Traitement de la confidentialité dans la réponse au règlement européen sur les recensements de la population et du logement





Heidi Koumarianos
Division « Méthodes et Traitements des Recensements » - INSEE

Insee
Mesurer pour comprendre

24/06/2014

Plan

Contexte et problématique

Les traitements mis en œuvre

Les résultats

Contexte et problématique

Le cadre réglementaire européen (1/2)

Premier règlement européen concernant les données de recensement : l'année de référence est 2011

60 hypercubes : tableaux croisant plusieurs variables, selon plusieurs niveaux de détail (au total 478 millions de cases pour la France)

Jusqu'à 10 variables par hypercube (et jusqu'à 4 niveaux de détail pour certaines variables)

Le recensement français est l'unique source de données pour répondre au règlement : 5 enquêtes annuelles de recensement sont utilisées, soit 45 millions d'individus

Le cadre réglementaire européen (2/2)

Pas de préconisation concernant le traitement de la confidentialité

« Les États membres prennent toutes les mesures nécessaires afin de se conformer aux exigences de la protection des données. Les dispositions relatives à la protection des données en vigueur dans les États membres ne sont pas affectées par le présent règlement. »

Règlement n°CE 763/2008

Mais une forte attente européenne sur le respect du code de bonnes pratiques

Questionnaire pour les États membres, métadonnées

Le cadre réglementaire français

Arrêté de diffusion des données du recensement (19 juillet 2007)

Il signale quelques variables sensibles :

Les informations relatives à la nationalité et aux migrations, date d'arrivée en France

Cet arrêté porte uniquement sur des niveaux géographiques fins (communes ou quartiers).

Comment font les autres ?

Des exigences différentes :

- Définition d'observations ou de variables sensibles
- Toutes les petites cases

Des techniques différentes :

- Suppression de cellules
- Swap (échange d'informations entre individus)
- Micro-agrégation
- Arrondis

Des résultats différents :

- Perte d'information
- Ou perturbation des données

Quelle problématique de confidentialité ?

Les données des hypercubes respectent les conditions de l'arrêté de diffusion du recensement en France

… mais les cases de petit effectif sont très nombreuses, même à des niveaux géographiques agrégés (régions)

Les tableaux sont liés entre eux, l'élaboration à partir d'une source unique assure la cohérence des informations

Ils comportent de nombreuses variables

=> très grande difficulté technique de traitement post-tabulation

Les traitements

Les traitements mis en œuvre pour la réponse au règlement européen (1/6)

Le traitement de la confidentialité pour le recensement européen va au-delà de l'arrêté de diffusion français

- Les variables diffusées au niveau communal dans les hypercubes européens ne sont pas considérées comme sensibles au sens de l'arrêté de diffusion français
- Les variables considérées comme sensibles sont présentes dans des hypercubes de niveau régional

La complexité du recensement français introduit une grande incertitude sur l'identification effective d'individus

La présence de nombreuses cases de petit effectif incite à faire davantage

- Limiter le risque d'identification et la divulgation de caractéristiques sensibles

Les traitements mis en œuvre pour la réponse au règlement européen (2/6)

On a retenu une technique de perturbation des données individuelles :
le swap c'est-à-dire un échange d'informations entre les individus

Dans un double objectif : limiter le risque d'identification, tout en dégradant le moins possible la qualité

Cela permet de conserver la cohérence des informations, et rend les travaux réalisables : une fois le swap réalisé, la construction des hypercubes ne pose plus de difficulté liée à la confidentialité

Plusieurs interrogations :

- Quelles informations swape-t-on ?
- Comment sélectionner les individus swapés ? Dans quelle mesure ?
- Comment ?

Les traitements (3/6) : Quelles informations ?

Les variables identifiées comme sensibles

Un couple de variables : le pays de naissance et la nationalité

Précautions :

Variables liées : l'année d'arrivée en France

Pas de « problème » de géographie : toutes les données sont en géographie du 1er janvier 2012

Corrélations entre variables swapées et les autres ?

