Estimation de la variance liée au nouveau plan de sondage de l'Enquête Emploi en Continu (EEC)

Lionel Delta

DG Insee - Département des méthodes statistiques - Division Sondages

Journées de méthodologie statistique de l'Insee (JMS2022)

Sommaire

Chapitre 1

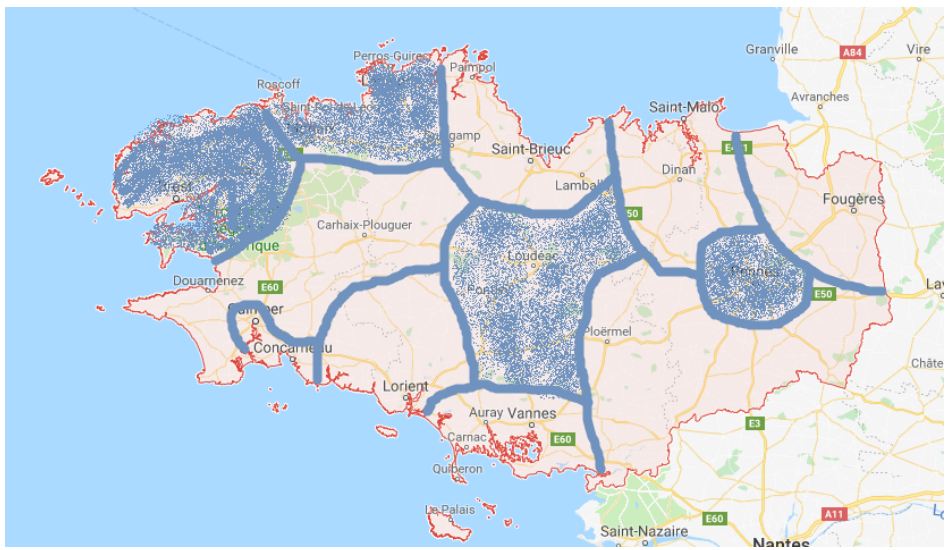
Le plan de sondage de l'EM et celui de l'EEC

Quelques éléments de contexte - Le principe de l'EM

- D'un point de vue pratique, 4 étapes :
  - Créer une partition du territoire ;

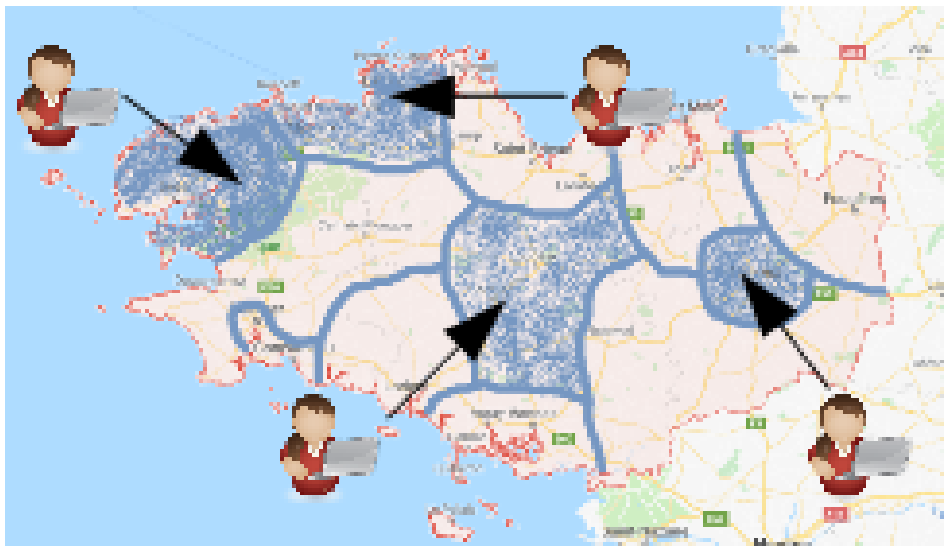
Quelques éléments de contexte - Le principe de l'EM

- D'un point de vue pratique, 4 étapes :
  - Créer une partition du territoire ;
  - Sélectionner *intelligemment* les zones ;



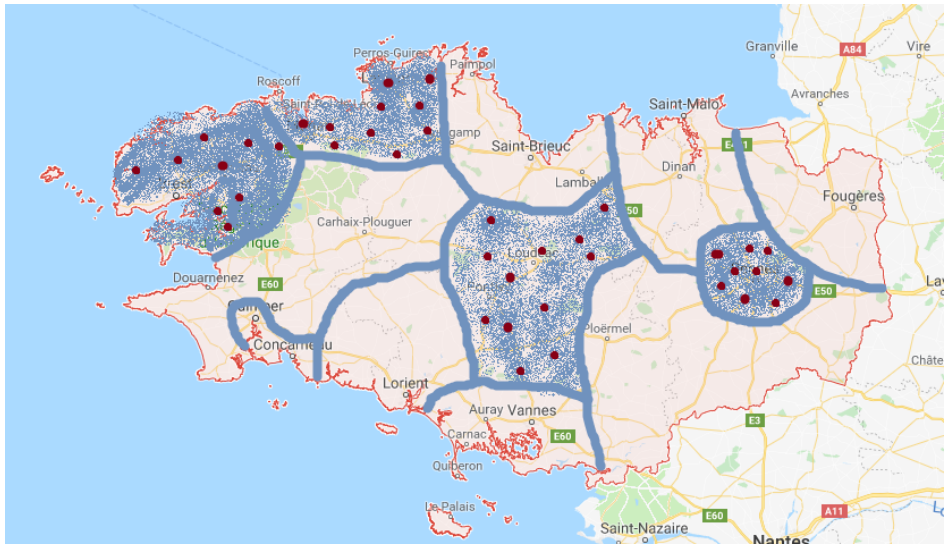
Quelques éléments de contexte - Le principe de l'EM

- D'un point de vue pratique, 4 étapes :
  - Créer une partition du territoire ;
  - Sélectionner *intelligemment* les zones ;
  - Affecter un enquêteur à chaque zone ;



Quelques éléments de contexte - Le principe de l'EM

- D'un point de vue pratique, 4 étapes :
  - Créer une partition du territoire ;
  - Sélectionner *intelligemment* les zones ; tirage équilibré sur plusieurs variables auxiliaires et spatialement équilibré
  - Affecter un enquêteur à chaque zone ;
  - Tirer des logements dans chaque zone, pour chaque enquête, à partir d'une base de sondage.



