LA CONSTRUCTION DU NOUVEL ÉCHANTILLON DE L'ENQUÊTE EMPLOI EN CONTINU À PARTIR DES FICHIERS DE LA TAXE D'HABITATION.

(Version provisoire)

Vincent LOONIS (*)

(*) Insee, Unité Méthodes Statistiques

Introduction

Pour alimenter le débat social autour des statistiques du chômage, l'Insee a expertisé, en 2007, l'ensemble des procédures liées à l'Enquête Emploi en Continu (EEC). Les travaux méthodologiques ont concerné l'échantillonnage, la correction de la non-réponse, les redressements et le calcul de précision (D. Place [1]). Si ce n'est un suivi de la construction neuve qui pourrait être amélioré, ces investissements n'ont pas mis en évidence, compte tenu des contraintes de collecte, de défauts structurels. Ils ont cependant souligné, avec le rapport IGF-IGAS [2], la faiblesse relative de la taille de l'échantillon français comparée à celle des autres pays européens : 75 000 individus interrogés contre 100 000 au Royaume-Uni, 150 000 en Italie ou 180 000 en Espagne. C'est dans ce contexte que l'Institut a décidé, en décembre 2007, de procéder à une augmentation de 50 % de la taille de l'échantillon de l'EEC. Cette décision était assortie de quelques recommandations :

- l'échantillon actuel doit continuer à être mobilisé sans modifications par rapport à sa durée de vie initialement prévue,
- le schéma de rotation et d'interrogation de l'échantillon actuel doit être reconduit,
- l'objectif de 50 % doit être atteint mi-2010. Pour parvenir à cet objectif, la mise en place du nouvel échantillon doit être progressive à compter de janvier 2009. Ce calendrier implique que les Directions Régionales de l'Insee soient informées dès la fin du premier trimestre 2008 de la localisation et du nombre de zones impactées.
- le gain en précision attendu doit être, dans la mesure du possible, plus important que celui que l'on observerait lors d'une augmentation de 50 % de l'échantillon, toutes choses égales par ailleurs.

Ces recommandations ont constitué la feuille de route d'un groupe de travail piloté par l'auteur et auquel ont participé *Nicole Cadenel, Guillaume Chauvet, Marc Christine, Sébastien Durier, Sébastien Hallépée, Anne Flipo et Dominique Place*. Dans cet article, on ne présente pas toutes les options étudiées par le groupe mais seulement celles qui ont finalement été retenues. Dans une première partie, on rappelle les grandes lignes de l'échantillon actuel. Sachant qu'elles sont globalement reconduites, on invite le lecteur intéressé à consulter *M. Christine* [3] pour en apprécier les justifications et les détails. La principale spécificité du nouvel échantillonnage est le recours aux fichiers de la Taxe d'Habitation (TH) comme base de sondage. Si ces fichiers ont déjà été mobilisés pour des enquêtes ponctuelles[1], c'est la première fois qu'ils sont utilisés pour une opération d'échantillonnage de l'ampleur de l'EEC. Dans une seconde partie, on justifie ce recours. L'information géographique disponible dans les fichiers de la TH autorise la création automatique des secteurs et « grappes », éléments structurants de l'échantillon EEC. On présente dans une troisième partie les principes suivis pour cette construction et les résultats obtenus. Une conséquence de cette automatisation est le passage de la *contigüité* territoriale des logements constitutifs des *aires* de l'échantillon actuel à la *proximité* des logements dans les *grappes*[2] du nouvel échantillon. Dans la suite de l'article, le terme « grappe » fera référence au nouvel échantillon, alors que « aire » qualifiera l'échantillon actuel. Ce glissement sémantique n'est pas neutre sur la collecte. Dans le nouvel

[1] Notamment dans le cadre de l'enquête « Prime Pour l'Emploi » de la Drees.
[2] Pour bien marquer la différence, on utilise préférentiellement le terme de « grappes » dans le nouvel échantillon.

échantillon, les enquêteurs se voient remettre des listes de logements à enquêter, incluant la construction neuve. Dans l'échantillon actuel, les enquêteurs ont une description physique de l'aire à enquêter. Ils « ratissent » l'aire afin de repérer tous les logements s'y trouvant, y compris les logements neufs. L'EEC s'inscrit dorénavant dans le cadre standard des enquêtes de l'INSEE auprès des ménages. La quatrième partie traite de la répartition de l'échantillon par région et du calendrier d'introduction. Le mode de sélection fera l'objet d'une cinquième partie. Dans une sixième partie, on verra comment ont été prises en compte des contraintes de collecte, qui n'existaient pas pour l'échantillon actuel, afin de lisser dans l'espace et le temps la charge de travail des enquêteurs. Une fois l'échantillon principal sélectionné, des opérations supplémentaires ont dû être menées pour contrôler le nombre de logements non principaux dans l'échantillon ou pour peaufiner la structure géographique des grappes. Elles sont présentées dans une septième partie. Enfin, la dernière partie présentera les procédures de mise à jour de l'échantillon, notamment la question des logements nouveaux.

1. L'échantillon actuel de l'EEC

Comme pour toutes les enquêtes auprès des ménages l'échantillon de l'EEC a été construit pour répondre aux objectifs de l'enquête en « maximisant » la précision tout en respectant des contraintes de collecte.

Les objectifs de l'EEC sont le suivi trimestriel de la situation, et de son évolution, sur le marché du travail : chômage et activité. En termes de sondage, ces objectifs se traduisent par l'obtention d'estimations transversales et longitudinales. L'outil adapté est l'échantillonnage rotatif. Il consiste à tirer régulièrement des échantillons panélisés pour une durée commune à tous les panels. A chaque date d'enquête, un panel entre dans le système tandis qu'un panel sort. Le panel entrant est représentatif de la population à la date d'entrée. Les avantages principaux de l'échantillonnage rotatif sont en fait ceux du panel : gain en précision pour estimer les évolutions et réduction de l'erreur d'observation, sans en subir les principaux inconvénients, d'une part l'attrition est limitée puisque la charge d'enquête d'un individu est limitée à la durée commune, d'autre part l'existence d'un échantillon entrant permet de pratiquer une exploitation transversale sans devoir recourir à un échantillonnage complémentaire pour mettre à jour les autres échantillons. Dans le cas de l'EEC c'est surtout l'attrition[3] qui importe puisqu'une procédure de mise à jour de tous les échantillons par la construction neuve est prévue. L'arbitrage entre précision longitudinale et limitation de l'attrition a conduit à fixer à 6 trimestres la durée de vie du panel entrant pour l'EEC.

Les contraintes de collectes sont doubles : respect de la notion de « semaine de référence » et lissage temporel de la charge de travail des enquêteurs. La collecte de l'information dans l'EEC est répartie uniformément sur un trimestre. Quand les membres d'un ménage sont interrogés, les questions relatives à leur situation sur le marché du travail font référence à une semaine qui a été fixée au préalable et est appelée «semaine de référence ». Les logements attribués à un enquêteur ont tous la même semaine de référence. Par ailleurs, pour limiter les effets mémoire des enquêtés, la collecte de l'information doit avoir lieu au maximum 2 semaines et 2 jours après la fin de la dite semaine de référence. Ce délai restreint impose que l'on est plus exigeant sur les questions de proximité géographique dans l'EEC que pour les autres enquêtes de l'Insee auprès des ménages. La période de collecte peut s'étaler, en standard, sur plusieurs mois. A la limite, pour l'EEC, si tous les logements sont contigus, l'enquêteur pourra passer l'intégralité de son temps de travail dans la prise de rendez-vous et la passation du questionnaire sans avoir à se déplacer. C'est pourquoi l'échantillon de l'EEC a mis en avant la notion[4] d' « aire » qui est le regroupement d'une vingtaine de logements dans un espace géographiquement cohérent. La taille des « aires » a été calibrée de manière à assurer suffisamment de travail à un enquêteur sans le surcharger pour autant. Dans le cadre de l'EEC les échantillons panels sont donc des échantillons aréolaires.

Au bout de 6 trimestres un échantillon sortant doit être remplacé par un échantillon entrant. Si on ne s'impose pas de contraintes, il n'y a pas de raison que les aires entrantes soient proches géographiquement des aires sortantes. Il faut alors renouveler partiellement le réseau d'enquêteurs à chaque date, avec la perte d'expérience qui en découle. Pour assurer la proximité des échantillons entrant et sortant, on introduit la notion de « secteur ». Un secteur est un regroupement d'aires

[3] Même si ce phénomène persiste dans un panel rotatif.
[4] Cette notion est héritée des Enquêtes Emploi Annuelles.

proches. Quand une aire est sortante, on la remplace par une aire du même secteur. La taille du secteur, en nombre d'aires, dépend de la durée de vie souhaitée de l'échantillon. Pour l'EEC, elle a été fixée à 9 ans, soit 36 trimestres. Chaque aire étant enquêtée 6 trimestres, il faut que les secteurs aient 6 aires. Enfin, on notera que, indépendamment de la question des logements neufs, pour assurer cette cohérence d'ensemble l'échantillon doit être sélectionné en une seule fois.

Au final, pour construire l'échantillon actuel, dans un premier temps des secteurs ont été sélectionnés. Ces secteurs ont été répartis aléatoirement en 6 sous-échantillons, numérotés de 1 à 6. Chaque sous-échantillon est représentatif de la situation des logements au RP99. Afin d'assurer la rotation, les secteurs sont entrés progressivement dans l'échantillon à partir du troisième trimestre 2001. Dans chaque secteur, les « aires » se sont vues attribuées aléatoirement un rang d'entrée dans l'échantillon de 1 à 6. Les aires de rang 1 sont mobilisées en premier pour une période de 6 trimestres après quoi elles sont remplacées par les aires de rang 2. Quand une aire est entrante l'opération de ratissage permet de rendre le sous-échantillon représentatif de la date d'entrée[5]. Chaque secteur, et donc chaque aire, s'est vu attribué une semaine de référence de 1 à 13 de sorte que ces semaines soient réparties uniformément, en régime de croisière, par région. Le régime de croisière va du 4ème trimestre 2002 au 2ème trimestre 2010. Au-delà de cette date, l'échantillon décroît progressivement jusqu'au T3-2011 inclus. Le tableau 1 résume l'ensemble du processus.

La compréhension de ce processus est importante. Étant reconduit pour le nouvel échantillon, il en conditionne le rythme d'entrée. Les marges de manœuvre et les éventuels gains de précision porteront sur la construction des secteurs et des grappes, leur mode de sélection, les traitements avals et la qualité des informations fournies aux enquêteurs pour le repérage des logements.

