Consolidation de jeux de données pour la prospective :

la génération d'une population synthétique pour les communes de Belgique

Philippe Toint (with J. Barthélemy)

Centre de Recherche en Systèmes Complexes (NAXYS),
FUNDP

(philippe.toint@fundp.ac.be)

IWEPS, Mai 2010

Introduction

Contexte : la prospective

- prospective : association de contextes complémentaires en vue de prévisions intégrées à moyen/long terme
- requiert typiquement l'utilisation simultanée de jeux de données différents (sources, méthodes, dates)

Réconciliation des jeux de données ?

Difficultés :

- inconsistence entre données (légères $\longrightarrow$ fortes)
- contraintes légales (protection de la vie privée)

Notre objectif : la population belge

Notre problème :

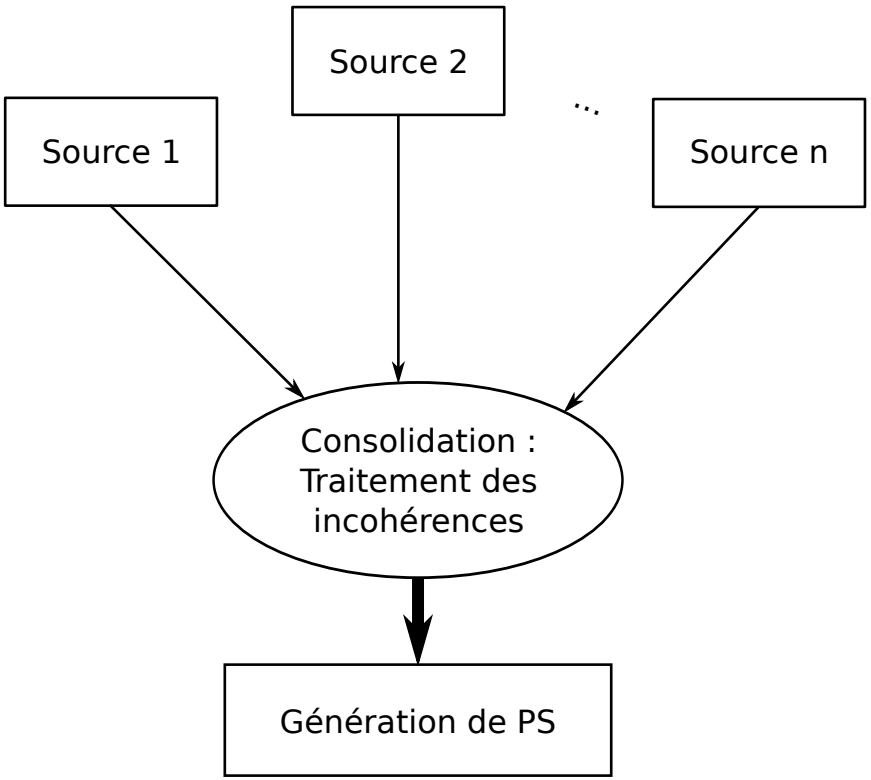
reconstruire une version synthétique de la population des 589 communes belges

Mais . . .

Tab. contingence	Source	Tot. Marginaux	Prop.
commune × sexe × age	GéDAP, 2001	405.491	1,00
commune × type mén.	GéDAP, 2001	380.653	0,94
commune × diplôme	GéDAP, 2001	426.372	1,05
commune × status	GéDAP, 2001	396.594	0,97
arrond. × type mén. × age	INS, 2001	357.884	0,88
arrond. × diplôme	INS, 2001	398.582	0,98

Exemples d'incompatibilités pour l'arrondissement de Charleroi

Un peu plus précisement...



But :

Générer une PS représentant les individus et les ménages au niveau communal (NUTS-5).

Sources des données :

- INS, GéDAP (UCL), MOBEL
- 2 niveaux d'aggrégation :
  - communes (*COM*)
  - arrondissements (*ARR*)

Génération de la population synthétique
idée générale

Philosophie du générateur

Construire les individus et ménages synthétiques en tirant aléatoirement les caractéristiques / membres dans les distributions adéquates au niveau le + désaggrégé disponible tout en préservant les corrélations connues.