Particularités de l'Enquête Emploi en Continu

- L'échantillon est « aréolaire » : il n'est pas issu directement d'un tirage de logements, mais d'un tirage de groupes de logements proches.
- Ces groupes, ou « grappes », sont des ensembles géographiquement compacts d'une vingtaine de logements.
- L'échantillon est construit selon une logique ascendante : construction de grappes, puis de secteurs comportant en général 6 grappes qui seront mises en collecte successivement, chacune 6 trimestres d'affilée, sur un cycle de 9 ans.
- Tirer l'échantillon EEC au sein de l'EM risque de saturer la contrainte d'un tirage maximum de chaque logement sur la durée de vie de l'EM (pour limiter le fardeau de réponse)

L'échantillon EEC : tirage coordonné au-delà de l'EM

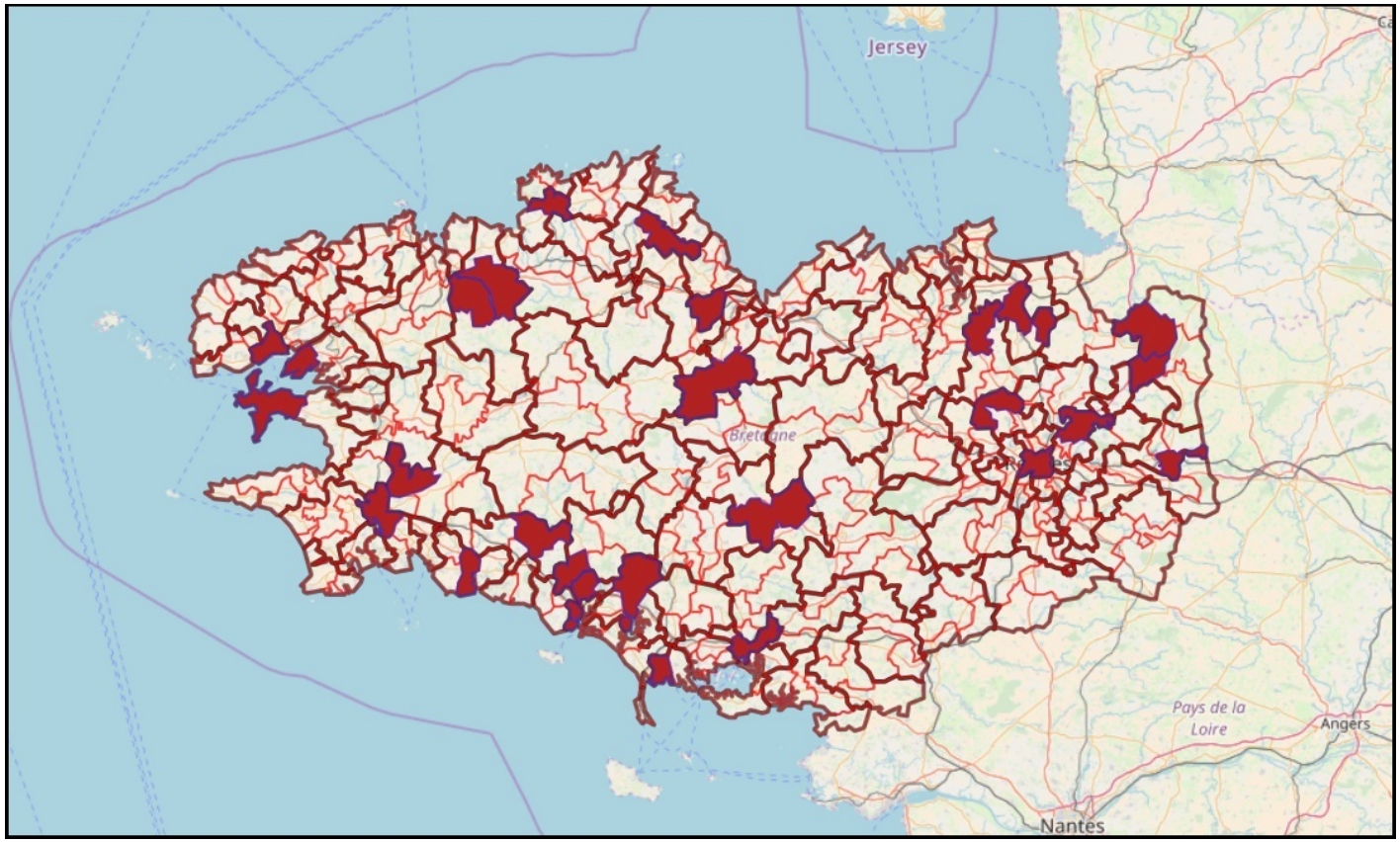
Pour pallier le risque de saturation de l'EM, le choix est fait de tirer les secteurs emplois dans un échantillon de zones plus étendues que les UP.

Cet échantillon spécifique intègre des zones où sont déjà affectés des enquêteurs pour l'EM : cela permet de limiter les déplacements enquêteur.

