Consolidation de jeux de données pour la prospective : la génération d'une population synthétique pour les communes de Belgique

Philippe Toint (with J. Barthélemy)

Namur Center for Complex Systems (naXys), University of Namur, Belgium

(philippe.toint@fundp.ac.be)

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 $\rightarrow$ fortes)
- contraintes légales (protection de la vie privée)

Notre objectif : la population belge

Notre problème :

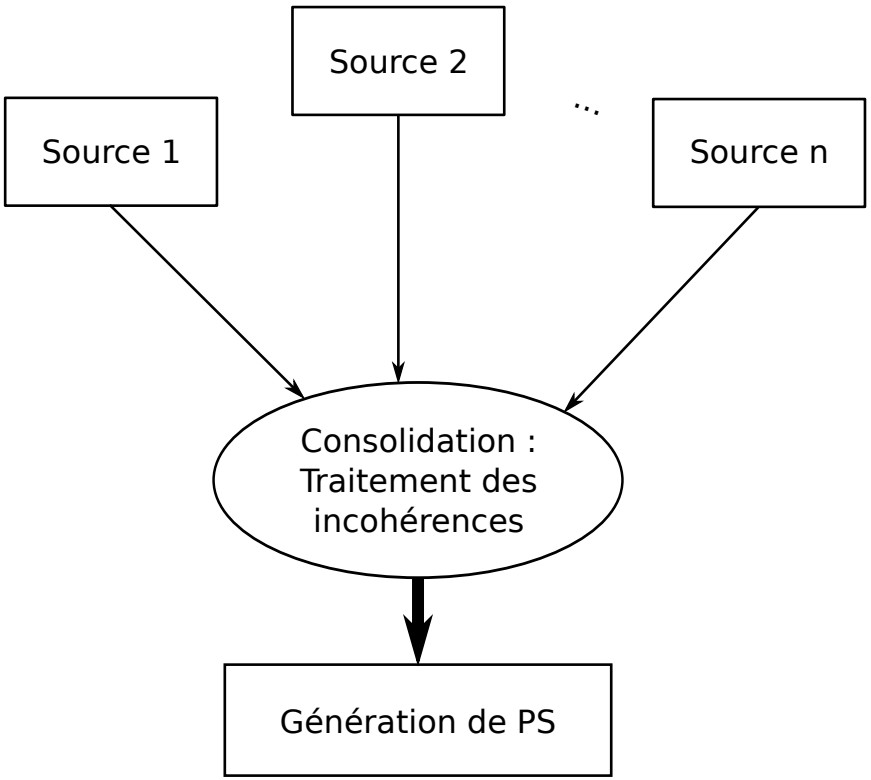
reconstruire une version synthétique de la population des 589 communes belges