2. Les fichiers de la Taxe d'Habitation comme base de sondage.

La taxe d'habitation[6] est un impôt dû par toute personne disposant d'un logement au 1er janvier, que le statut d'occupation soit propriétaire, locataire ou à titre gratuit, que la catégorie du logement soit principale ou secondaire. Les logements vacants, bien que non assujettis à la TH, sont également présents dans le fichier. Les locaux soumis à la taxe d'habitation sont les locaux meublés affectés à l'habitation ainsi que leurs dépendances immédiates (chambres de service, garage). Certains locaux sont exonérés et non gérés dans la TH : résidences universitaires, casernes. D'autres sont exonérés mais gérés dans la TH : bâtiments servant aux exploitations rurales. Au total, les fichiers de la TH sont des fichiers de locaux, dont les résidences universitaires et les casernes sont les seuls locaux d'habitation non couverts. Cette limitation n'est pas gênante en soit puisque le champ de l'EEC est celui des ménages ordinaires dont sont exclues ces deux catégories.

Les fichiers de la TH couvrent un champ plus large que celui des seuls logements ordinaires. On ne peut pas passer directement de la notion de local à celle de logement. On utilise pour cela un filtre qui a été étalonné dans le cadre des travaux du recensement rénové. Le fichier filtré comporte légèrement plus d'observations que ce que l'on peut observer au recensement. La différence est en partie due à la gestion des communautés qui sont pour partie intégrées dans la source. Cette sur-exhaustivité se traduit plus par un problème de collecte que de statistique. En effet, elle se répercutera sur le terrain par un nombre plus important de déchets sans pour autant affecter le biais, ni la taille globale de l'échantillon. Il est en effet calibré sur le nombre de résidences principales qui est quant à lui globalement comparable à celui observé dans le recensement.

La catégorisation des logements à la TH (résidence principale, secondaire, logement vacant) répond à une règle fiscale qui peut être différente de celle retenue dans les définitions Insee. Dans la mesure où les fichiers ne sont pas utilisés pour établir directement des statistiques, mais comme base de sondage, cette différence n'est pas en soit problématique. D'une part elle est limitée : 10 % des ménages par exemple disposent d'une résidence secondaire. Les éventuelles divergences de déclaration ne porteront que sur cette population. Par ailleurs, ce qui importe c'est surtout la catégorisation par l'enquêteur sur le terrain. En ce sens, la situation n'est pas différente de celle des

[5] L'opération de ratissage a lieu en première et sixième interrogations.

[6] La description des fichiers de la taxe d'habitation est très largement inspirée des documents fournis par Françoise Dupont et utilisés dans le cadre des formations du Recensement Rénové de la Population.

autres enquêtes ménages de l'Insee. Elles sont sélectionnées dans l'échantillon maître 1999, selon la catégorie du logement à cette date qui peut être différente de celle observée sur le terrain.

Les fichiers de la TH vérifient ainsi la première des propriétés demandées à une base de sondage, à savoir *l'exhaustivité*. Cette exhaustivité s'accompagne de la *fraîcheur* puisque l'Insee reçoit l'information relative à la situation au 1er janvier de l'année *N* à la fin de la même année. La fraîcheur est un atout important. L'échantillon étant sélectionné en une seule fois dans la TH 2006 pour toute sa durée de vie, la fraîcheur permettra de le mettre à jour régulièrement. Par mise à jour, on entend le double processus d'enrichissement par les logements nouveaux et de rafraîchissement des informations nécessaires au repérage (Nom, adresse…).

Tableau 1 : Schéma de rotation de l'échantillon actuel de l'EEC.

	Sous-échantillon de secteurs					
	1	2	3	4	5	6
Date d'observation	Date d'entrée					
	T3 2001	T4 2001	T1 2002	T2 2002	T3 2002	T4 2002
T3-2001	1					
T4-2001	1	1				
T1-2002	1	1	1			
T2-2002	1	1	1	1		
T3-2002	1	1	1	1	1	
T4-2002	1	1	1	1	1	1
T1-2003	2	1	1	1	1	1
T2-2003	2	2	1	1	1	1
T3-2003	2	2	2	1	1	1
T4-2003	2	2	2	2	1	1
T1-2004	2	2	2	2	2	1
T2-2004	2	2	2	2	2	2
T3-2004	3	2	2	2	2	2
T4-2004	3	3	2	2	2	2
T1-2005	3	3	3	2	2	2
T2-2005	3	3	3	3	2	2
T3-2005	3	3	3	3	3	2
T4-2005	3	3	3	3	3	3
T1-2006	4	3	3	3	3	3
T2-2006	4	4	3	3	3	3
T3-2006	4	4	4	3	3	3
T4-2006	4	4	4	4	3	3
T1-2007	4	4	4	4	4	3
T2-2007	4	4	4	4	4	4
T3-2007	5	4	4	4	4	4
T4-2007	5	5	4	4	4	4
T1-2008	5	5	5	4	4	4
T2-2008	5	5	5	5	4	4
T3-2008	5	5	5	5	5	4
T4-2008	5	5	5	5	5	5
T1-2009	6	5	5	5	5	5
T2-2009	6	6	5	5	5	5
T3-2009	6	6	6	5	5	5
T4-2009	6	6	6	6	5	5
T1-2010	6	6	6	6	6	5
T2-2010	6	6	6	6	6	6
T3-2010		6	6	6	6	6
T4-2010			6	6	6	6
T1-2010				6	6	6
T2-2011					6	6
T3-2011						6

Note de lecture : Au T1 2010 sont interrogées les aires de rang 6 des sous-échantillons 1 à 5 et de rang 5 du sous-échantillon 6. Dans le sous-échantillon 5, les aires sont entrantes. Dans le sous-échantillon 6, les aires sont interrogées pour la dernière fois.

Cette mise à jour est rendue possible par la disponibilité d'un *identifiant pérenne* des logements. A une date donnée, les autres informations relatives au repérage sont le nom et le prénom du dernier occupant connu et l'adresse postale. Une originalité est l'accès aux *références cadastrales*. Ces dernières constituent un identifiant de la propriété foncière. Même si les informations d'adressage ne sont pas suffisantes, les bâtiments échantillonnés peuvent être retrouvés en lisant l*es plans cadastraux*. Ces derniers sont accessibles gratuitement en ligne depuis janvier 2008, sur le site

http://www.cadastre.gouv.fr. Les références cadastrales joueront par ailleurs un rôle déterminant dans la construction des secteurs et des grappes.

D'autres informations sont disponibles dans les fichiers, on les qualifie d'information auxiliaire. Elles sont importantes non pas dans le processus de construction de l'échantillon mais pour les traitements aval : correction de la non-réponse. On dispose de caractéristiques fraîches du logement : logement social, collectif, nombre de pièces, superficie, de caractéristiques des ménages : âge du chef du ménage, composition, revenus. Les définitions de ces variables sont fiscales. Elles ne peuvent pas être utilisées directement dans un but de description. On rappelle à ce titre que dans un modèle de correction de la non-réponse ce n'est pas tant la qualité de la variable qui importe que son homogénéité sur l'ensemble des logements.

Exhaustivité, fraîcheur, identifiabilité, présence d'information auxiliaire sont les qualités demandées à une source pour constituer une bonne base de sondage (P. Ardilly [4]). Il faut y adjoindre l'absence de double compte. Cette dernière propriété n'a pas été directement vérifiée, elle découle en partie des autres propriétés. Par ailleurs, compte tenu de la finalité des fichiers de la TH : payer un impôt, on a supposé que peu de contribuables étaient prêts à payer deux fois (ou plus) le même impôt. Dans le cadre de l'échantillonnage de l'EEC, ces propriétés ne sont pas encore suffisantes pour qualifier la TH. Il faut en effet pouvoir construire les secteurs et les grappes et donc disposer d'information géographique fine et révélatrice des distances entre logements. Cette information est contenue dans les références cadastrales.

3. La construction des secteurs et des grappes

Les secteurs et les grappes (ou aires) sont les éléments structurants de l'échantillonnage de l'EEC. Dans l'échantillon actuel, ils ont été construits manuellement selon une logique descendante. Dans le futur échantillon, ils l'ont été automatiquement selon une logique ascendante.

3.1. Les secteurs et les aires

Les contraintes budgétaires ont conduit à fixer le nombre d'aires à interroger par trimestre dans l'échantillon actuel à 2550. Comme chaque trimestre une seule aire est mobilisée par secteur, ce nombre est également le nombre de secteurs à échantillonner. Pour les obtenir, sans avoir à découper manuellement l'ensemble du territoire, la logique descendante a été retenue.

Parmi les zonages administratifs ou d'étude disponibles, des unités primaires (UP) ont été construites de manière à être les plus petites possibles tout en ayant au moins 120 logements (6 aires de 20 logements). Les unités primaires sont constituées de communes, d'IRIS, de districts, ou de regroupements de ces entités. Parmi l'ensemble des UP, 2 550 ont été sélectionnées proportionnellement à leur taille, en nombre de logements. Chaque UP sélectionnée à été découpée, sur la base des documents cartographiques associés au RP 99, en secteurs regroupant de 120 à 240 logements contigus. Dans chaque UP, 1 secteur a été sélectionné. Il a fait l'objet ensuite d'un découpage en aires d'une vingtaine de logements contigus. Le nombre d'aires constituées par secteur pouvant varier de 6 à 13, 6 aires ont été sélectionnées par sondage aléatoire simple sans remise. On voit tout l'intérêt de cette logique puisque seules les UP échantillonnées sont découpées manuellement en secteurs, et seuls les secteurs échantillonnés sont découpés en aires. Cette opération a néanmoins nécessité une année et 140 000 heures de travail dans les directions régionales de l'Insee. Pour le nouvel échantillon, compte tenu du calendrier très contraint, il n'était pas envisageable de reconduire cette procédure. Le seul recours a été l'automatisation.

