Estimation de variance pour l'Échantillon-Maître Octopusse

Guillaume Chauvet (Crest, Ensai)

Journées de Méthodologie Statistique
Paris, 26/01/2012

Plan de l'exposé

Principe de l'échantillon-maître

Estimation de variance

Etude par simulations

Principe de l'échantillon-maître

Principe

L'Echantillon-Maître (EM) est un échantillon de zones, utilisé comme réserve de logements pour les enquêtes auprès des ménages. L'Echantillon-Maître de 1999 (EM99) a été utilisé pour les enquêtes réalisées entre 1999 et 2009.

Pour l'EM99, ces zones ont été sélectionnés selon un plan de sondage stratifié selon le degré d'urbanisation, et à plusieurs degrés. On a sélectionné des communes ou des groupes de communes dans le rural, des districts dans l'urbain, ... (Ardilly, 2006).

Chacune des zones était confiée à un enquêteur "stable dans le temps et localisé à proximité" (Christine et Faivre, 2009). On parle de Zones d'Action Enquêteur (ZAE).

Principe

Pour l'EM99, une liste à jour des logements était fournie par le RP99 et par la Base de Sondage des Logements Neufs (BSLN).

Le passage depuis 2004 à des Enquêtes de Recensement a nécessité de modifier le système de tirage de l'EM, puisqu'on ne dispose plus à une date donnée de la connaissance du parc de logements complet. Autre objectif : s'affranchir du coût de la BSLN.

Grandes lignes de l'EM Octopusse :

- principe des ZAE conservé,
- séparation ZAE Grandes Communes et ZAE Petites Communes,
- sélection de l'échantillon de logements pour une enquête l'année $t+1$ dans les logements recensés l'année t.

Le Nouveau Recensement

Dans chaque grande commune (+ de 10,000 habitants au RP99) :

- stratification selon le type d'adresse,
- répartition des adresses en 5 groupes de rotation,
- une année donnée, enquête auprès de 8 % environ des logements d'un groupe de rotation.

Pour les petites communes :

- stratification par région,
- répartition des communes en 5 groupes de rotation par tirage équilibré selon la méthode du Cube (Deville et Tillé, 2004),
- une année donnée, enquête auprès de l'ensemble des logements des petites communes d'un groupe de rotation.

L'EM Octopusse

Au niveau des Grandes Communes :

- 1 ZAE = 1 GC,
- tirage d'un échantillon de ZAE-GC (méthode du Cube),
- pour une enquête l'année $t+1$, tirage d'un échantillon de logements parmi ceux enquêtés l'année t

$\Rightarrow$ tirage à 2 degrés, avec tirage en 2 phases au 2nd degré

Au niveau des petites communes :

- 1 ZAE = regroupement de PC contigues, contenant au moins 300 résidences principales de chaque groupe de rotation,
- tirage d'un échantillon de ZAE-PC (méthode du Cube),
- pour une enquête l'année $t+1$, tirage d'un échantillon de logements dans les communes recensées l'année t

$\Rightarrow$ tirage à 2 degrés ; 1er degré conditionnel au Recensement.

L'EM Octopusse

Au niveau des Grandes Communes :

- 1 ZAE = 1 GC,
- tirage d'un échantillon de ZAE-GC (méthode du Cube),
- pour une enquête l'année $t+1$, tirage d'un échantillon de logements parmi ceux enquêtés l'année t

$\Rightarrow$ tirage à 2 degrés, avec tirage en 2 phases au 2nd degré

Au niveau des petites communes :

- 1 ZAE = regroupement de PC contigues, contenant au moins 300 résidences principales de chaque groupe de rotation,
- tirage d'un échantillon de ZAE-PC (méthode du Cube),
- pour une enquête l'année $t+1$, tirage d'un échantillon de logements dans les communes recensées l'année t

$\Rightarrow$ tirage à 2 degrés ; 1er degré conditionnel au Recensement.

L'EM Octopusse

Au niveau des Grandes Communes :

- 1 ZAE = 1 GC,
- tirage d'un échantillon de ZAE-GC (méthode du Cube),
- pour une enquête l'année $t+1$, tirage d'un échantillon de logements parmi ceux enquêtés l'année t

