,40 = 62,12% comparé à 40,77% est significatif

Observé \ Estimé	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

Matrice de confusion sur attente 7

Classement sur les pourcentages marginaux

Pour le thème « ce que vous attendez d'EDF »
exceptées 5 et 9, les attentes sont mieux reconstituées par la méthode séquentielle, que par la méthode multivariée

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

Test statistique sur les pourcentages marginaux

Taux de bien classés

Résultats du thème « ce que vous attendez d'EDF » montrent que la méthode séquentielle est meilleure dans ce cas que la méthode multivariée

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

Test statistique sur le pourcentage de bien classés

Reconstitution des croisements

La méthode multivariée préserve davantage les distributions croisées que la méthode séquentielle

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)**

Test statistique sur les distributions croisées

Critère individuel de validation

Pour le thème « ce que vous pensez de votre chauffage électrique ». Niveaux de confiance individuels associés aux individus du fichier receveur

Moyenne des niveaux de confiance individuels supérieur à 0,75

Proportion d'individus du fichier ayant un niveau de confiance individuel supérieur à 0,75

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

Les proportions de clients que l'on pourrait retenir sur des actions marketing sont plus élevées pour la méthode séquentielle

Plan de la présentation

Contexte

Définition et Notations

Nouvelles méthodes de fusion statistique

Validation du modèle

Application à deux enquêtes

Application dans le domaine de la connaissance clientèle

Validation opérationnelle

Conclusion

Validation opérationnelle

Objectif : Fournir une aide à la décision à l'expert de terrain

Le LIFT, un outil d'aide à la décision, il donne le pourcentage de clients prédits correctement relativement à ceux ordonnés par probabilité décroissante

L'expert dispose de plusieurs critères de validation :

- Le gain obtenu grâce au modèle, pour chaque seuil de sélection d'individus estimés appartenant à une cible
- L'indice de GINI, une mesure globale de la qualité, permettant de comparer différents modèles sur l'ensemble d'une courbe LIFT

