Plans par grappes et à plusieurs degrés

Sommaire

1. Principe, justification et premiers exemples

a. Principe

On accède aux unités échantillonnées par l'intermédiaire de regroupements de ces unités.

Exemple d'une enquête auprès des ménages :

- *sélection d'un échantillon de communes,*
- *puis on retient :*
  - *tous les ménages des communes retenues : plans par grappes*
  - *un échantillon de ménages dans les communes retenues : plans à 2 degrés où les communes s'appellent « unités primaires » et les ménages « unités secondaires »*
  - *un échantillon d'individus dans les ménages et communes retenus aux degrés précédents : plans à 3 degrés*

b. Justifications

Avantages :

- Absence de base de sondage ou mauvaise qualité : seule la connaissance exhaustive des unités primaires (par ex, des villes) est nécessaire.

- Economie

- Gain de temps

Inconvénients :

- Perte de précision par rapport à un sondage aléatoire simple sans remise de même taille

  Effets de grappes : les unités statistiques regroupées dans une même unité primaire ont souvent tendance à se ressembler.

  ↳ Il faut donc soigner le premier degré de tirage.

c. Premiers exemples

- En industrie, contrôle par lots : produits de série conditionnés par caisses.

- Etudes médicales : on accède aux patients ou à des prescriptions en interrogeant des médecins ou des laboratoires d'analyses médicales

- Sondages électoraux « sortie des urnes » réalisés auprès d'électeurs à la sortie de certains bureaux de vote.

- Enquête Emploi et enquêtes ménages INSEE

Les notations :

- Population : $U = \{1,...,k,...,N\}$

- *M* unités primaires (UP) ou *M* grappes : $U = \bigcup_{i=1}^{M} U_i$

  Dont *m* sont échantillonnées dans S_G ou S_{UP}

- *N* unités secondaires (US) : $N = \sum_{i=1}^{M} N_i$ où $N_i = card(U_i)$

  Choix de n_i unités parmi N_i dans les *m* UP *i* tirées dans S_I

- Total de la variable *Y* :

$$t_Y = \sum_{i=1}^{M} t_{Yi} = \sum_{k=1}^{N} Y_k \qquad \text{Où} \qquad t_{Yi} = \sum_{k \in U_i} Y_k$$

- Moyenne de la variable *Y* :

$$\mu_Y = \frac{1}{N} \sum_{k=1}^{N} Y_k = \frac{1}{N} \sum_{i=1}^{M} N_i \mu_{Yi} = \frac{t_Y}{N}$$

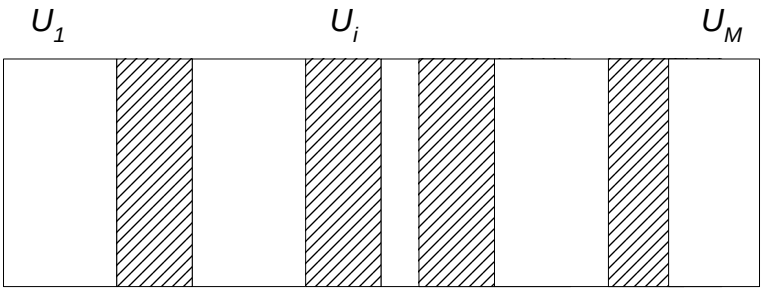
- Dispersion du total de la variable Y entre les grappes ou les UP

$$S^2_T = \frac{1}{M-1} \sum_{i=1}^{M} \left(t_{Yi} - \frac{t_Y}{M} \right)^2$$

2. Plans par grappes

Principe :

- Sélection d'un échantillon S_G de grappes
- Recensement dans chaque grappe retenue



- Probabilités d'inclusion des individus et des grappes:

$$\pi_k = \pi_i, \ \forall k \in U_i$$

$$\pi_{kl} = \begin{cases} \pi_i, \text{ si } k,l \in U_i \\ \pi_{ij}, \text{ si } k \in U_i \text{ et } l \in U_j \end{cases}$$

