Calcul de précision pour une région ayant procédé à une extension régionale et locale de l'enquête Logement 2006

Guillaume Chauvet

École Nationale de la Statistique et de l'Analyse de l'Information

Rencontres de Statistique Appliquée de l'INED
Plans de sondages complexes
20/01/2011

En résumé

L'Enquête Logement 2006 est une enquête auprès des ménages, qui a donné lieu à une extension régionale et à plusieurs extensions locales au niveau de la région Bretagne notamment. Un complément d'échantillon a également été sélectionné dans des bases externes.

Un plan de sondage et une technique d'estimation complexes ont été nécessaires pour la mise en commun et l'exploitation des différents sous-échantillons, la prise en compte de la non-réponse et le redressement des estimateurs.

Un outil SAS de calcul de précision basé sur les formules d'estimation de variance a été mis au point pour les partenaires de l'Enquête et les chargés d'étude de la Direction Régionale.

Présentation de l'enquête

Présentation

L'Enquête Logement est une des plus grosses enquêtes réalisées par l'Insee auprès des ménages. Elle a lieu environ tous les quatre ans (dernières éditions en 2002 et 2006).

Le champ est celui des logements résidences principales en 2006, accessibles à l'aide du RP99 et de la Base de Sondage de Logements Neufs (BSLN).

Objectifs de l'enquête :

- connaître le parc de logements (ancienneté de la construction, nombre de maisons individuelles/appartements, nombre de propriétaires/locataires,...),
- décrire les conditions de vie des ménages (mobilités et causes de mobilité, confort du logement, emprunts,...).

Sélection de l'Enquête Logement

L'échantillon est sélectionné en quatre temps :

- Sélection de l'échantillon national dans l'Echantillon Maître de 99 (RP99, BSLN),
- Sélection d'une extension régionale dans l'EMEX, pour les régions concernées,
- Sélection d'extensions d'échantillon au niveau local, pour les régions concernées,
- Sélection d'échantillons complémentaires dans des bases externes.

L'EM 99

L'échantillon maître de 1999 (EM99) est une réserve de logements destinée à servir de base de sondage pour les enquêtes auprès des ménages.

Il est obtenu par un tirage stratifié (selon le degré d'urbanisation) et à plusieurs degrés. On sélectionne des communes dans le rural, des districts (pâtés de maisons) dans l'urbain, ... (Ardilly, 2006).

Les échantillons destinés aux enquêtes ménages seront ensuite tirés dans les zones sélectionnées. Dans ces zones, une liste à jour de logements est fournie par le RP99 et la BSLN.

L'EMEX

Pour les extensions régionales, il existe un échantillon-maître spécifique : l'EMEX (Bourdalle et al., 2000), constitué selon des principes voisins de l'EM :

- tirage stratifié (selon le degré d'urbanisation) et à plusieurs degrés,
- même système de rotation des logements situés dans l'EMEX,
- disjonction par rapport à l'EM.

A quoi servent les extensions ? La précision est liée à la taille d'échantillon : plus le domaine d'estimation est petit, plus la précision se détériore. La sélection d'un échantillon dans l'EMEX vise à assurer de meilleures estimations au niveau régional.

Extensions locales d'échantillon

Un complément d'échantillon peut être sélectionné autour de zones particulières pour lesquelles on souhaite produire des estimations fiables.

En Bretagne, c'est le cas des 6 principales aires urbaines (Brest, Lorient, Quimper, Rennes, Saint-Brieuc et Vannes).

Cet échantillon est sélectionné en excluant les logements précedemment échantillonnés dans l'EM ou l'EMEX au titre de l'Enquête Logement.

Bases externes

Enfin, pour surreprésenter des sous-populations particulières, des échantillons ont été sélectionnés dans des fichiers externes :

- Base des adresses situées dans les Zones Urbaines Sensibles (ZUS),
- Base d'allocataires de prestations.

Cette sélection n'est pas disjointe de celle des autres sous-échantillons.

Au niveau de la Bretagne, seul l'échantillon ZUS a été utilisé (fusion des autres échantillons très problématique).

