Simulations de tirages de zones d'action pour les enquêtes-ménages de l'Insee

UMS, Insee, Juin 2007

Fabien Guggemos

Université de Neuchâtel, Suisse

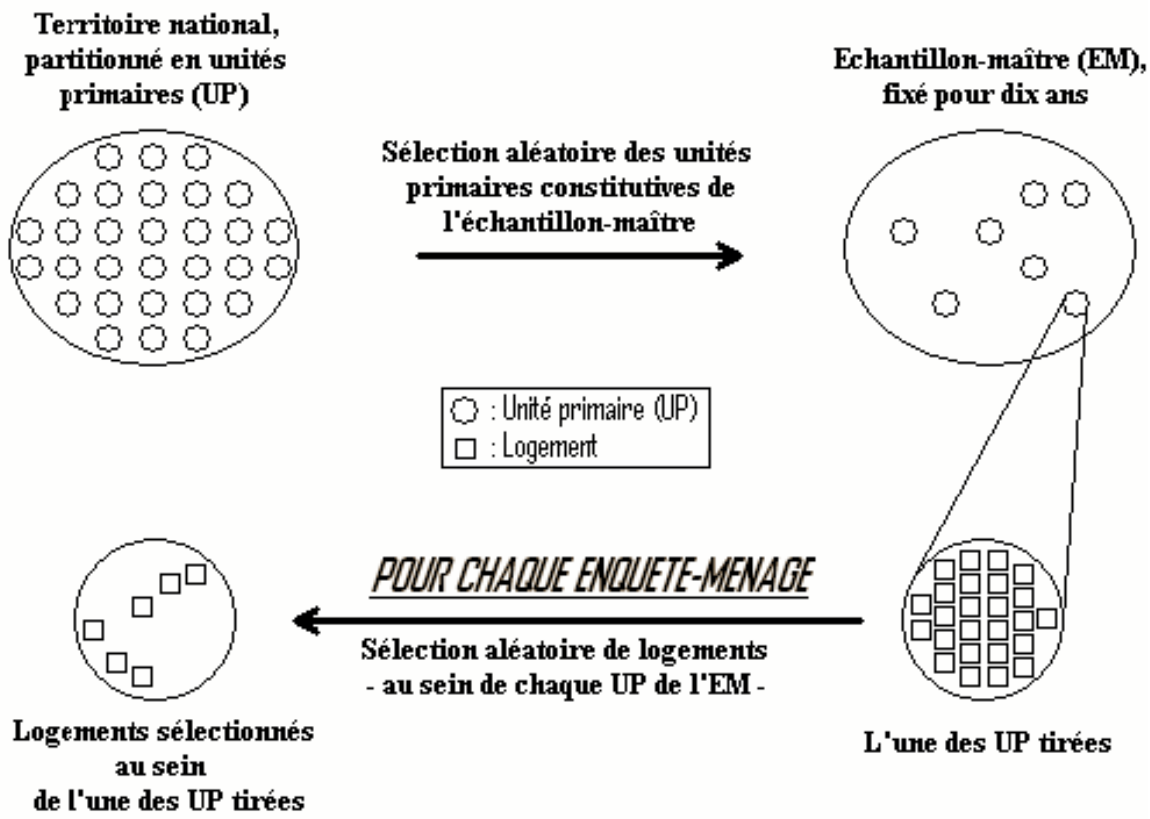
Journées de Méthodologie Statistique de l'Insee 2009
24 mars 2009

Lignes directrices

Lignes directrices

Echantillonnage pour les enquêtes-ménages de l'Insee

UP = Communes ou regroupements de communes



Le nouveau recensement (depuis 2004)

- Petites Communes (PC) (<10000 habitants) : réparties en 5 groupes de rotation ; 1 groupe recensé chaque année.
- Grandes Communes (GC) (>10000 habitants) : enquête annuelle de recensement par sondage, disjonction des échantillons enquêtés sur un cycle de 5 ans.

Création des Zones Action Enquêteurs (ZAE)

Unités primaires reconstruites pour contenir systématiquement des logements recensés l'année précédente.