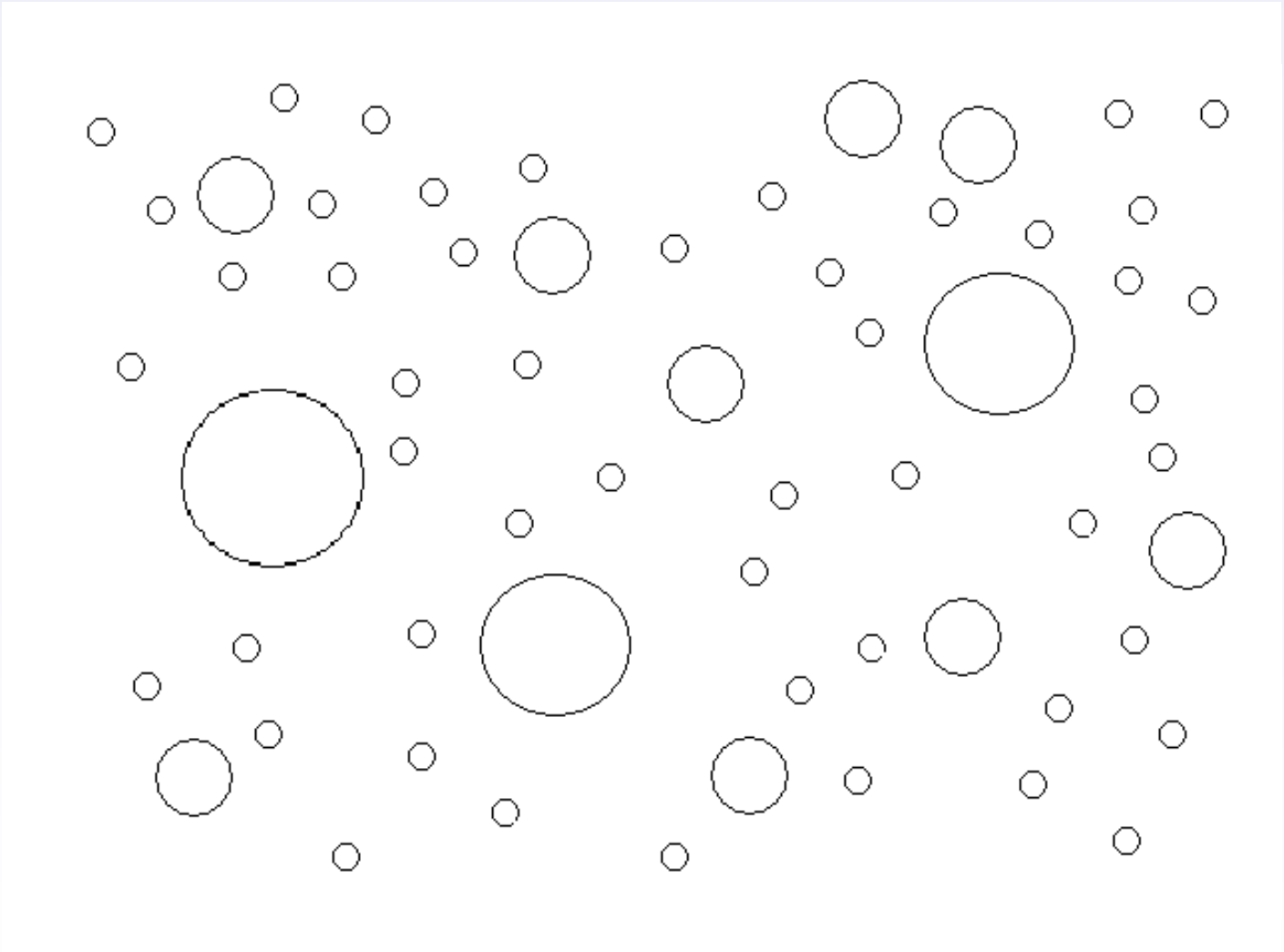
$\Rightarrow$ tirage à 2 degrés, avec tirage en 2 phases au 2nd degré

Au niveau des petites communes :

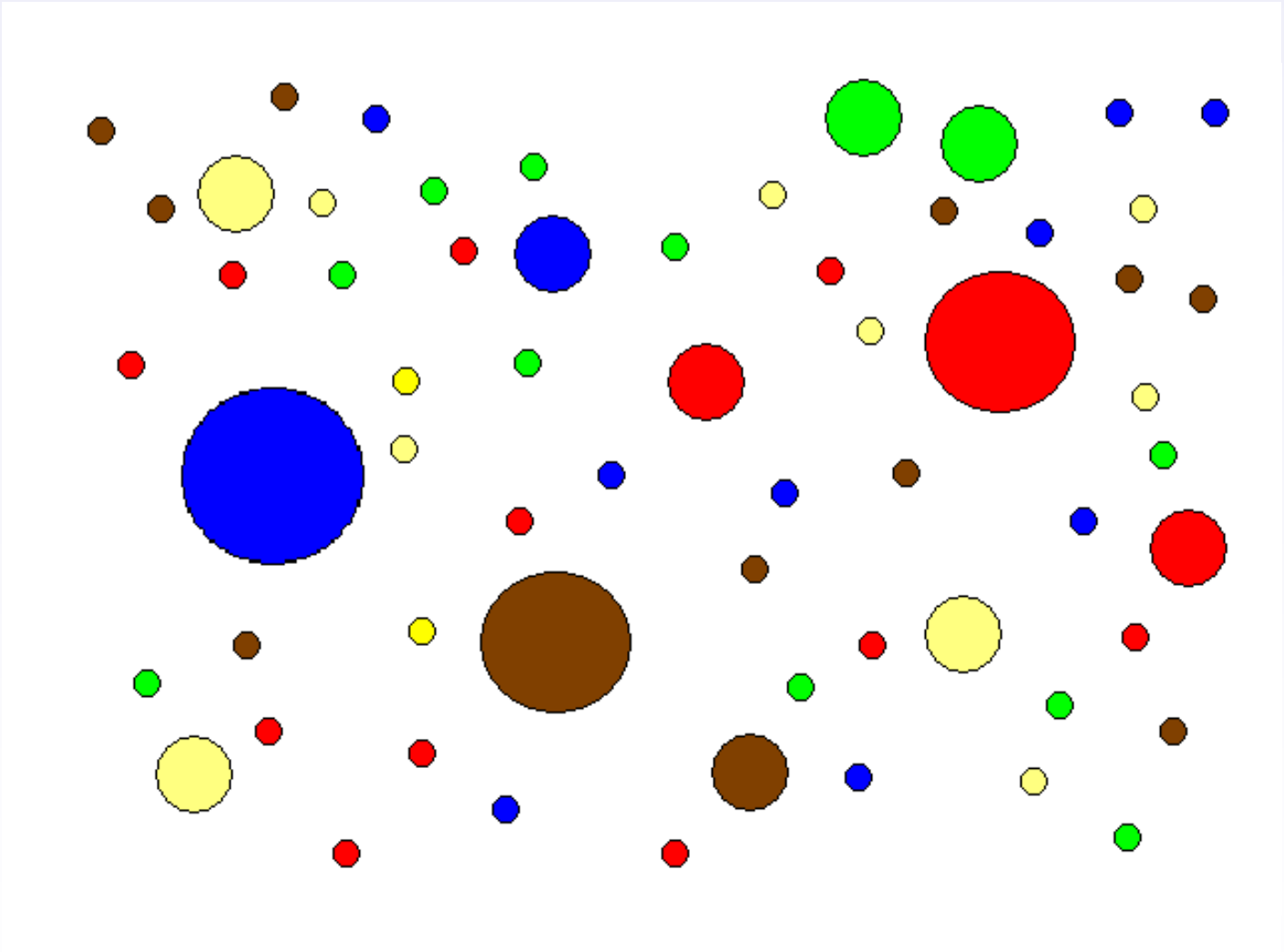
- 1 ZAE = regroupement de PC contigues, contenant au moins 300 résidences principales de chaque groupe de rotation,
- tirage d'un échantillon de ZAE-PC (méthode du Cube),
- pour une enquête l'année $t+1$, tirage d'un échantillon de logements dans les communes recensées l'année t

$\Rightarrow$ tirage à 2 degrés ; 1er degré conditionnel au Recensement.