- Taille d'échantillon aléatoire :

$$n_s = \sum_{k \in S} 1 = \sum_{i \in S_G} = N_i$$

a. Cas général : tirage de grappes à probabilités inégales

Estimateur de Horvitz-Thompson du total

$$\hat{t}_y = \sum_{i \in s_G} \frac{t_{yi}}{\pi_i}$$

- Estimateur sans biais

- De variance :

$$Var\left(\hat{t}_y\right) = \sum_{i,j \in U_I} \sum \Delta_{ij} \frac{t_{yi}}{\pi_i} \frac{t_{yj}}{\pi_j}$$

ou Sen-Yates-Grundy

- Estimateur sans biais de la variance si $\pi_{ij} > 0$ pour toutes les grappes i et j :

$$\hat{Var}\left(\hat{t}_y\right) = \sum_{i,j \in S_G} \sum \frac{\Delta_{ij}}{\pi_{ji}} \frac{t_{yi}}{\pi_i} \frac{t_{yj}}{\pi_j}$$

ou Sen-Yates-Grundy

Estimation de la moyenne

Si N connu ,

Estimateur de Horvitz-Thompson

$$\hat{u}_y = \frac{\hat{t}_y}{N} = \frac{1}{N}\sum_{i \in s_G} \frac{t_{y_i}}{\pi_i}$$

- Estimateur sans biais
- De variance : $\frac{1}{N^2} Var\left(\hat{t}_y\right)$
- estimée par : $\frac{1}{N^2} \hat{V}ar\left(\hat{t}_y\right)$

Si N inconnu ,

Estimateur de Hajek :

$$\hat{u}_{y,Hajek} = \frac{\hat{t}_y}{\hat{N}} = \frac{\sum_{i \in s_G} \frac{t_{y_i}}{\pi_i}}{\sum_{i \in s_G} \frac{N_i}{\pi_i}}$$

- Estimateur biaisé
- Plus précis en présence d'un effet taille :

$$Var\left(\hat{u}_{y,Hajek}\right) \cong \frac{1}{N^2} \sum_{i,j} \sum_{=1,\ldots,M} \Delta_{ij} \frac{N_i\left(\mu_{Yi} - \mu_Y\right)}{\pi_i} \times \frac{N_j\left(\mu_{Yj} - \mu_Y\right)}{\pi_{lj}}$$

$$Var\left(\hat{u}_{y,Hajek}\right) \cong \frac{1}{N^2} \sum_{i,j} \sum_{\in S_G} \frac{\Delta_{ij}}{\pi_{ij}} \frac{N_i\left(\mu_{Yi} - \hat{u}_{y,Hajek}\right)}{\pi_i} \times \frac{N_j\left(\mu_{Yj} - \hat{u}_{y,Hajek}\right)}{\pi_{lj}}$$

b. Tirage des grappes proportionnellement à leur taille

$$\pi_i = m\frac{N_i}{N} = \pi_k\,,\ \forall k \in U_i$$

En moyenne, la taille d'échantillon vaut :

$$E(n_s) = \sum_{i\in U_I} N_i m\frac{N_i}{N} = \frac{m}{N}\sum_{i\in U_I} N_i^2$$

Estimateur de Horvitz-Thompson du total :

$$\boxed{\hat{t}_y = \frac{N}{mN_i}\sum_{i\in s_G} t_{y\,i}}$$

Estimateur de Horvitz-Thompson de la moyenne :

$$\boxed{\hat{\mu}_y = \frac{1}{m}\sum_{i\in s_G} \mu_{y\,i}}$$

c. Tirage des grappes à probabilités égales

$$\pi_i = \frac{m}{M} = \pi_k, \quad \forall i = 1,...M, \forall k \in U$$

En moyenne, la taille d'échantillon vaut :

$$E(n_s) = \sum_{i \in U_I} N_i \frac{m}{M} = \frac{Nm}{M}$$

Estimateurs de Horvitz-Thompson du total et de la moyenne :

$$\hat{t}_y = \frac{M}{m} \sum_{i \in s_G} t_{yi} \qquad \hat{\mu}_y = \frac{M}{Nm} \sum_{i \in s_G} N_i \mu_{yi}$$