Schéma récapitulatif

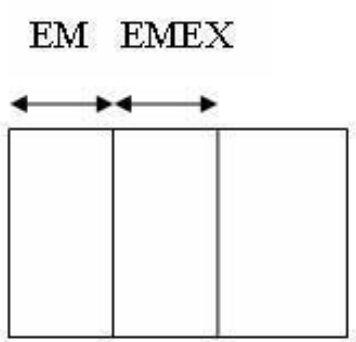


Schéma récapitulatif

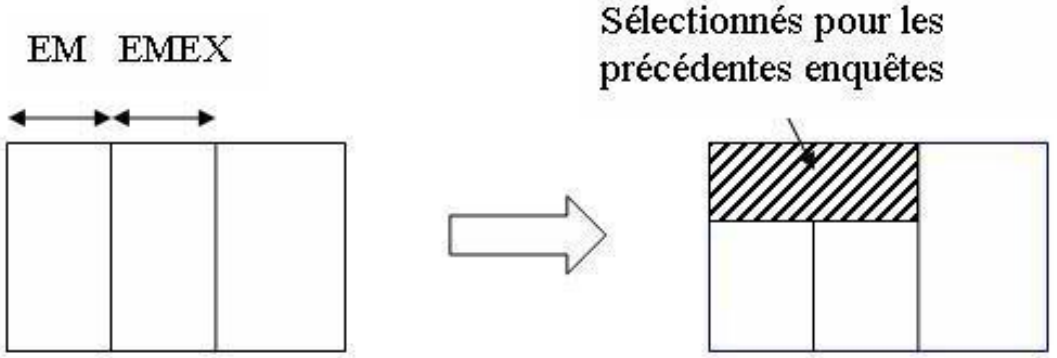