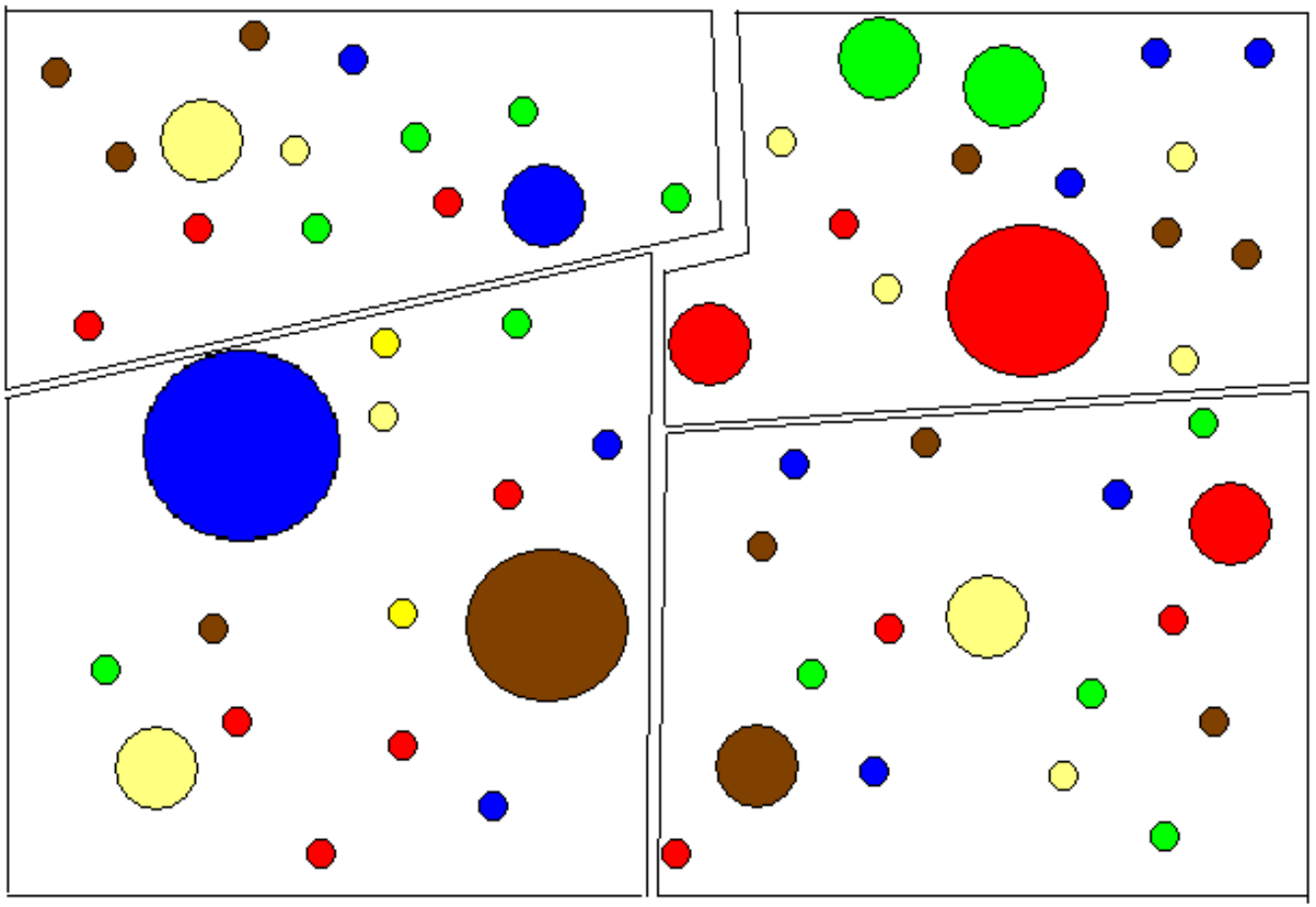
Tirage de l'EM : illustration



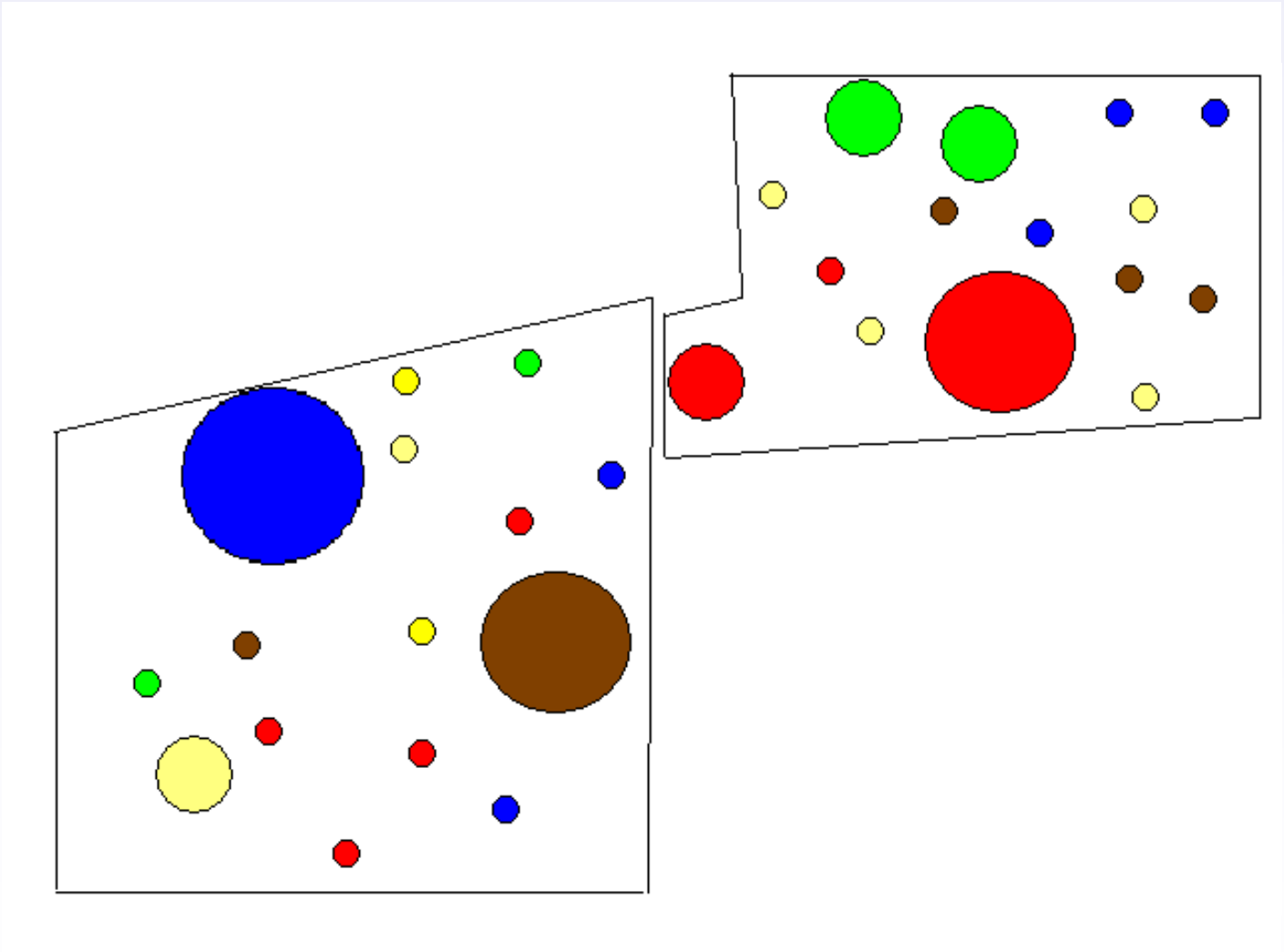
Découpage des PC en 5 groupes de rotation