- Estimateurs sans biais

- De variance pour le total :

$$Var(\hat{t}_y) = M^2 \left(1 - \frac{m}{M}\right) \frac{S_T^2}{m} \quad \text{Où} \quad S^2_T = \frac{1}{M-1} \sum_{i=1}^{M} \left(t_{Yi} - \frac{t_Y}{M}\right)^2$$

- Estimée sans biais par :

$$\hat{Var}(\hat{t}_y) = M^2 \left(1 - \frac{m}{M}\right) \frac{s_T^2}{m} \quad \text{où} \quad s_T^2 = \frac{1}{m-1} \sum_{i \in s_G} \left(t_{Yi} - \frac{\hat{t}_Y}{M}\right)^2$$

↳ Bonne précision quand s_T^2 est faible et *m* grand

d. Comparaison avec un *SAS* de même taille

Supposons $N_i = N_o = N/M$ pour toutes les grappes

Alors,

$n = m\,N_o$ et $m/M = n/N$

$$\hat{\mu}_y = \frac{1}{m}\sum_{i\in s_G} \mu_{yi}$$

$$Var(\hat{\mu}_y) = \left(1-\frac{m}{M}\right)\frac{M}{M-1}\frac{\sigma^2_{Y\,inter}}{m} \quad \text{avec} \quad \sigma^2_{Y\,inter} = \sum_{i=1}^{M}\frac{N_o}{N}(\mu_{Yi}-\mu_Y)^2$$

Car :

$$Var(\hat{\mu}_y) = \left(\frac{M}{N}\right)^2\left(1-\frac{m}{M}\right)\frac{1}{m}\frac{1}{M-1}\sum_{i=1}^{M}\left(N_o\mu_{Yi}-\frac{t_Y}{M}\right)^2$$

$$= M^2\left(1-\frac{m}{M}\right)\frac{1}{m}\frac{1}{M-1}\sum_{i=1}^{M}\left(\frac{N_o}{N}\mu_{Yi}-\frac{N_o}{N}\mu_Y\right)^2$$

Avec un ***SAS*** sans remise de même taille *n*, on aurait eu :

$$Var\,(\hat{\mu}_{Ysas}) = \left(1-\frac{m}{M}\right)\frac{1}{n}S_y^2$$

- **Perte de précision du plan par grappes due aux « effets de grappes » : « qui se ressemble s'assemble »**
  - *clientèles typées selon les médecins, leur lieu d'exercice*
  - *hétérogénéité des bureaux de vote,etc…*

Plus précisément, on mesure l'**effet de sondage** du plan par grappes :

$$Deff = \frac{Var(\hat{\mu}_y)}{Var(\hat{\mu}_{y,sas})} \cong 1 + (N_o - 1)\rho_{\text{intra}}$$

et $Deff > 1$ si $\rho_{\text{intra}} > 0$

$Deff < 1$ si $\rho_{\text{intra}} < 0$

qui s'obtient en écrivant :

$$Var(\hat{\mu}_y) \cong \left(1 - \frac{m}{M}\right)\frac{1}{mN_o} S_y^2 \left[1 + (N_o - 1)\rho_{\text{intra}}\right]$$

avec le ***coefficient de corrélation intra-grappe :***

$$\rho_{\text{intra}} = \frac{\sum_{i=1}^{M} \sum_{j \neq j' \in U_k} \sum (Y_{ij} - \mu_Y)(Y_{ij'} - \mu_Y)}{(N_0 - 1)(N - 1)S_Y^2} \cong \frac{N_o \frac{\sigma_{Y\,\text{int}\,er}^2}{\sigma_Y^2} - 1}{N_o - 1}$$

e. Considérations pratiques

Comment construire les grappes ?

Combien de grappes faut-il construire ?