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

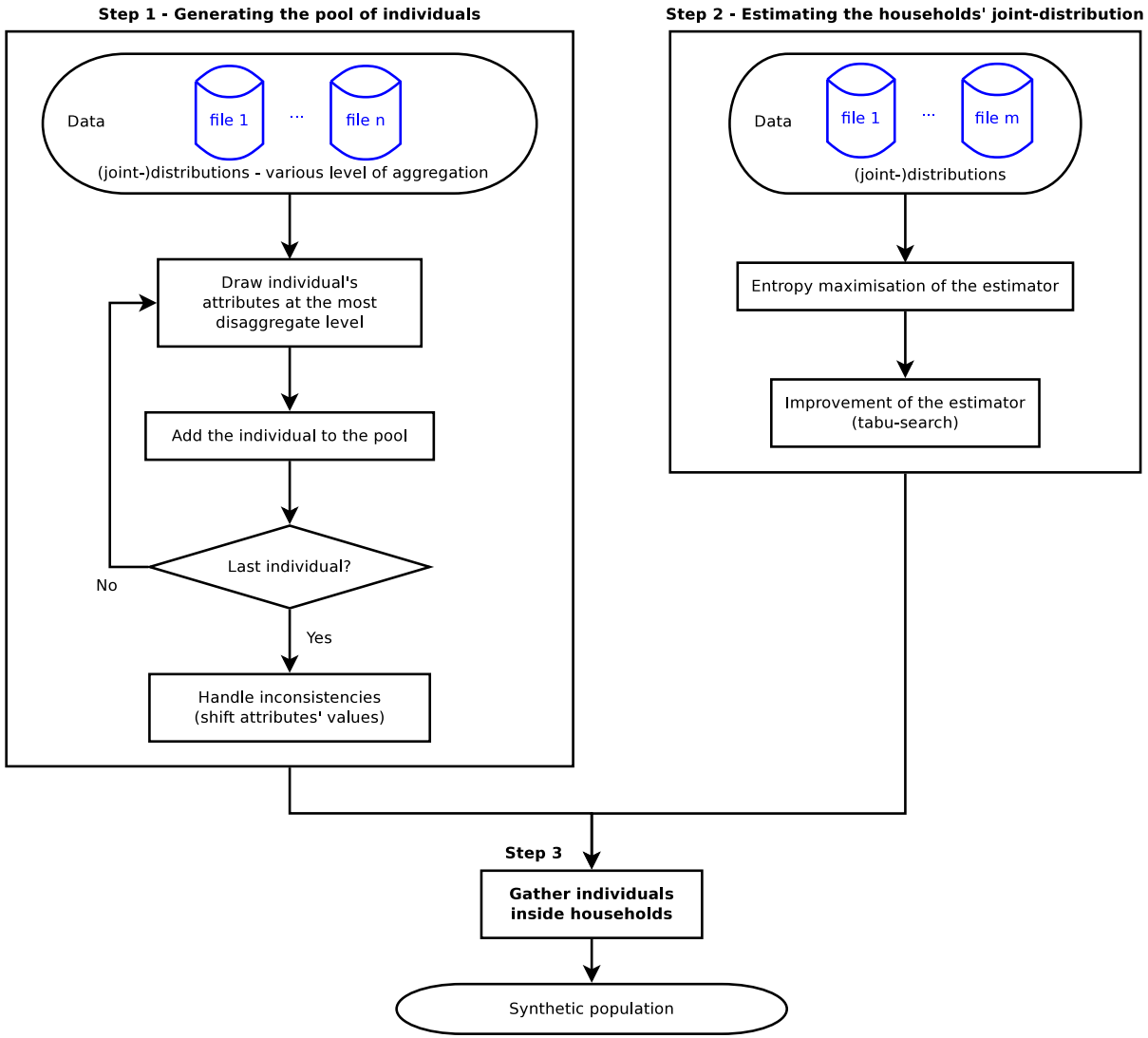
Construire les individus et ménages synthétiques en tirant aléatoirement les caractéristiques / membres dans les distributions adéquates au niveau le + désaggrégé disponible tout en préservant les corrélations connues.

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. ($\rightarrow$ *Ind*$'$ et *Men*$'$)