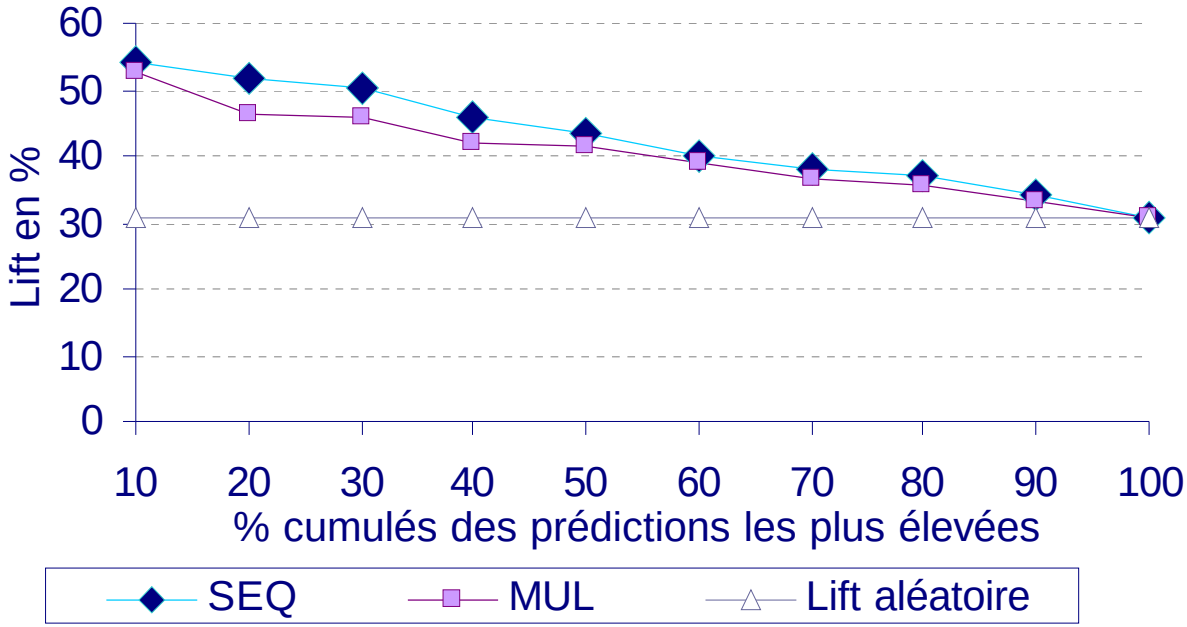
Validation opérationnelle

Illustration en terme de ciblage marketing

Définition d'une cible : clients insatisfaits de leur système de chauffage

54% méthode SEQ
53% méthode MUL
31% aléatoire

10% des prédictions les plus élevées



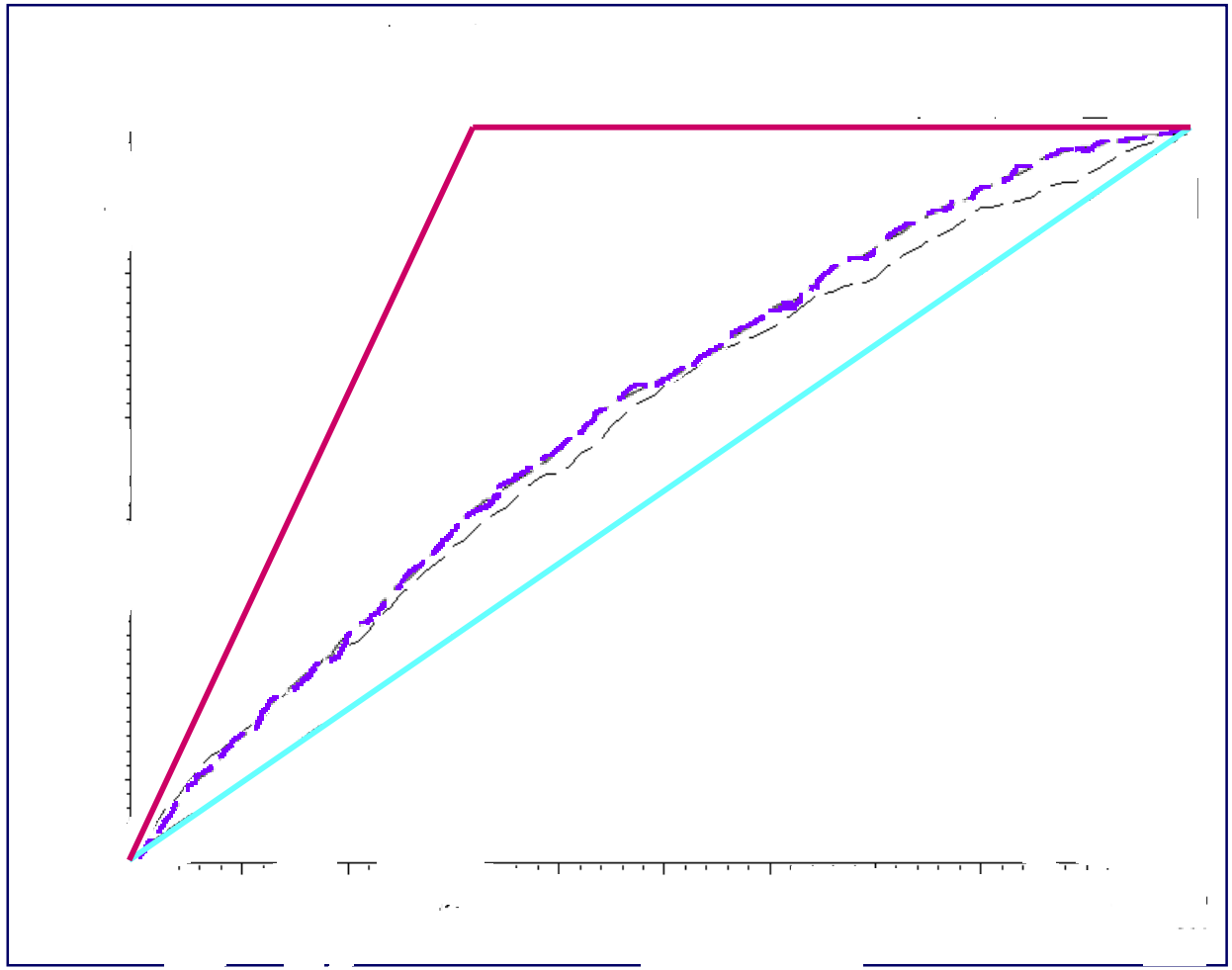
Le modèle statistique fait 54/31 = 1,74 fois mieux qu'un tirage aléatoire