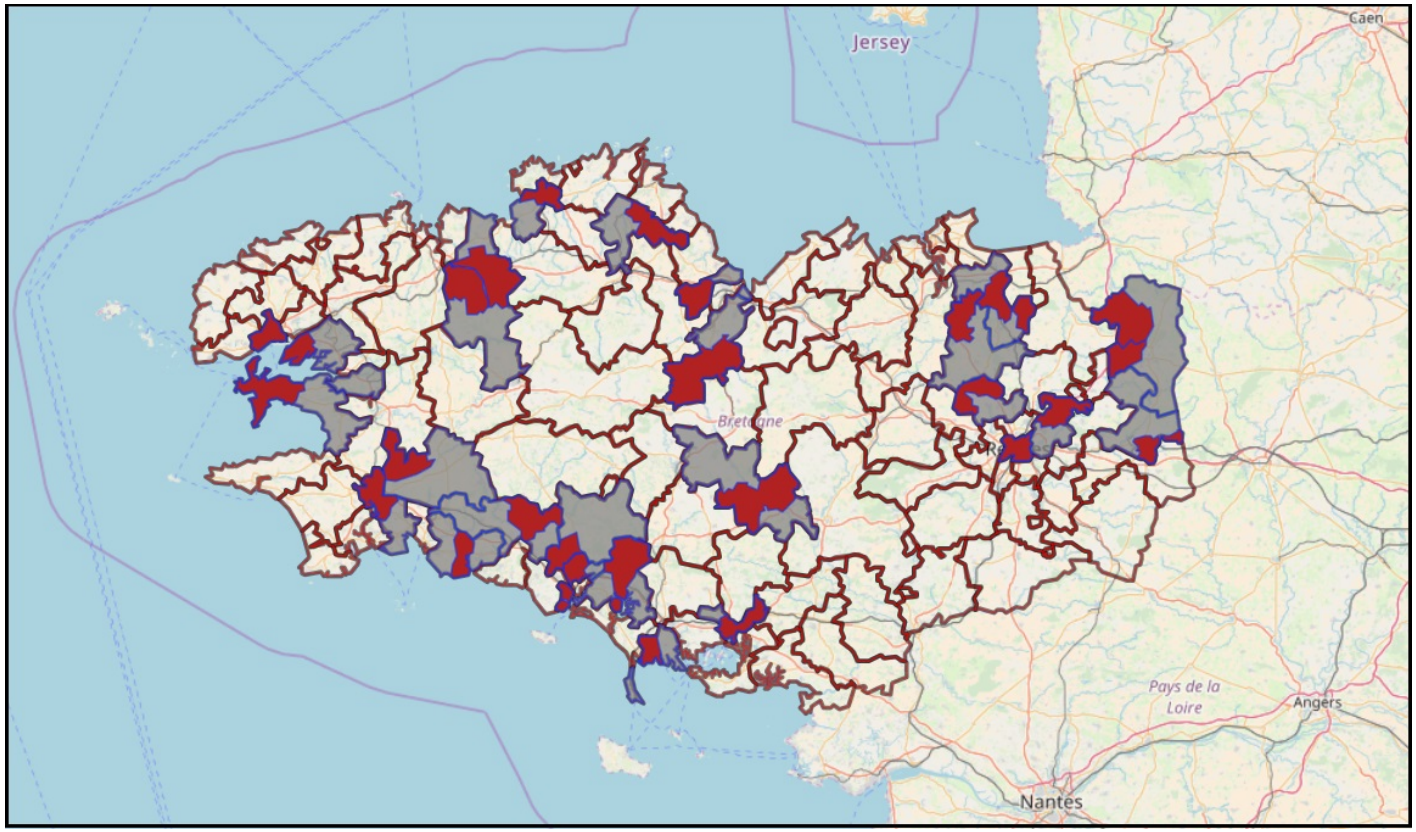
La coordination de l'EEC et de l'EM

- Les UC (unités de coordination) sont des regroupements d'unités primaires
- Une UC doit contenir au moins 10 000 logements. Dans la majorité des cas, une UC est le regroupement de 1 à 4 UP.
- Tirage indirect des UC
  - 1e étape : Tirage des UP
  - 2e étape : Récupération des UC associées aux UP tirées
  - 3e étape : Tirage des secteurs dans les UC tirées

