Amélioration du redressement de la non-réponse des communautés dans le recensement

Lise Reynaert

Direction des Statistiques Démographiques et Sociales, Insee

JMS 2015

Principaux objectifs du recensement

Déterminer **chaque année** la population légale de chaque commune et décrire les caractéristiques de la population et des logements

Les résultats se basent sur les **cinq enquêtes annuelles les plus récentes**

Fortes contraintes : culture du "chiffre exact" et lourdeur du processus de production

Parmi 65 millions d'habitants en France, le recensement distingue :

- la population des ménages (97,7 %)
- la population des communautés (2,1 %)
- les autres populations (0,2 %)

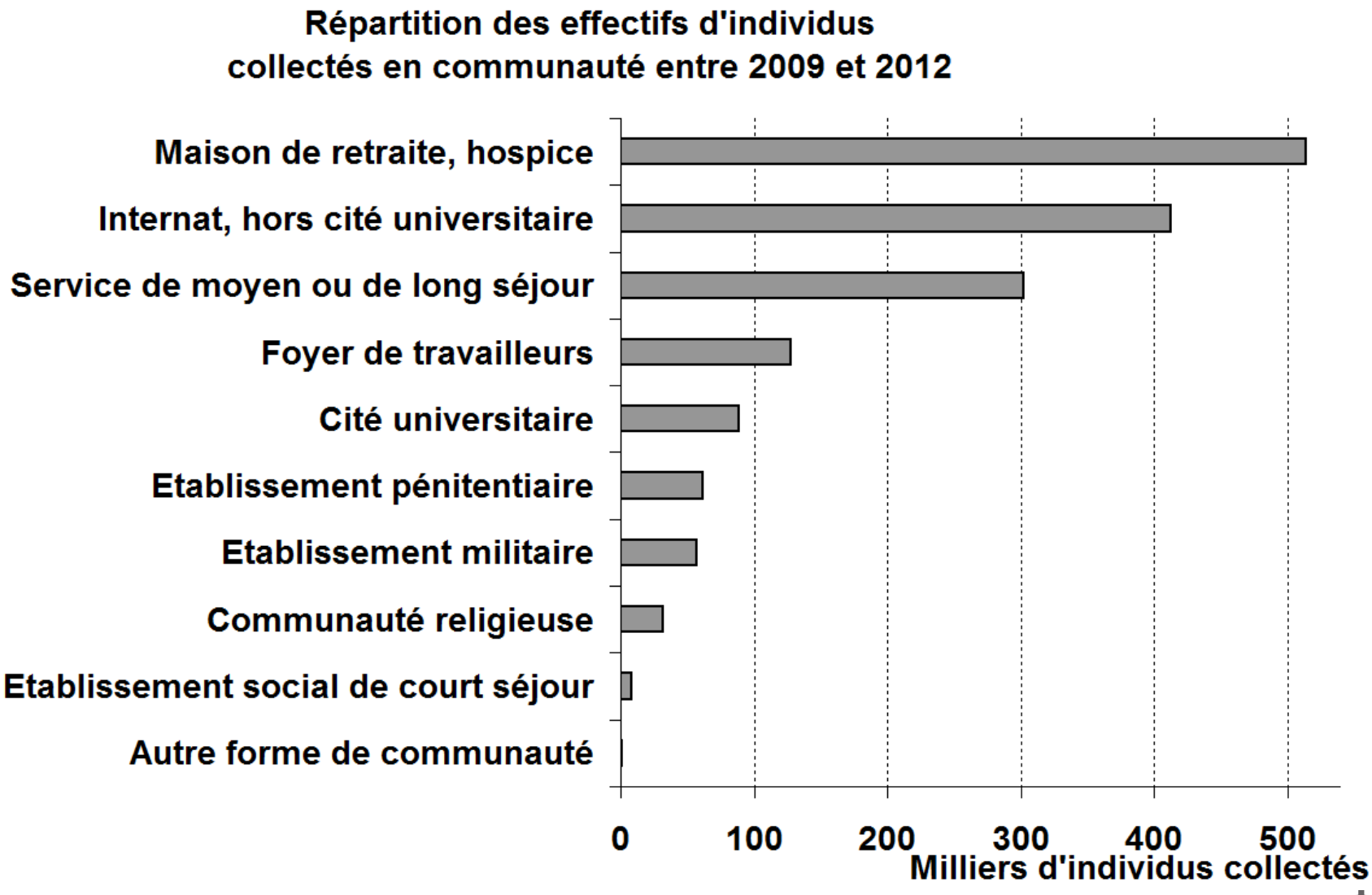
1,6 million d'individus dans **plus de 30 000 communautés**