Quelques aspects importants :

i. Grappes homogènes en **inter** et hétérogènes en **intra** *(logique contraire à celle de la stratification, analogue à celle du systématique).*

ii. Grappes de tailles voisines ou faible dispersion de tailles des grappes.

iii. Grappes de faibles tailles (ou beaucoup de grappes constituées).

iv. Un maximum de grappes tirées

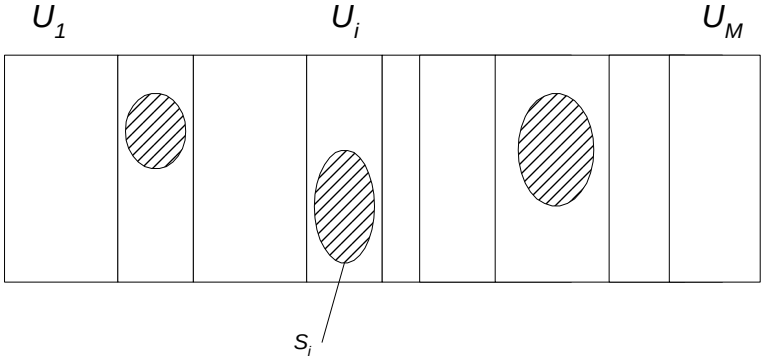
Remarque :

Le tirage systématique est un cas particulier du plan par grappes (m=1).

Impossible d'estimer sans biais la variance de l'estimateur du total.

3. Plans à plusieurs degrés

a. Principe général



i. Un échantillon d'*UP*, s_{UP} , est tiré dans la population des *UP* selon un plan de sondage P_{Up}

ii. Dans chaque U_i sélectionnée ($i \in S_{Up}$), un échantillon s_i d'*US* est tiré selon un plan de sondage P_{U_i} , indépendamment d'une UP à l'autre.

iii. L'échantillon complet est : $S = \bigcup_{i \in s_{UP}} S_i$

Sa taille est aléatoire.

Généralisation à plusieurs degrés

- Probabilités d'inclusion des UP dans $\boldsymbol{S_{UP}}$:

$$\pi_{Ii},\ \pi_{Iij}$$

- Probabilités d'inclusion des individus ***k*** et ***l*** de l'UP ***i*** dans $\boldsymbol{S_i}$, sachant l'UP ***i*** tirée :

$$\pi_{k/i},\ \pi_{kl/i}$$

- Probabilités d'inclusion des individus ***k*** et ***l*** dans l'échantillon complet ***S*** :

$$\pi_k = \pi_{Ii}\pi_{k/i}, \text{ si } k \in U_i$$

$$\pi_{kl} = \begin{cases} \pi_{Ii}\pi_{kl/i}, & \text{si } k \text{ et } l \in U_i \\ \pi_{Iij}\pi_{k/i}\pi_{l/j}, & \text{si } k \in U_i \text{ et } l \in U_j \end{cases}$$

Estimateur d'Horvitz-Thompson d'un total

$$\hat{t}_y = \sum_{i \in s_{UP}} \frac{\hat{t}_{y_i}}{\pi_{Ii}} \qquad \text{avec} \qquad \hat{t}_{yi} = \sum_{k \in s_i} \frac{y_k}{\pi_{k/i}}$$

soit : $\hat{t}_y = \sum_{i \in s_{UP}} \frac{1}{\pi_{Ii}} \sum_{k \in s_i} \frac{y_k}{\pi_{k/i}}$

- Estimateur sans biais

Estimation d'une moyenne

Si N connu ,

Estimateur de Horvitz-Thompson

$$\hat{u}_y = \frac{\hat{t}_y}{N} = \frac{1}{N} \sum_{i \in s_{UP}} \frac{\hat{t}_{y_i}}{\pi_{Ii}}$$