Quelques chiffres sur les ZAE

Création (UMS, automne 2006) d'un total de 3743 ZAE, dont :

- 850 ZAEGC. 1 ZAEGC = 1 grande commune
- 2893 ZAEPC. 1 ZAEPC = regroupement d'au moins une commune de chaque groupe de rotation

Question

Comment sélectionner les ZAE pour la constitution du nouvel échantillon-maître, opérationnel de 2009 à 2019 ?

Questions suscitées par la construction des ZAE

1. Comment assurer le tirage d'un échantillon qui soit représentatif pour chaque année du cycle de rotation ?
2. La qualité d'un échantillon tiré se dégrade-t-elle au fil des années ?
3. Comment assurer le tirage d'un échantillon représentatif des différents types d'espace, rural, périurbain et urbain ?
4. Utiliser des plans de sondage distincts d'une région à l'autre ?

Questions suscitées par la construction des ZAE

1. Comment assurer le tirage d'un échantillon qui soit représentatif pour chaque année du cycle de rotation ?
2. La qualité d'un échantillon tiré se dégrade-t-elle au fil des années ?
3. Comment assurer le tirage d'un échantillon représentatif des différents types d'espace, rural, périurbain et urbain ?
4. Utiliser des plans de sondage distincts d'une région à l'autre ?

Questions suscitées par la construction des ZAE

1. Comment assurer le tirage d'un échantillon qui soit représentatif pour chaque année du cycle de rotation ?
2. La qualité d'un échantillon tiré se dégrade-t-elle au fil des années ?
3. Comment assurer le tirage d'un échantillon représentatif des différents types d'espace, rural, périurbain et urbain ?
4. Utiliser des plans de sondage distincts d'une région à l'autre ?

Questions suscitées par la construction des ZAE

1. Comment assurer le tirage d'un échantillon qui soit représentatif pour chaque année du cycle de rotation ?
2. La qualité d'un échantillon tiré se dégrade-t-elle au fil des années ?
3. Comment assurer le tirage d'un échantillon représentatif des différents types d'espace, rural, périurbain et urbain ?
4. Utiliser des plans de sondage distincts d'une région à l'autre ?

Lignes directrices

Principe général des simulations Monte-Carlo

- ▷ Variables du RP 99
- ▷ Variables fiscales 1996 et 2004

} disponibles sur toutes les communes françaises.

Pour un plan de sondage donné :

1. Tirage de *M* échantillons indépendants selon ce plan,
2. Pour chaque échantillon, estimations des totaux (nationaux et régionaux) de variables d'intérêt préalablement choisies,
3. Analyse des distributions empiriques de ces totaux ; (Moyenne, biais, variance, EQM, CV empiriques)

▷ Comparaison des plans avec les résultats de l'étape 3.

Comment tirer un échantillon s_{ZAE} de ZAE ?

- **Choix des paramètres :**

$$\begin{cases} \text{Taux de sondage : } \tau = 1/2000, \\ \text{Nombre de fiches-adresses par enquêteur : } e = 20. \end{cases}$$

- **Stratification régionale.**
- **Probabilité d'inclusion** π_k de la ZAE k dans l'échantillon :

  $\pi_k \propto nres_k$, nombre de résidences principales de la ZAE k.

  - 37 ZAE exhaustives : Communes t.q. $nres \geq e/\tau = 40000$.
  - 1 enquêteur par ZAE non exhaustive de l'échantillon, $\left[\frac{nres_k \times \tau}{e}\right]$ enquêteur(s) par ZAE exhaustive.

Comment tirer un échantillon "représentatif" de ZAE ?

- **Equilibrage sur des variables X choisies au préalable :** Sélection aléatoire de l'échantillon parmi tous ceux pour lesquels l'estimateur de Horvitz-Thompson du total de X coincide avec le vrai total de X :

$$\sum_{k \in s_{ZAE}} \frac{X_k}{\pi_k} = \sum_{k \in U_{ZAE}} X_k.$$

Equilibrage réalisé par la méthode du CUBE, développée par J.-C. Deville et Y. Tillé.

Question

Quelles variables d'équilibrage choisir ?