3.2. Les grappes et les secteurs

3.2.1. Les grappes

La constitution des nouvelles grappes a été l'occasion de s'interroger sur leur nature. En effet, dans l'échantillon actuel, les regroupements ont été effectués sur la base d'un nombre total de logements. Dans certaines zones, notamment touristiques, il se peut que les aires actuelles soient entièrement constituées de résidences secondaires. L'enquêteur affecté à cette aire n'est pas mobilisé sur une période de 6 trimestres. Outre les conséquences négatives sur la gestion du réseau, un critère fondé

sur le nombre total de logements conduit à une variabilité de la taille effective des aires en termes de nombre de logements enquêtables. Cette variabilité est un facteur de détérioration de la précision dans le cadre de sondages en grappe où la constance de taille est recherchée. C'est pourquoi il a été décidé que les nouvelles grappes contiendraient une 20^{aine} de résidences principales. Il en résulte un léger accroissement de la taille des grappes qui n'a pas été jugé incompatible avec la charge de travail d'un enquêteur. Au total, les principes de constitution des grappes et des secteurs ont été les suivants :

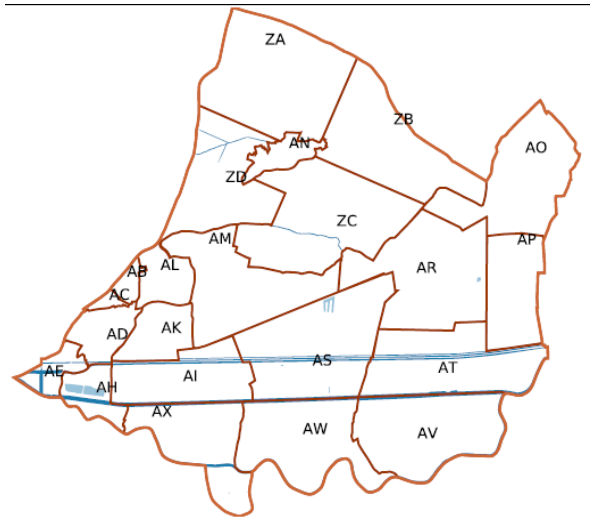
- Une grappe est un regroupement de logements proches.
- Tous les logements, y compris les logements non principaux, sont affectés à une grappe.
- Dans la limite du respect des frontières communales, les grappes ont un nombre de résidences principales le plus proche possible de 20.
- Dans les immeubles collectifs, les logements d'un même étage sont inclus dans la même grappe.
- Afin de réduire les degrés de tirage, on s'efforce de constituer directement un maximum de secteurs de taille 6 :
    - Dans les communes où plus de 95 grappes sont constituées, les secteurs sont construits à l'intérieur des communes. On est ainsi assuré d'avoir au moins 2 fois plus de secteurs de taille 6 que de secteurs de taille 7 : 95 = 10*6 + 5*7.
    - Pour les autres communes, dans la limite du respect des frontières régionales, les secteurs peuvent avoir des grappes appartenant à des communes différentes mais limitrophes.
- Après échantillonnage des secteurs, si dans une grappe, on observe plus de 10 logements non principaux, on procède à un échantillonnage des logements non principaux.

Ces principes ont pu être respectés grâce à la présence et au type de codification des références cadastrales dans les fichiers de la TH. Les références cadastrales sont constituées de deux éléments : la **section** et la **parcelle**.

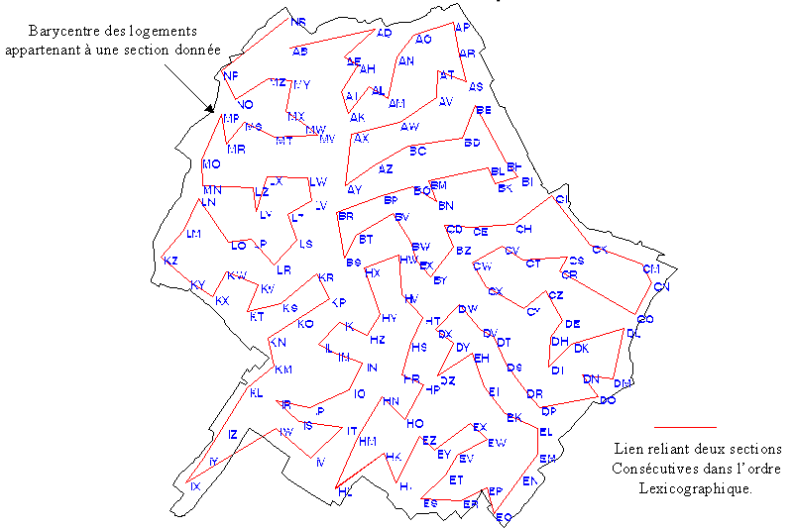
- « La **section** est une fraction du territoire communal déterminée de façon à faciliter l'établissement et la consultation des documents cadastraux. Son périmètre est constitué, dans la mesure du possible, par des limites naturelles présentant un caractère suffisant de fixité (voie de communications, cours d'eau, …). La commune est décomposée en un nombre minimal de sections. » Toutes les communes de France sont découpées en sections cadastrales (*graphique 1*). Elles sont identifiées par un code alphanumérique. L'ordre lexicographique suivi par ce code respecte la contiguïté des sections : la section *j* est limitrophe de la section *j+1* (*graphique 2*).

- « La **parcelle** est une portion de terrain d'un seul tenant, située dans un même lieu-dit, appartenant à un même propriétaire ou à une même indivision et constituant une unité foncière indépendante selon l'agencement donné à la propriété. Le numérotage parcellaire est effectué, à l'origine, sans interruption et par sections (graphique 3). *Toute parcelle nouvelle ou modifiée reçoit un nouveau numéro pris à la suite du dernier attribué dans la section ; le numéro de la parcelle primitive n'est jamais réutilisé mais il permet de localiser la nouvelle parcelle créée qui fait référence à la parcelle primitive. » (graphique 4).*

Graphique 1 : Toutes les communes françaises sont découpées en sections cadastrales



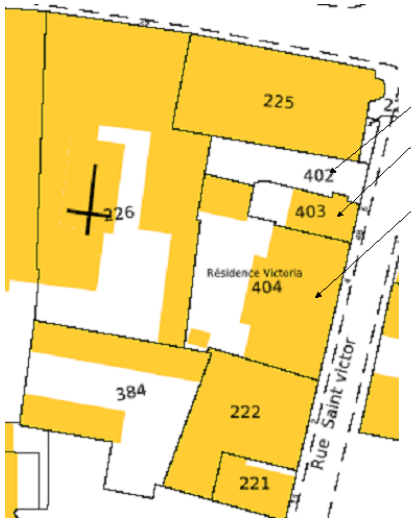
Graphique 2 : Organisation géographique et codification des sections cadastrales à l'intérieur d'une commune. L'exemple de Roubaix



Note de lecture : *Deux sections d'une même commune ayant des codes qui se suivent selon l'ordre lexicographique seront limitrophes*

Graphique 3 : à l'intérieur des sections, les parcelles sont numérotées de manière a priori contiguë.

Graphique 4 : Parfois cela ne fonctionne pas



On corrige, dans la mesure du possible, le numéro de parcelle de manière à recréer de la contiguïté à partir du nom de la voie et du numéro.

On ne peut pas tout corriger automatiquement, ce qui explique que l'on passe à de la contiguïté à la proximité.

Muni de ces éléments la constitution des grappes devient relativement aisée :

- A partir du nombre total de résidences principales dans la section, on détermine le nombre de grappes à créer : si N est ce total, le nombre de grappes est alors n=E(N/20), où E est la partie entière. La taille des grappes est alors t=N/n.
- On agrège les logements par étage. Pour chaque étage, on connaît le nombre de résidences principales et de logements non principaux[7]. Par convention, on appelle également étage les maisons individuelles. Tous les étages ont par ailleurs le même numéro de parcelle.
- On trie les étages de la section par numéro de parcelle. On lit le fichier séquentiellement en cumulant le nombre de résidences principales des étages. Dès que ce cumul est supérieur ou égal à t, une grappe est constituée. On notera qu'avec ce processus les logements non principaux sont également agrégés. On passe à la grappe suivante.
- Selon la procédure précédente, deux parcelles proches, mais dont l'une a subi une modification, peuvent se retrouver dans des grappes différentes (graphique 5). En utilisant le nom de la rue et les numéros dans la voie, on corrige ces ruptures de continuité (graphique 6). En réalité, cette opération a lieu dès le début du processus. C'est parce que l'on n'a pas pu corriger toutes les ruptures que la contiguïté ne peut plus être assurée dans le nouvel échantillon.
- Enfin, le nombre de résidences principales d'une section peut être inférieur à 20. Dans ce cas, si on trouve une suite de sections, indexées par leur code alphanumérique, et ayant toutes moins de 20 résidences principales, on les agrège en une pseudo-section dans laquelle sont construites les grappes. L'ordre des parcelles, après correction, est alors celui de la section d'origine et du numéro de parcelle corrigé. Cette étape d'agrégation a également lieu en début de processus.

3.2.2. Les secteurs

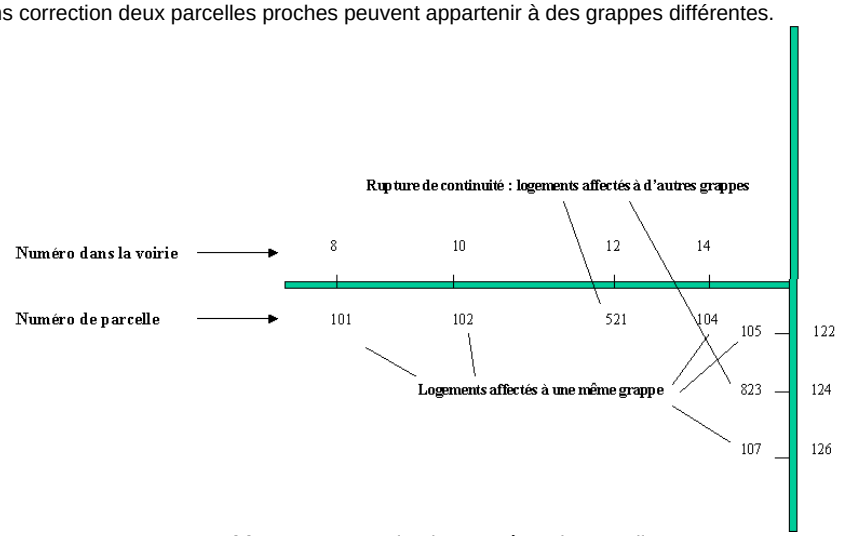
Dans les communes ayant plus de 95 grappes constituées, le processus précédent assure que la $i^{ème}$ grappe construite sera proche de la $i+1^{ème}$. Si *X* est le nombre total de grappes de la commune, alors on pourra y construire *(p-r)* secteurs de taille 6 et *r* secteurs de taille 7, où *p* est le quotient de la division euclidienne de *X* par 7, et *r* le reste de cette division. Les 6*(p-r)* premières grappes sont agrégées selon l'ordre de construction en *(p-r)* secteurs de taille 6, alors que les 7*r* restantes le sont dans des secteurs de taille 7. On notera que comme *r* est inférieur ou égal à 5, plus X est grand, plus on aura proportionnellement (modulo 6) de secteurs de taille 6. Cette remarque, peut-être triviale, est à l'origine du mode de constitution des secteurs dans les autres communes.