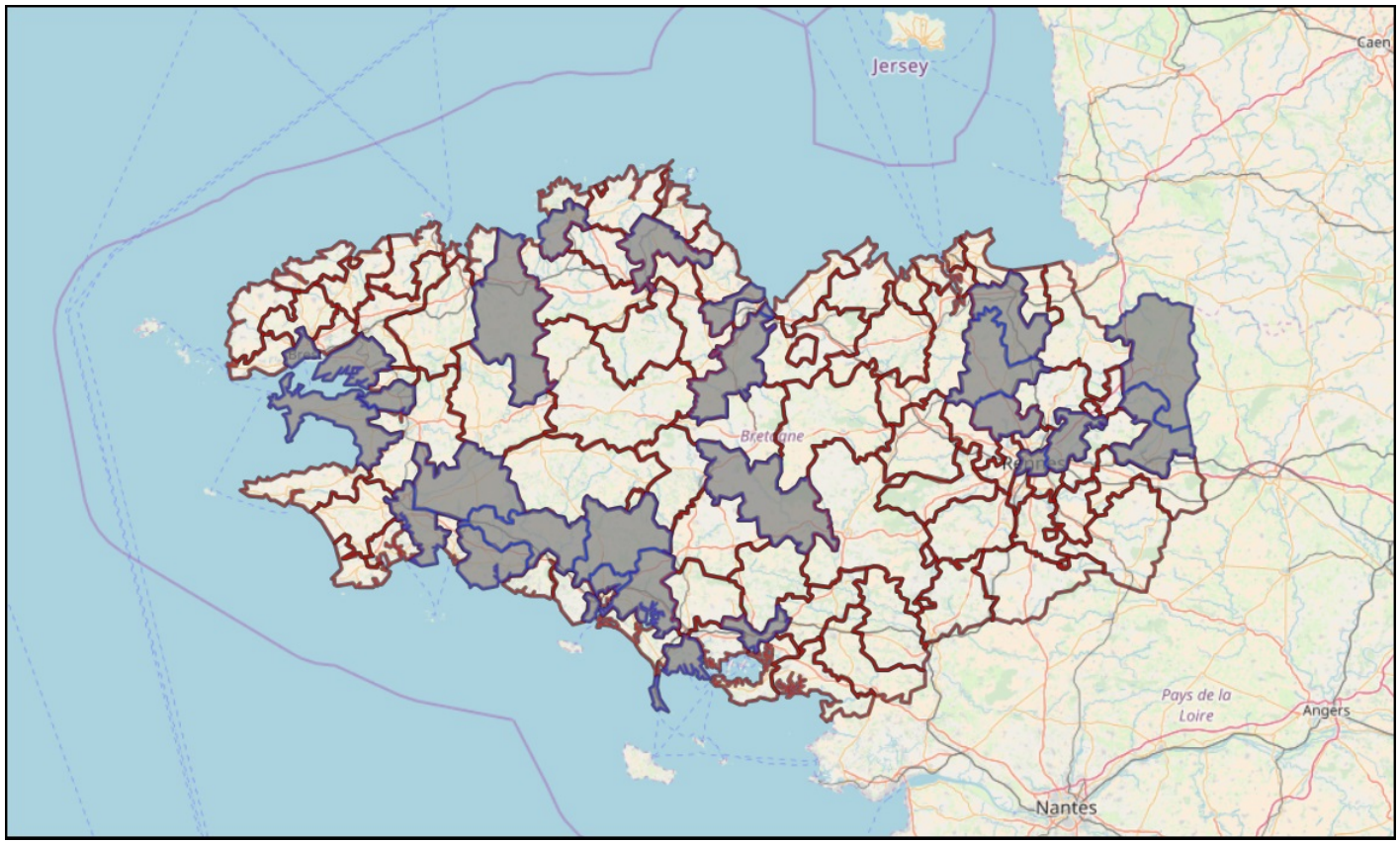
1e étape : Tirage des UP



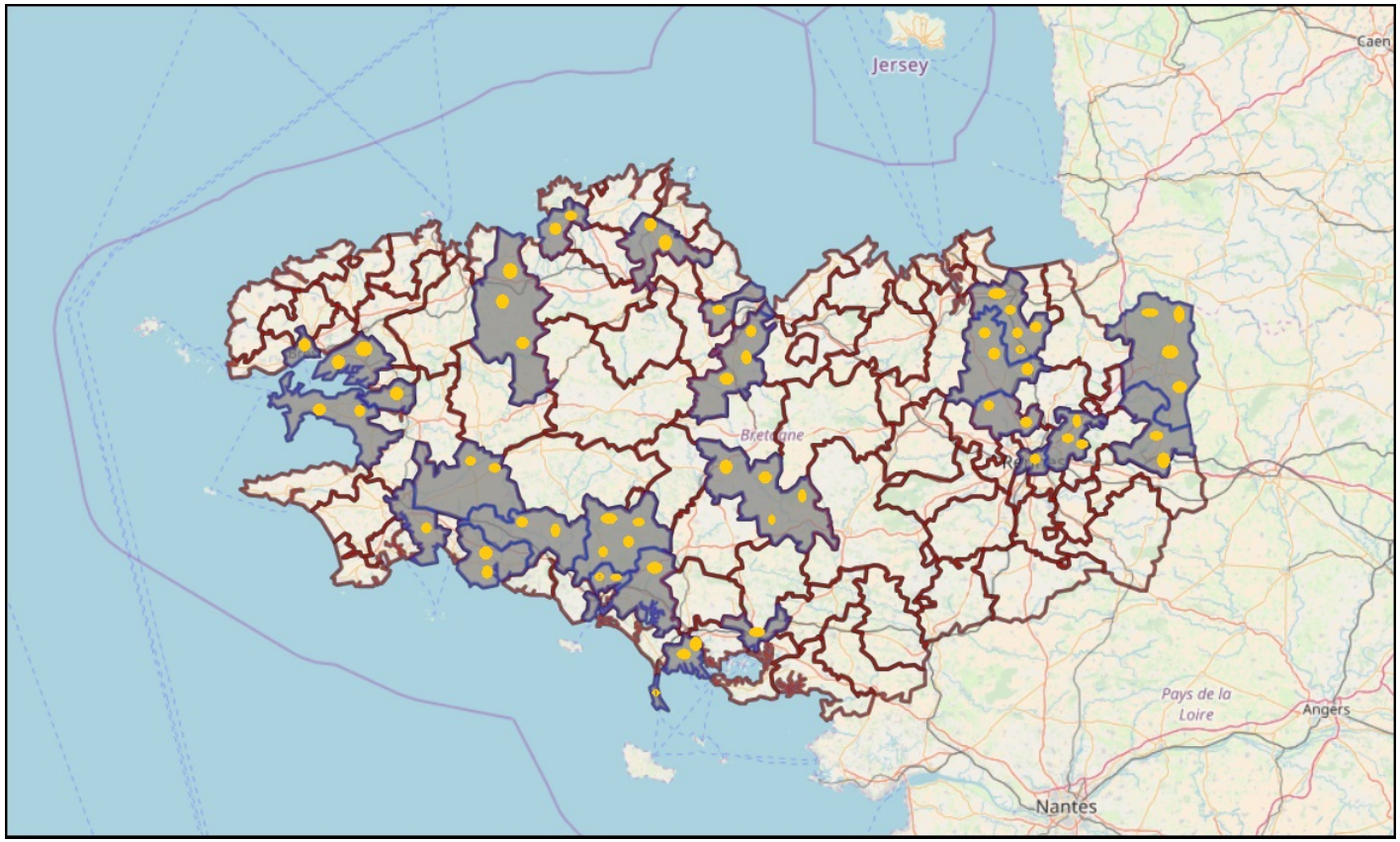
2e étape : Récupération des UC (1/2)



2e étape : Récupération des UC (2/2)



3e étape : Tirage des secteurs



Plan de sondage de l'enquête emploi en continu

1. Tirage de 541 UP : tirage doublement équilibré (sur 57 variables d'équilibrage et équilibrage spatial) selon π^{UP}, proportionnes à la taille en nombre de résidences principales, suivant un plan de sondage doublement équilibré, stratifié par régions (26 strates de 20 UP en moyenne) avec atterrissage national.