Gain pour un mailing ciblant 1000 clients infidèles :

1000/0,54 = 1852 courriers
1000/0,31 = 3226 courriers

Courbes de lift sur l'insatisfaction pour les trois méthodes

Critère global de validation sur la courbe LIFT



Une mesure globale de la qualité permettant de comparer les modèles entre eux sur l'ensemble de la courbe LIFT : l'indice de Gini. Il se calcule par rapport d'aires sous les courbes de concentration.

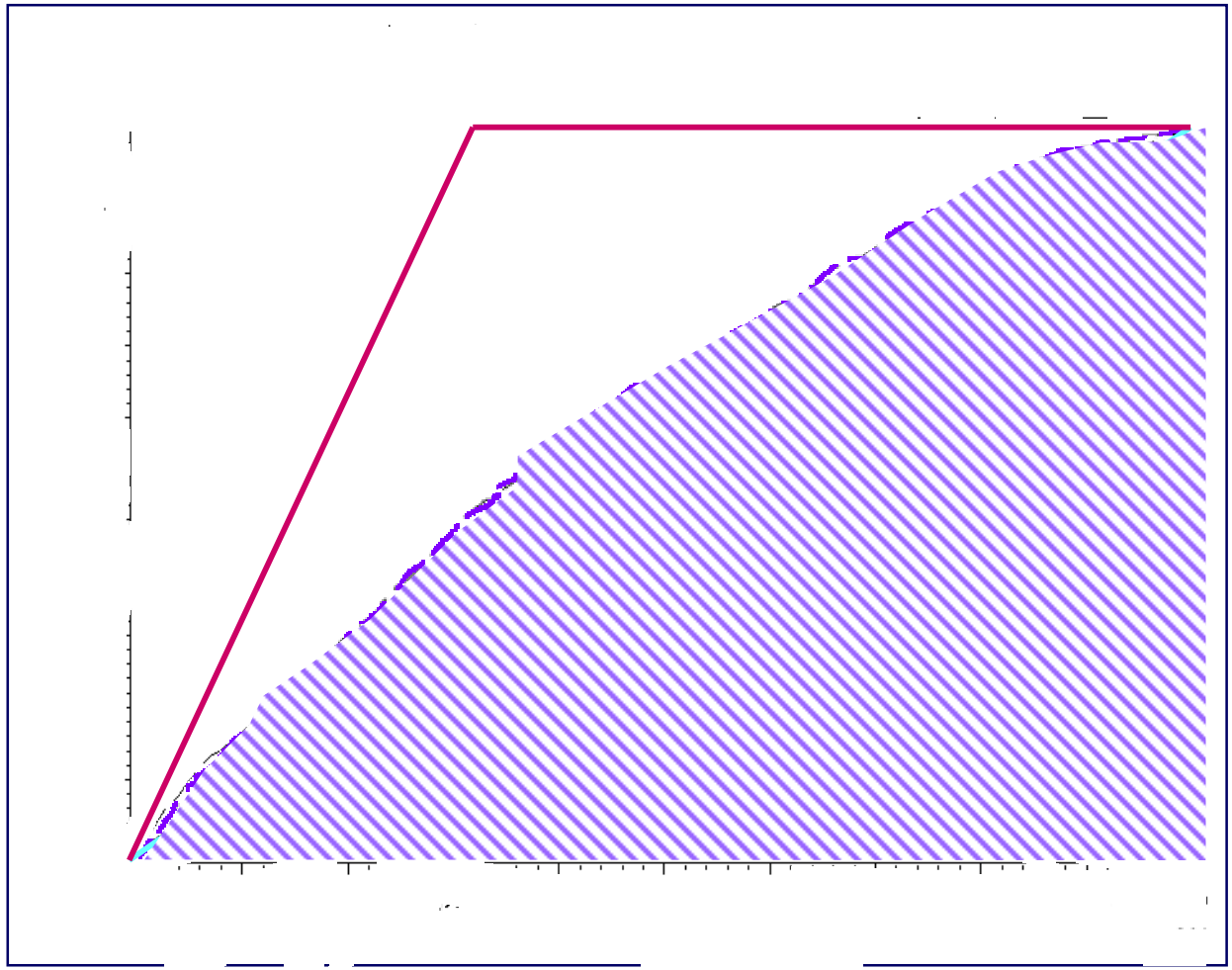
Lift estimé SEQ

Lift estimé MUL

Lift max

Lift aléatoire

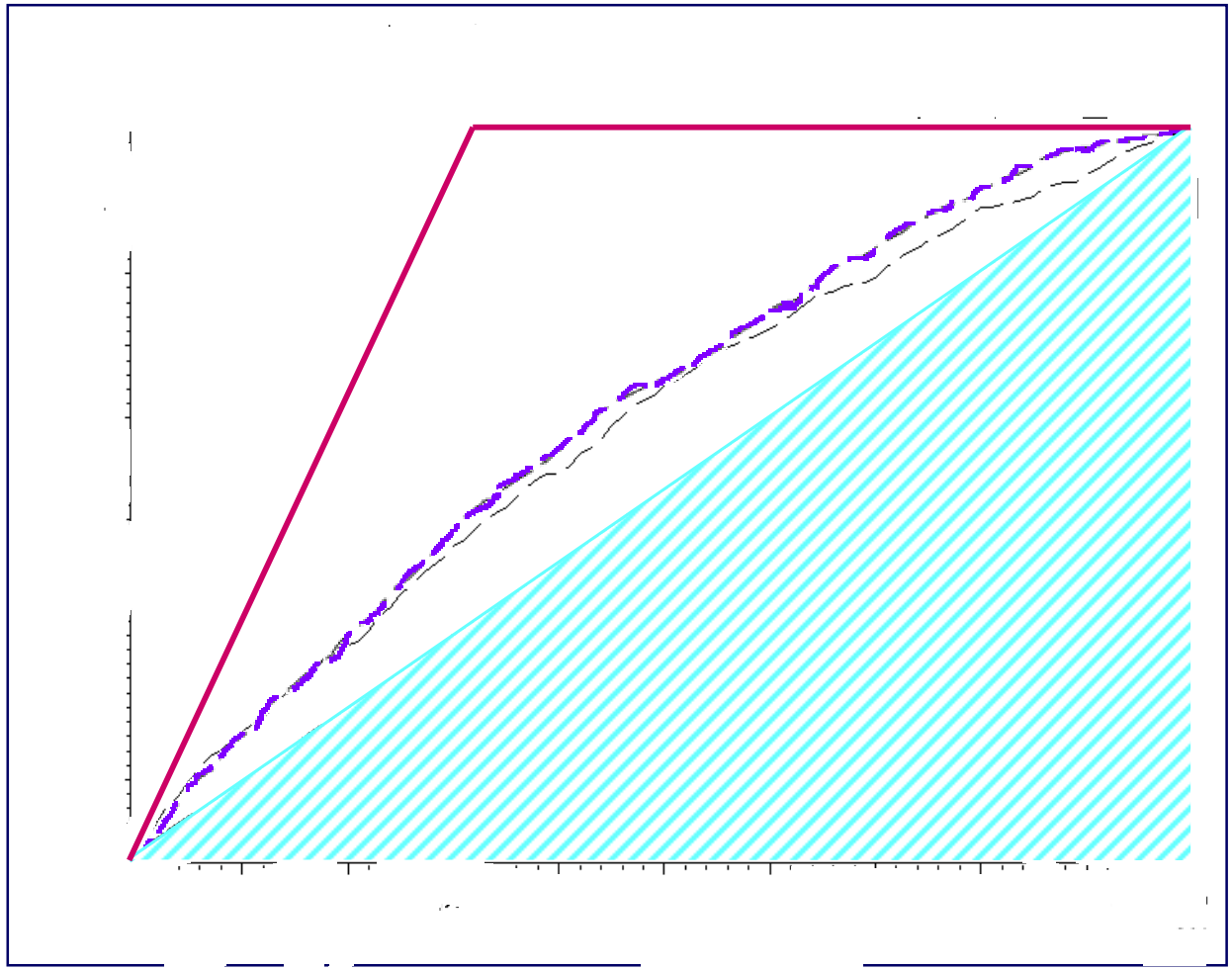
Critère global de validation sur la courbe LIFT