Si N inconnu ,

Estimateur de Hajek

$$\hat{u}_{y,Hajek} = \frac{\hat{t}_y}{\hat{N}} = \frac{\sum_{i \in s_{UP}} \frac{\hat{t}_{y_i}}{\pi_{Ii}}}{\sum_{i \in sUP} \frac{\hat{N}_i}{\pi_{Ii}}}$$

b. Tirage SAS à chaque degré

$$\hat{t}_y = \frac{M}{m}\sum_{i\in s_{UP}} \hat{t}_{yi} \qquad \text{avec} \qquad \hat{t}_{yi} = \frac{N_i}{n_i}\sum_{k\in s_i} y_k$$

soit : $$\hat{t}_y = \frac{M}{m}\sum_{i\in s_{UP}} \frac{N_i}{n_i}\sum_{k\in s_i} y_k$$

Remarque : inutile de connaître *N* pour estimer le total

Propriétés :

- Estimateur sans biais
- De variance :

$$Var\left(\hat{t}_y\right) = M^2\left(1-\frac{m}{M}\right)\frac{S_T^2}{m} + \frac{M}{m}\sum_{i=1}^{M} N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{S_{Yi}^2}{n_i}$$

variance inter-UP + variance intra-UP

Où $S^2_T = \frac{1}{M-1}\sum_{i=1}^{M}\left(t_{Yi} - \frac{t_Y}{M}\right)^2$ et $S^2_{Yi} = \frac{1}{N_i - 1}\sum_{k=1}^{Ni}\left(y_k - \mu_{Yi}\right)^2$

- Estimée sans biais par :

$$\hat{Var}\left(\hat{t}_y\right) = M^2\left(1-\frac{m}{M}\right)\frac{s_T^2}{m} + \frac{M}{m}\sum_{i\in S_{UP}} N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{s_{Yi}^2}{n_i}$$

où $s_T^2 = \frac{1}{m-1}\sum_{i\in s_{UP}}\left(\hat{t}_{Yi} - \frac{\hat{t}_Y}{M}\right)^2$ et $s^2_{Yi} = \frac{1}{n_i - 1}\sum_{k\in si}\left(y_k - \hat{\mu}_{Yi}\right)^2$

c. Tirages autopondérés

Toutes les unités statistiques de l'échantillon (US) possèdent la même pondération, ou ce qui revient au même, les probabilités d'inclusion des US sont constantes

i. Tirage à probabilités égales des UP et des US

ii. Tirage des UP proportionnel à leur taille suivi d'un tirage d'un nombre fixe d'US dans chaque UP retenue

Dans ce cas, l'estimateur de la moyenne est la moyenne simple calculée sur l'ensemble des US tirées.

Exemple : enquêtes ménages à l'INSEE

d. Considérations pratiques

Quand utiliser des sondages à plusieurs degrés ?

- Absence de base de sondage ou de mauvaise qualité
- Motifs économiques

Comment améliorer la précision ?

- Avant tout, construire des UP le plus ressemblantes possible entre elles pour limiter les effets de grappes.

Exemple : l'unité ménage est intéressante pour estimer des variables comme le sexe, l'activité, l'âge, etc, mais elle est moins efficace pour étudier le niveau d'instruction, la CS , etc.

- Privilégier le nombre d'UP enquêtées plutôt que le nombre d'US

- Tirer les UP à probabilités inégales

- Stratifier au niveau des UP

4. Exemples des enquêtes ménages INSEE

A- Les grands principes des enquêtes ménages

A.1. Objectif des enquêtes ménages

- Une des missions de l'INSEE
- Rôle du CNIS

Exemples :
Logement, Budget, Revenus et Conditions de Vie, Santé,…

A.2. L'unité échantillonnée

- On construit des échantillons de logements ordinaires
- On atteint les ménages et/ou les individus par l'intermédiaire de leur logement

A3. Principes d’échantillonnage

- Echantillons probabilistes de logements
- Tous les logements principaux possèdent la même probabilité de participer à l’enquête
- Un même logement ne peut-être interrogé pour des enquêtes distinctes entre deux recensements consécutifs