2. Sélection indirecte des UC par identification des UC incluant une ou plusieurs UP tirées dans l'EM

Plan de sondage de l'enquête emploi en continu

3. Tirage des secteurs : plan de sondage doublement équilibré (spatialement et sur 17 variables auxiliaires) stratifié par UC (entre 4 et 9 secteurs tirés par UC) avec atterrissage régional. Les probabilités d'inclusion conditionnelles, notées π_{SE}, sont égales pour tous les secteurs appartenant à une même UC.

4. Tirage d'une grappe de logements en plusieurs phases au sein des différents secteurs : tirage à probabilités inégales au sein des secteurs.

Quel taux de sondage à chaque étape ?

1. Tirage de 541 UP parmi 5064, soit environ une UP sur dix
2. Le tirage de l'EM a conduit à retenir 524 UC parmi les 1646 soit environ une UC sur 3. Les UC de l'échantillon regroupent 100 039 secteurs : du point de vue des secteurs, le taux de sondage résultant du tirage indirect des UC s'élève à 43%.
3. Tirage de 2944 secteurs au sein des UC de l'échantillon. Conditionnellement à l'étape précédente cela fait un peu moins de 3%
4. Taux de sondage de 1/6 ou 1/7 pour le tirage des grappes au sein des secteurs.

Hors sélection des logements, le tirage de l'échantillon emploi conduit à un taux de sondage global de l'ordre de 1/500.

Chapitre 2

Méthodologie de recherche d'un estimateur de variance

Partie 1

Forme d'estimation de la variance

Principe de calcul de variance dans le cas d'un sondage à plusieurs phases (1/4)

Avec les notations :

- Π_1, le plan de sondage de première phase
- S_1 l'échantillon de première phase
- Π_2, le plan de sondage de deuxième phase
- S_2 l'échantillon de deuxième phase
- $\pi_k^{(2)}$ la probabilité d'inclusion de l'unité k de S_2 conformément au plan Π_2 et conditionnellement à l'échantillon S_1
- $\pi_i^{(1)}$ la probabilité d'inclusion de l'unité i de S_1 conformément au plan Π_1 (pour l'unité k de $\mathcal{S}_2$, $\pi_k^{(1)}$ désigne donc sa probabilité d'inclusion à $\mathcal{S}_1$)

L'estimateur classique par expansion s'écrit alors :

$$T(\hat{Y}) = \sum_{k \in S_2} \frac{y_k}{\pi_k^{(1)} \pi_k^{(2)}}$$

Principe de calcul de variance dans le cas d'un sondage à plusieurs phases (2/4)

La variance de l'estimateur du total est toujours donnée par :

$$\mathbb{V}_{\Pi_1,\Pi_2}(T(\hat{Y})) = \mathbb{V}_{\Pi_1}(\mathbb{E}_{\Pi_2}(T(\hat{Y})/\mathcal{S}_1)) + \mathbb{E}_{\Pi_1}(\mathbb{V}_{\Pi_2}(T(\hat{Y})/\mathcal{S}_1))$$

où $\mathbb{V}_{\Pi_1}(\mathbb{E}_{\Pi_2}(T(\hat{Y})/\mathcal{S}_1)) = \mathbb{V}_{\Pi_1}(\sum_{i\in\mathcal{S}_1} \frac{y_i}{\pi_i^{(1)}})$ est la variance résultant de la 1ère phase de sondage.

Le 2nd terme, $\mathbb{E}_{\Pi_1}(\mathbb{V}_{\Pi_2}(T(\hat{Y})/\mathcal{S}_1))$ correspondant à la variance additionnelle générée par la 2ème phase de sondage, est en général approximé selon :

$$\mathbb{E}_{\Pi_1}(\mathbb{V}_{\Pi_2}(T(\hat{Y})/\mathcal{S}_1)) \approx \mathbb{V}_{\Pi_2}(T(\hat{Y})/\mathcal{S}_1)$$

D'où la nécessité d'estimer la variance due à chaque phase de sondage.

Principe de calcul de variance dans le cas d'un sondage à plusieurs phases (3/4)

Pour estimer : $\mathbb{V}_{\Pi_1}(\sum_{i\in\mathcal{S}_1}\frac{y_i}{\pi_i^{(1)}})+\mathbb{V}_{\Pi_2}(\sum_{k\in\mathcal{S}_2}\frac{y_k}{\pi_k^{(1)}\pi_k^{(2)}}/\mathcal{S}_1)$ on peut donc avoir recours à deux estimateurs $\hat{\mathbb{V}}_1$ et $\hat{\mathbb{V}}_2$ pour chacun des termes.

$\hat{\mathbb{V}}_1$ est généralement de forme quadratique :

$$\hat{\mathbb{V}}_1=\sum_{i\in\mathcal{S}_1}\sum_{j\in\mathcal{S}_1}q_{ij}y_iy_j$$

En pratique, on n'observe pas l'échantillon de première phase mais uniquement celui de seconde phase. La solution théorique (Särndal, Swensson, 1987) est d'utiliser l'estimateur dérivé :

$$\tilde{\mathbb{V}}_1=\sum_{k\in\mathcal{S}_2}\sum_{l\in\mathcal{S}_2}\frac{q_{kl}}{\pi_{kl}^{(2)}}y_ky_l$$

Principe de calcul de variance dans le cas d'un sondage à plusieurs phases (4/4)

$$\tilde{\mathbb{V}} = \sum_{k \in \mathcal{S}_2} \sum_{l \in \mathcal{S}_2} \frac{q_{kl}}{\pi^{(2)}_{kl}} y_k y_l + \hat{\mathbb{V}}_2$$

est un estimateur de variance totale utilisable sous réserve :

- de mettre en place un estimateur sans biais de la variance de 1ere phase à partir de $\mathcal{S}_1$
- de mettre en place un estimateur résultant de la variance de 2eme phase conditionnellement à l'échantillon de 1ere phase $\mathcal{S}_1$
- d'obtenir des probabilités d'inclusions doubles $\pi^{(2)}_{kl}$ liées au tirage de la 2eme phase... sauf dans le cas particulier où les q_{kl} de la forme quadratique $\hat{\mathbb{V}}_1$ sont nuls pour $k \neq l$

Le principe peut être itéré dans le cas d'un sondage à plus de 2 phases.