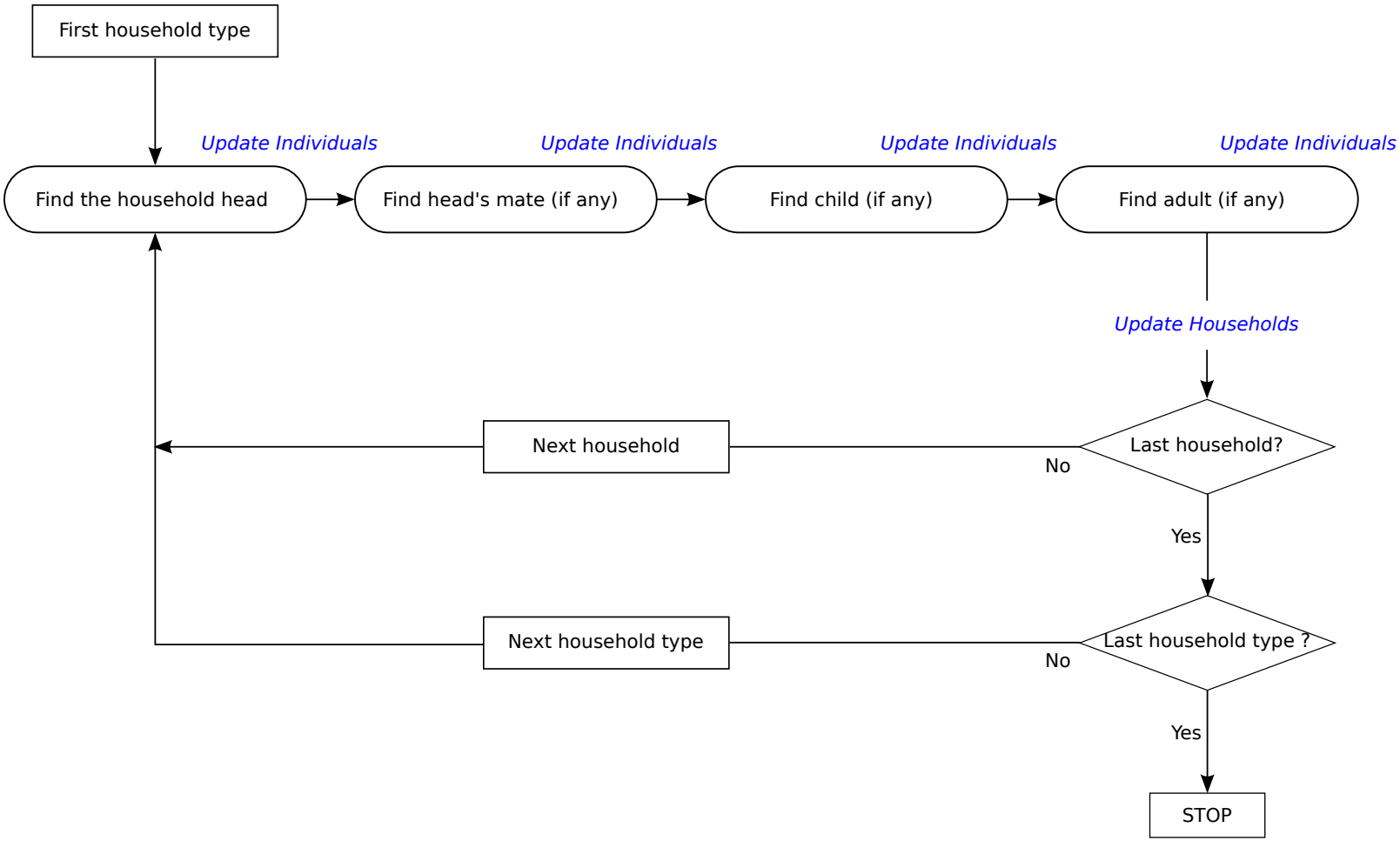
$\Rightarrow$ Etapes 1 et 2 : Consolidation des données disponibles

Générateur



Générateur de populations synthétiques

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 ; 18-39 ; 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	10.637.107	10.635.691	1.416	< 0.001
Ménages	4.334.281	4.333.448	933	< 0.001

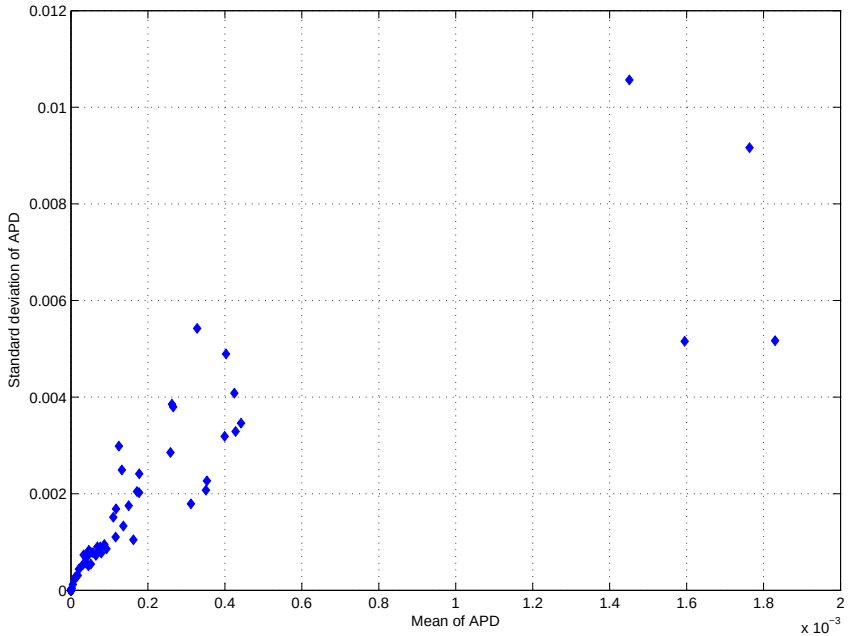
Population synthétique belge

Distribution	**Min**	**Max**	**Ecart-type**	**Moyenne**
Individus	0,000	0,005	0,013	0,020
Ménages	0,000	0,003	$< 0,001$	$< 0,001$

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

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

Application du générateur aux communes belges
Résultats (suite)