Une mesure globale de la qualité permettant de comparer les modèles entre eux sur l'ensemble de la courbe LIFT : l'indice de Gini. Il se calcule par rapport d'aires sous les courbes LIFT.

$A_{estimé}$

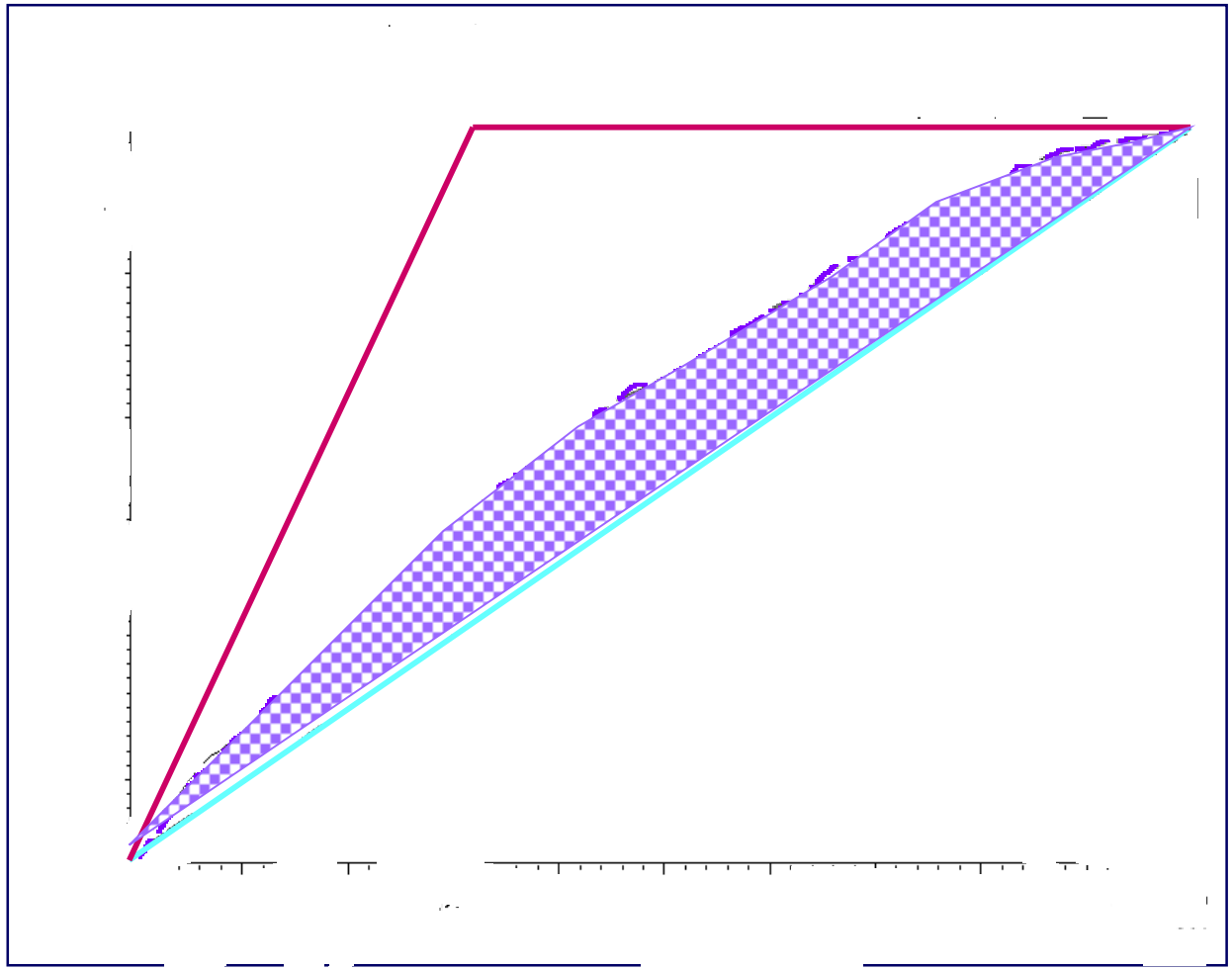
Critère global de validation sur la courbe LIFT