Simplification avec la formule de Rao

Cas particulier du sondage à deux degrés (ou plus). C'est un type de sondage dans lequel la première phase est un tirage par grappes (grappes = unités primaires) et on tire des unités secondaires lors de la deuxième phase indépendamment au sein de chaque unité primaire.

Dans ce cas l'estimateur de variance utilisé correspond à la formule dite de Rao :

$$\tilde{\mathbb{V}} = \sum_{i \in \mathcal{S}_1} \sum_{j \in \mathcal{S}_1} q_{ij} \hat{y}_i \hat{y}_j + \sum_{i \in \mathcal{S}_1} ((\frac{1}{\pi_i^{(1)}})^2 - q_i) \hat{\mathbb{V}}_{2,i}$$

où $\hat{y}_i = \sum_{k \in \mathcal{S}_{2,i}} \frac{y_k}{\pi_k^{(2)}}$ est l'estimateur Horvitz-Thompson du total de y_i à partir de l'échantillon d'unités secondaires issu de l'UP i et dont la variance est notée $\hat{\mathbb{V}}_{2,i}$

Le plan de sondage de l'EEC : plusieurs degrés ou plusieurs phases ?

En théorie, l'hypothèse d'indépendance, cruciale pour parler de sondages à plusieurs degrés, n'est pas respectée lors du tirage des secteurs au sein des UC.

Pour rappel, il s'agit d'un plan de sondage :

- équilibré spatialement
- équilibré sur données auxiliaires
- stratifié par unité de coordination...
- ... mais avec une phase d'atterissage régional

Au sein, d'une même région, le tirage des secteurs n'est pas totalement indépendant d'une UC à l'autre.

Le plan de sondage de l'EEC : plusieurs degrés ou plusieurs phases ?

En théorie, l'hypothèse d'indépendance, cruciale pour parler de sondages à plusieurs degrés, n'est pas respectée lors du tirage des secteurs au sein des UC.

Pour rappel, il s'agit d'un plan de sondage :

- équilibré spatialement
- équilibré sur données auxiliaires
- stratifié par unité de coordination...
- ... mais avec une phase d'atterissage régional

Au sein, d'une même région, le tirage des secteurs n'est pas totalement indépendant d'une UC à l'autre.

À quel point sommes-nous éloignés de la situation d'indépendance parfaite ?

Partie 2

Méthode de validation

Recours à des simulations de tirage

Les simulations ont plusieurs buts :

- estimer empiriquement la variance liée à chaque étape du plan de sondage (unités de coordination, UP, secteurs)
- estimer la variance liée à l'ensemble du plan de sondage
- estimer empiriquement les probabilités d'inclusion double des unités (non calculables théoriquement)
- calculer les estimateurs de variance sur tout échantillon simulé et estimer empiriquement l'espérance, le biais, la variance et autre indicateur portant sur les estimateurs envisagés.

Les campagnes de simulations réalisées

3 grandes campagnes de simulations de tirage d'échantillon ont été réalisées :

- Un scénario de tirage de la chaîne entière allant du tirage des UP jusqu'au tirage des grappes (un demi-million de tirages)
- Un scénario pour le tirage des UP et UC uniquement (environ 50 millions de tirages)
- Un scénario pour le tirage des secteurs conditionnellement à l'échantillon d'UC effectivement entré en service (3 millions de tirages)

Les critères de validation

Divers indicateurs ont été calculés sur la distribution des estimateurs d'erreur-type découlant de la variance propre à chaque étape (et variance globale, cas de la formule de Rao) :

- biais
- variance

- le taux de couverture de l'intervalle de confiance à 95% : $T(Y)$ doit être dans $\hat{IC}_{95} = \hat{T}_y \pm 2\hat{\sigma}$ dans environ 95% des échantillons possibles

Indicateurs emploi

L'enquête emploi en continu fournit chaque trimestre des indicateurs nationaux, le plus souvent en ratio, sur la population en âge de travailler. Si l'on sait estimer la variance d'un total on pourra estimer la variance d'un ratio via des techniques de linéarisation.

La répartition de la population en âge de travailler selon le statut d'activité y est principalement déclinée :

- par région
- par sexe
- par âge
- par croisement sexe/âge au niveau national

Idée : Construire des variables proxys de ces indicateurs pour tester la pertinence des estimateurs de variance.

Variables d'intérêt utilisées

En pratique deux grandes familles de variables d'intérêts ont été mobilisés

1. des variables issues de la source fiscale (fortement corrélées avec les variables d'équilibrage) :
   - indicatrice de perception de revenus du travail, revenus d'assurance chômage, de pensions de retraite
   - croisement avec l'âge, le sexe, le lieu de résidence
   - variables de revenus en montant
2. des variables issues de la source démographique (davantage assimilables aux futures variables d'enquêtes) :
   - variables en lien avec le marché du travail
   - autres variables, moins corrélées avec l'emploi, collectées au recensement

Chapitre 3

Estimateurs et résultats

Partie 1

Variance de première phase

Panorama des différents estimateurs de variance envisagés

Pour rendre compte de la variance propre à chaque étape, choix (parmi les estimateurs présentés dans la littérature) entre les estimateurs de :