Schéma récapitulatif

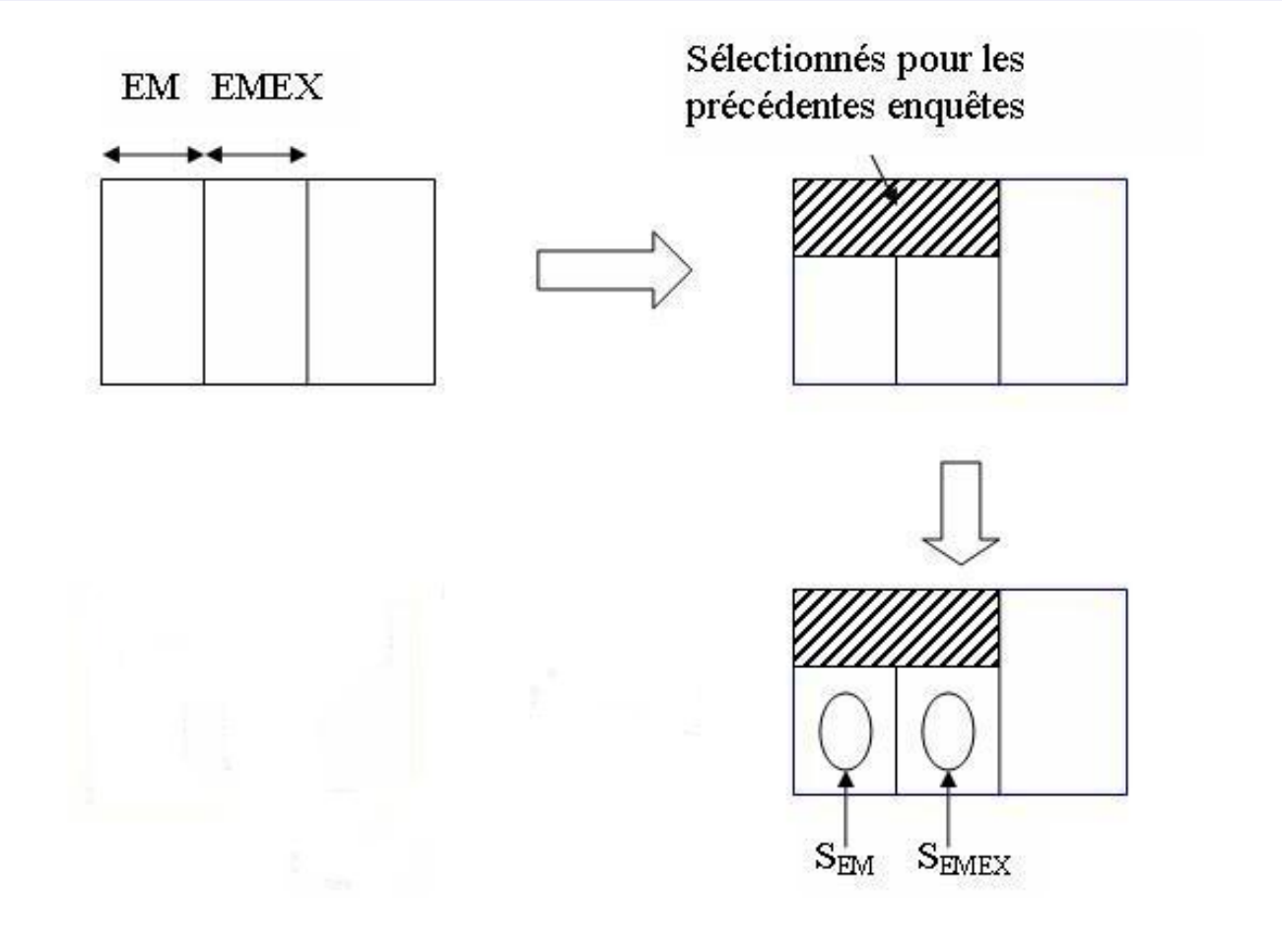


Schéma récapitulatif

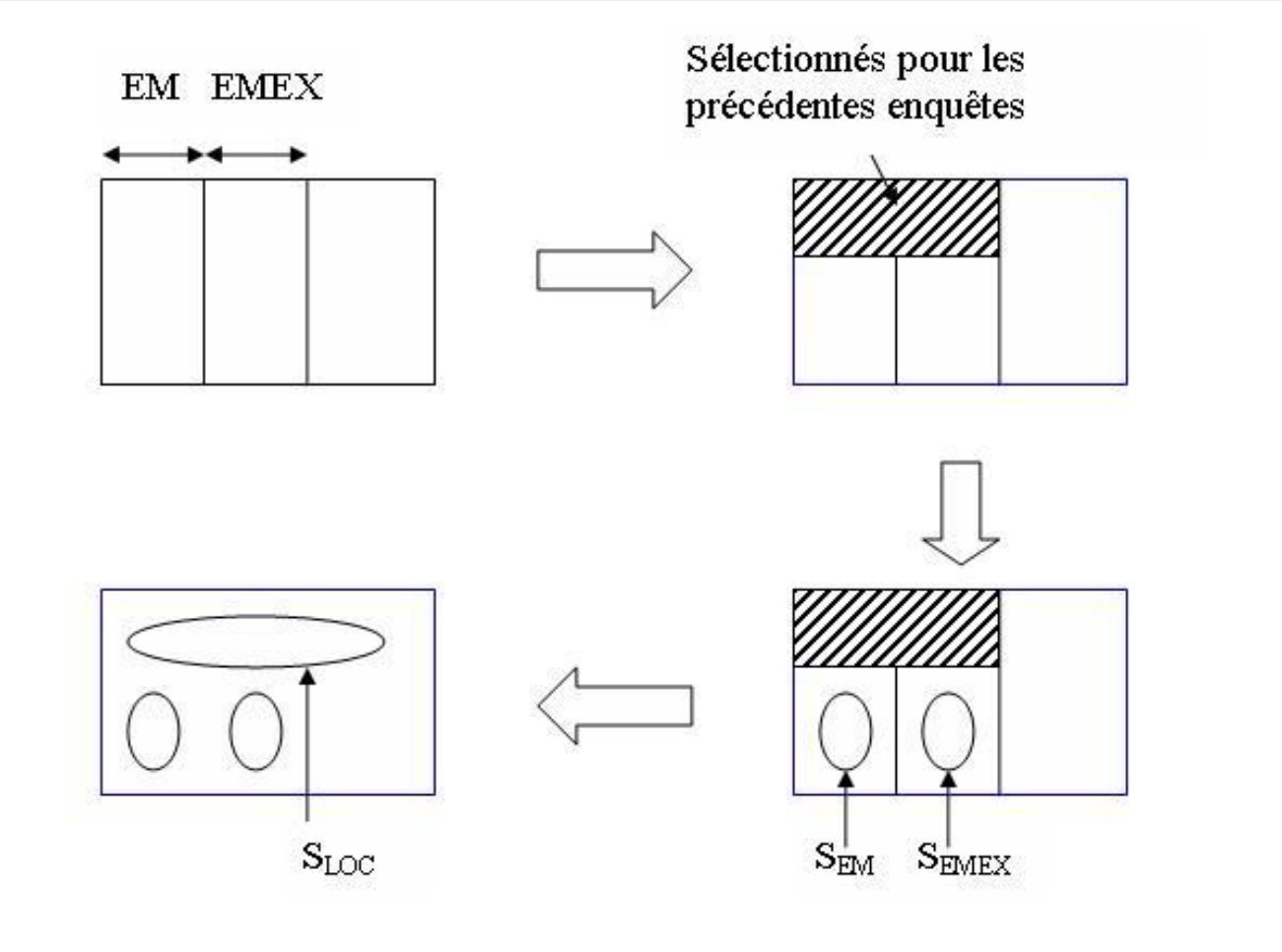