Collecte exhaustive répartie sur 5 ans

Décret 2003-485 relatif au recensement de la population

Une communauté est un ensemble de locaux d'habitation relevant d'une même autorité gestionnaire et dont les habitants partagent à titre habituel un mode de vie commun

Insee


Répartition des effectifs d'individus collectés en communauté entre 2009 et 2012
Maison de retraite, hospice
Internat, hors cité universitaire
Service de moyen ou de long séjour
Foyer de travailleurs
Cité universitaire
Etablissement pénitentiaire
Etablissement militaire
Communauté religieuse
Etablissement social de court séjour
Autre forme de communauté
0
100
200
300
400
500
Milliers d'individus collectés

- Répartition de l'ensemble des questionnaires dans une **vingtaine de lots de saisie** structurés par commune
- Traitements post-collecte **simultanés ménages / communautés**
- Redressement de la non-réponse partielle par **hot-deck séquentiel** :
  - Tri du lot de saisie selon : département, commune, iris, rang d'adresse, rang de logement, rang d'individus
  - Imputation par la réponse de l'individu précédent le plus proche dans le parcours du lot de saisie
  - Couplée éventuellement avec d'autres critères (tranche d'âge, sexe...)

- **Redressement variable par variable**, dans un ordre logique prédéfini :
  - Sexe
  - Âge
  - ...
  - Diplôme
- **Particularité de la non-réponse des communautés :**
  - Faible taux de non-réponse pour les variables « sexe » et « âge » (consignes de collecte)
  - Pour les autres variables : taux de non-réponse sensiblement plus élevés que pour les ménages
  - Phénomène de non-réponse massive de communautés entières

- **Méthode robuste et efficace** permettant le respect de la forte contrainte de délai du recensement
- Satisfaisante pour les ménages
- Mais **faiblesses ponctuelles** visibles sur les communautés
  - Redressement de communautés entières par un unique donneur
  - Deux individus qui se suivent dans le lot de saisie ne sont pas forcément similaires

Insee

Variables disponibles pour tous les individus :

- au niveau communauté : sous-catégorie, effectif, indicatrice "première collecte"
- au niveau communal : tranche et catégorie d'aire urbaine 2010

Fort pouvoir explicatif de la sous-catégorie de communauté sur les réponses et sur le comportement de réponse

Critère 1 : résolution des redressements aberrants de la méthode actuelle ?

Critère 2 : qualité prédictive en cas de non-réponse diffuse

Critère 3 : distorsion induite par le redressement en cas de non-réponse massive et concentrée

- Ecart moyen par rapport à la distribution avant simulation de la non-réponse

Critère 4 : nombre maximum d'utilisation d'une même réponse pour imputer des non-répondants

Proposition 1 : redéfinir les contraintes d'imputation

- Détermination des variables auxiliaires qui expliquent le mieux les variables à imputer

Proposition 2 : ajout d'un aléa d'imputation pour éviter les redressements massifs par un unique individu répondant

- Tirage aléatoire parmi les 3 donneurs potentiels précédents

Proposition 3 : regrouper les individus en communauté dans un seul lot de saisie

	Contraintes d'imputation	
Variable à imputer	**Hot-deck séquentiel actuel**	**Hot-deck séquentiel avec redéfinition des contraintes d'imputation**
Sexe		Sous catégorie de communauté
		Effectif de la communauté
Age		Sous catégorie de communauté
		Effectif de la communauté
		Sexe
Etat matrimonial légal		Sous catégorie de communauté
		Effectif de la communauté
		Sexe
		Tranche d'âge
Situation principale	Indicatrice "travaille actuellement"	Sous catégorie de communauté
	Sexe	Inscription dans un ét. d'enseign.
	Tranche d'âge	Tranche d'âge
Diplômes obtenus	Tranche d'âge	Sous catégorie de communauté
	Indicatrice de nationalité	Tranche d'âge