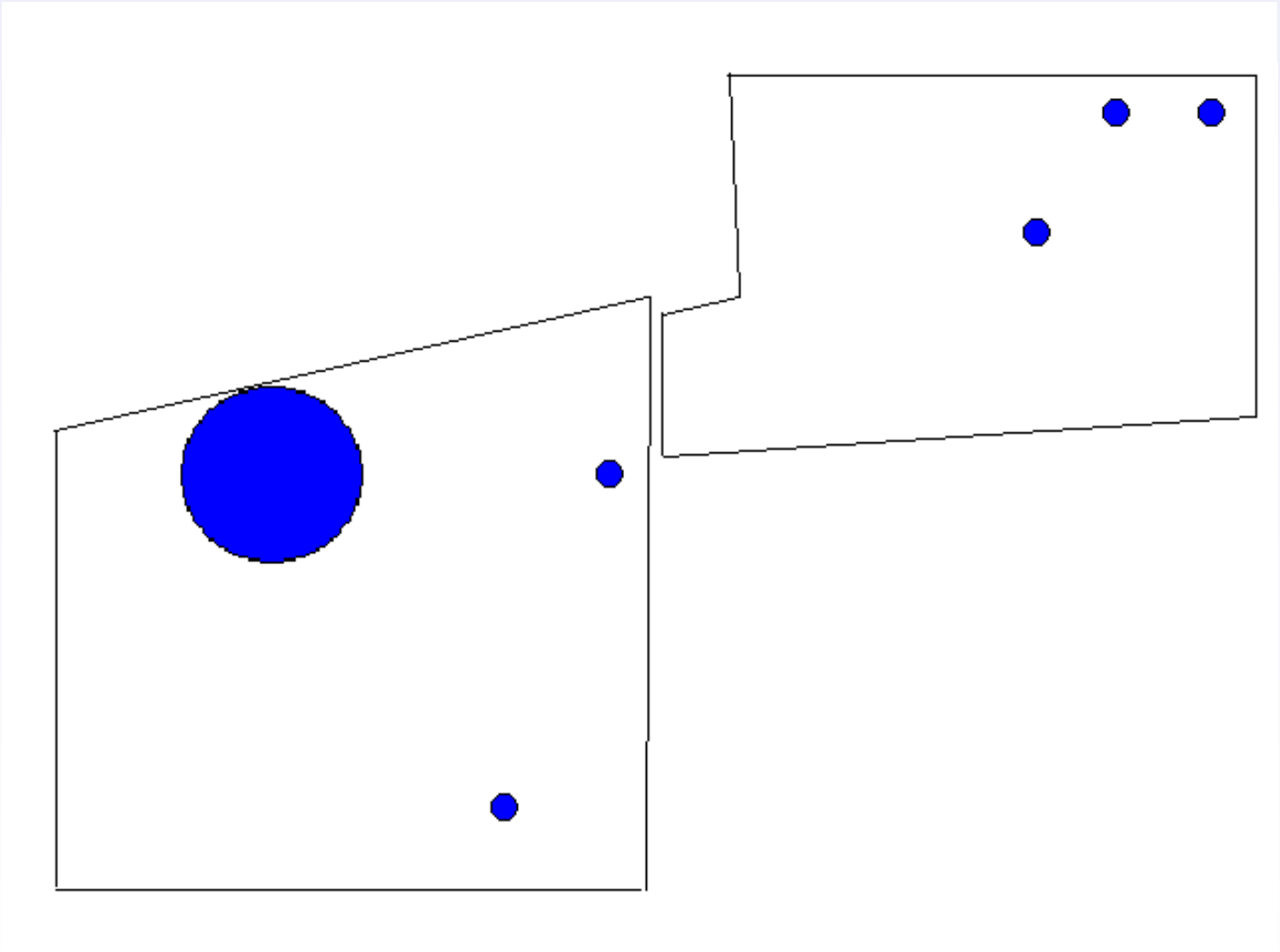
Constitution des PC-ZAE



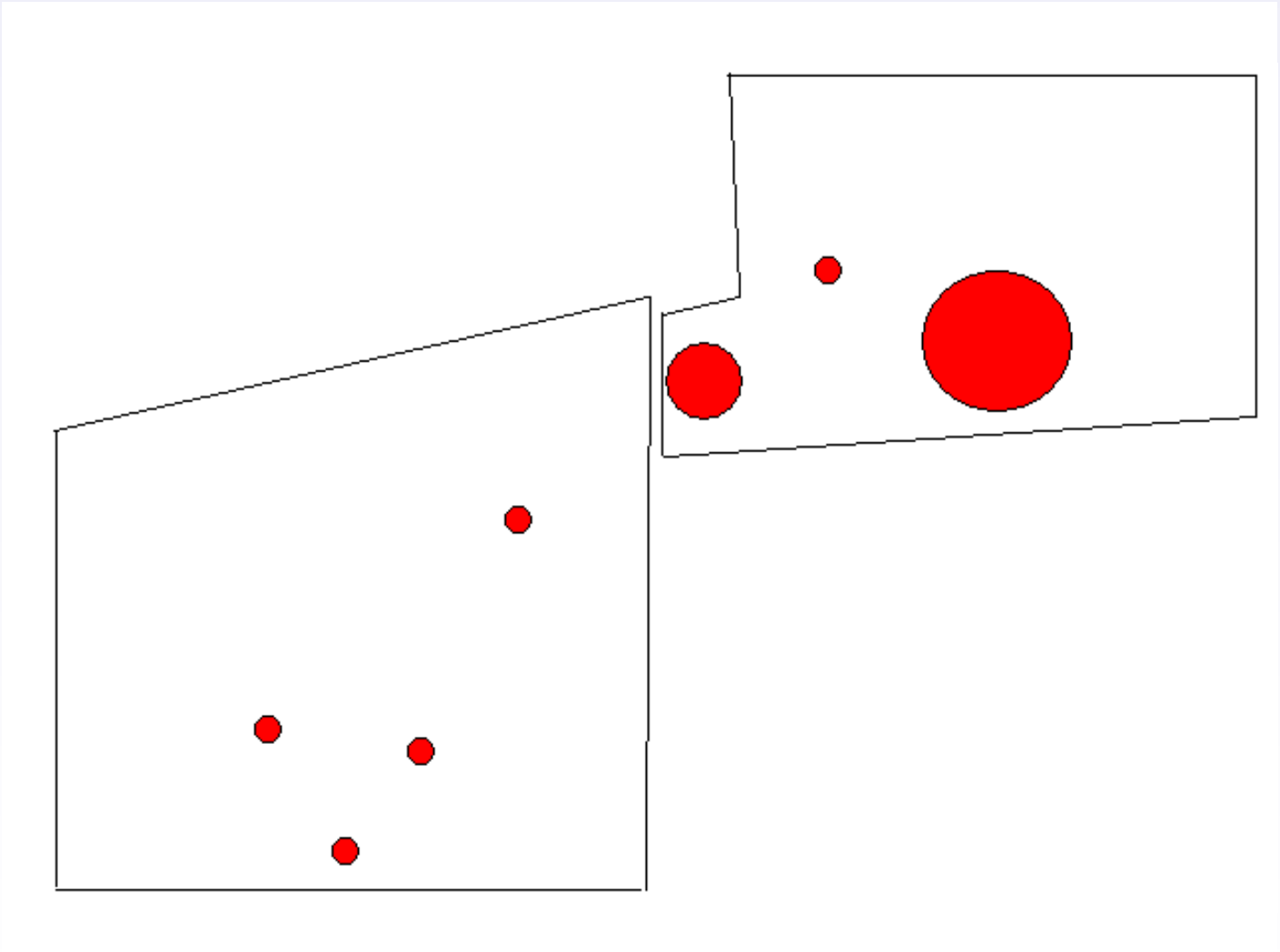
Tirage d'un échantillon de PC-ZAE



Petites communes du GR1 utilisées l'année 2



Petites communes du GR2 utilisées l'année 3



Estimation de variance

Estimateur par expansion

Soit U la population des Petites Communes d'une région. Les Petites Communes tirées pour les Enquêtes-ménage de l'année $t+1$ sont celles situées dans les ZAE-PC u_i sélectionnées et dans le groupe de rotation G_r enquêté l'année t :

$$S_r \quad = \quad \{k \in U;\ k \in u_i \in S_I \text{ et } k \in G_r\}.$$

Un total $t_y = \sum_{k \in U} y_k$ peut être estimé sans biais par l'estimateur par expansion

$$\hat{t}_{yr} \quad = \quad \sum_{k \in S_r} \frac{y_k}{\pi_{Ii}\ \alpha_{kr}},$$

avec

- π_{Ii} : proba de sélection d'une ZAE u_i dans l'échantillon S_I,
- α_{kr} : proba de sélection d'une PC k dans le groupe de rotation G_r.

Variance de l'estimateur par expansion

La variance de cet estimateur est donnée par

$$\begin{aligned} V(\hat{t}_{yr}) &= VE(\hat{t}_{yr}|G_1,\ldots,G_5) + EV(\hat{t}_{yr}|G_1,\ldots,G_5) \\ &= V_{NR} + V_{EM} \end{aligned}$$

avec

- V_{NR} : variance due au tirage du groupe de rotation G_r,
- V_{EM} : variance due au tirage de l'échantillon de ZAE-PC S_I.

Ces deux termes sont estimés séparément, en obtenant une estimation de la matrice de variance-covariance associée à chaque tirage.

Variance de l'estimateur par expansion

Le groupe de rotation G_r et l'échantillon de ZAE-PC S_I sont obtenus à l'aide d'un tirage équilibré selon la méthode du Cube.

On peut utiliser une approximation de variance de type Deville-Tillé (2005) pour un échantillonnage équilibré à entropie maximale :

$$v_{DT}(\hat{t}_{yr}) = v_{DT,NR}(\hat{t}_{yr}) + v_{DT,EM}(\hat{t}_{yr}).$$

Inconvénient : une partie de la variance n'est pas prise en compte.