Schéma récapitulatif

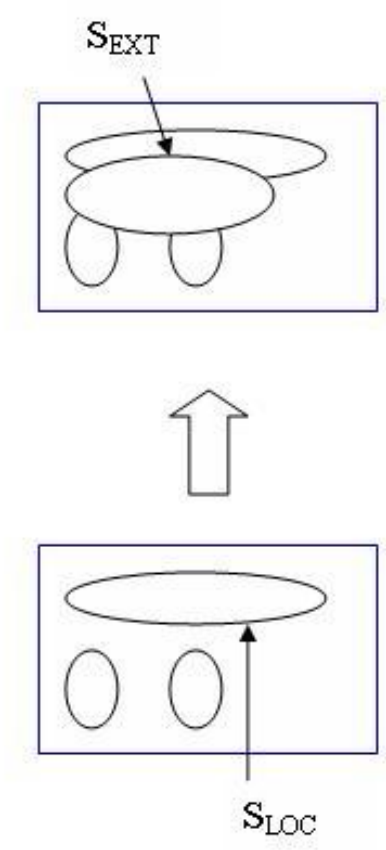
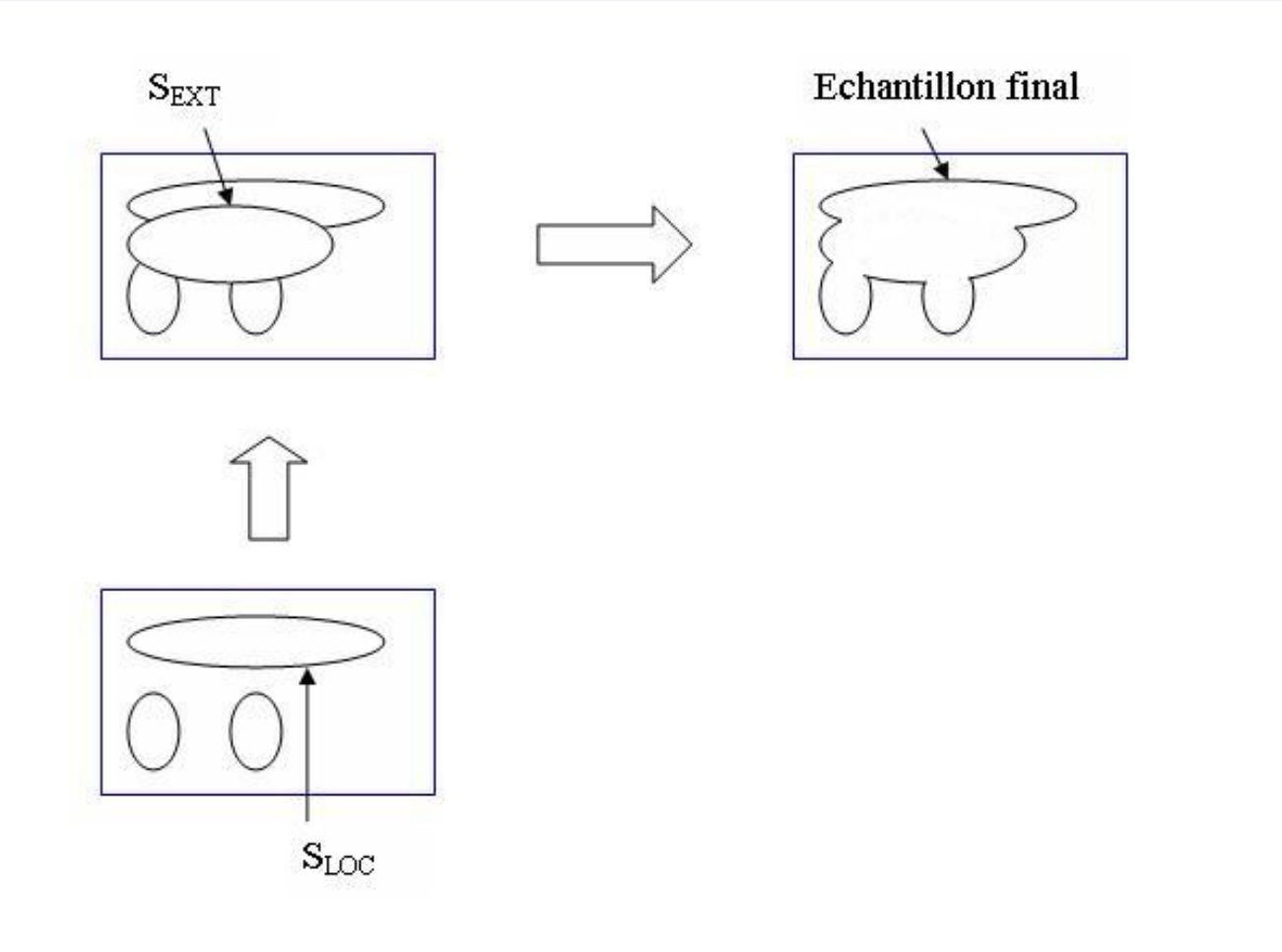