Une mesure globale de la qualité permettant de comparer les modèles entre eux sur l'ensemble de la courbe LIFT : l'indice de Gini. Il se calcule par rapport d'aires sous les courbes LIFT.

$A_{aléatoire}$

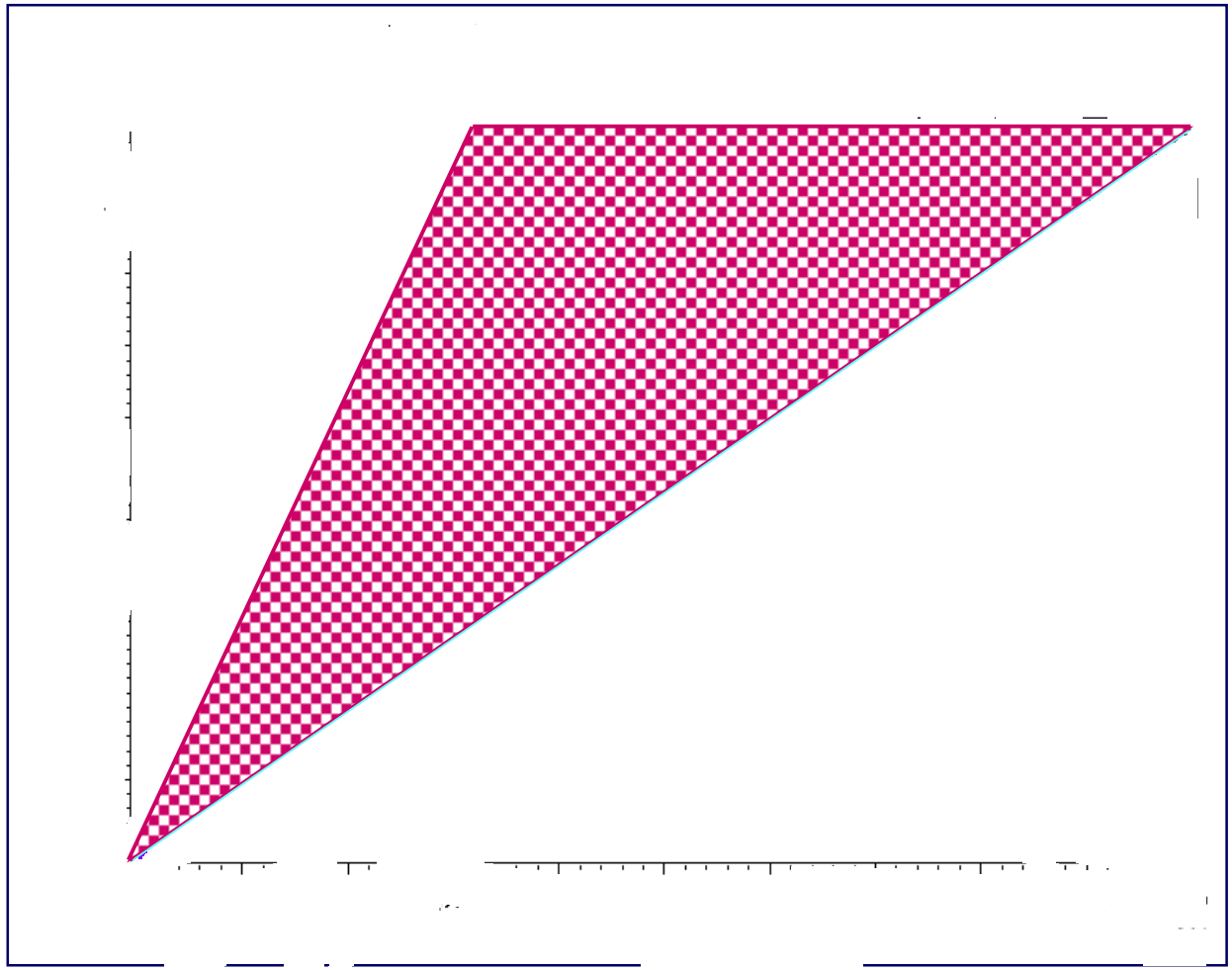
Critère global de validation sur la courbe LIFT