Etapes du générateur

1. Estimer le pool d'individus *Ind*, *i.e.* le tableau de contingence des caractéristiques des individus.
2. Estimer le tableau de contingence *Men* des variables "ménages".
3. Construire les ménages synthétiques en tirant aléatoirement leurs membres dans *Ind* et en respectant la distribution conjointe calculée à l'étape 2. ($\longrightarrow$ *Ind*$'$ et *Men*$'$)

$\Rightarrow$ Etapes 1 et 2 : Consolidation des données disponibles

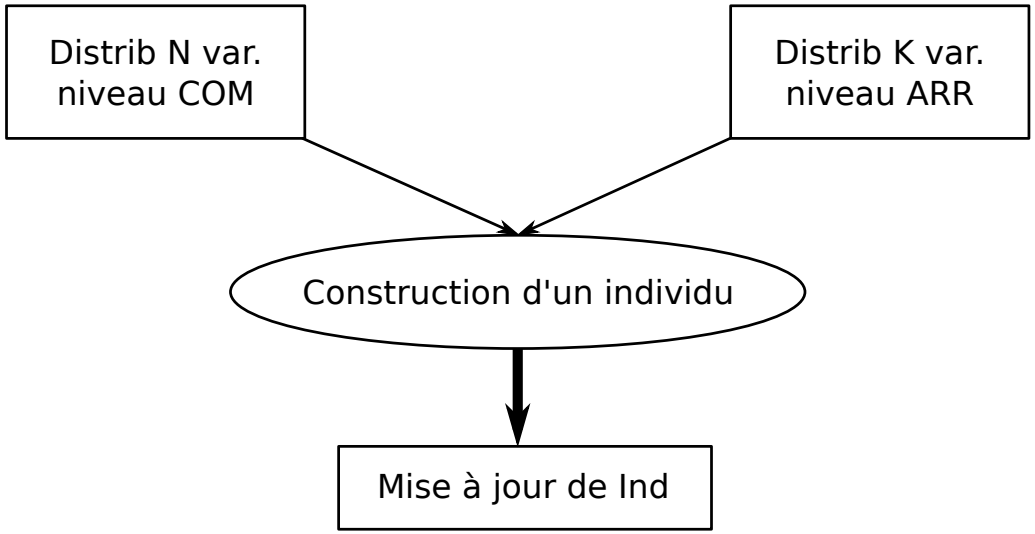
Génération de la population synthétique
Création du pool d'individus (1)

Etape 1 : Estimation

Individus caractérisés par $M = N + K$ variables :

$\Rightarrow$ tirage aléatoire

- de $N \leq M$ variables dans les distributions conjointes disponibles communal
- des K caractéristiques manquantes dans les distributions conjointes au niveau de l'arrondissement