Schéma récapitulatif



Mise en commun des sous-échantillons

La difficulté consiste à produire un estimateur sans biais en gérant les trois sous-échantillons et leur intersection. Il y a essentiellement deux problèmes :

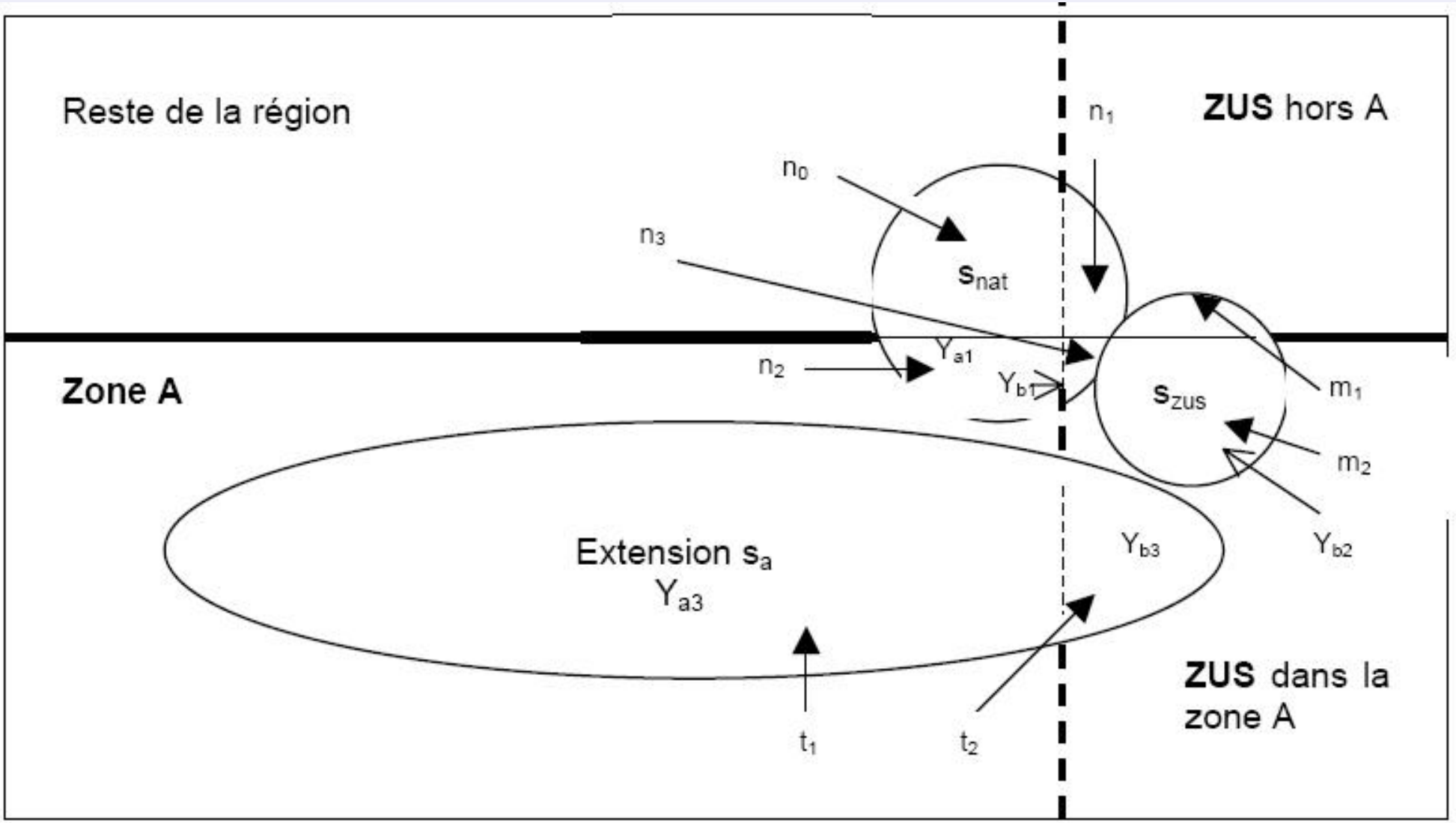
- Des sous-échantillons même disjoints peuvent représenter une même sous-population. On a donc un risque de biais dû à des doubles ou triples comptes.
- Certains logements sont sélectionnés dans deux échantillons différents.

