L'HOMOGAMIE DES COUPLES

MISE EN ŒUVRE DE DIVERSES METHODES DE TRAITEMENT DE LA NON-REPONSE ET ANALYSE DE LEURS EFFETS SUR LA MESURE DE L'HOMOGAMIE

JMS 2005

Mélanie VANDERSCHELDEN
INSEE, département de la démographie

Contexte

Etude sur le thème de l'homogamie et du choix du conjoint à partir de l'enquête « Etude de l'Histoire Familiale » de 1999.

Pourquoi corriger la non-réponse?

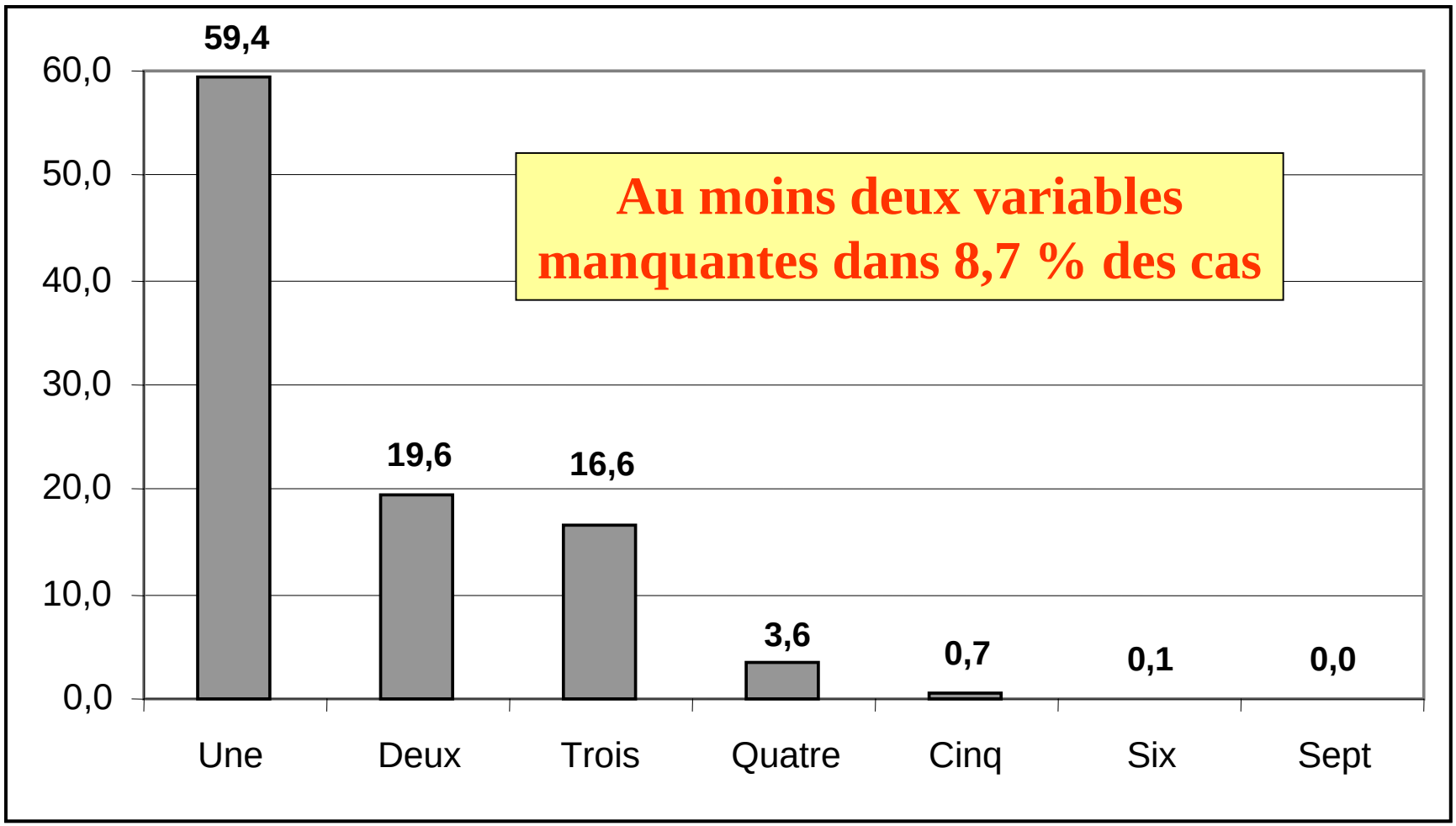
- Obtenir une matrice complète de données.
- Corriger le biais lié à la non-réponse.