Pour les autres communes, on a procédé à une renumérotation de manière à ce que la commune numérotée *k* soit limitrophe de la commune *k+1*. L'objectif est de trier le fichier des grappes constituées par ce nouveau numéro puis par ordre de création des grappes à l'intérieur de la commune. On n'a plus qu'à agréger les grappes dans cet ensemble par paquet de 6 (ou de 7 pour les dernières) dans l'ordre du fichier. Ainsi un secteur sera soit à l'intérieur d'une commune, les grappes seront alors proches les unes des autres dans la commune, soit à cheval sur plusieurs communes. Dans ce cas les communes seront limitrophes. A l'intérieur d'une commune les grappes seront proches.

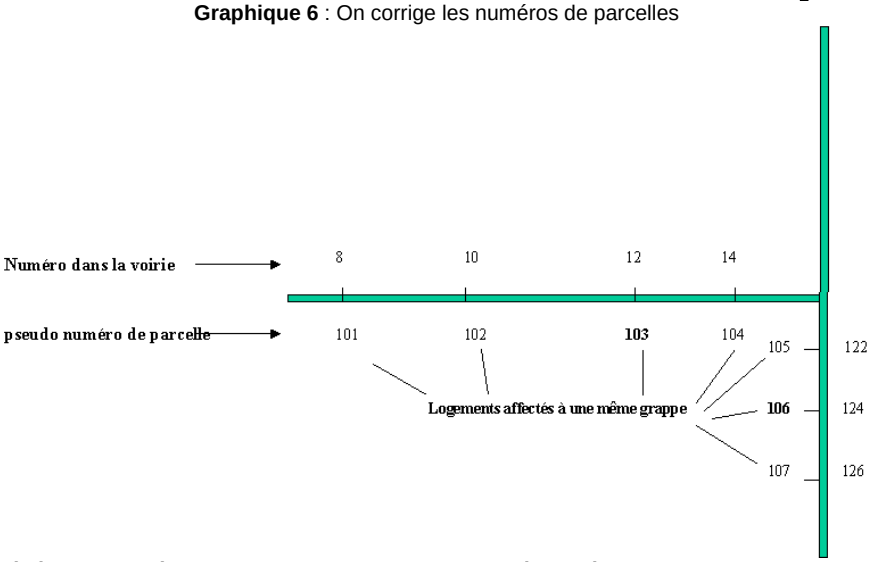
Le problème que l'on cherche à résoudre est de trouver un chemin d'un seul trait qui passe par toutes les communes une et une seule fois. On parcourt ce chemin dans sa totalité depuis le début en incrémentant à chaque fois le numéro de commune. L'échantillon final étant tiré par région, la création de ces chemins s'entend par région. Par ailleurs, compte tenu de l'aspect quelque peu irrégulier des frontières communales et régionales, trouver un tel chemin peut s'avérer chronophage, à supposer qu'il existe. Notre ambition étant plus modeste, on s'est borné à créer des « balades » à savoir des chemins d'un seul trait les plus longs possibles ne passant qu'une seule fois dans un ensemble de communes limitrophes. Les secteurs ont été crées par « balade ». Les « balades » ont été créées algorithmiquement à partir de la matrice de contiguïté des communes françaises. On ne présente pas les subtilités de cet algorithme mais un résultat obtenu sur la région Basse-Normandie (*graphique 7*).

[7] Ces nombres peuvent être éventuellement nuls.

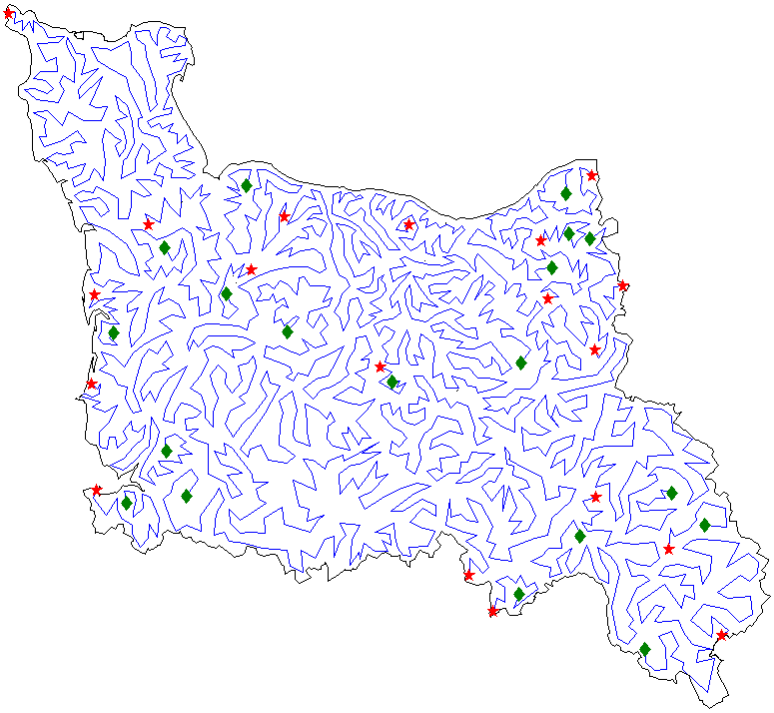
Graphique 5 : Sans correction deux parcelles proches peuvent appartenir à des grappes différentes.



Graphique 6 : On corrige les numéros de parcelles



Graphique 7 : 'balades normandes' ou comment passer par un *maximum* de communes sans repasser deux fois par la même afin de construire les secteurs de l'EEC.



Note de lecture : les étoiles rouges constituent le début d'une balade, les losanges verts la fin (ou inversement).

3.2.3. Taille des grappes et des secteurs créés.

Au 1er janvier 2006, on dénombre dans les fichiers de la TH 32,2 millions de logements dont 25,4 millions sont des résidences principales. Ces logements ont été regroupés dans 1,2 million de grappes, soit en moyenne 21,1 résidences principales par grappe (tableau 2). Les grappes ont été regroupées dans 199 500 secteurs dont 96 % ont exactement 6 grappes et 4 % en ont 7. Pour l'échantillon actuel, 89 % des secteurs créés avaient au moins 7 grappes (tableau 3).

Tableau 2 : Bilan de la création automatique des grappes

Taille des grappes	Nombre total de					
	grappes		résidences principales		logements	
	N	%	N	%	N	%
15	45 430	3,8	681 450	2,7	955 429	3,0
16	15 334	1,3	245 344	1,0	353 331	1,1
17	13 873	1,2	235 841	0,9	340 154	1,1
18	14 884	1,2	267 912	1,1	377 126	1,2
19	23 812	2,0	452 428	1,8	598 909	1,9
20	324 893	27,0	6 497 860	25,5	7 998 531	24,8
21	366 436	30,4	7 695 156	30,2	9 591 002	29,8
22	173 937	14,4	3 826 614	15,0	4 805 335	14,9
23	91 083	7,6	2 094 909	8,2	2 651 749	8,2
24	53 161	4,4	1 275 864	5,0	1 631 778	5,1
25	28 728	2,4	718 200	2,8	945 365	2,9
26	19 199	1,6	499 174	2,0	670 664	2,1
27	13 191	1,1	356 157	1,4	480 435	1,5
28	10 848	0,9	303 744	1,2	413 620	1,3
29	9 846	0,8	285 534	1,1	391 227	1,2
30	183	0,0	5 490	0,0	6 498	0,0
Ensemble	1 204 838	100	25 441 677	100	32 211 153	100

Tableau 3 : taille des secteurs.

Nombre d'aires ou de grappes Par secteur	6	7	8	9	10	11	12	13
% des secteurs actuels	11,6	21,6	18,9	18,9	13,6	11	4,2	0,2
% des secteurs du futur échantillon	96	4						

3.2.4. Dimension géographique des grappes créées.

Compte tenu des enjeux liés aux propriétés géographiques des grappes créées, une attention particulière a été portée à la distance moyenne parcourue par un enquêteur au sein d'une grappe. Pour les résidences principales des communes de plus de 10 000 habitants, on dispose des coordonnées X,Y. On peut mesurer directement cette distance théorique, elle est donnée par le tableau 4. Avec une médiane à 34 mètres, ces résultats ont été de nature à conforter le processus de création des grappes. Ils ne portent cependant que sur les grandes communes. Pour les petites communes, des tests ont été menés dans les Directions Régionales du Nord-Pas-de-Calais, du Centre, de Bretagne et en PACA. Ces tests ont également conclu, qu'une fois admise l'absence de contiguïté, les distances à parcourir restaient compatibles avec une durée de collecte de 2 semaines et 2 jours.

4. La sélection des secteurs et des grappes.

4.1. La taille globale de l'échantillon

L'augmentation de 50 % de la taille de l'échantillon peut s'entendre de plusieurs façons selon qu'elle s'applique aux nombres de logements repérés, de résidences principales repérées, de résidences principales répondantes… A été privilégié ici le nombre de résidences principales repérées. En moyenne par trimestre depuis 2003, ce nombre est de l'ordre de 44 900. Le calibrage de la taille du

nouvel échantillon est de 67 350 résidences principales par trimestre. Par ailleurs, la taille moyenne des grappes étant de 21 résidences principales, chaque trimestre 67350/21 ≈ 3211 grappes seront enquêtées. Enfin, un trimestre donné, une seule grappe par secteur est mobilisée. Le nombre 3211 correspond également au nombre de secteurs qu'il faut sélectionner dans l'échantillon. La question principale qui se pose est alors la répartition de ces secteurs par région.

Tableau 4 : Répartition de la distance moyenne parcourue par un enquêteur au sein d'une grappe (en mètre) Communes de + de 10 000 habitants.