Les variables d'équilibrage retenues

Variables d'équilibrage	Plan de sondage				
	No 1	No 2	No 3	No 4	No 5
Nombre de Résidences principales RP 99	1	1	1	1	1
Nb de Résidences princ. 99 dans le groupe de rotation 1	2	2	2	2	2
Nb de Résidences princ. 99 dans le groupe de rotation 2	3	3	3	3	3
Nb de Résidences princ. 99 dans le groupe de rotation 3	4	4	4	4	4
Nb de Résidences princ. 99 dans le groupe de rotation 4	5	5	5	5	5
Nb de Résidences princ. 99 grandes communes	-	6	6	-	-
Nb de Résidences princ. 99 en zone rurale	-	-	-	6	6 ou 11
Nb de Résidences princ. 99 en zone périurbaine	-	-	7	7	7 ou 12
Revenu fiscal 2004 dans le groupe de rotation 1	-	-	8	8	8 ou 6
Revenu fiscal 2004 dans le groupe de rotation 2	-	-	9	9	9 ou 7
Revenu fiscal 2004 dans le groupe de rotation 3	-	-	10	10	10 ou 8
Revenu fiscal 2004 dans le groupe de rotation 4	-	-	11	11	11 ou 9
Revenu fiscal 2004 dans le groupe de rotation 5	-	-	12	8	12 ou 10

Prise en compte du cycle quinquennal

Estimateur du total de la variable X dans le groupe de rotation i

$$\widehat{T}_{X_i} = \sum_{k \in s_{ZAE}} \frac{\widehat{X}_k}{\pi_k} \quad \text{avec} \quad \widehat{X}_k = X_{k,i} \frac{nres_k}{nres_{k,i}}$$

Le double indice k, i désignant le groupe de rotation i de la ZAE k.

- Estimateur légèrement biaisé.
  $\widehat{X}_k$, estimateur par le ratio du total de X dans la ZAE k.
  Meilleur (en termes d'EQM) que l'estimateur sans biais $\widehat{X}_k = 5 \cdot X_{k,i}$.
- Estimation exacte (biais et variance nuls) pour la variable de repondération *nres*.

Les différents types d'estimation

1. Estimations par années de rotation du cycle quinquennal :
$$\widehat{T}_{X_i},\ i = 1, \ldots, 5.$$
2. Estimations sans distinction des groupes de rotation :
$$\widehat{T}_X = \left(\sum_{i=1}^{5} \widehat{T}_{X_i}\right) / 5.$$
3. Estimations sur données des recensements antérieurs :
RP 99, 90, 82, 75, 68, 62.
4. Estimations par type d'espace : urbain, périurbain, rural.

Les différents types d'estimation

1. Estimations par années de rotation du cycle quinquennal :
$$\widehat{T}_{X_i},\ i = 1, \ldots, 5.$$
2. Estimations sans distinction des groupes de rotation :
$$\widehat{T}_X = \left(\sum_{i=1}^{5} \widehat{T}_{X_i}\right) /5.$$
3. Estimations sur données des recensements antérieurs : RP 99, 90, 82, 75, 68, 62.
4. Estimations par type d'espace : urbain, périurbain, rural.

Les différents types d'estimation

1. Estimations par années de rotation du cycle quinquennal :
$$\widehat{T}_{X_i},\ i = 1, \ldots, 5.$$
2. Estimations sans distinction des groupes de rotation :
$$\widehat{T}_X = \left(\sum_{i=1}^{5} \widehat{T}_{X_i}\right) / 5.$$
3. Estimations sur données des recensements antérieurs :
RP 99, 90, 82, 75, 68, 62.
4. Estimations par type d'espace : urbain, périurbain, rural.

Les différents types d'estimation

1. Estimations par années de rotation du cycle quinquennal :
$$\widehat{T}_{X_i},\ i = 1, \ldots, 5.$$
2. Estimations sans distinction des groupes de rotation :
$$\widehat{T}_X = \left(\sum_{i=1}^{5} \widehat{T}_{X_i}\right) / 5.$$
3. Estimations sur données des recensements antérieurs :
RP 99, 90, 82, 75, 68, 62.
4. Estimations par type d'espace : urbain, périurbain, rural.