Au moins une variable manquante pour 21,5 % (64 000) des observations

Taux de non-réponse par variable

Variable	Taux de non-réponse (en %)
Catégorie socioprofessionnelle du père	8,4
Catégorie socioprofessionnelle du conjoint	7,9
Etat matrimonial antérieur du conjoint	5,5
Lieu de naissance du conjoint (pays ou département)	5,2
Année de naissance du conjoint	4,8
Niveau d'études du conjoint	4,4
Nationalité du conjoint	3,0
Niveau d'études	1,9
Lieu de naissance (pays ou département)	0,9
Nationalité	0,1

Répartition des observations concernées par la non-réponse partielle selon le nombre de non-réponses



Méthode la plus adaptée : imputation par substitution.

Hot-deck : méthode qui consiste à remplacer, pour une observation appelée ***receveur***, une valeur manquante sur une variable donnée par une valeur observée sur la même variable pour un individu répondant choisi au hasard et appelé ***donneur***.

Imputations indépendantes par hot-deck séquentiel

Principe

- Choisir des variables auxiliaires (qui expliquent la variable à imputer) par :
  * Tris croisés
  * Modélisation (modèles polytomiques non ordonnés).

- Trier le fichier selon les variables auxiliaires sélectionnées.

- Imputer la valeur de l'observation précédente.

Imputations indépendantes par hot-deck séquentiel

Limites

- Il est nécessaire de choisir des variables de tri peu corrélées à la non-réponse pour éviter les duplications en chaîne.

- L'imputation étant réalisée indépendamment pour chaque variable, on néglige les liens qui peuvent exister entre les différentes variables imputées.

Imputations simultanées des valeurs prises par un unique donneur choisi au hasard

Principe

- Un unique donneur est utilisé pour imputer simultanément toutes les variables à blanc d'une même observation.

- Utilisation du hot-deck aléatoire ➔ le donneur est choisi au hasard.

- Des classes d'imputation sont constituées à l'aide des variables auxiliaires et le donneur est choisi dans la classe du receveur.

- Il faut définir les donneurs.

Imputations simultanées des valeurs prises par un unique donneur choisi au hasard

Application

Sexe de la pers.	Groupe social de la pers.	Age de la pers. déb. union	Cs conj.	Ecart âge	Lieu naiss. conj.	Lieu naiss. pers.	Nat. pers.	Niv. ét. pers.	Cs père	Nat. conj.	Niv. ét. conj.	Etat matri.
1	1	25	1	3	21	21	01	2	1	01	1	1
1	1	25	1	0	59	62	01	3	6	01	2	3
1	6	20	5	-11	92	75	01	2	3	01	3	3
1	6	20	5	1	63	75	01	1	3	01	1	2
1	6	20	6	2	92	75	01	2	5	01	3	2

Imputations simultanées des valeurs prises par un unique donneur choisi au hasard

Limites

- Information auxiliaire sous-exploitée.

- Choix des variables de classes difficile.

  N.B. : Les classes peuvent être constituées à l'aide des méthodes de classification, qui permettent de prendre en compte plus d'information auxiliaire, mais les résultats ne sont pas satisfaisants ici.

Imputations simultanées des valeurs prises par le donneur le plus proche du receveur

Principe

Hot–deck métrique : consiste à remplacer une valeur manquante par la valeur observée pour l'individu le plus proche, au sens d'une « distance » à définir et calculée à partir des variables auxiliaires.

Imputations simultanées des valeurs prises par le donneur le plus proche du receveur

Définition de la distance

Nombre de variables auxiliaires

$$D(r,d) = \sum_{j=1}^{p} \omega_j \delta_{j,rd}$$

Poids = V de Cramer

Distance partielle qui vaut:

- 0

si la variable auxiliaire j est renseignée et prend la même modalité pour le receveur r et le donneur d

- 1

sinon

Imputations simultanées des valeurs prises par le donneur le plus proche du receveur

Définition de la distance

Cs conj.	Ecart âge	Lieu naiss. conj.	Lieu naiss. pers.	Nat. pers.	Niv. ét. pers.	Cs père	Nat. conj.	Niv. ét. conj.	Etat matri.
1	3	21	21	01	2	1	01	1	1
.	0	59	62	01	3	6	01	2	3
5	-11	92	75	01	2	3	01	3	3
1	1	63	21	01	1	3	01	1	2
6	2	.	.	01	2	5	01	.	2