1er décile	0
2ème décile	5.9
3ème décile	12.4
4ème décile	21.3
Médiane	34.0
6ème décile	49.3
7ème décile	67.7
8ème décile	93.2
9ème décile	140.3

4.2. La question de la répartition par région

La question de l'allocation par région est certainement l'une des plus importantes car elle conditionne, a priori, la précision nationale, les précisions par région et la charge de travail par région. Pour y répondre, on s'est intéressé dans un premier temps à la précision. On a alors mesuré les conséquences sur la charge de travail par région, ce qui a pu amener à modifier l'allocation initiale.

Les questions de précision sont directement liées aux objectifs de l'enquête. Les débats méthodologiques sur les statistiques du chômage ont essentiellement porté sur la précision du taux de chômage (ou de son évolution) au niveau national et par trimestre. Il s'agit d'un objectif essentiel. Toutefois, l'enquête emploi s'inscrit également dans le cadre d'un règlement européen qui fixe des objectifs en termes de précision d'estimation d'évolution trimestrielle au niveau national, mais aussi d'estimation annuelle moyenne au niveau régional.

Pour l'échantillon actuel, les objectifs nationaux et régionaux sont antagonistes (*M. Christine* [3]). Compte tenu de l'aspect rotatif trimestriel de l'échantillon, la précision annuelle régionale sera d'autant meilleure que la corrélation du phénomène étudié est faible entre deux trimestres alors que c'est le contraire pour la précision de variation trimestrielle nationale. L'échantillon actuel a d'abord privilégié les contraintes nationales, pour intégrer ensuite les contraintes régionales. On a suivi également cette démarche. Évolution trimestrielle nationale et estimation trimestrielle sont reliées par un coefficient de proportionnalité. Il est ainsi équivalent de s'intéresser à une estimation trimestrielle. On a retenu, comme pour l'échantillon actuel, une optimisation de la précision d'une estimation nationale du nombre de chômeurs comme critère principal en mesurant à chaque fois les conséquences en termes de précision régionale et de charge de travail.

L'allocation utilisée dans l'échantillon actuel est issue d'une allocation de Neyman qui, sous certaines hypothèses, est théoriquement optimale. Cette allocation se justifie dès lors que la variable de stratification, ici la région, explique fortement le phénomène étudié et que, par ailleurs, les quantités utilisées dans le calcul de l'allocation, l'écart-type du nombre de chômeurs par grappe, sont estimées avec une grande précision. On rappelle ici quelques résultats liés à l'utilisation de la stratification en sondage.

- Si la région n'est pas un bon critère de stratification, le gain en précision sera faible pour l'estimation du nombre de chômeurs. L'allocation pourra donner, pour d'autres variables, des estimations moins précises que si l'on avait procédé à un sondage aléatoire simple (SAS).
- Si les grandeurs intervenant dans le calcul de l'allocation de Neyman sont estimées avec trop peu de précision, cette allocation pourra conduire à des estimations qui, même pour le nombre de chômeurs, seront moins précises qu'un SAS.
- L'EEC étant un panel rotatif, l'allocation est définie sur la base d'informations disponibles au moment de la sélection de l'échantillon. Au moment de l'observation, les modifications sur le

marché du travail peuvent conduire à une allocation qui n'est plus optimale et donc encore à une précision moindre qu'un SAS.

- Enfin avec une allocation de l'échantillon proportionnelle à la taille, on est assuré d'avoir des estimations au moins aussi précises que celles issues d'un sondage aléatoire simple.
- Pour comprendre les subtilités liées à l'usage de la stratification, on invite le lecteur à consulter P. Ardilly [4], par exemple.

Pour vérifier ces points, on a calculé[8] les différentes allocations de Neyman que l'on obtiendrait en utilisant des données annuelles estimées en moyenne trimestrielle sur les années 2003 à 2007, ainsi que l'allocation moyenne sur cette période. Pour chacune de ces allocations, on a estimé, selon la date d'observation, la précision que l'on aurait obtenue pour une estimation nationale du nombre de chômeurs. On a par ailleurs estimé la précision qui découlerait de l'utilisation d'un SAS ou d'une allocation proportionnelle. Cette dernière a servi de référence dans le tableau 5 qui répertorie les résultats. De ces calculs, il ressort les constats suivants :

- Les allocations par région peuvent varier beaucoup selon l'année de référence choisie pour le calcul. Dans le Nord-Pas-de-Calais, on peut obtenir un nombre de secteurs variant de 230 à 280, de 136 à 187 en Aquitaine, de 108 à 170 en Midi-Pyrénées, de 250 à 310 en PACA[9].
- Une année donnée, le gain en précision obtenu en utilisant une allocation de Neyman plutôt qu'une allocation proportionnelle est de l'ordre de 1% sur l'écart-type du nombre de chômeurs estimés (tableau 5). Ce résultat provient du faible pouvoir explicatif de la région sur le taux de chômage et de la relativité de la notion d'optimalité. Le gain est à comparer avec celui d'un calage qui peut être de l'ordre de 25 %.
- Dans la pratique, il faudrait calculer une seule allocation une année donnée et l'utiliser pour toute la durée de vie de l'échantillon, le gain relatif pourrait alors être moindre. Par exemple, une allocation calculée sur la base des données de 2003 aurait effectivement conduit à gain « maximal » pour l'estimation du nombre de chômeurs en 2003, ce gain aurait été moindre pour des estimations portant sur les autres années (tableau 5).
- Au total, il ressort que l'usage de l'allocation de Neyman ne se justifie pas pleinement pour l'estimation du nombre de chômeurs. Compte tenu du risque potentiel, même faible, qu'elle fait porter sur l'estimation d'autres variables d'intérêt de l'enquête, le principe d'une répartition proportionnelle de l'échantillon par région a été retenu.
- Pour tenir compte des attentes d'Eurostat en matière de précision régionale, si l'allocation proportionnelle conduit à un nombre de secteurs dans une région donnée plus faible que celui actuellement observé, alors le nombre actuel a été reporté. Le reliquat d'échantillon a été réparti dans les autres régions proportionnellement à la taille. On a vérifié que cet écart à la proportionnalité n'affectait pas la précision nationale.

Tableau 5 : Valeur relative de la précision des différentes allocations sur l'estimation nationale du nombre de chômeurs par rapport à l'allocation proportionnelle à la taille selon l'année d'observation.

Type d'allocation		Année d'observation					
		2003	2004	2005	2006	2007	2003-2007
Année de référence pour les allocations de Neyman	2003	98,7	99,6	99,3	98,8	99,2	99,1
	2004		99,2	99,0	99,0	99,2	99,1
	2005			98,3	98,9	99,8	99,4
	2006				98,3	99,4	99,1
	2007					98,6	99,2
	2003-2007	98,9	99,4	98,7	98,5	98,9	98,9
Sondage Aléatoire Simple		101,0	101,1	101,4	101,0	101,3	101,2
Allocation proportionnelle à la taille		100,0	100,0	100,0	100,0	100,0	100,0

Note de lecture : l'écart-type du nombre de chômeurs est de l'ordre de 44 000 France entière. Si pour 2007, on avait utilisé une allocation de Neyman calculée sur la base des résultats de 2005, le gain en précision par rapport à une allocation du type proportionnelle aurait été de l'ordre de 0,2 %.

[8] Ces travaux ont été effectués par Dominique Place de la Division Echantillonnage et Traitement Statistique des Données (ETSD) de l'UMS.
[9] On n'a pas reporté ici l'ensemble des résutats par région.

4.3. La répartition finale de l'échantillon

L'échantillon comportant actuellement 2 554 secteurs, l'augmentation en nombre de secteurs est de 25 % en moyenne nationale. Elle est moindre que celle du nombre de résidences principales du fait que les secteurs et les grappes avec la solution TH contiennent en moyenne plus de résidences principales que les secteurs et les aires de l'échantillon actuel (respectivement 21 contre 17). Ces évolutions affectent de manière différenciée les régions (*cf.* tableau 6).

A l'augmentation générale qui concerne toutes les régions se superposent des effets spécifiques

1. pour les régions actuellement sous-représentées s'ajoutent les corrections apportées aux distorsions par rapport à la proportionnalité. Ces régions connaissent au final une augmentation supérieure à la moyenne nationale.
2. pour les régions actuellement surreprésentées se déduisent les corrections apportées aux distorsions par rapport à la proportionnalité. Ces régions connaissent au final une augmentation inférieure à la moyenne nationale.

Ces résultats précédents caractérisent l'échantillon en régime de croisière, c'est à dire à partir d'Octobre 2011, date à laquelle l'échantillon actuel sera entièrement épuisé (rappel : voir tableau 1). En régime transitoire, la montée en charge tient compte de la coexistence des deux types d'échantillon.

Tableau 6 : structure théorique[10] du futur échantillon de l'EEC issu de la TH par région administrative.

Région administrative	Nombre de résidences principales au 1er janvier 2006 (TH)	Échantillon actuel			Échantillon futur			Évolution en nombre de	
		Nombre de secteurs	Nombre de résidences principales	Coefficient de sur représentation	Nombre de secteurs	Nombre de résidences principales	Coefficient de surreprésentation	secteurs	résidences principales
Île-de-France	4 670 219	432	7926,3	0,96	572	12012	0,97	132,4	151,5
Champagne-Ardenne	561 065	91	1631,7	1,64	99	2079	1,39	108,8	127,4
Picardie	736 658	81	1457,7	1,12	90	1890	0,96	111,1	129,7
Haute-Normandie	732 268	83	1519,2	1,17	90	1890	0,97	108,4	124,4
Centre	1 059 609	92	1649,4	0,88	129	2709	0,96	140,2	164,2
Basse-Normandie	602 842	74	1260,9	1,18	85	1785	1,11	114,9	141,6
Bourgogne	700 851	91	1576,5	1,27	91	1911	1,02	100,0	121,2
Nord-Pas de Calais	1 563 045	185	3468,6	1,25	191	4011	0,96	103,2	115,6
Lorraine	955 024	82	1559,1	0,92	116	2436	0,96	141,5	156,2
Alsace	728 701	65	1260,9	0,98	93	1953	1,01	143,1	154,9
Franche-Comté	479 059	63	1166,7	1,38	79	1659	1,30	125,4	142,2
Pays de la Loire	1 426 732	122	2176,5	0,86	174	3654	0,96	142,6	167,9
Bretagne	1 310 301	91	1563	0,67	161	3381	0,97	176,9	216,3
Poitou-Charentes	744 240	84	1398,3	1,06	91	1911	0,97	108,3	136,7
Aquitaine	1 313 776	108	1876,5	0,81	161	3381	0,97	149,1	180,2
Midi-Pyrénées	1 172 516	93	1721,4	0,83	143	3003	0,96	153,8	174,5
Limousin	324 361	71	1190,7	2,08	76	1596	1,85	107,0	134,0
Rhône-Alpes	2 468 344	213	3645	0,83	302	6342	0,97	141,8	174,0
Auvergne	588 090	69	1143,6	1,10	72	1512	0,97	104,3	132,2
Languedoc-Roussillon	1 064 572	123	1916,1	1,02	130	2730	0,96	105,7	142,5
PACA	2 036 587	224	3501,3	0,97	249	5229	0,97	111,2	149,3
Corse	107 727	17	229,5	1,20	17	357	1,25	100,0	155,6
France	**25 346 587**	**2 554**	**44 838,9**	**1**	**3 211**	**67 431**	**1**	**125,7**	**150,4**

[10] Dans la pratique, les contraintes de tirage font que l'on pourra s'écarter très légèrement de cette structure.