Les sous-échantillons sont mis en commun à l'aide de la technique d'estimation composite. On note

$$\hat{t}_{y,d} = \sum_{k \in S} d_k y_k$$

l'estimateur obtenu, avec S la réunion des sous-échantillons et d_k le poids du logement k.

Schéma récapitulatif (Le Guennec, 2009)



Estimation de précision

Les étapes de l'enquête

Les différentes étapes de traitement de l'enquête (Le Guennec, 2009) sont :

1. Sélection des sous-échantillons,
2. Mise en commun des sous-échantillons,
3. Redressement de la non-réponse partielle (imputation),
4. Redressement de la non-réponse totale (méthode des groupes homogènes de réponse),
5. Calage sur une information externe.

L'estimation de variance réalisée ne prend pas en compte la variance d'imputation.

Calcul de variance (1)

On note

$$\hat{t}_{y,w} = \sum_{k \in S} w_k y_k$$

l'estimateur obtenu avec les poids calés w_k. On a :

$$V\left[\hat{t}_{y,w}\right] \simeq V\left[\hat{t}_{e,d}\right]$$

avec e_k le résidu de régression de y sur les variables de calage.

Cette variance se décompose en

$$V\left[\hat{t}_{e,d}\right] = V_p\left[\hat{t}_{e,d}\right] + V_{nr}\left[\hat{t}_{e,d}\right],$$

où la variance due à l'échantillonnage $V_p[\cdot]$ et la variance due à la non-réponse $V_{nr}[\cdot]$ sont estimées séparément.

Calcul de variance (2)

L'estimateur $\hat{t}_{e,d}$ peut se décomposer sur les trois sous-échantillons tirés nationalement (N), localement (L) ou dans les ZUS (Z) :

$$\hat{t}_{e,d} = \hat{t}^N_{\tilde{e},d} + \hat{t}^L_{\tilde{e},d} + \hat{t}^Z_{\tilde{e},d}$$

avec $\tilde{e}_k = f(e_k)$ la variable synthétique associée à la technique d'estimation composite.

En utilisant l'indépendance (réelle ou approchée) des trois sous-échantillons :

$$V_p\left[\hat{t}_{e,d}\right] \simeq V_p\left[\hat{t}^N_{\tilde{e},d}\right] + V_p\left[\hat{t}^L_{\tilde{e},d}\right] + V_p\left[\hat{t}^Z_{\tilde{e},d}\right].$$

Un exemple de calcul de précision

On souhaite estimer, sur l'ensemble des résidences principales de l'aire urbaine de Rennes :

- La structure des logements selon le nombre de chambres,
- La surface moyenne par habitant.

On est dans le cas d'une estimation sur un **domaine** D, c'est à dire sur une sous-population de U. Cette estimation ne pose pas de problème particulier, en remarquant que

$$t_{yD} = \sum_{k \in D} y_k = \sum_{k \in U} y_k \ 1_{k \in D}$$

où $1_{k \in D}$ vaut 1 si le logement k est dans le domaine, et 0 sinon.

On va donc estimer :

- des effectifs (nombre de logements ne comptant aucune chambre, comptant 1 chambre, ...),
- un ratio (surface totale rapportée au nombre d'habitants).