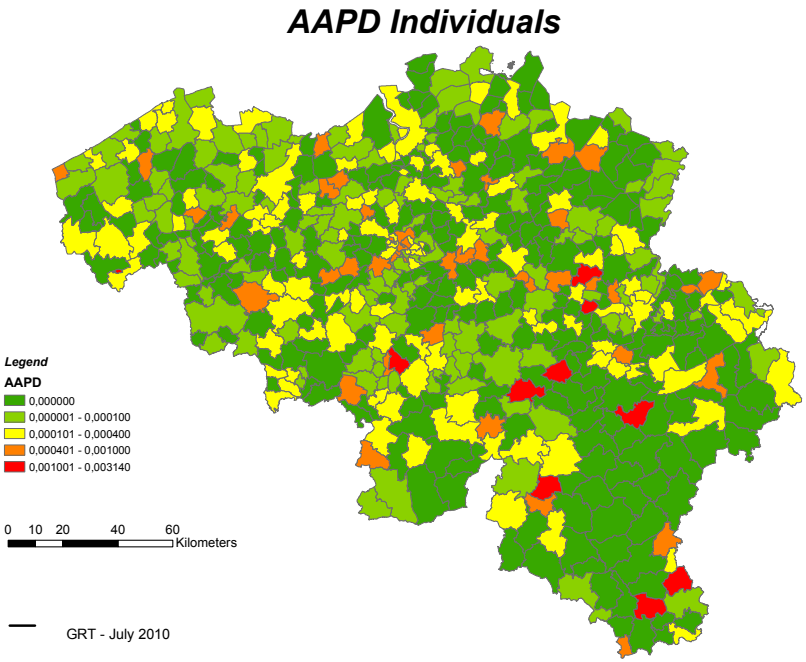


Moyenne et écart-type des APD pour chaque type d'individus

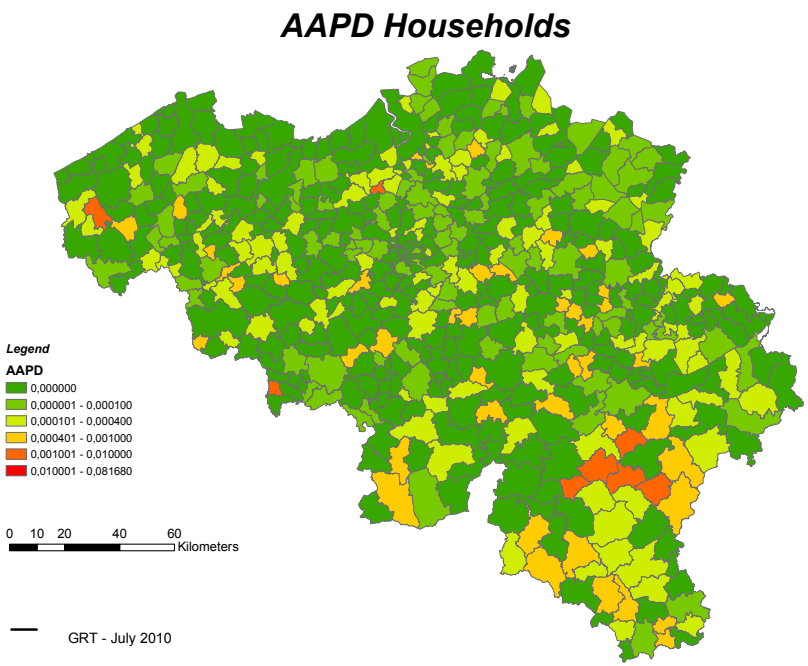
Erreur

- acceptable en moyenne ($< 0.2\%$)
- relativement stable (écart-type $< 1.5\%$)