Une mesure globale de la qualité permettant de comparer les modèles entre eux sur l'ensemble de la courbe LIFT : l'indice de Gini. Il se calcule par rapport d'aires sous les courbes LIFT.

$$A_{estimé} - A_{Aléatoire}$$

Critère global de validation sur la courbe LIFT



Une mesure globale de la qualité permettant de comparer les modèles entre eux sur l'ensemble de la courbe LIFT : l'indice de Gini. Il se calcule par rapport d'aires sous les courbes LIFT.

$$A_{max} - A_{Aléatoire}$$

Indice de Gini

$G = (A_{estimé} - A_{aléatoire}) / (A_{max} - A_{aléatoire})$

G varie entre -1 et +1

Question	SEQ	MUL
Satisfaction	0,3487	0,2678
Choix	0,3914	0,3739
Conseil	0,2934	0,2954

La méthode séquentielle offre une meilleure qualité de résultats selon cet indice

Plan de la présentation

Contexte

Définition et Notations

Nouvelles méthodes de fusion statistique

Validation du modèle

Application à deux enquêtes

Application dans le domaine de la connaissance clientèle

Validation opérationnelle

Conclusion

Conclusion

La fusion statistique, une alternative :

Bons résultats sur les taux de bien classés

➔ Effet de bonne prédiction (bon résumé des données)

Bons résultats sur les marginales et les croisements

➔ Fort effet de lissage

Références

- Aluja-Banet T., ius R., Juarez C. (2002) Data fusion by PLS regression XXXIV Journées de Statistique JSBL-2002. Bruxelles, Louvain-la-Neuve.
- Fischer N. (2004) Fusion Statistique de Fichiers de Données. Thèse de Doctorat, CNAM, Paris.
- Rassler S. (2002) Statistical Matching, collection Lecture in Statistics, Springer.
- Rubin D.B. (1987) Multiple imputation for nonresponse in Surveys, Wiley.
- Saporta G. (2002) Data fusion and data grafting . Computational Statistics and Data Analysis, 38(4),465-473
- Santini G. (2002) Méthode de fusion procustéenne. Traitement des données d'enquêtes. XXXIV Journées de Statistique JSBL-2002. Bruxelles, Louvain-la-Neuve.
- Tenenhaus M. (2000) La Régression Logistique PLS, Journées d'Etudes en Statistique, Modèles Statistiques pour données Qualitatives, CIRM, Luminy.
- Wold H. (1983) Partial Least Square, Encyclopedia of Statistical Sciences, vol. 6, Kotz S. & Johnson N.L., John Wiley & Sons, pp. 581-591