Dans le premier cas, on estime un total (variable indicatrice pour un effectif).

Dans le second cas, on estime un ratio de totaux : l'estimation de variance se fait par linéarisation (Deville, 1999).

Résultats obtenus

Paramètre	Estim.	Var.	CV (%)	BI (95%)	BS (95%)	DEFF	DCAL	NR (%)
Surf. moy.	$38,27$	$0,23$	$1,25$	$37,34$	$39,21$	$0,48$	$0,42$	$21,79$
% Log.								
0 cha.	$0,08$	$4,3\ 10^{-5}$	$7,75$	$0,07$	$0,10$	$0,46$	$0,99$	$17,34$
1 cha.	$0,18$	$1,1\ 10^{-4}$	$5,69$	$0,16$	$0,20$	$0,59$	$0,90$	$21,80$
2 cha.	$0,26$	$2,1\ 10^{-4}$	$5,68$	$0,23$	$0,29$	$0,91$	$0,96$	$19,15$
3 cha.	$0,26$	$3,0\ 10^{-4}$	$6,64$	$0,23$	$0,29$	$1,26$	$0,98$	$18,06$
4 cha.	$0,18$	$1,7\ 10^{-4}$	$7,43$	$0,15$	$0,20$	$0,97$	$0,80$	$18,82$
5 cha.	$0,04$	$8,5\ 10^{-5}$	$23,6$	$0,02$	$0,06$	$1,86$	$0,97$	$19,96$
6 cha.	$3\ 10^{-3}$	$1,7\ 10^{-6}$	$48,5$	10^{-4}	$5,2\ 10^{-3}$	$0,52$	$1,01$	$17,36$
+ 6 cha.	$4\ 10^{-4}$	$3,3\ 10^{-7}$	152	-10^{-3}	$1,5\ 10^{-3}$	$0,72$	$1,01$	$17,21$

Résultats obtenus

Si on se fixe (par exemple) un seuil à 10% pour le coefficient de variation, la précision est insuffisante pour les logements comportant un grand nombre de chambres (peu de logements correspondant $\Rightarrow$ imprécision importante).

On va donc réaliser un regroupement pour définir une variable à 5 modalités :

- Logements ne comportant pas de chambre,
- Logements comportant 1 chambre,
- Logements comportant 2 chambres,
- Logements comportant 3 chambres,
- Logements comportant 4 chambres et plus.

Résultats obtenus (suite)

Paramètre	Estim.	Var.	CV (%)	BI (95%)	BS (95%)	DEFF	DCAL	NR (%)
Surf. moy.	$38,27$	$0,23$	$1,25$	$37,34$	$39,21$	$0,48$	$0,42$	$21,79$
% Log.								
0 cha.	$0,08$	$4,3\ 10^{-5}$	$7,75$	$0,07$	$0,10$	$0,46$	$0,99$	$17,34$
1 cha.	$0,18$	$1,1\ 10^{-4}$	$5,69$	$0,16$	$0,20$	$0,59$	$0,90$	$21,80$
2 cha.	$0,26$	$2,1\ 10^{-4}$	$5,68$	$0,23$	$0,29$	$0,91$	$0,96$	$19,15$
3 cha.	$0,26$	$3,0\ 10^{-4}$	$6,64$	$0,23$	$0,29$	$1,26$	$0,98$	$18,06$
+ 4 cha.	$0,22$	$2,0\ 10^{-4}$	$6,49$	$0,19$	$0,25$	$0,97$	$0,81$	$19,94$

Références

- Ardilly, P. (2006), *Les techniques de Sondage*, Technip
- Bourdalle, G., Christine, M., Wilms, L. (2000), *Echantillons maître et emploi*, Journées de Méthodologie Statistique.
- Chauvet, G. (2009), *Calcul de précision pour une région ayant procédé à des extensions régionale et locale de l'Enquête logement 2006*, Journées de Méthodologie Statistique.
- Deville, J-C. (1999), *Variance estimation for complex statistics and estimators : linearization and residual techniques*, Survey Methodology, 25, p. 193-204.
- Le Guennec, J. (2009), *Enquête Logement 2006, des extensions d'échantillon pour les estimations locales : échantillonnage et redressement*, Journées de Méthodologie Statistique.

Pour finir ...

7E COLLOQUE FRANCOPHONE SUR LES SONDAGES

26 > 28 JUIN 2012
Rennes | Bretagne | France