AAPD

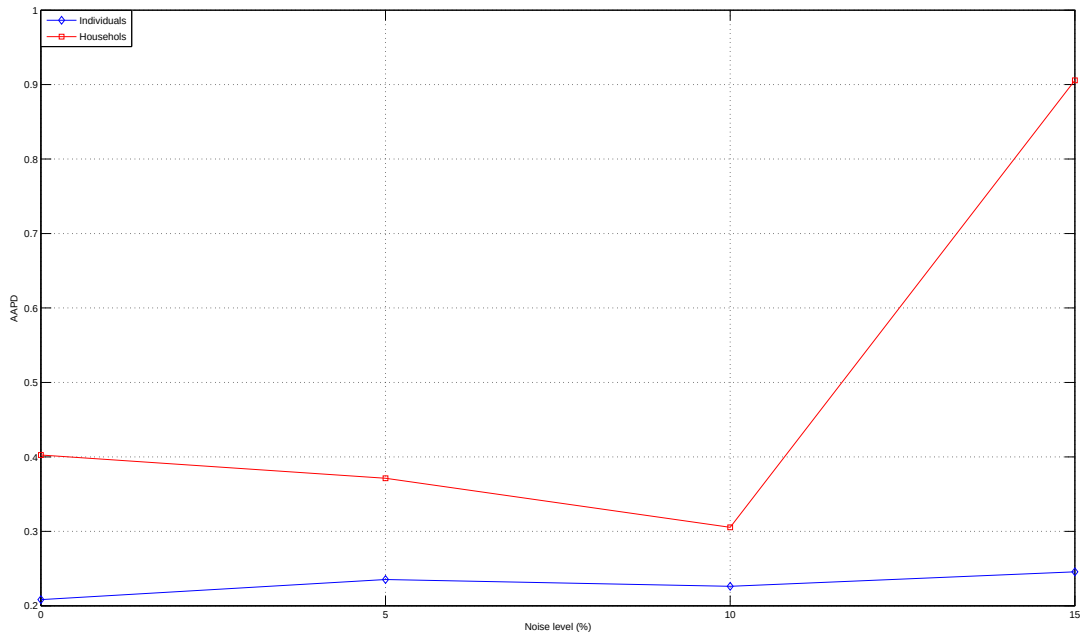


Individus



Ménages

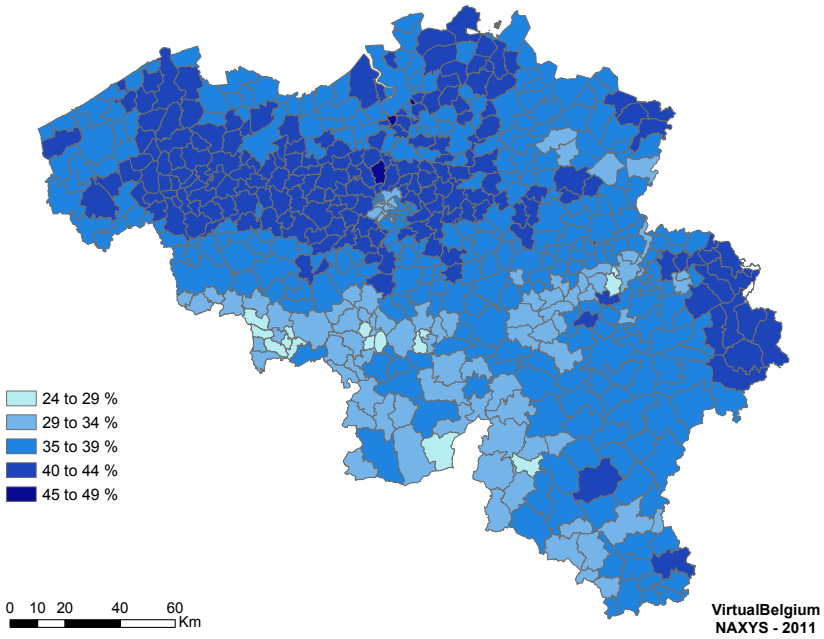
Analyse de sensibilité



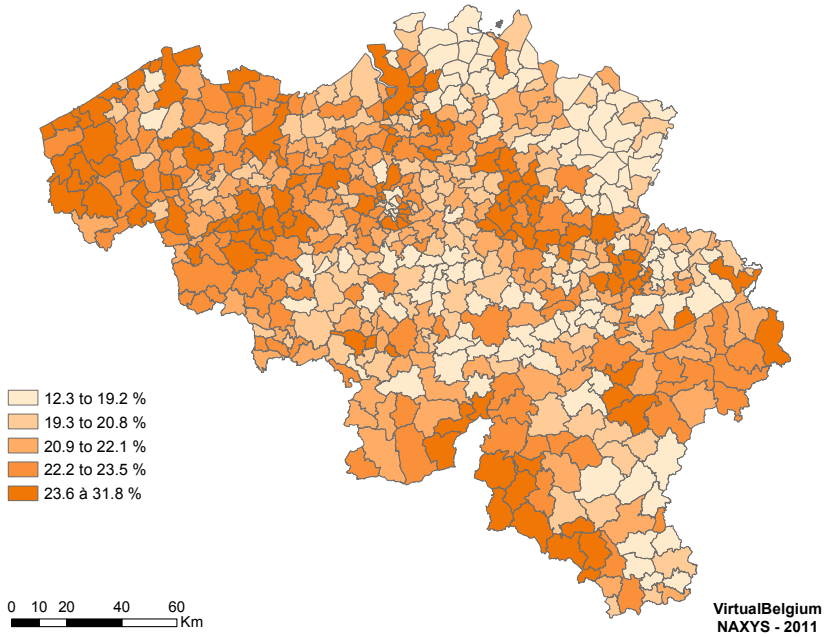
AAPD en fonction du degré d'incohérence des données