Lignes directrices

Estimations par année du cycle quinquennal

Quelques résultats, pour le plan de sondage préconisé *in fine*.

Année de rotation	Biais relatif (%)	cv (%)
1	0.683	0.311
2	0.566	0.320
3	0.538	0.321
4	0.700	0.324
5	0.850	0.340
SANS	0.667	0.270

Population sans double compte, RP 99

Année de rotation	Biais relatif (%)	cv (%)
1	0.000	0.000
2	-0.023	0.067
3	-0.029	0.075
4	0.000	0.000
5	0.000	0.000
SANS	-0.010	0.021

Nombre de résidences principales, RP 99

Année de rotation	Biais relatif (%)	cv (%)
1	1.407	0.438
2	1.538	0.467
3	1.459	0.449
4	1.488	0.457
5	1.447	0.441
SANS	1.468	0.311

Revenu fiscal 2004

Estimations sur données des recensements antérieurs

RP	Population sans double compte		Nombre de résidences principales		Nombre de naissances depuis RP précédent	
	Biais relatif (%)	cv (%)	Biais relatif (%)	cv (%)	Biais relatif (%)	cv (%)
1962	1.559	1.349	1.353	1.419	2.119	1.470
1968	0.395	1.104	0.331	1.186	-0.127	1.207
1975	-0.522	0.776	-0.556	0.851	-2.129	1.165
1982	-0.185	0.481	-0.450	0.524	-2.696	1.071
1990	0.380	0.334	-0.188	0.248	-1.368	0.839
1999	0.661	0.327	-0.011	0.021	-0.644	0.717

Estimations par type d'espace

Type d'espace	Plan No 1		Plan No 2		Plan No 3		Plan No 4		Plan No 5	
	Biais relatif (%)	cv (%)	Biais relatif (%)	cv (%)	Biais relatif (%)	cv (%)	Biais relatif (%)	cv (%)	Biais relatif (%)	cv (%)
Population sans double compte RP 99										
urbain	0.030	3.050	0.111	2.181	0.028	2.099	0.039	0.899	0.001	0.904
périurbain	-0.080	6.017	-0.348	5.147	-0.171	4.503	0.015	1.824	0.001	1.865
rural	0.011	6.155	0.141	5.854	0.093	5.008	-0.103	2.127	0.037	2.060
Nombre de résidences principales RP 99										
urbain	0.040	3.051	0.095	2.059	0.027	2.009	0.038	0.793	-0.011	0.791
périurbain	-0.084	5.989	-0.336	5.117	-0.165	4.524	0.011	1.789	0.003	1.804
rural	-0.010	6.162	0.116	5.819	0.106	4.956	-0.107	2.093	0.024	2.022
Revenu fiscal 2004										
urbain	-0.045	3.109	0.149	2.431	0.035	2.084	0.048	0.979	0.010	0.976
périurbain	-0.086	6.161	-0.369	5.146	-0.169	4.447	0.037	2.025	-0.029	2.050
rural	-0.024	6.130	0.128	5.848	0.069	5.201	-0.076	2.229	0.018	2.175

Conclusion

Des simulations pour guider le choix du plan de sondage du nouvel échantillon-maître.

- Estimations de bonne précision dès le premier plan testé, malgré la prise en compte des groupes de rotation.
- Bon comportement des estimations vis-à-vis des évolutions temporelles de la base de sondage
- Nécessité d'équilibrer sur des variables caractéristiques des types d'espace.

Lectures complémentaires I


Faivre, S.
Le projet Octopusse de nouvel échantillon-maître de l'Insee.
Communication JMS 2009.


Tillé, Y.
Théorie des sondages, échantillonnage et estimation en populations finies.
Dunod Paris, 2001.


Deville, J.-C. and Tillé, Y.
Efficient balanced sampling : The cube method.
Biometrika, 91 :893–912, 2004.


Chauvet, G. and Tillé, Y.
A fast algorithm of balanced sampling.
Computational Statistics, 21 :53–61, 2006.

Merci de votre attention.