



IMPUTATIONS DANS L'ENQUÊTE STRUCTURELLE

Séance d'information RFP
10 octobre 2013

Christian Panchard, Daniel Kilchmann

OFS / METH



Contenu

Principes de l'imputation par plus proches voisins

Mise en oeuvre de l'imputation par plus proches voisins

Résultats

Suite des travaux



Principes de l'imputation par plus proches voisins



Imputation par plus proches voisins

► Imputation :

- attribuer à une valeur manquante une valeur existante
- rendre cohérentes les valeurs incohérentes
- modifier un minimum de valeurs incohérentes

→ Règles de contrôle pour détecter les valeurs manquantes et les valeurs incohérentes

► Plus proches voisins

- choix de l'observation qui "donne" des valeurs à une observation qui ne vérifie pas au moins une règle, pour la rendre complète et cohérente ("donneur" → "receveur")
- choix de donneurs semblables au receveur (proche) et sans erreurs

→ Définition d'une distance



Mise en oeuvre de l'imputation par plus proches voisins



Enquête structurelle 2010

- ▶ 317'221 enregistrements pour les questionnaires individuels
- ▶ 689'914 enregistrements pour les questionnaires ménage et logement

→ 317'221 observations pour la procédure d'imputation (traitement des ménages encore à voir)

→ Total de 197 variables (toutes ne sont pas imputées)



Choix des variables à imputer

4 paquets de variables à imputer séparément :

1. Formation et profession :

- ▶ Plus haute formation achevée
- ▶ Formation en cours
- ▶ Statut sur le marché du travail
- ▶ Situation dans la profession

2. Langues, migration et logement

3. Mobilité

4. Ménage



Création de variables d'aide pour l'imputation

But : maintenir la cohérence des imputations entre les différents paquets.

- ▶ Nombre de réponses aux questions concernant le travail (langue habituelle, adresse, trajet)
- ▶ Nombre de réponses aux questions concernant la formation (langue habituelle, adresse, trajet)



Règles de contrôle

- ▶ Sélection des règles qui touchent uniquement les variables à imputer pour le paquet considéré
- ▶ Ajout de règles concernant les variables d'aide

→ Règles de contrôle

- ▶ concernant les valeurs manquantes
 - ▶ plus haute formation achevée = .
- ▶ concernant les valeurs incohérentes
 - ▶ $\text{age} < 60$ et statut d'activité = "retraité"
- ▶ concernant les variables d'aide (règles "souples")
 - ▶ nombre de réponses concernant le travail > 3 et statut d'activité \neq "actif occupé"



Paquet 1 : nombre d'observation avec erreurs

- ▶ 11'793 observations contenant au moins une valeur manquante (3.7%)
- ▶ 11'234 observations contenant au moins une valeur incohérente (3.5%)

→ 22'653 observations requièrent des imputations (7.1%)



Délimitation du voisinage

1. Définition d'une distance :

- ▶ Basée sur les variables à imputer, les variables d'aide et les variables de registre
- ▶ Compare chaque variable du receveur avec celles du donneur
 - ▶ distance discrète : par exemple $\{0, 1\}$ pour le sexe
 - ▶ distance continue : par exemple $[0, 1]$ pour l'âge

→ Combinaison pondérée des distances par variable



Délimitation du voisinage

2. Choix des classes d'imputation :

- ▶ Croisement des cantons et des tailles de ménage (1, ..., 5, 6 et plus)
- ▶ S'il n'y a pas de donneur dans la classe d'imputation → recherche dans l'ensemble des classes d'imputation



Choix d'un donneur

- ▶ Séparation entre donneurs potentiels et receveurs
- ▶ Calcul d'une distance entre chaque donneur potentiel et chaque receveur
- ▶ Choix aléatoire d'un donneur par receveur, avec probabilité proportionnelle à l'inverse de la distance



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Département fédéral de l'intérieur DFI
Office fédéral de la statistique OFS

Résultats

Statistique suisse



Comparaison avant vs après imputation

- ▶ Tous les receveurs ont été imputés de manière cohérente
- ▶ Plus haute formation achevée :

		<i>après imputation</i>						
		<i>aucune</i>	<i>obligatoire</i>	<i>professionnelle</i>	<i>culture générale</i>	<i>supérieure</i>	<i>universitaire</i>	<i>total</i>
avant imputation	.	1 016	988	1 413	240	284	390	4 331
	<i>aucune</i>	27 485						27 485
	<i>obligatoire</i>		54 397					54 397
	<i>professionnelle</i>			121 926				121 926
	<i>culture générale</i>				24 879			24 879
	<i>supérieure</i>	7	3	1		36 402		36 413
	<i>universitaire</i>	15	7	2			47 766	47 790
	<i>total</i>	28 523	55 395	123 342	25 119	36 686	48 156	317 221



Comparaison de totaux pondérés

- ▶ Plus haute formation achevée :

	<i>avant imputation</i>		<i>après imputation</i>		<i>différence</i>	
	<i>nombre</i>	<i>%</i>	<i>nombre</i>	<i>%</i>	<i>nombre</i>	<i>%</i>
.	88 663	-	-	-	-	-
<i>aucune</i>	575 057	8.9%	596 385	9.1%	21 328	0.2%
<i>obligatoire</i>	1 161 038	18.1%	1 181 962	18.1%	20 924	0.1%
<i>professionnelle</i>	2 578 565	40.1%	2 607 458	40.0%	28 893	-0.1%
<i>culture générale</i>	502 858	7.8%	507 838	7.8%	4 980	0.0%
<i>supérieure</i>	741 753	11.5%	747 249	11.5%	5 496	-0.1%
<i>universitaire</i>	871 319	13.5%	878 361	13.5%	7 042	-0.1%
<i>total</i>	6 430 590	100.0%	6 519 253	100.0%		



Mesures de qualité

- Nombre de donneurs potentiels par receveur :

Nombre de donneurs	Nombre de receveurs
1 - 10	2'941
11 - 50	6'785
51 - 100	4'368
101 -	8'559

- Nombre de fois qu'un même donneur est réutilisé :

Nombre d'utilisation	Nombre de receveurs
1	20'268
2	1'109
3	53
4	2



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Département fédéral de l'intérieur DFI
Office fédéral de la statistique OFS

Suite des travaux

Statistique suisse



- ▶ Imputations des paquets 2 à 4
- ▶ Imputations pour les enquêtes structurelles 2011 et suivantes