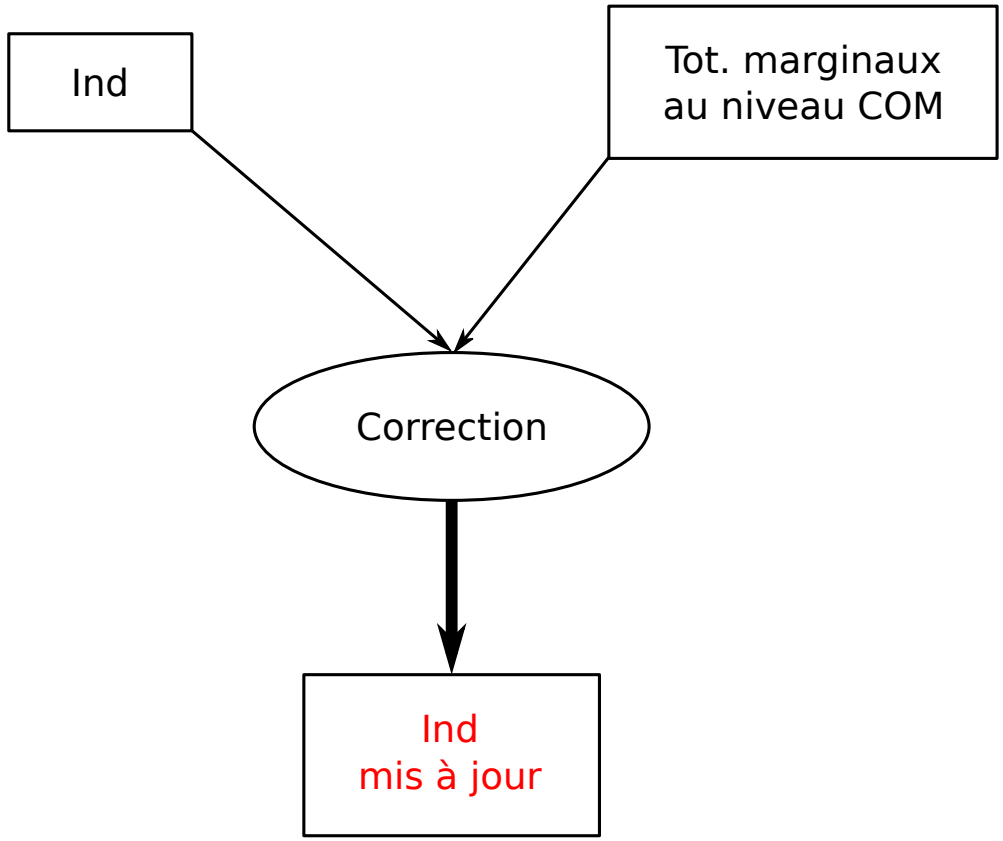
Génération de la population synthétique
Création du pool d'individus (2)

Etape 2 : Corrections

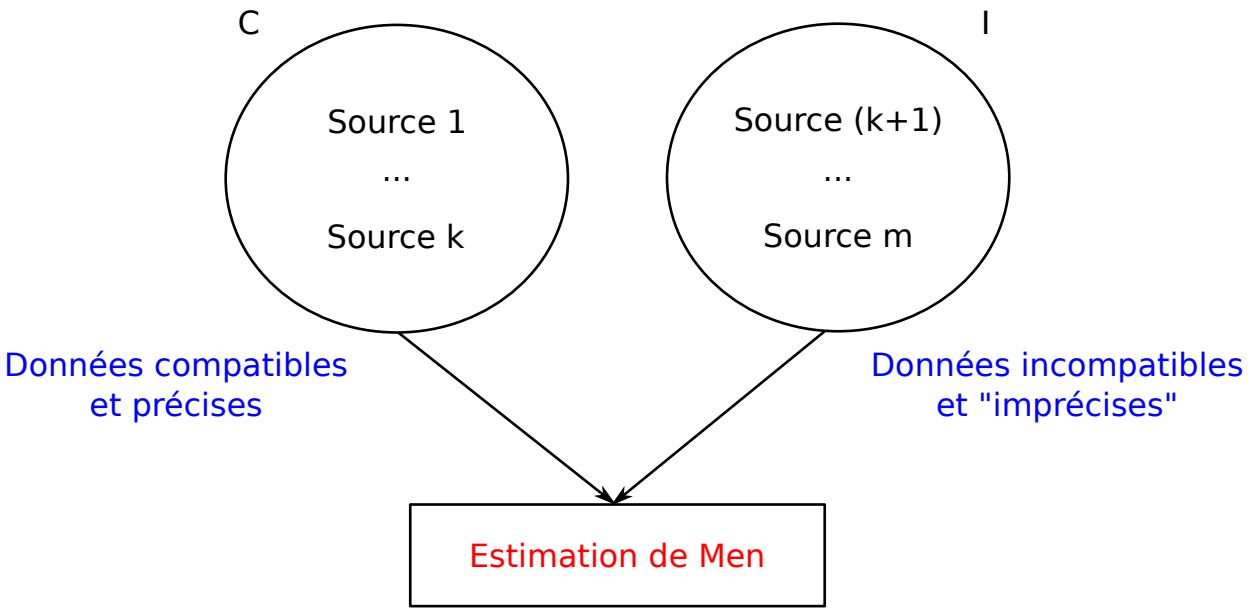
Totaux marginaux au niveau communal pour $L \leq K$ variables tirées au niveau de l'arrondissement.