- Deville : adapté pour les tirages à probabilités inégales
- Deville-Tillé : tirage équilibré sur données auxiliaires (néglige la phase d'atterissage) ; variance des termes résiduels
- Grafström-Tillé : tirage doublement équilibré (dérive de l'estimateur de Deville-Tillé) ; termes évalués localement.
- Sen-Yates-Grundy : tirage sans remise et à taille fixe (nécessite de connaître les probabilités d'inclusion doubles des unités)
- Horvitz-Thompson : tout type de plan de sondage (probabilités d'inclusion doubles)

Quel échantillon pour la variance de premier niveau ? (1/2)

Il est possible de considérer directement (avantage de la simplicité) l'échantillon des UC (unités au sein desquelles on tire directement les secteurs) pour rendre compte de la variance de 1ere phase

Inconvénients :

- Plan de sondage indirect difficilement modélisable : équilibré ? stabilité ?
- Probabilités de tirage inconnues
- plan de sondage à taille non fixe (empiriquement, taille quasi fixe autour de 523 UC)

Estimateurs utilisables sous certaines hypothèses :
Deville, Horvitz-Thompson et Sen-Yates-Grundy.

Quel échantillon pour la variance de premier niveau ? (2/2)

Alternative : Utiliser l'échantillon d'UP car tout estimateur de total de niveau UC peut s'exprimer, après partage de poids, comme un estimateur de niveau UP à partir de variables d'intérêts dites transformées.
L'inconvénient principal a priori est la complexité que cela rajoute dans les chaînes de calcul de variance.

Avantages :

- échantillon tiré directement donc plan de sondage connu : échantillon à peu près équilibré, spatialement réparti, etc ;
- plan de sondage à taille fixe
- Les probabilités d'inclusion simple sont connues (mais pas les probabilités d'inclusion multiple)

Estimateurs utilisables a priori facilement : Deville, Deville-Tillé et Grafström-Tillé et, sous certaines hypothèses :
Horvitz-Thompson et Sen-Yates-Grundy.

Variance rendant compte du tirage des UP

Au niveau national et sur les variables d'intérêts issues de la source fiscale, on obtient les résultats suivants (% en moyenne sur les variables) :

Estimateurs	Biais relatif	CV	taux de couverture IC95
Deville-Tillé*	13.3	7.3	95.9
Sen-Yates-Grundy	62.9	72.4	99.7

Dans le cas de l'estimateur de Deville, on obtient un CV relativement faible (du même ordre que Deville-Tillé) mais une très forte surestimation de la variance (jusqu'à plusieurs dizaines de fois supérieure à la variance attendue pour certaines variables).

Compléments

- L'estimateur de variance de Horvitz-Thompson donnent des résultats encore moins bons
- La formule de variance $Var(\hat{T}_y) = -\frac{1}{2}\sum_{i\in U}\sum_{j\in U, j\neq i}(\pi_{ij} - \pi_i\pi_j)(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2$ permet bien de retrouver la variance empirique mais c'est l'estimateur $\hat{V}(\hat{T}_y) = -\frac{1}{2}\sum_{i\in S}\sum_{j\in S, j\neq i}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2$ qui est biaisé et instable
- L'estimateur est-il adapté à un tel plan de sondages ? (forte répulsion entre certaines unités)

Hypothèses d'utilisation de l'estimateur de Deville-Tillé

L'estimateur de Deville-Tillé, dont les résultats sont présentés ici, utilise 27 variables d'équilibrage. C'est cet estimateur qui semble le plus adapté pour le niveau UP (les résultats semblent moins bons avec Grafström-Tillé)

En pratique, le plan de sondage équilibré ne conduit pas à une estimation parfaite du total des variables d'équilibrage utilisées. Ces estimations présentent une certaine variance, relativement faible pour la moitié des variables, et nettement plus élevée pour d'autres.

Hypothèse simplificatrice : on assimile le plan de sondage à un tirage équilibré sur un nombre plus restreint de variables d'équilibrage (CV d'estimation inférieur à un certain seuil).

Résultats mitigés sur les variables transformées

Lorsque l'on utilise directement des variables de niveau UC, transformées au niveau UP, on observe une dégradation des résultats obtenus avec l'estimateur de Deville-Tillé (** voir dernière ligne)

Estimateurs	Biais relatif	CV	taux de couverture IC95
Deville-Tillé *up	13.3	7.3	95.9
Sen-Yates-Grundy	62.9	72.4	99.7
Deville-Tillé **up-T	29.6	6.9	98.4

L'utilisation de la formule de Deville-Tillé est d'autant moins adaptées aux variables transformées de niveau UC, que ces variables sont corrélées aux variables mobilisées pour l'équilibrage.

Variance rendant compte du tirage des UC (1/2)

La recherche d'estimateurs de variance directement appliquables aux variables d'intérêt de niveau UC constitue l'autre voie majeure explorée pour le calcul de variance.

Compte-tenu du plan de sondage permettant le tirage final d'un échantillon d'UC, les différents estimateurs envisagés apparaissent beaucoup moins naturels que précédemment et la plupart d'entre eux conduisent à des résultats encore insatisfaisants :

- L'estimateur de Deville surestime encore fortement la variance
- L'estimateur de Deville-Tillé tend à la sous-estimer selon les variables introduites (risque amplifié avec Grafström-Tillé)

Variance rendant compte du tirage des UC (2/2)

Option finalement retenue : l'utilisation de l'estimateur de Sen-Yates-Grundy pour lequel on obtient les résultats suivants :

Type de variables	Biais relatif	CV	taux de couverture IC95
Fiscales	10.2	20	95.8
Démographiques (emploi)	4.1	17	95.1
Autres démographiques	5.3	19	95.2

Partie 2

Variance de seconde phase

Utilisation de la formule de Rao (1/3)

En dépit de l'écart à la situation d'indépendance théorique de tirage des secteurs au sein des UP, on a cherché à valider la possibilité d'utiliser la formule de Rao pour rendre compte des 2 premières phases successives de tirage. Rappel de la formule :

$$\tilde{\mathbb{V}} = \sum_{i \in \mathcal{S}_1} \sum_{j \in \mathcal{S}_1} q_{ij} \hat{y}_i \hat{y}_j + \sum_{i \in \mathcal{S}_1} ((\frac{1}{\pi_i^{(1)}})^2 - q_i) \hat{\mathbb{V}}_{2,i}$$