Imputations simultanées des valeurs prises par le donneur le plus proche du receveur

Définition de la distance

Nombre de variables auxiliaires

Nombre de variables à imputer

$$D(r,d) = \frac{\sum_{j=1}^{p} \omega_j \delta_{j,rd} + \sum_{k=1}^{q} \omega_k \delta_{k,rd}}{\sum_{j=1}^{p} \omega_j + \sum_{k=1}^{q} \omega_k}$$

Distances partielles

Poids

Imputations simultanées des valeurs prises par le donneur le plus proche du receveur

Avantages :

- Le lien entre les variables à imputer est préservé.
- Toute l'information auxiliaire est prise en compte.
- Pas de hiérarchie entre les variables imputées.
- Sans valeurs « supposées vraies ».
- Le donneur est proche du receveur.

Limite :

Temps de calcul important (surtout quand l'utilisation du produit cartésien n'est pas possible).

Imputations simultanées des valeurs prises par le donneur le plus proche du receveur

Améliorations

- Hot-deck métrique par classes ou combinaison du hot-deck hiérarchisé et du hot-deck métrique pour réduire le nombre de donneurs.
- Variante moins déterministe.
- Choix de distances partielles plus « fines ».
- Imputations simultanées de variables qualitatives et quantitatives possibles.

Impact de la correction de la non-réponse sur les résultats de l'étude

Nombre d'utilisations d'un même donneur

Variable cat. soc. du conjoint

Méthode	1	2	3 et plus
Séquentiel	90%	9%	1%
Aléatoire	93%	7%	1%
Métrique	91%	8%	1%

Impact de la correction de la non-réponse sur les résultats de l'étude

Mesures globales de l'homogamie

Mesures de l'homogamie	Avant imputation		Séquentiel	Aléatoire	Métrique	Métrique + calage
	Proportion	Intervalle de confiance				
% couples 2 pers. même groupe social	30,0	[29,8;30,2]	30,1	30,0	29,9	29,9
% d'ouvriers avec un conjoint ouvrier	35,2	[35,0;35,4]	34,9	34,9	35,0	34,8
% couples 2 pers. nées une même année civile	10,5	[10,4;10,6]	10,4	10,5	10,5	10,5
% couples 2 pers. même âge +/- un an	29,0	[28,8;29,2]	28,9	28,9	28,9	29,0

Impact de la correction de la non-réponse sur les résultats de l'étude

Mesures globales de l'homogamie

Mesures de l'homogamie	Avant imputation		Séquentiel	Aléatoire	Métrique	Métrique + calage
	Proportion	Intervalle de confiance				
% couples 2 pers. nées en France	81,4	[81,3;81,5]	80,8	81,0	81,2	81,4
% pers. françaises avec un conj. français	98,2	[98,2;98,2]	98,2	98,1	98,2	98,2
% pers. étrangères avec un conj. de même nationalité	66,3	[66,1;66,5]	63,3	63,0	66,6	66,1
% pers. étrangères avec un conj. français	29,9	[29,7;30,1]	30,6	33,1	29,7	30,2

Impact de la correction de la non-réponse sur les résultats de l'étude

Mesures globales de l'homogamie

	Parmi les données non imputées	Données pour lesquelles la nationalité ou le lieu de naissance de la personne ont été imputées		
		Après hot-deck séquentiel	Après hot-deck aléatoire	Après hot-deck métrique
% pers. de nat. portugaise parmi celles nées au Portugal	76	29	11	84
% pers. nées dans le même pays que leur nationalité parmi les pers. nées à l'étranger	40	15	2	65

Impact de la correction de la non-réponse sur les résultats de l'étude

Mesures « fines » de l'homogamie

Proportion de femmes non mariées, dont l'union a commencé dans les années 1990, de niveau d'études école primaire, vivant avec un conjoint de niveau d'études :			
	Séquentiel	Aléatoire	Métrique
Ecole primaire	43,4	39,6	46,2
Collège, CAP ou BEP	43,4	43,1	41,5

Conclusion

- Hot-deck métrique = méthode la plus adaptée mais difficile à mettre en œuvre pour de gros volumes de données.

- Impact faible mais non négligeable sur les les résultats de l'étude.

- Conditionnellement aux taux de non-réponse, au nombre de variables à imputer, à la taille de l'échantillon et à la corrélation entre variables.

<u>Ce qu'il faut retenir :</u>

L'imputation de plusieurs variables corrélées est risquée.