$\Downarrow$

Modifications de variables pour certains individus afin de retrouver ces totaux marginaux.

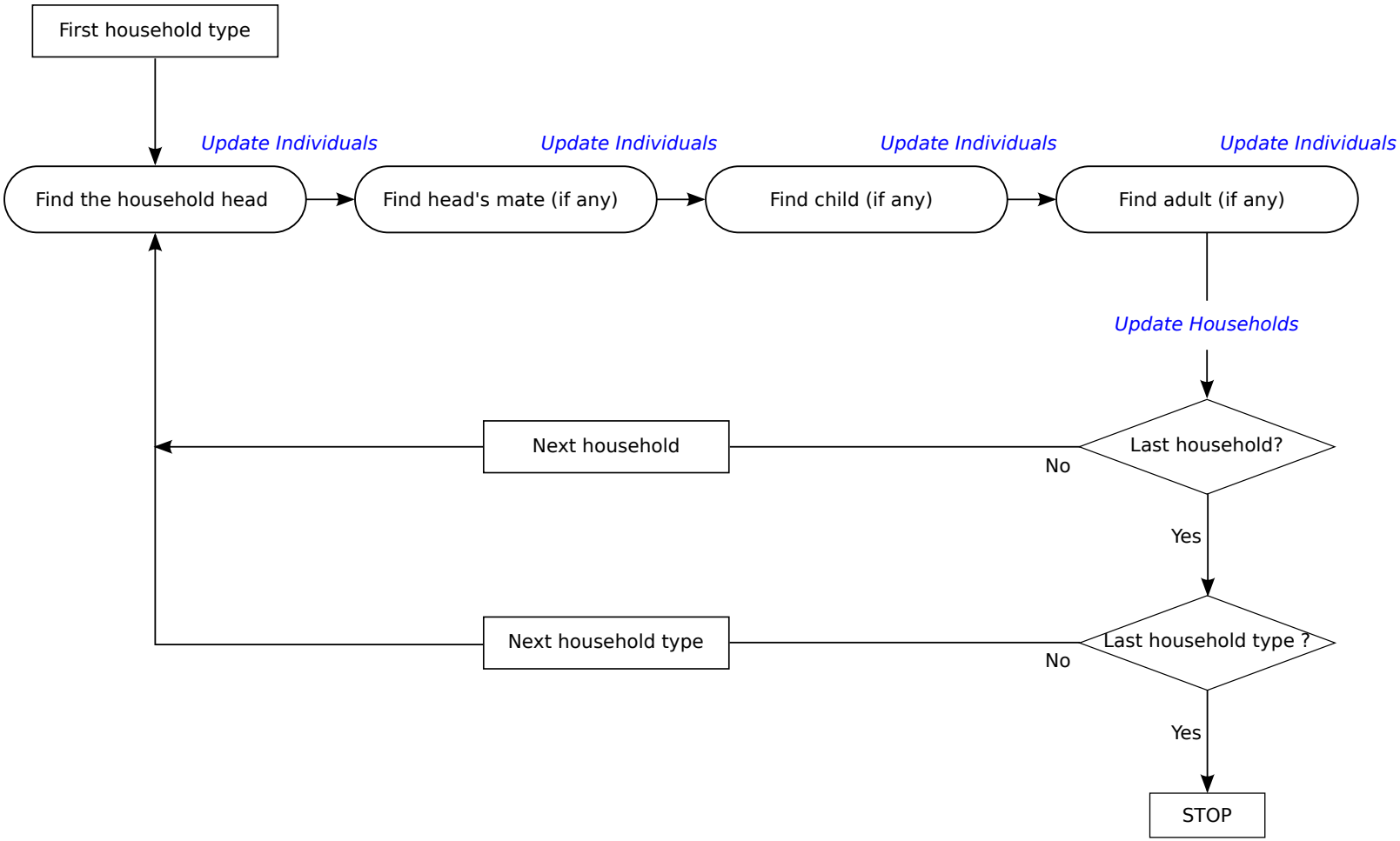


Générateur de populations synthétiques
Estimation de la distribution conjointe des ménages



- C et I : contraintes pour l'estimation de Men.
  mais problème sous-déterminé...
- Solution : optimisation d'un critère de qualité pour l'estimateur telle que C soit satisfait et I soit le plus satisfait possible.
- Critère choisi : maximisation de l'entropie, *i.e.* $-\sum_i (x_i \ln x_i - x_i)$.