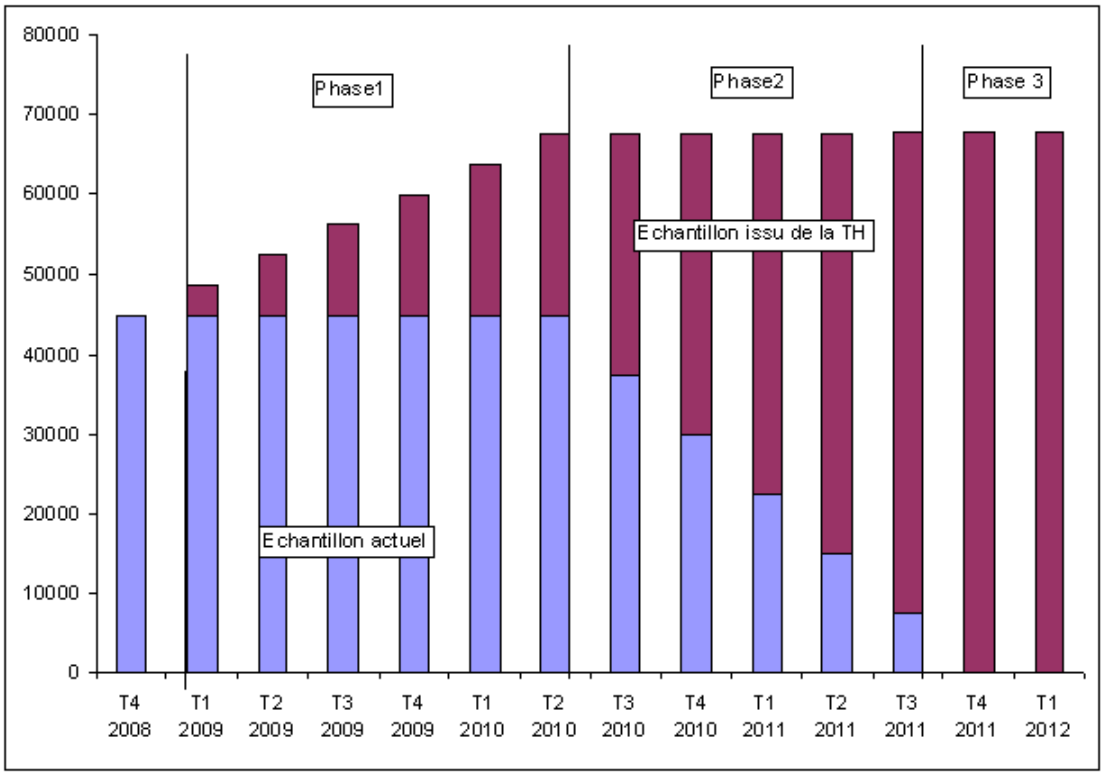
4.4. La question du calendrier d'introduction

Le calendrier d'introduction doit tenir compte des deux contraintes qui sont d'une part le maintien du calendrier de l'échantillon actuel tel qu'il a été initialement prévu et d'autre part l'objectif d'augmentation de 50 % de la taille de l'échantillon en nombre de résidences principales repérées qui doit être atteint progressivement mi-2010. Le graphique 8 montre comment tenir compte de ces contraintes, il s'agit essentiellement de distinguer 3 phases :

1. La première phase du T1 2009 au T2 2010 inclus se caractérise par le maintien de l'échantillon actuel dans son rythme de croisière. L'échantillon TH est introduit progressivement pour parvenir à 67 000 résidences principales au T2 2010.
2. A partir du T3 2010, l'échantillon actuel commence à décliner progressivement par 6ème. L'échantillon TH le remplace tout en maintenant constant le nombre de résidences principales repérées.
3. A partir du T4 2011, l'échantillon EEC n'est constitué que de l'échantillon TH.

L'introduction doit s'opérer sur 12 trimestres consécutifs. Les 6 premiers sont des trimestres d'augmentation, les 6 derniers de remplacement. Dans la première phase 1/3 de l'échantillon TH[11] doit être introduit soit 1/18ème par trimestre. Dans la seconde phase, les 2/3 restants sont mobilisés soit 1/9ème par trimestre. On en déduit le schéma de rotation du nouvel échantillon. La fixation des paramètres de dimensionnement et de calendrier permet, en théorie, de passer au tirage effectif de l'échantillon.

Graphique 8 : répartition du nombre de résidences principales repérées selon le type d'échantillon



[11] 1/3 de 66 000 = 22 000 = 50 % de l'échantillon actuel.

Tableau 7 : Schéma de rotation de l'échantillon issu de la TH

Date d'observation	Sous-échantillon de secteurs											
	Phase 1						Phase 2					
	1	2	3	4	5	6	7	8	9	10	11	12
	Date d'entrée											
	2009				2010				2011			
	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4
T1-2009	1											
T2-2009	1	1										
T3-2009	1	1	1									
T4-2009	1	1	1	1								
T1-2010	1	1	1	1	1							
T2-2010	1	1	1	1	1	1						
T3-2010	2	1	1	1	1	1	1					
T4-2010	2	2	1	1	1	1	1	1				
T1-2011	2	2	2	1	1	1	1	1	1			
T2-2011	2	2	2	2	1	1	1	1	1	1		
T3-2011	2	2	2	2	2	1	1	1	1	1	1	
T4-2011	2	2	2	2	2	2	1	1	1	1	1	1
T1-2012	3	2	2	2	2	2	2	1	1	1	1	1
T2-2012	3	3	2	2	2	2	2	2	1	1	1	1
T3-2012	3	3	3	2	2	2	2	2	2	1	1	1
T4-2012	3	3	3	3	2	2	2	2	2	2	1	1
T1-2013	3	3	3	3	3	2	2	2	2	2	2	1
T2-2013	3	3	3	3	3	3	2	2	2	2	2	2
T3-2013	4	3	3	3	3	3	3	2	2	2	2	2

5. Sélection de l'échantillon[12]

Le schéma global de sélection de l'échantillon est descendant : Sélection de l'ensemble des secteurs par région, répartition des secteurs en 6 groupes de même taille, scission de chaque groupe en 2 sous-échantillons le premier comportant 1/3 des secteurs le second les 2/3, dans chaque sous-échantillon affectation aléatoire d'un numéro de 1 à 6 pour chaque grappe d'un secteur donné.

5.1. Sélection des secteurs par région

Dans un premier temps, l'ensemble des secteurs correspondant au tableau 6 ont été sélectionnés par région, de manière équilibrée, proportionnellement au nombre de résidences principales. On a eu recours aux principes théoriques d'équilibrage définis par *J-C Deville et Y Tillé* [5] et accessibles grâce aux macros SAS développées par *G. Chauvet* [6], notamment la macro *FASTCUBE* qui met en œuvre un algorithme plus rapide que les macros habituelles : *G. Chauvet et Y.Tillé* [7]. Pour guider le choix des variables d'équilibrage, ont été réalisées des régressions par aire à partir de l'enquête actuelle. La variable expliquée est le nombre de chômeurs. Parmi les variables disponibles dans la base de sondage, le statut d'occupation du logement : locataire privé ou HLM est le critère qui apparaît le plus discriminant. Les autres ne contribuent pas à une augmentation significative du R^2. Toutefois, compte tenu de l'étendue des thèmes abordés par l'EEC, il a semblé que des critères de revenus, de typologie rural-urbain, de construction neuve, d'âge étaient également de bons candidats à l'équilibrage. Au final, les variables retenues sont, par ordre décroissant d'importance :

- La répartition des résidences principales par type d'espace selon le Zonage en Aires Urbaines : pôles urbains, couronnes périurbaines, communes multi polarisées, communes rurales,
- La répartition des résidences principales par quintile de revenus,

[12] La programmation et le tirage effectif de l'échantillon a été réalisé par Sébastien Hallépée alors en poste à la division ETSD.

- Le nombre de locataires,
- La répartition selon la date d'achèvement, essentiellement pour les logements récents,
- Le nombre de logements sociaux,
- Le nombre de logements collectifs,
- Le nombre de résidences principales dont le chef de ménage a plus de 55 ans,
- Le nombre de résidences principales,
- La probabilité d'inclusion pour assurer un sondage de taille fixe.

Dans l'optique d'un éventuel accroissement de la taille de l'échantillon, le tirage a été effectué de manière à ce que le complémentaire de l'échantillon tiré soit également équilibré.

5.2. Répartition des secteurs en 6 groupes

L'introduction de 12 sous-échantillons de secteurs est quelque peu artificielle. Le rythme qui importe en régime de croisière pour l'EEC est celui du rang d'interrogation et donc du mode de collecte. On rappelle que les premières et dernières interviews ont lieu en face à face, alors que les interviews intermédiaires sont téléphoniques. Dans ce régime, les grappes des sous-échantillons 1 et 7 seront dans la même phase d'interrogation, de même celles des 2 et 8… C'est pourquoi les secteurs échantillonnés en première approche n'ont pas été directement répartis en 12 sous-échantillons mais en 6 groupes.