A.4. Modes de collecte

- Enquêtes généralement en face à face
- Quelques enquêtes par téléphone

Coûts de collecte

Stabilité du réseau d'enquêteurs

↳ ***localisation des échantillons***

A.5. Procédé : échantillonnage des enquêtes ménages en 2 temps

i. 1ère phase de localisation des enquêtes

Tirage de bases de sondage intermédiaires

ii. 2ème phase de tirage des échantillons des enquêtes ***:***

Choix des logements à interroger dans ces bases

Tirages autopondérés, systématiques, stratifiés

A.6. Sources des bases de sondage

i. Le dernier recensement de la population (*source : INSEE*)

ii. Les permis de construire pour les logements construits après le dernier RP (*source : Ministère de l'Equipement)*

En l'absence de liste exhaustive de logements à une date quelconque ou de registres de population

B. La construction de l'Echantillon–Maître (EM99)

B.1. Entre contraintes pratiques et désir de précision

- Etablir une réserve de logements localisée pour alimenter la plupart des enquêtes ménages hors DOM, enquête Emploi, enquêtes locales,…

- Optimiser la précision des résultats nationaux

B.2. Taille de l'EM99 : 2 022 889 logements

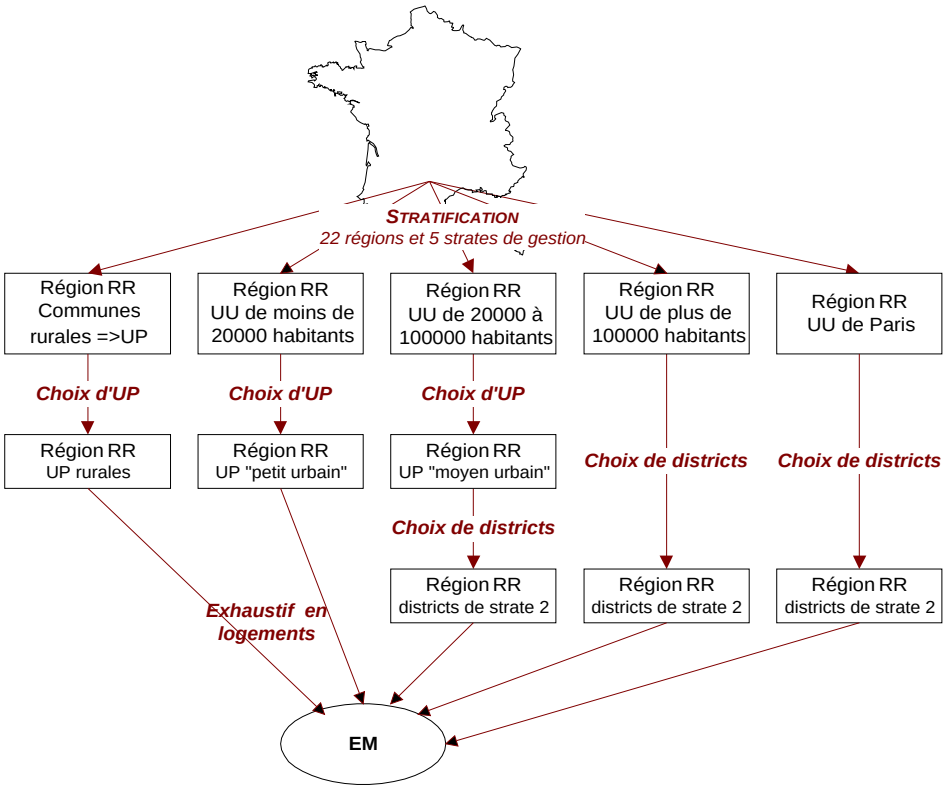
= 7 % des logements recensés en mars 99

Réserve calibrée dans l'optique :

- du nombre moyen d'enquêtes nationales par année,
- de la taille moyenne des enquêtes ménages,
- de la durée de vie de cette réserve

B.3. Constitution de l'EM99

Selon un tirage stratifié à 1 ou 2 degrés selon la strate



a. Stratification

Quadrillage du territoire en régions × densité d'habitat

- Collecte des enquêtes en Direction Régionale
- Le degré d'urbanisation explique assez bien le comportement des ménages

↳ ***Cette stratification est un élément de qualité***.

b. Les unités primaires

Définition :

- Dans les communes rurales, en petit et moyen urbain, une UP correspond au rayon d'action d'un enquêteur.
- Dans les UU de plus de 100 000 habitants, une UP est une UU.