Construction des ménages (étape 3)



Génération des menages synthétiques pour une commune

Choix des variables :

Individus et ménages caractérisés par des variables influençant la mobilité

Variable	Valeurs possibles
Sexe	H ; F
Classe d'age	0-5 ; 6-17 ; 40-59 ; 60+
Diplôme	aucun ; primaire ; secondaire ; supérieur
Status socio-pro	actif : inactif ; étudiant
Permis de conduire	oui ; non

Caractéristiques des individus

Application du générateur aux communes belges

Caractéristiques des ménages et communes

Variable	Valeurs possibles
Type	H célibataire seul F célibataire seul H avec enfant(s) (+ adultes) F avec enfant(s) (+ adultes) couple vivant seul couple avec enfant(s) (+ adultes)
Nombre d'enfants	0 à 5
Nombre d'adultes (conjoint non compris)	0 à 2

Caractéristiques des ménages

Variable	Valeurs possibles
Type d'urbanisation	urbain banlieue migratoire rural

Niveau d'aménagement du territoire par commune

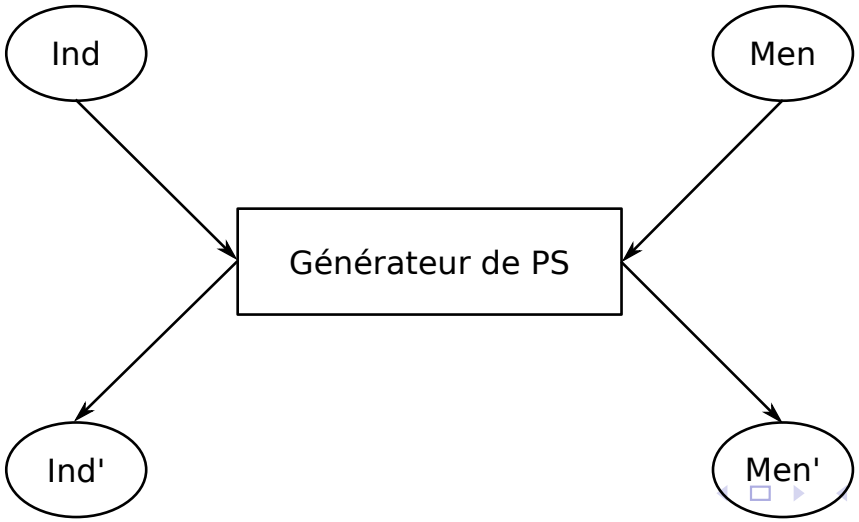
Application du générateur aux communes belges
Résultats

Erreur relative absolue (APD) :

$$APD(x, y) = \left| \frac{x - y}{x} \right|$$

où

- x = valeur désirée ;
- y = valeur approchée / estimée.



Application du générateur aux communes belges
Résultats (suite)

	Estimés	Générés	Différence	APD
Individus	11.060.573	10.638.112	422.461	0.039
Ménages	4.333.769	4.333.762	7	< 0.001