La sélection des groupes a été effectuée de manière emboîtée dans l'échantillon de secteur : le premier dans l'ensemble avec une probabilité de 1/6, le second dans le complément avec une probabilité de 1/5… A chaque sélection, on a eu recours à l'équilibrage. Sachant que la population dans laquelle on sélectionne est de faible taille par région[13], les ambitions ont été revues à la baisse sur le nombre de critères d'équilibrage. Parmi la liste précédente n'ont été conservés que : la population des pôles urbains, les nombres de logements neufs, sociaux, collectifs et principaux. Par ailleurs si l'équilibrage a eu lieu par région, les phases d'atterrissage ont été mises en commun pour une estimation nationale (G. Chauvet [8])

5.3. Répartition des groupes en sous-échantillon

Enfin, les sous-échantillons ont été construits à partir de chaque groupe. On a sélectionné au taux de 1/3 les secteurs par groupe et région en les équilibrant sur les mêmes variables que les groupes et mise en commun des phases d'atterrissage. On dispose ainsi des 12 sous-échantillons, dont le rythme d'entrée est donné par le tableau 8.

5.4. Attribution d'un numéro par grappe

L'attribution des numéros de grappe reprend globalement la méthodologie précédente. Pour chaque sous-échantillon, on a sélectionné de manière emboîtée un échantillon de taille 1 par secteur : les secteurs constituent les strates. Les phases d'atterrissage sont toujours mises en commun. Le nombre de variables d'équilibrages décroît avec le numéro de la grappe selon la logique suivante :

1. probabilités d'inclusion, nombre de résidences principales, nombre de chef de ménage de 55 à 65 ans, nombre de logements collectifs, nombre de logements sociaux,
2. probabilités d'inclusion, nombre de résidences principales, nombre de chef de ménage de 55 à 65 ans, nombre de logements collectifs
3. probabilités d'inclusion, nombre de résidences principales, nombre de chef de ménage de 55 à 65 ans
4. probabilités d'inclusion, nombre de résidences principales
5. probabilités d'inclusion.

[13] Il s'agit de celles données par le tableau 8.

Tableau 8 : nombre de secteurs entrants par trimestre et région de gestion (échantillon effectivement tiré dans la TH)													
	Trimestre d'entrée												Ensemble
Région de gestion	T1 2009	T2 2009	T3 2009	T4 2009	T1 2010	T2 2010	T3 2010	T4 2010	T1 2011	T2 2011	T3 2011	T4 2011	
Île-de-France	22	20	21	25	26	22	49	47	49	40	42	48	**411**
Champagne-Ardenne	8	12	14	8	8	8	18	10	18	22	23	10	**159**
Picardie	0	2	6	7	5	6	8	13	11	10	13	9	**90**
Haute-Normandie	8	10	5	8	10	7	14	14	17	18	13	16	**140**
Centre	9	12	7	3	14	11	19	22	18	22	21	22	**180**
Basse-Normandie	9	6	6	6	4	4	5	9	9	8	10	10	**86**
Bourgogne	4	2	6	7	5	5	13	11	9	7	10	12	**91**
Nord-Pas de Calais	11	9	10	9	11	12	21	21	23	24	20	20	**191**
Lorraine	8	5	4	6	6	11	15	12	15	17	9	8	**116**
Alsace	8	3	5	6	6	3	8	10	8	8	14	13	**92**
Franche-Comté	6	7	4	4	5	1	10	8	4	12	9	9	**79**
Pays de la Loire	12	11	12	9	9	10	19	18	17	22	17	19	**175**
Bretagne	8	8	8	10	11	9	17	20	17	17	20	17	**162**
Poitou-Charentes	8	6	4	2	4	6	9	12	10	10	8	12	**91**
Aquitaine	9	8	10	12	10	5	18	19	17	13	20	20	**161**
Midi-Pyrénées	6	7	10	9	8	6	19	15	16	17	15	15	**143**
Limousin	3	5	5	2	3	6	9	9	9	8	8	10	**77**
Rhône-Alpes	17	20	18	18	13	17	34	31	34	34	34	33	**303**
Auvergne	4	4	4	4	4	5	9	9	5	8	8	9	**73**
Languedoc-Roussillon	9	7	7	6	5	10	13	15	17	15	12	14	**130**
PACA	11	15	13	14	12	14	30	27	30	25	30	30	**251**
Corse	0	0	2	2	2	1	1	4	0	0	0	4	**16**
Ensemble	**180**	**179**	**181**	**177**	**181**	**179**	**358**	**356**	**353**	**357**	**356**	**360**	**3217**

6. Prise en compte de la charge de travail des enquêteurs

6.1. les bisecteurs...

Le schéma précédent est celui qui a été effectivement appliqué à la différence prêt que l'on a sélectionné, jusqu'en phase 3, des bisecteurs et non pas directement des secteurs. En effet, la sélection des secteurs peut conduire in fine à ce que, dans des zones peu denses, un seul secteur soit échantillonné. Un trimestre donné, si la charge de travail associée à un secteur et donc à une grappe, est suffisante pour une durée de collecte de 2 semaines et 2 jours, elle peut paraître faible sur une année donnée. Il peut potentiellement y avoir des difficultés de recrutement ou de fidélisation des enquêteurs. C'est pourquoi les responsables de la collecte ont émis le souhait d'éviter de sélectionner des secteurs isolés. Pour répondre à cette attente, a été mise en place la notion de bisecteur. Un bisecteur est un regroupement de 2 secteurs dans le même canton ou fraction de canton.

Dans un premier temps, on a estimé la probabilité que dans l'environnement proche d'une commune au moins 3 secteurs soient échantillonnés, dans l'hypothèse où l'on sélectionne directement des secteurs. L'environnement proche a été défini comme étant l'ensemble des communes que l'on trouve autour d'une commune par un critère de contiguïté d'ordre 2 : la commune elle-même, les communes limitrophes, et les communes limitrophes de l'ensemble ainsi défini. Le graphique *9* représente l'environnement proche d'une commune.

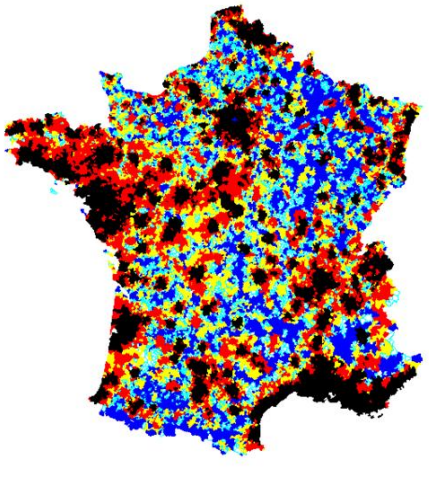
Par simulation du plan de sondage précédent, au niveau secteur, et limité à la première phase, on a estimé cette probabilité. Le graphique 10 donne la représentation des résultats obtenus pour la France métropolitaine.

Chaque région a été découpée en deux strates : les communes où la probabilité d'avoir au moins 3 secteurs est supérieure à 0.95, les communes où cette probabilité est inférieure à 0.95. Dans ces dernières, les secteurs de la base de sondage ont été regroupés aléatoirement par deux dans le

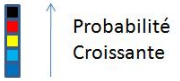
Graphique 9 : l'environnement proche d'une commune : contiguïté d'ordre 2

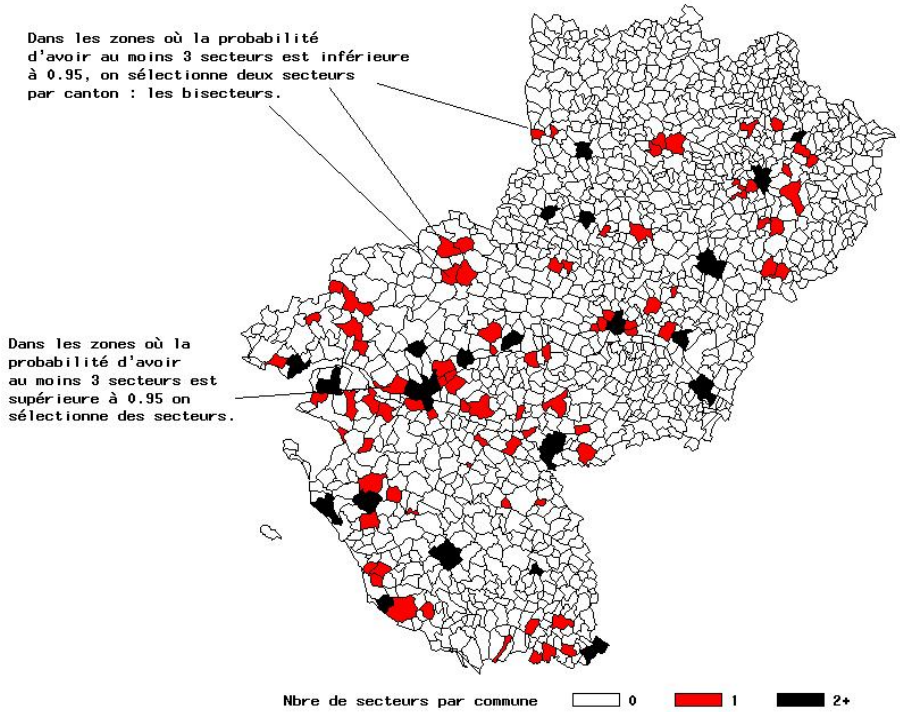
Graphique 10 : Probabilité par commune d'avoir au moins 3 secteurs dans l'environnement proche

Degré d'isolement des communes : sondage stratifié par région, répartition proportionnelle au nombre de résidences principales, probabilité d'inclusion proportionnelle à la taille, équilibré.



Pour une commune donnée, on calcule, par simulation, la probabilité de trouver dans un environnement proche au moins 3 secteurs pour un plan de sondage donné.





cadre des frontières cantonales. Si le canton a un nombre impair de secteurs, un trisecteur a été créé. La procédure de sélection a été appliquée avec en entrée la base de sondage composée des secteurs des communes non isolées et des bisecteurs des communes isolées. Le seul changement est que la contrainte de taille n'a pas pu être respectée. Au final, sur les 3211 secteurs prévus, 3217 ont été échantillonnés en pratique (tableau 8). On remarquera que l'on est assuré d'avoir, pour les zones isolées, au moins deux secteurs proches dans le même canton (graphique 11) et que ces secteurs appartiendront au même sous-échantillon : ils rentreront dans l'échantillon à la même date. Une autre demande des responsables collecte étaient d'éviter d'avoir à proposer à un enquêteur un secteur entrant en janvier 2009 et un autre en octobre 2011 !!!

Par la concentration géographique qu'ils impliquent, les bisecteurs peuvent avoir un effet négatif sur la précision. On notera que cette situation prévalait déjà pour l'enquête annuelle où, dans la strate rurale, les aires étaient de 40 logements, c'est à dire une concentration plus forte encore. Ici, pour introduire un peu d'hétérogénéité, la contrainte de proximité a été limitée au canton. Afin de vérifier que l'impact du regroupement sur la précision nationale était faible, on a estimé la précision dans les deux plans de sondage (avec et sans regroupement). Les conclusions montrent que la détérioration due au regroupement reste faible (G. Chauvet [9]).