Regroupements : UP de taille voisine, de profis moyens voisins

Mode de tirage : à probabilités inégales, proportionnelles à la taille

Tirage équilibré selon la méthode du Cube

c. Les unités secondaires (districts)

Tirage uniquement dans les UU de plus de 20000 habitants

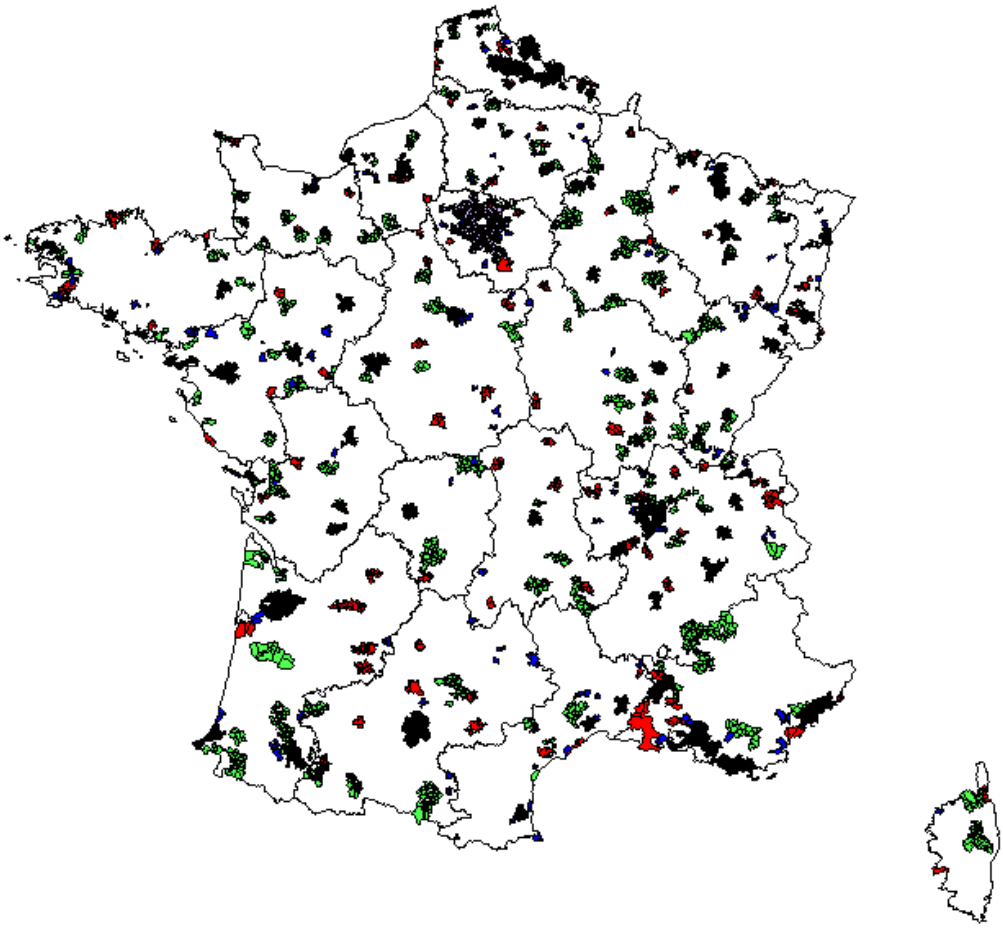
Tirage stratifié par groupes de communes

Tirage à probabilités égales des districts dans l'UU

Tirage équilibré sur l'UU par la méthode du Cube

d. Constitution de l'EM99

Ensemble de tous les logements des UP et des US sélectionnées



Source : INSEE