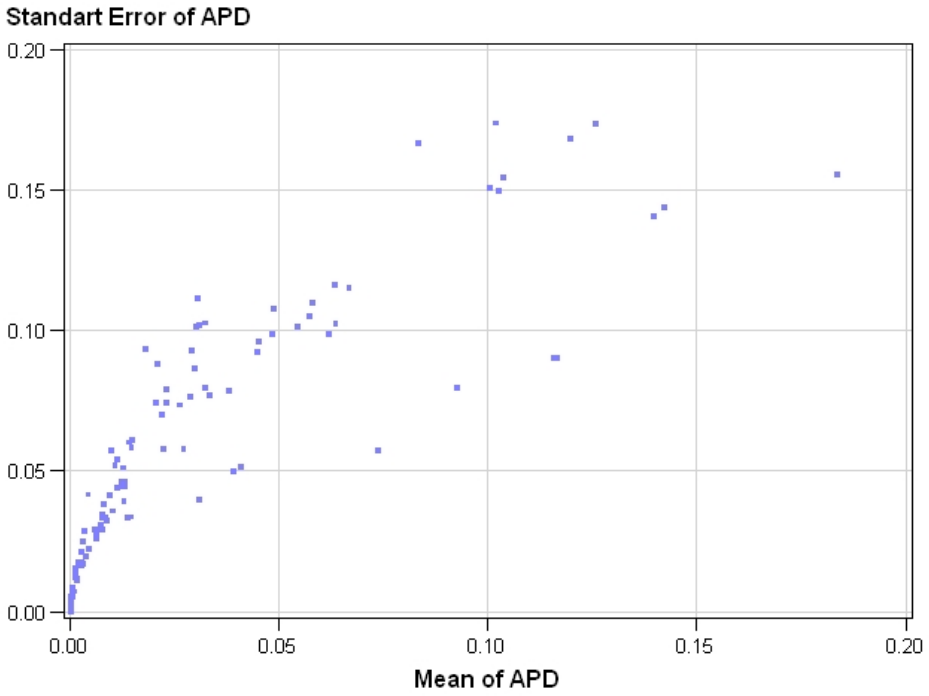
Population synthétique belge

Distribution	Min	Max	Ecart-type	Moyenne
Individus	0,000	0,065	0,013	0,020
Ménages	0,000	0,006	$< 0,001$	$< 0,001$

Statistiques de la moyenne des *APD* pour les types d'agents (*AAPD*)

$\Rightarrow$ Analyse plus fine des APD pour les individus

Application du générateur aux communes belges
Résultats (suite)



Moyenne et écart-type des APD pour chaque type d'individus

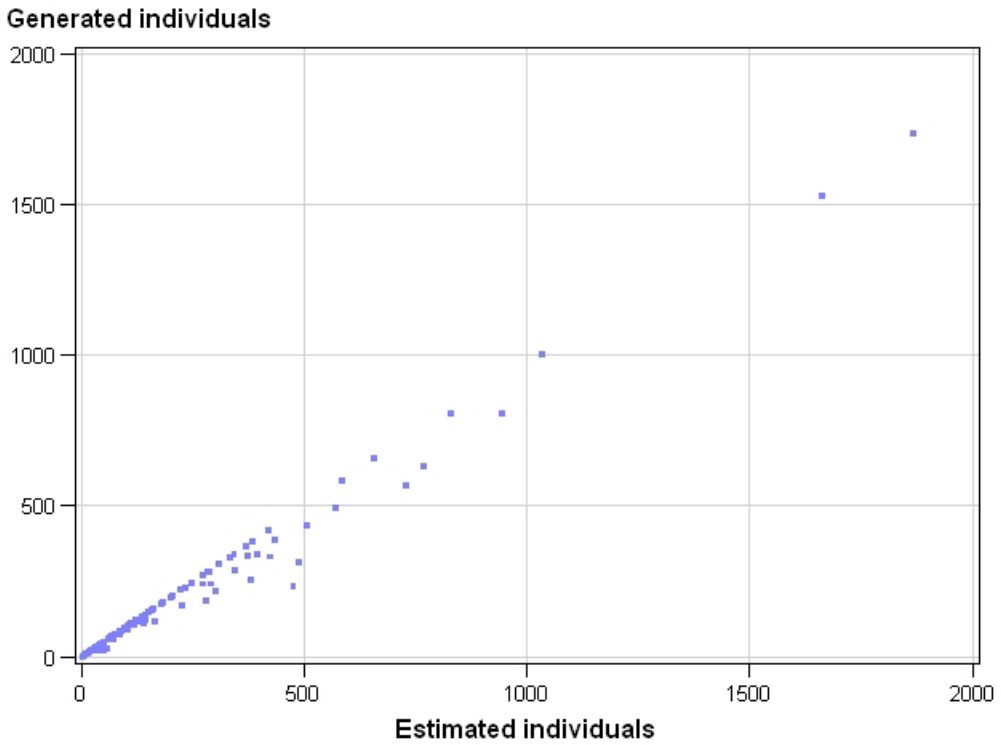
Erreur

- acceptable en moyenne ($< 20\%$)
- relativement stable (écart-type $< 20\%$)