$\Rightarrow$ AAPD semble stable pour des incohérence $< 15\%$

Résultat : Virtual Belgium

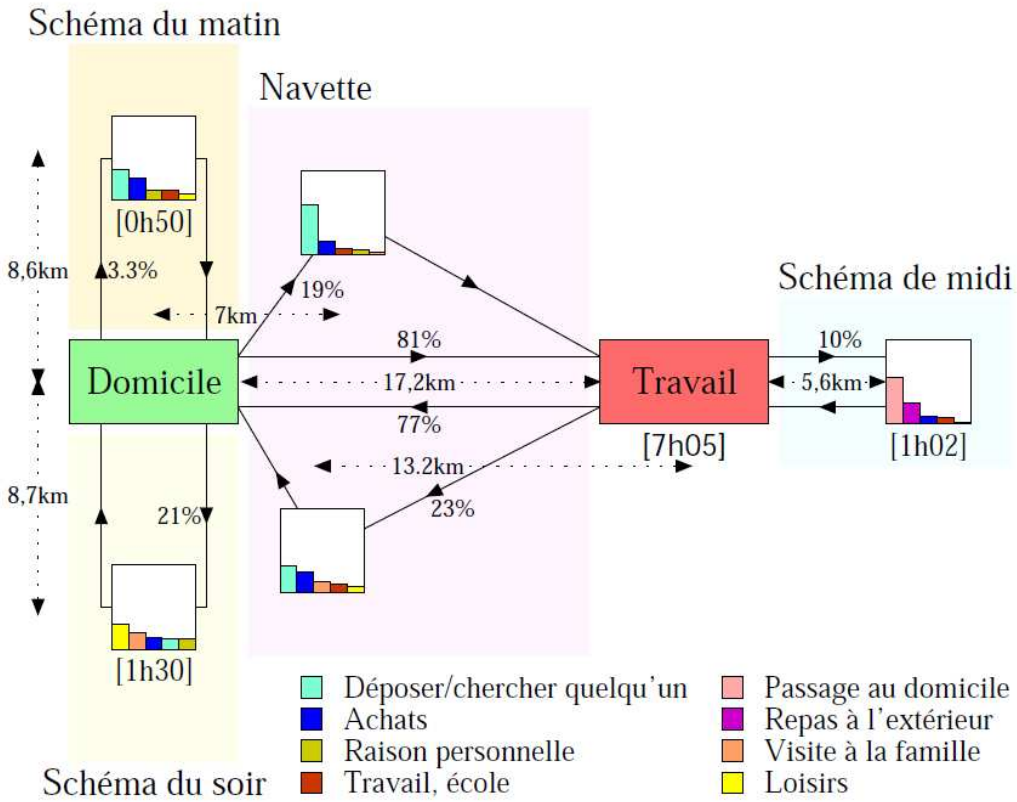


Actifs



60+ ans

Chaînes d'activités : Contexte



Chaines d'activité (MOBEL, 2001)

Déplacement = conséquence des activités

⇒ Assignation d'une chaine d'activités (CA) aux individus

Chaînes d'activités : Méthode

13 motifs

déposer / reprendre qq	aller a la maison	visite travail
travail	école	repas extérieur
courses	raison personnelle	visite famille / amis
se promener	loisir	autre