Imputations indépendantes par hot-deck séquentiel

Application

Groupe social de la personne	Age de la personne à la fin des études	Age de la personne	Cs conj.	Ecart âge	Lieu naiss. conj.	Lieu naiss. pers.	Nat. pers.	Niv. ét. pers.	Cs père	Nat. conj.	Niv. ét. conj.	Etat matri.
1	18	42	1	3	21	21	01	2	1	01	1	1
1	19	25	1	0	59	62	01	3	6	01	2	3
1	19	36	5	-11	87	75	01	2	3	01	3	3
1	20	85	5	1	63	.	.	1	3	01	1	2
2	23	23	6	2	.	.	01	2	5	01	.	2

Imputations indépendantes par hot-deck séquentiel

Adaptations

Méthode 1:

N°1	N°2	N°3	Cs conj.	Ecart âge	Lieu naiss. conj.	Lieu naiss. pers.	Nat. pers.	Niv. ét. pers.	Cs père	Nat. conj.	Niv. ét. conj.	Etat matri.
1	25	7	1	3	21	21	01	2	1	01	1	1
1	25	8	1	0	59	62	01	3	6	01	2	3
1	30	5	5	-11	87	75	01	2	3	01	3	3
1	30	8	5	1	63	75	01	1	3	01	1	2
2	25	3	6	2	63	75	01	2	5	01	1	2

Imputations indépendantes par hot-deck séquentiel

Adaptations

Méthode 2:

Nationalité de la personne (française/étrangère)	Cs conj.	Ecart âge	Lieu naiss. conj.	Lieu naiss. pers.	Nat. pers.	Niv. ét. pers.	Cs père	Nat. conj.	Niv. ét. conj.	Etat matri.
0	5	-11	87	75	01	2	3	01	3	3
0	1	3	21	21	01	2	1	01	1	1
0	.	0	59	62	01	3	6	01	2	3
0	.	1	63	.	01	1	3	01	1	2
0	6	2	.	.	01	2	5	01	.	2

Imputations indépendantes par hot-deck séquentiel

Méthode 2:

Adaptations

Lieu naiss. pers.	Dép. rés. pers	*Nat. pers*	Cs conj.	Ecart âge	Lieu naiss. conj.	Lieu naiss. pers.	Niv. ét. pers.	Cs père	Nat. conj.	Niv. ét. conj.	Etat matri.
1	59	*01*	.	0	59	62	3	6	01	2	3
1	63	*01*	.	1	63	62	1	3	01	1	2
1	77	*01*	5	-11	87	75	2	3	01	3	3
1	92	*01*	1	3	21	21	2	1	01	1	1
0	93	*01*	6	2	.	21	2	5	01	.	2

Imputations simultanées des valeurs prises par le donneur le plus proche du receveur

Améliorations

- Imputations simultanées de variables qualitatives et quantitatives.

  Distances partielles possibles (variables quantitatives):

$$* \quad \delta_{j,rd} = \frac{\left| x_{j,r} - x_{j,d} \right|}{R_j}$$

$$* \ \delta_{j,rd} = 0 \ \text{ si } \left| x_{j,r} - x_{j,d} \right| \leq T$$

T est un seuil fixé et tel que $T \leq R_j$

$= 1$ sinon

Impact de la correction de la non-réponse sur les résultats de l'étude

Répartition des variables

Catégorie socioprofessionnelle du conjoint	Avant imputation		Séquentiel	Aléatoire	Métrique	Métrique + calage
	Proportion	*Intervalle de confiance*				
Agriculteurs exploitants	*4,8*	*[4,7;4,9]*	4,8	4,8	4,9	4,9
Artisans, commerçants, chefs d'entreprise	*7,4*	*[7,3;7,5]*	7,4	7,4	7,3	7,3
Cadres et professions intellectuelles supérieures	*8,7*	*[8,6;8,8]*	8,6	**8,5**	**8,5**	**8,4**
Professions intermédiaires	*18,8*	*[18,7;18,9]*	**18,6**	18,7	**18,5**	**18,5**
Employés	*29,7*	*[29,5;29,9]*	29,9	29,9	29,8	29,8
Ouvriers	*26,1*	*[25,9;26,3]*	26,0	26,2	26,3	26,3
Sans activité professionnelle	*4,5*	*[4,4;4,6]*	**4,7**	4,6	**4,7**	**4,9**
Ensemble	*100,0*		100,0	100,0	100,0	100,0