Application du générateur aux communes belges

Résultats (suite) : le cas de Tubize (le plus défavorable)

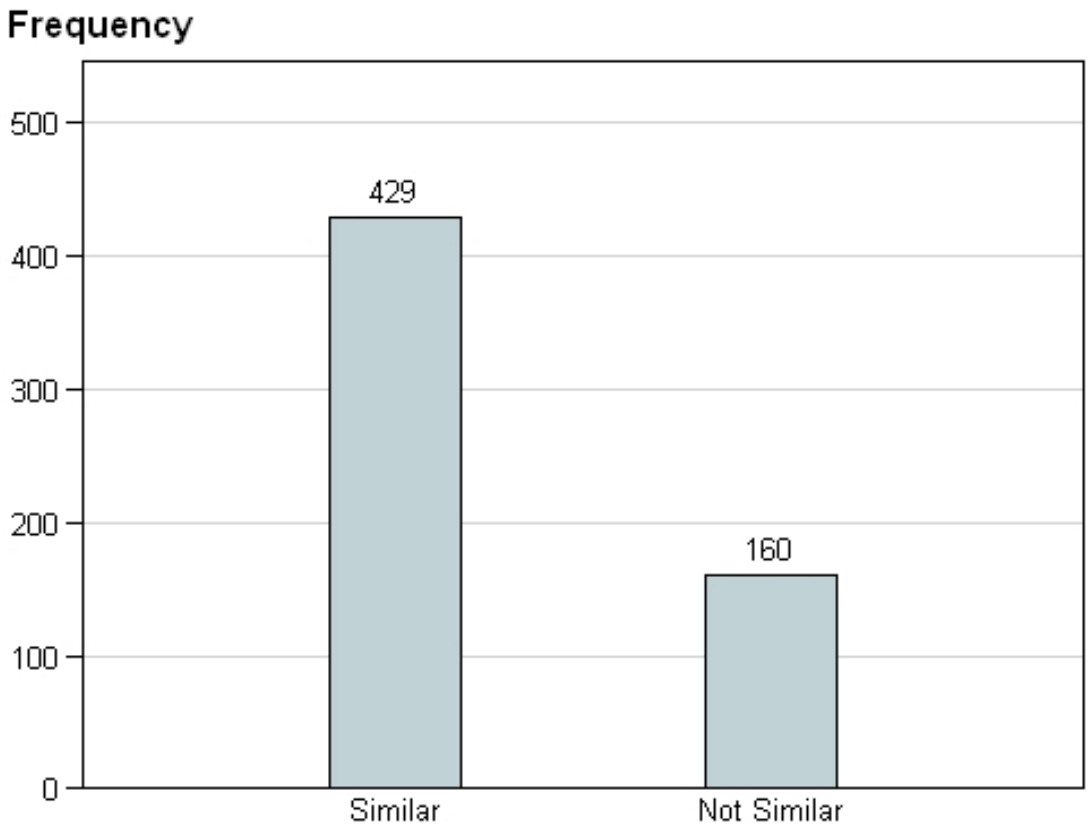
	Valeur
Estimés	24.441
Générés	22.170
Différence	2.271
APD(Est.,Gén.)	0,093
$AAPD(Ind')$	0,065



Agents de chaque type générés vs estimés

Application du générateur aux communes belges

Résultats (suite et fin) : Nombre de communes statistiquement similaires



Test du χ^2 entre *Ind* et *Ind'* ($\alpha = 0.05$)

$\Rightarrow$ $72,8\%$ de communes générées statistiquement similaires à leur estimation de départ.

Conclusions

Nouvelle approche de génération de populations synthétiques pour la fusion de données

- accepte données (modérement) incompatibles
- pas d'échantillon représentatif au niveau le plus désaggrégé
- permet l'incorporation de nouvelles variables
- respecte la vie privée
- requiert une mise à jour dans le temps
- premières expériences prometteuses...
- très nombreuses applications !

Merci de votre attention !