Les traitements (4/6) : quels individus ?

Il faut swaper les « bons » individus :

- L'objectif « limiter le risque d'identification » conduit à swaper prioritairement les individus appartenant à des cases de petit effectif
- L'objectif de conservation de la qualité incite à minimiser le nombre d'individus swapés, et en particulier à ne pas swaper inefficacement dans les cases d'effectif suffisamment grand

Comment cibler les individus à swaper ?

- Après plusieurs essais, on a choisi :
- De calculer une probabilité de swap, pour chaque individu, selon la case du cube à laquelle il appartient
- Une probabilité fonction décroissante de la taille de la case

Les traitements (5/6) : Quels groupes de swap ?

Tout en contrôlant a minima la perte d'information :

- Les cubes concernés sont de niveau région : on va donc swaper au sein d'une même région, afin de ne pas (trop) perturber les marges

Contrainte pour respecter la cohérence avec la variable « Date d'arrivée en France »

Les traitements (6/6)

Plus concrètement :

- Création des groupes de swap
- Calcul de la probabilité de swap selon la case du cube à laquelle appartient l'individu
- Génération d'un aléa pour chaque individu
- Sélection des individus qui seront swapés si l'aléa est inférieur à la probabilité de swap
- Constitution des fichiers d'individus à swaper

Un exemple (1/2) : les agrégats avant

	Avant	Sexe		
		F	M	Total
Pays de naissance (continent)	Afrique	3	1	4
	Amérique du Nord	1		1
	Amérique du Sud	1		1
	Asie	6		6
	Europe (hors UE)		4	4
	Europe (UE)	2		2
	Océanie	3		3
	Total	**16**	**5**	**21**

Un exemple (2/2) : les agrégats après

	Après	Sexe		
		F	M	Total
Pays de naissance (continent)	Afrique	4		4
	Amérique du Nord	1		1
	Amérique du Sud	4		4
	Asie		**5**	5
	Europe (hors UE)	**1**		1
	Europe (UE)	4		4
	Océanie	3		3
	Total	**16**	**5**	**21**

Les résultats

Les résultats (1/3)

On conserve l'ensemble des informations autres que pays de naissance et nationalité

On conserve par région le nombre d'individus non pondéré pour chaque croisement pays de naissance X nationalité…

Mais pas l'estimation car les individus ne sont pas équipondérés au sein d'une région !

=> on aurait pu rajouter une décomposition supplémentaire des groupes de swap entre petites et grandes communes, ou même poids de 1 environ, vs poids différent de 1, mais cela diminue fortement la taille des groupes de swap

Les résultats (2/3) : Une faible modification des marges

Nationalité	Nombre de personnes	Impact du swap	Coefficient de variation
Afrique	1 544 200	0,00%	0,14%
Amérique du Nord	45 200	0,04%	0,65%
Amérique du Sud	184 700	0,00%	0,25%
Asie	531 700	0,00%	0,26%
Europe (hors UE)	166 900	-0,01%	0,39%
Europe (UE)	1 339 300	0,00%	0,12%
France (pays de résidence)	60 794 400	0,00%	0,01%
Océanie	5 600	0,13%	1,82%

Données France entière

Les résultats (3/3) : Impact sur les petites cases

Une majorité des cases comprenant 1 à 3 individus ont été perturbées

Les cases d'effectif important sont très faiblement modifiées

Idéalement, on aurait pu…

Contrôler la présence ou l'absence de corrélations entre les variables swapées et les autres

=> mais la proportion d'individus swapés est faible

Contrôler la pondération dans la constitution de groupes de swap

=> diminue la taille des groupes de swap et difficulté de constitution des groupes de swap

Comment peut-on optimiser la perturbation sur plusieurs cubes?

Conclusion

Un fort impact sur les petites cases

Une détérioration des marges très acceptable

Une facilité à construire les tableaux

Merci de votre attention !

Contact
Mme Heidi Koumarianos
Tél. : 01 41 17 63 15
Courriel : heidi.koumarianos@insee.fr

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr
Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00