Critère 1 : les redressements aberrants sont évités

Critère 2 : qualité prédictive en cas de non-réponse diffuse similaire à la méthode actuelle

Critère 3 : distorsion plus faible

Ecart moyen par rapport à la distribution avant imputation (%)

	Méthode actuelle simulée	Hot-deck avec redéfinition des contraintes	Hot-deck avec redéfinition des contraintes et aléa d'imputation
Tranche d'âge	14,0	5,0	4,5
Situation principale	7,5	4,0	4,4
Situation matrimoniale	10,1	6,0	5,8
Diplôme	7,4	6,2	5,1

Critère 4 : nombre de réplications légèrement limité par l'aléa d'imputation

- **Hypothèse :** comportement de réponse homogène à l'intérieur de classes de population
- Constitution de **classes homogènes** par rapport au comportement de réponse
- **Tirage aléatoire** d'un donneur parmi les répondants de la classe du receveur
- Choix du nombre de classes : **compromis entre homogénéité et robustesse**

Critère 1 : les redressements aberrants sont évités

Critère 2 : qualité prédictive en cas de non-réponse diffuse inférieure à la méthode actuelle

	Taux de bien classés (en %)	
Variable à imputer	Méthode actuelle simulée	Hot-deck par classe
Tranche d'âge	76%	66%
Situation principale	90%	70%
Situation matrimoniale	76%	67%
Diplômes obtenus	50%	35%

Critère 3 : amélioration de la mesure de distorsion

Ecart moyen par rapport à la distribution avant imputation (%)

	Méthode actuelle simulée	Hot-deck séquentiel avec redéfinition des contraintes et aléa d'imputation	Hot-deck par classe
Tranche d'âge	14,0	4,5	2,5
Situation principale	7,5	4,4	3,8
Situation matrimoniale	10,1	5,8	4,4
Diplôme	7,4	5,1	4,1

Critère 4 : nombre de réplications limité de par l'aléa d'imputation

- Imputation par un des donneurs les plus proches **au sens d'une mesure de similarité basée sur le V de Cramer** :

$$S = \frac{\sum_{j=1}^{p} \omega_j \delta_{j,rd} + \sum_{k=1}^{q} \omega_k \delta_{k,rd}}{\sum_{j=1}^{p} \omega_j + \sum_{k=1}^{q} \omega_k}$$

p : nombre de variables auxiliaires
ω_j *: le poids de la variable auxiliaire j, calculé à partir du V de Cramer*
ω_k *: le poids de la variable imputable k, calculé à partir du V de Cramer*
$\delta_{j,rd}(\text{resp.}\delta_{k,rd})$ *vaut 0 si la variable j (resp. k) prend la même modalité pour le receveur r et le donneur d, 1 sinon*

- Donneur choisi aléatoirement parmi les répondants maximisant cette mesure

Critère 1 : les redressements aberrants sont évités

Critère 2 : qualité prédictive en cas de non-réponse diffuse similaire à la méthode actuelle

Critère 3 : amélioration de la mesure de distorsion

Ecart moyen par rapport à la distribution avant imputation (%)

	Méthode actuelle simulée	Hot-deck séquentiel avec redéfinition des contraintes et aléa d'imputation	Hot-deck métrique
Tranche d'âge	14,0	4,5	1,2
Situation principale	7,5	4,4	6,4
Situation matrimoniale	10,1	5,8	3,3
Diplôme	7,4	5,1	3,7

Critère 4 : nombre de réplications atténué

- **Principe :**
  Modification du poids des répondants pour compenser la présence de non-réponse
  Le poids initial de chaque entité répondante est augmenté de l'inverse de sa probabilité de réponse (à estimer)

- **Problème :**
  Faible taux ou absence de répondants dans certaines communes
  Populations légales et résultats statistiques établis à un niveau communal

Méthode par repondération inadaptée

Méthode actuelle **globalement satisfaisante** malgré des cas limités de redressements aberrants

Avantages : **robustesse et efficacité** (contrainte forte de délai)

Amélioration à court terme possible en :

- redéfinissant les contraintes d'imputation (ajout de la sous-catégorie de communauté)
- regroupant les individus en communauté dans un unique lot de saisie

Parmi les méthodes alternatives testées, quel compromis entre **amélioration des résultats** d'une part et **coût** d'autre part ?

Insee