On peut également utiliser une approximation de la matrice de variance-covariance basée sur les propriétés de martingale de la méthode du Cube (Breidt et Chauvet, 2011) :

$$v_{MD}(\hat{t}_{yr}) = v_{MD,NR}(\hat{t}_{yr}) + v_{MD,EM}(\hat{t}_{yr}).$$

Inconvénient : temps de calcul + estimateur plus instable.

Etude par simulations

Cadre de l'étude

L'étude par simulations est réalisée sur la population des $1,235$ petites communes de Bretagne. On dispose de 16 variables d'intérêt fournies par le RP99, relatives à la localisation, au sexe, à l'emploi,...

On utilise un plan de sondage proche de celui d'Octopusse, répété $B = 1,000$ fois :

- découpage des PC en 5 groupes de rotation $G_{1b}, \ldots, G_{5b}$,
- constitution de la population U_{Ib} des ZAE-PC ,
- tirage d'un échantillon S_{Ib} de ZAE-PC.

On utilise l'échantillon

$$S_{1b} \quad = \quad \{k \in U;\ k \in u_i \in S_{Ib} \text{ et } k \in G_{1b}\},$$

pour obtenir l'estimateur par expansion $\hat{t}_{y1,b}$ du total t_y.

Cadre de l'étude

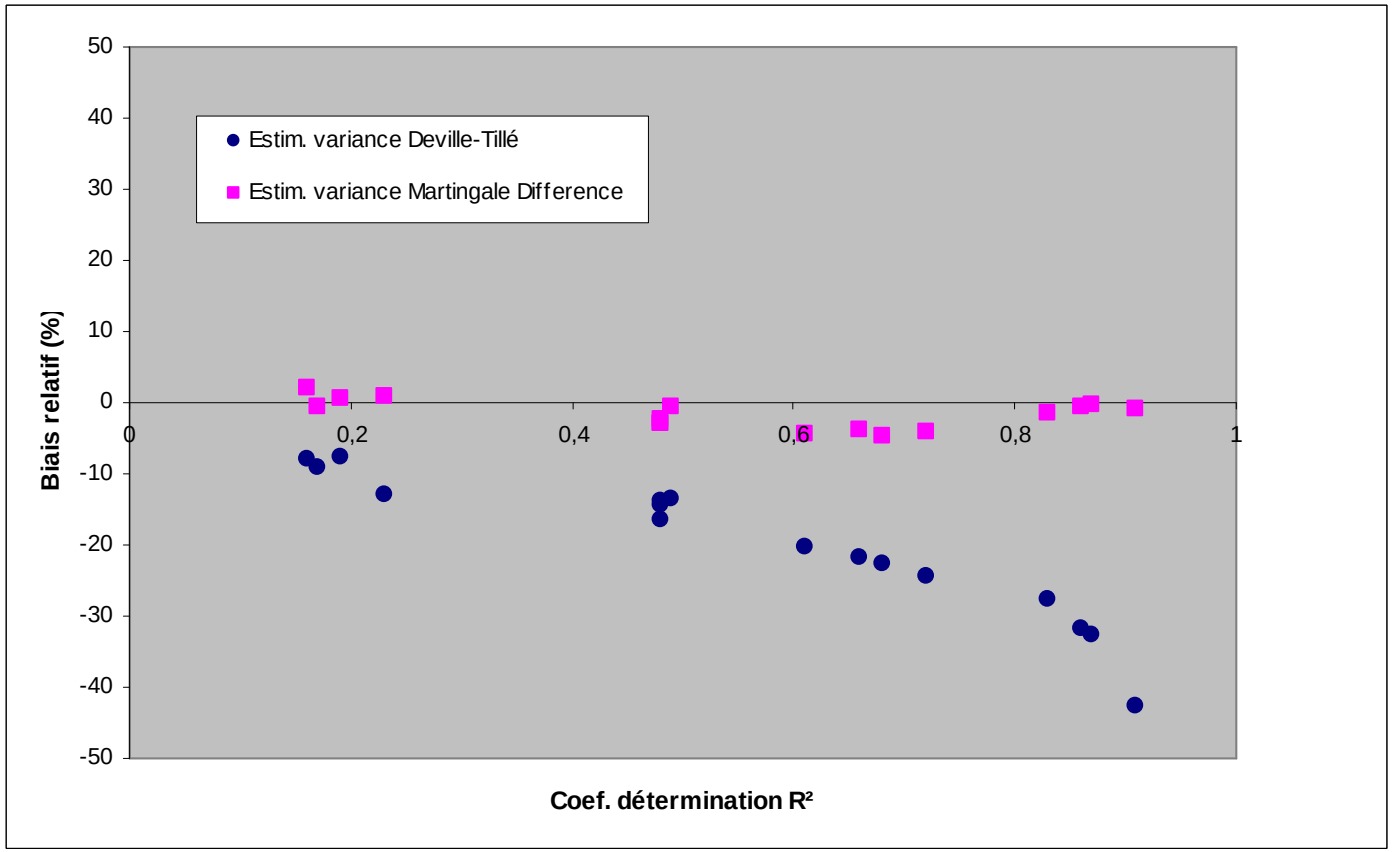
On compare les performances des estimateurs de variance basés (i) sur les formules Deville-Tillé et (ii) sur une approximation par simulations de la matrice de variance-covariance.