$\Rightarrow$ +/- 1500 combinaisons possibles

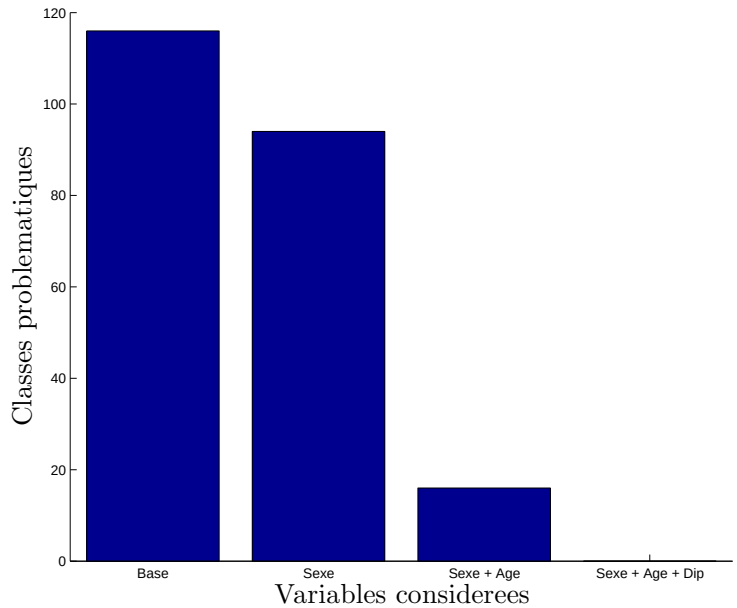
Méthode

1. générer un échantillon de minimum 5 CA pour chaque classe d'individus
2. tirage aléatoire d'une chaîne pour chaque individus
3. localisation des activités de la chaîne

Chaînes d'activités : Echantillon

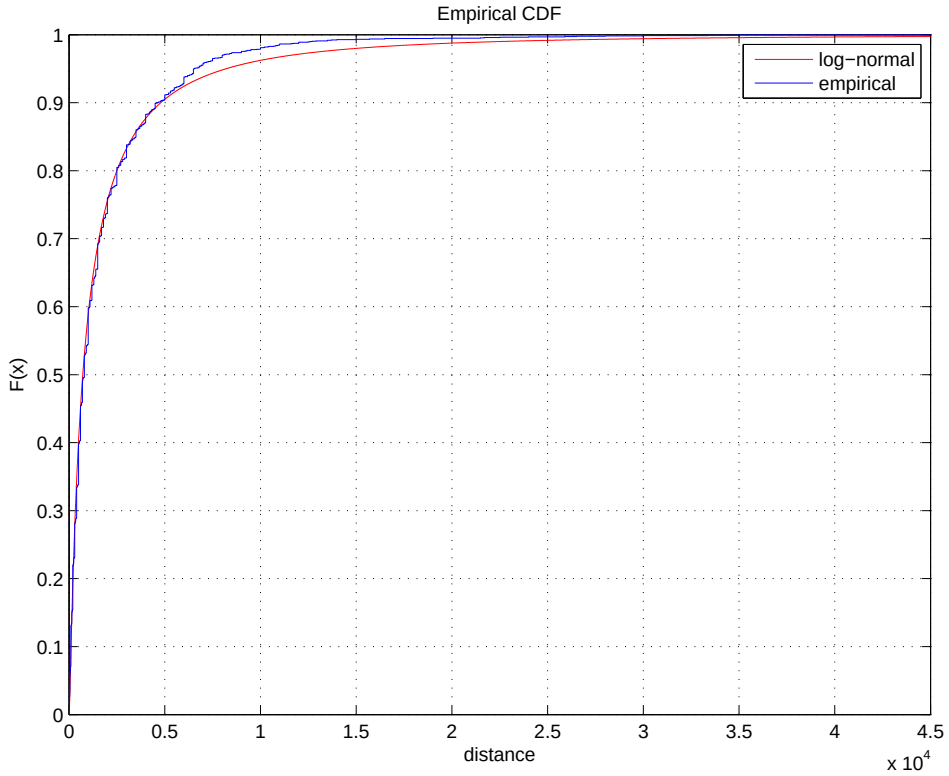
Génération de l'échantillon pour la classe I

1. Echantillon E_I tiré de MOBEL (2001)
2. Si $\#E_I < 5$ alors, CA tirées dans $E_{I'}$, avec $I' =$ classe voisine de I, obtenue en modifiant successivement les variables :
   - Sexe
   - Classe d'âge
   - Diplôme (uniquement $O \rightarrow P$ et $P \rightarrow S$)



Chaînes d'activités : Motif Travail

Activités caractérisées par une distance



Répartitions log-normale et empirique des distances

⇒ Bon fitting (conclusions similaires pour les autres motifs)

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...
- chaines d'activités
- très nombreuses applications !

Merci de votre attention !