Analyse empirique du biais relatif en utilisant les niveaux attendus de variance conditionnelle

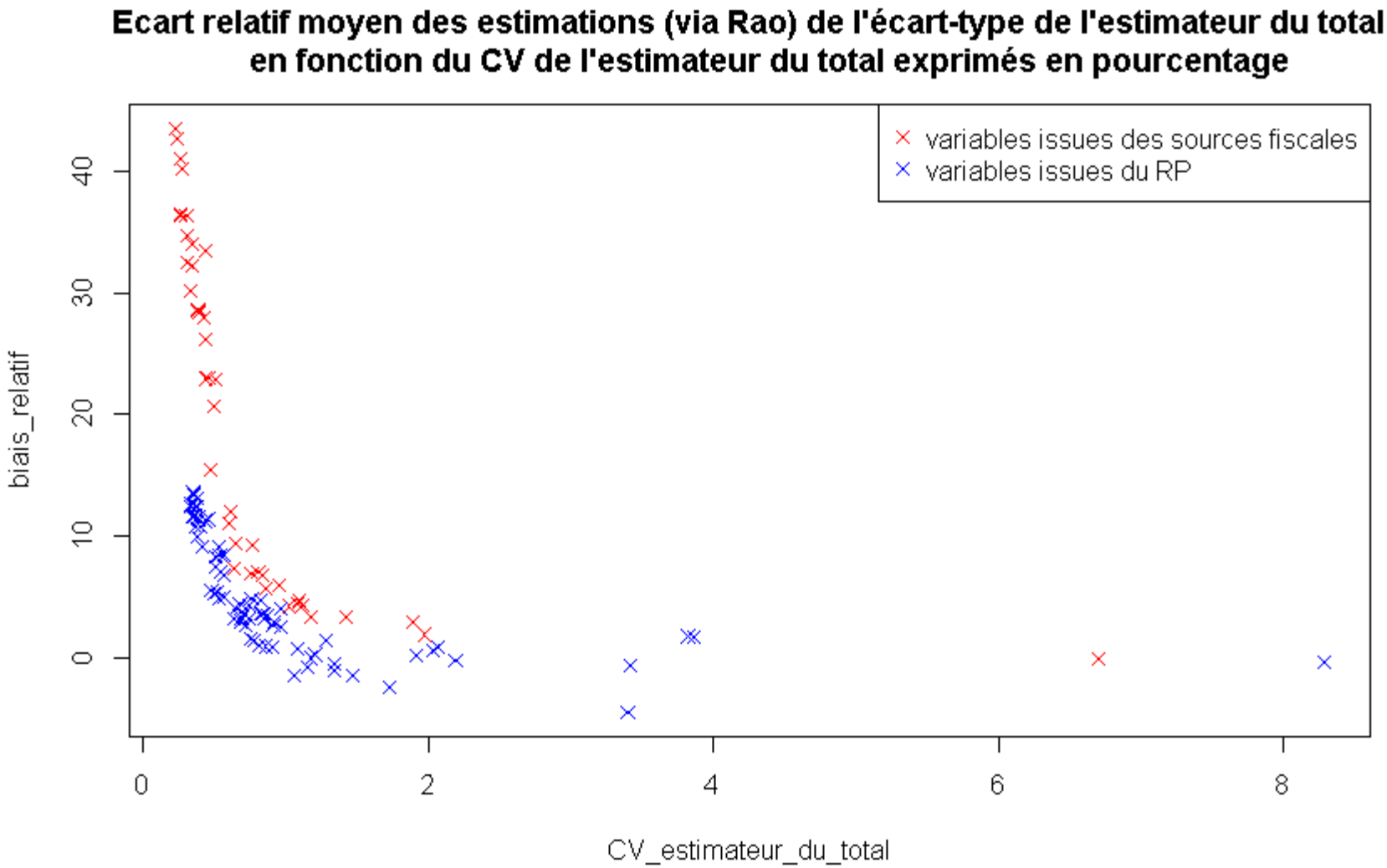
Utilisation de la formule de Rao (2/3)

Analyse empirique du biais relatif en utilisant les niveaux attendus de variance conditionnelle

Type de variables	écart relatif absolu	CV
Fiscales	19.6	6
Démographiques	5.6	10
Toutes variables	10.6	8.5

Compte-tenu des interactions avec les variables d'équilibrage liées à l'utilisation des variables fiscales, le profil obtenu pour les variables démographiques s'apparente plus à celui des futures variables de collecte.

Utilisation de la formule de Rao (3/3)



Variance conditionnelle de tirage des secteurs

Malgré un tirage stratifié, doublement équilibré, il n'est pas possible de pleinement tenir compte de l'équilibrage au sein de chaque UC (4 variables d'équilibrages prises en compte au maximum).

Les analyses empiriques portant sur l'ensemble des UC des écarts-relatifs moyens de variances conditionnelles a conduit à retenir le choix de l'estimateur de Deville-Tillé sur un nombre restreint de variables auxiliaires, à l'exception des UC avec petits échantillons (moins de 6 secteurs tirés, estimateur de Deville).

Ce choix d'estimateurs a également été validé dans le cadre d'analyses empiriques avec la formule de Rao, ce qui a permis de confirmer les premiers résultats portant sur les variances empiriques.

Pour finir

La prise en compte des étapes suivantes de tirage est inchangée dans le calcul de variance : sondage en grappes, tirage d'étages, prise en compte de la non-réponse.

Pour ces étapes également, recours à la formule de Rao.

Premiers résultats diffusés pour l'estimation de variance au 4ème trimestre 2020 (premier trimestre pour lequel l'échantillon provient entièrement du nouveau plan de sondage).