Ces deux estimateurs sont évalués en termes de biais relatif :

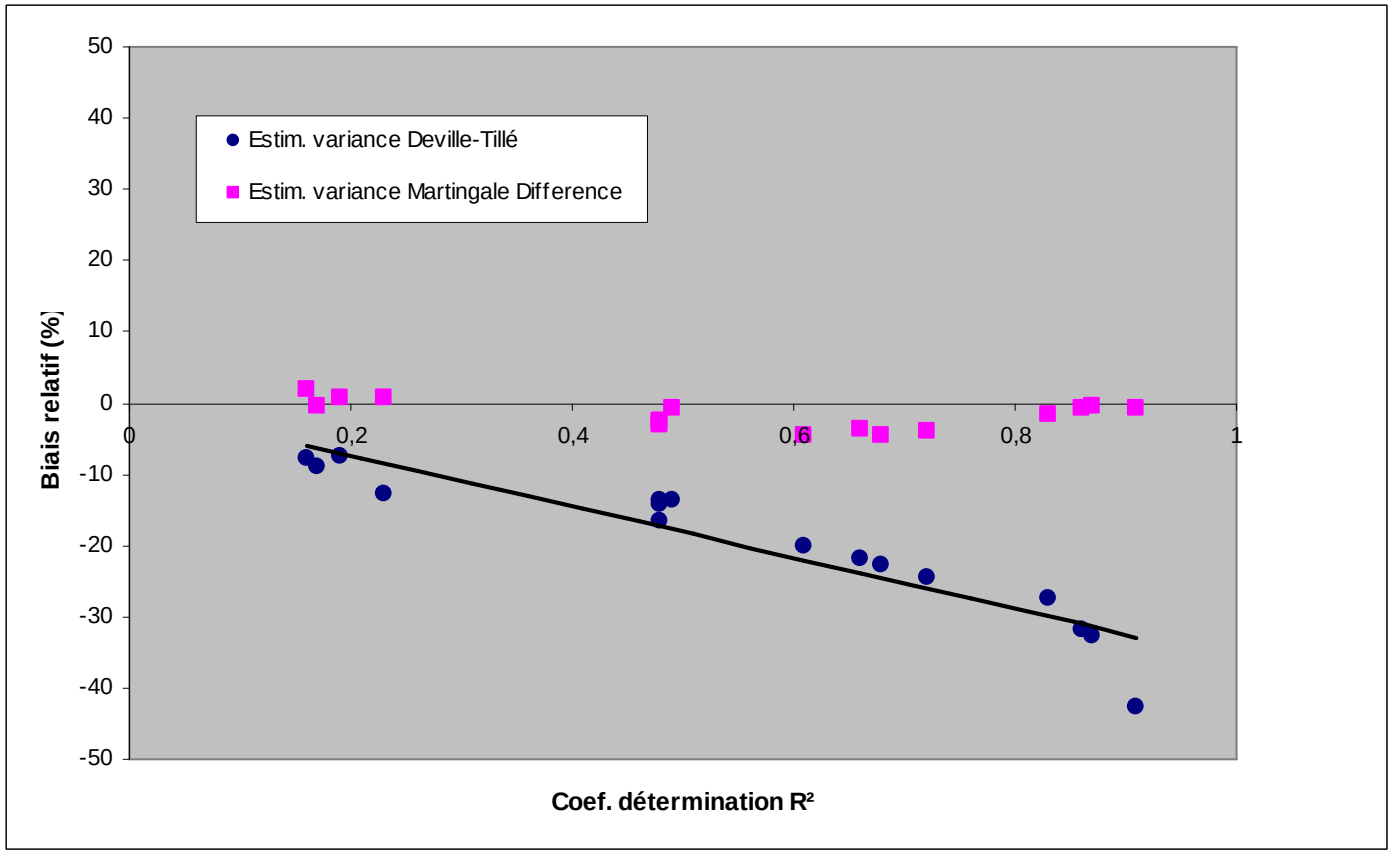
$$RB_{MC}(\hat{\theta}) \quad = \quad 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{\theta}_{(b)} - \theta}{\theta}.$$

Les graphiques suivants donnent, pour chaque composante de la variance, le biais relatif d'un estimateur de variance en fonction du coefficient de détermination (R^2) obtenu en prédisant la variable d'intérêt par les variables utilisées lors de l'équilibrage.

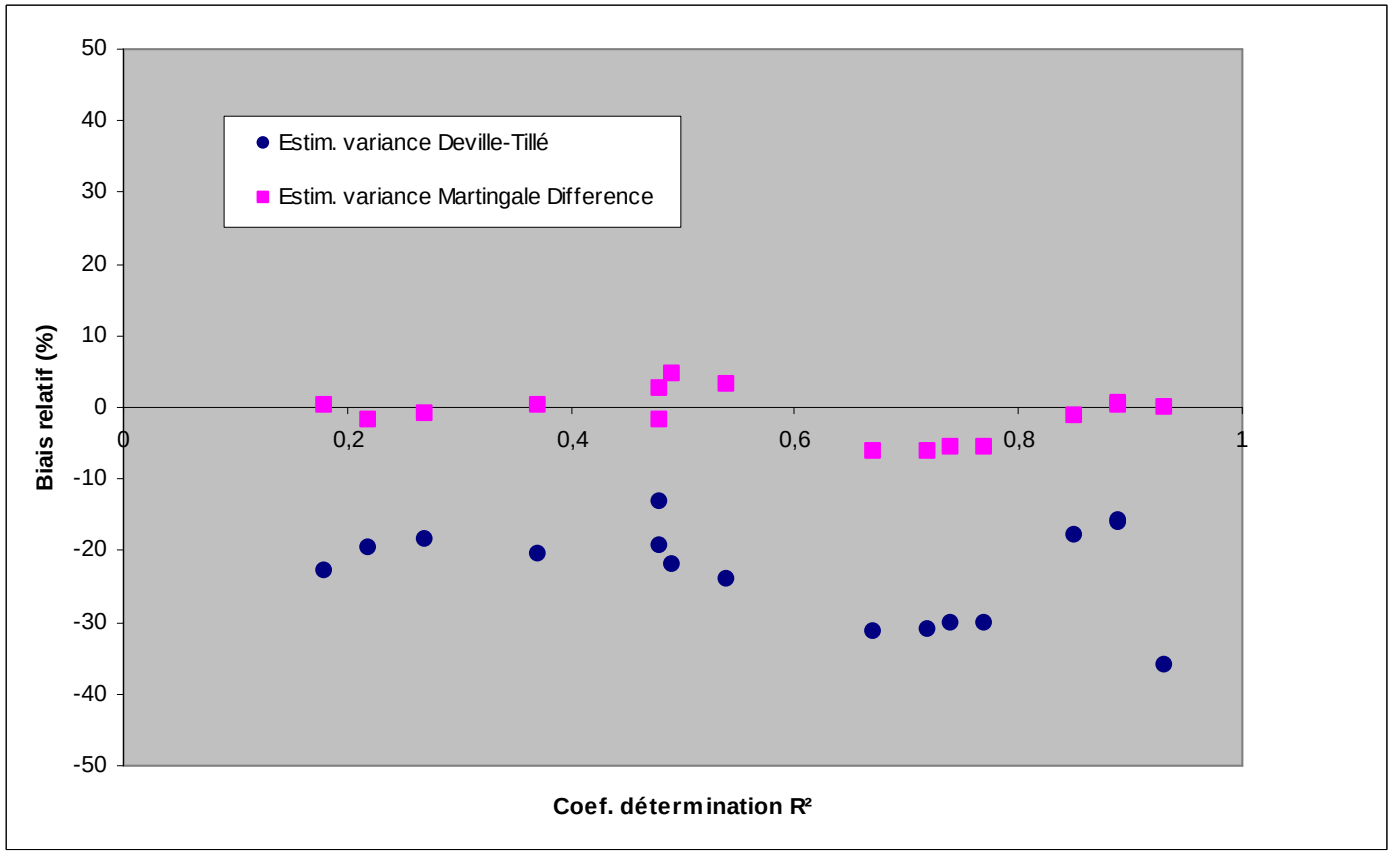
Estimation de la variance EM



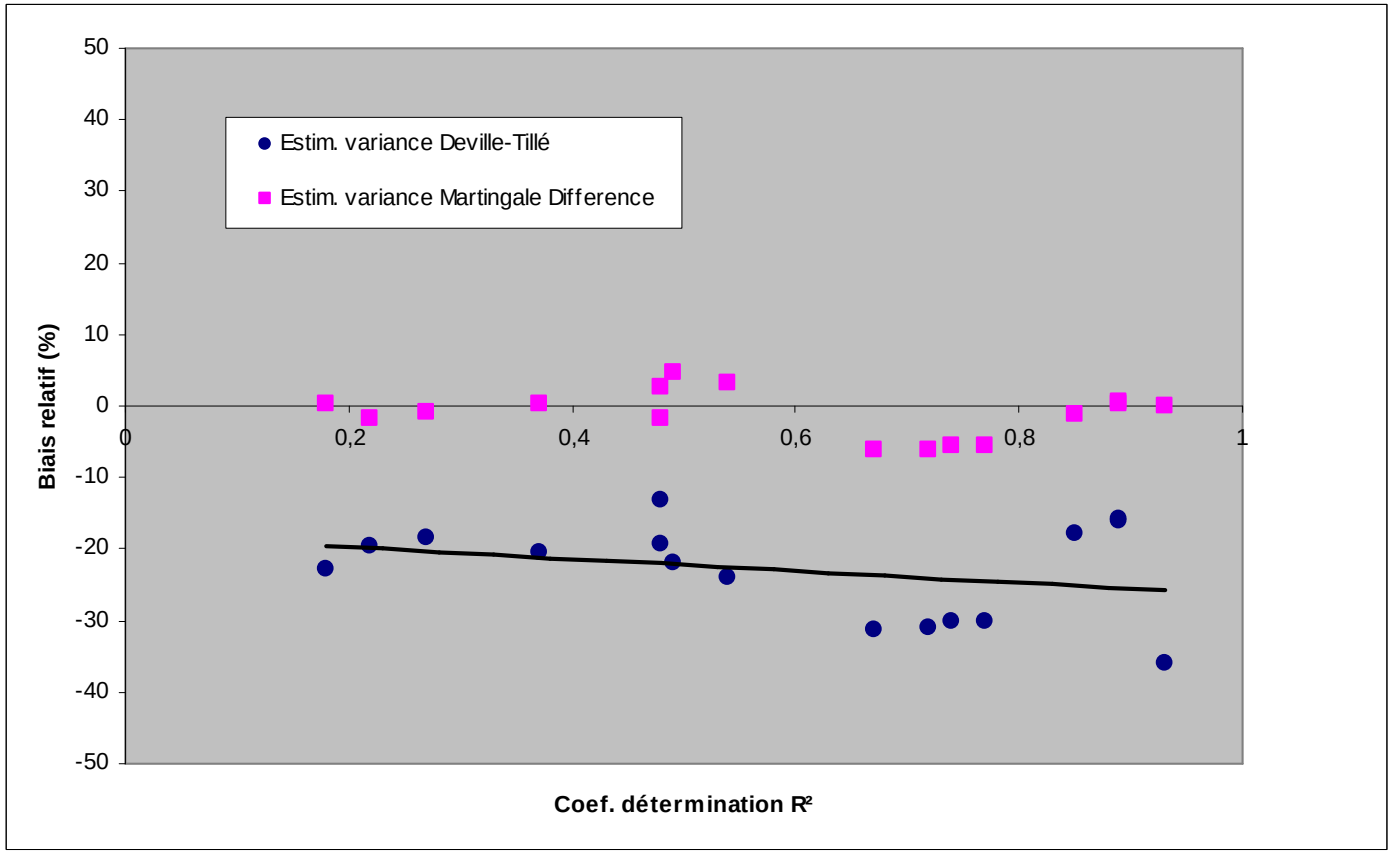
Estimation de la variance EM



Estimation de la variance Recensement



Estimation de la variance Recensement



Bibliographie

Ardilly, P. (2006). *Les techniques de Sondage*. Paris, Technip.

Breidt, F.J., and Chauvet, G. (2011). *Improved variance estimation for balanced samples drawn via the Cube method*. Journal of Statistical Planning and Inference, 141, 479-487.

Chauvet, G. (2011). *On variance estimation for the French Master Sample*. Journal of Official Statistics, 27, pp. 651-668.

Christine, M., and Faivre, S. (2009). *OCTOPUSSE : un système d'Echantillon-Maître pour le tirage des échantillons dans la dernière EAR*. JMS, Paris.

Deville, J.-C., and Tillé, Y. (2004). *Efficient balanced sampling : the cube method*. Biometrika, 91, pp. 893–912.

Deville, J-C., and Tillé, Y. (2005). *Variance approximation under balanced sampling*. Journal of Statistical Planning and Inference, 128, 569-591.