6.2. Les semaines de référence

La collecte de l'information sur la situation du marché du travail dans l'EEC s'effectue de manière continue par semaine de référence au cours d'un trimestre. Il importe que ces semaines soient réparties uniformément. Dans l'échantillon actuel, la répartition est uniforme, en régime de croisière, par région. Elle ne l'est pas par sous-échantillon, alors que la plupart des traitements avals s'effectuent par sous-échantillon. Dans le nouvel échantillon, on a conservé la répartition uniforme par région mais on a rajouté comme contrainte l'uniformité également par sous-échantillon.

Une contrainte de collecte supplémentaire s'est imposée. On sait que dans les zones peu denses des bisecteurs ont été sélectionnés de manière à ce que deux secteurs échantillonnés soient proches géographiquement et qu'ils aient la même date d'entrée dans l'échantillon. Pour ne pas surcharger les enquêteurs une semaine donnée, a été également émis le souhait que deux secteurs d'un même bisecteur ait des semaines de référence espacées d'au moins 3 semaines. Cette contrainte a été étendue aux Aires Urbaines ayant moins de 5 secteurs en tout. Pour ces dernières, la date d'entrée des différents secteurs peut varier, mais au final les semaines de référence seront espacées d'au moins 2 semaines.

Au final, le problème à résoudre est pour le moins compliqué ou du moins non trivial[14]. Il faut imputer la semaine de référence de manière à ce que :

1. la distribution de la semaine sachant la région est uniforme,
2. la distribution de la semaine sachant le sous-échantillon est uniforme,
3. la répartition marginale par région et par sous échantillon est fixée,
4. dans un bi-secteur l'écart entre les semaines de référence de chaque secteur est supérieur ou égal à trois semaines, sachant que le sous-échantillon est le même pour les deux secteurs,
5. dans les aires urbaines ayant moins de 5 secteurs échantillonnés, l'écart entre les semaines de référence est supérieur ou égal à 2 semaines.

La résolution de ce type de problème s'apparente à la résolution automatique du jeu de SUDOKU. On ne la présente pas ici. On montre cependant avec les tableaux 8 et 9 que sur deux dimensions le problème est bien résolu.

[14] L'algorithme a été mis au point par D. Place.

Tableau 8 : répartition des semaines du nouvel échantillon EEC selon la semaine de référence et le sous-échantillon

Sous échantillon	Semaine de référence													Ensemble
	1	2	3	4	5	6	7	8	9	10	11	12	13	
1	14	14	14	14	13	14	14	14	13	14	14	14	14	**180**
2	14	14	14	14	14	13	13	14	14	14	14	13	14	**179**
3	14	14	13	14	14	14	14	14	14	14	14	14	14	**181**
4	14	14	14	14	14	14	13	13	13	13	13	14	14	**177**
5	13	14	14	14	14	14	14	14	14	14	14	14	14	**181**
6	14	14	14	14	14	13	13	13	14	14	14	14	14	**179**
7	27	27	28	28	28	28	28	27	28	27	28	27	27	**358**
8	27	27	27	27	28	28	28	28	27	27	27	28	27	**356**
9	27	27	28	27	27	27	27	27	28	27	27	27	27	**353**
10	28	27	27	27	27	27	28	28	28	28	28	27	27	**357**
11	28	27	28	27	27	27	28	27	28	27	27	27	28	**356**
12	28	28	27	27	28	28	28	28	27	28	28	28	27	**360**
Ensemble	**248**	**247**	**248**	**247**	**248**	**247**	**248**	**247**	**248**	**247**	**248**	**247**	**247**	**3217**

Tableau 9 : répartition des semaines du nouvel échantillon EEC selon la semaine de référence et la région

Région	Semaine de référence													Ensemble
	1	2	3	4	5	6	7	8	9	10	11	12	13	
Île-de-France	44	44	44	44	44	44	44	44	44	44	44	44	44	**572**
Champagne-Ardenne	8	7	8	7	8	7	8	7	8	7	8	8	8	**99**
Picardie	7	7	7	7	7	7	7	7	7	7	7	7	6	**90**
Haute-Normandie	6	7	7	7	7	7	7	7	7	7	7	7	7	**90**
Centre	9	10	10	10	10	10	10	10	10	10	10	10	10	**129**
Basse-Normandie	6	6	6	7	6	7	7	7	7	7	7	7	6	**86**
Bourgogne	7	7	7	7	7	7	7	7	7	7	7	7	7	**91**
Nord-Pas de Calais	15	15	15	14	15	14	15	14	15	14	15	15	15	**191**
Lorraine	9	9	9	9	9	9	8	9	9	9	9	9	9	**116**
Alsace	8	7	7	7	7	7	7	7	7	7	7	7	7	**92**
Franche-Comté	7	6	6	6	6	6	6	6	6	6	6	6	6	**79**
Pays de la Loire	13	14	13	14	13	14	13	14	13	14	13	14	13	**175**
Bretagne	13	13	12	13	12	12	13	12	12	13	12	12	13	**162**
Poitou-Charentes	7	7	7	7	7	7	7	7	7	7	7	7	7	**91**
Aquitaine	13	12	12	12	12	13	13	13	12	12	12	12	13	**161**
Midi-Pyrénées	11	11	11	11	11	11	11	11	11	11	11	11	11	**143**
Limousin	5	6	6	6	6	6	6	6	6	6	6	6	6	**77**
Rhône-Alpes	23	23	23	24	23	24	23	24	23	24	23	23	23	**303**
Auvergne	6	6	6	5	6	5	6	5	6	5	6	5	6	**73**
Languedoc-Roussillon	10	10	10	10	10	10	10	10	10	10	10	10	10	**130**
PACA	19	19	20	19	20	19	19	19	20	19	20	19	19	**251**
Corse	2	1	2	1	2	1	1	1	1	1	1	1	1	**16**
Ensemble	**248**	**247**	**248**	**247**	**248**	**247**	**248**	**247**	**248**	**247**	**248**	**247**	**247**	**3217**

7. Procédures post-échantillonnage

A ce stade, on dispose d'un échantillon de secteurs, pour lesquels on dispose de la date d'entrée. Les secteurs sont constitués de 6 grappes, pour lesquelles on dispose du rang d'interrogation. Cet échantillon, sélectionné dans la TH 2006, n'est pas encore directement exploitable, il convient de contrôler le nombre de logements non principaux, de repérer d'éventuelles aberrations géographiques dans la constitution des secteurs et des grappes, d'assurer la disjonction avec les autres enquêtes ménages et d'enrichir l'échantillon par les logements nouveaux. Cette partie se propose de présenter ces points.

7.1. La sélection des logements non principaux

Les logements non principaux n'interviennent dans la procédure de construction de l'échantillon que dans la mesure où l'on impose qu'ils soient affectés à une grappe. Ces dernières sont cependant calibrées sur le nombre de résidences principales, et quand sélection proportionnelle il y a, elle s'effectue également selon le nombre de résidences principales. Au total, une grappe de l'échantillon

peut avoir un nombre variable, et donc potentiellement grand, de logements non principaux qui lui sont associés. On rappelle que ces logements doivent faire partie de l'échantillon pour intégrer les logements vacants ou résidences secondaires au sens de la TH, à la date de sélection, qui sont en fait des résidences principales au sens de l'Insee au moment de l'enquête. Pour conserver la charge de travail des enquêteurs dans des limites raisonnables, une procédure d'échantillonnage a été mise en place sur les logements non principaux, selon les règles suivantes :

- De 1 à 10 logements non principaux dans une grappe, tous les logements sont inclus dans l'échantillon,
- De 11 à 40 logements non principaux dans une grappe, 10 logements sont inclus aléatoirement dans l'échantillon,
- De 41 à 100 logements non principaux dans une grappe, 1 logement sur 4 est inclus aléatoirement dans l'échantillon,
- Plus de 100 logements non principaux dans une grappe, 25 logements sont inclus aléatoirement dans l'échantillon.

Ce cadre a été amendé pour tenir compte des contraintes intégrées dans la construction des grappes,

- Si dans une grappe échantillonnée, un logement non principal est situé au même étage que des résidences principales, il est systématiquement inclus dans l'échantillon. Ces logements sont qualifiés de « logements obligés ». On conserve ici le principe qu'un étage entier est inclus dans l'échantillon.
- Ce dernier principe s'applique également pour la sélection des autres logements non principaux. L'unité d'échantillonnage sera un étage entier.
- Si n est le nombre de logements non principaux à sélectionner dans une grappe et n_{ob} le nombre de logements obligés, on montre que l'on peut trouver une procédure d'échantillonnage qui assure à tous les logements une probabilité non nulle d'être sélectionnés que si le plus grand des étages constitués uniquement de logements non principaux a un nombre de logements au plus égal à $n- n_{ob}$.

7.1.1. Sélection dans le cas 1

Si le plus grand des étages a moins de $n-n_{ob}$ logements, on regroupe les étages par proximité géographique dans le bâtiment ou dans la rue, de manière à disposer de regroupements ayant au plus $n- n_{ob}$ logements. Cette procédure part du principe que si il y a beaucoup de résidences secondaires dans une grappe, il est préférable de demander à un enquêteur de ne visiter que des résidences proches. En effet, la plupart du temps, la résidence secondaire sera effectivement une résidence secondaire ne donnant pas lieu à enquête, multiplier les déplacements pour constater des hors champs n'est pas alors considéré comme un facteur de motivation.

Dans ce regroupement, on procède de manière itérative et emboîtée. On sélectionne un groupe proportionnellement au nombre de logements, on continue dans le reliquat tant que le nombre total sélectionné ne dépasse pas la valeur cible $n- n_{ob}$. Quand ce critère est vérifié, à l'étape k, on conserve les logements sélectionnés jusqu'à l'étape k-1.

Une telle procédure, avec critère d'arrêt et sélection emboîtée proportionnellement à un critère de taille, n'est pas de nature à faciliter le calcul des probabilités d'inclusion. Pour les obtenir, on procède par simulation. C'est à dire que l'on répète la procédure un grand nombre de fois, disons 1000, on compte à la fin le nombre de fois où un logement ou un regroupement a été sélectionné. La probabilité d'inclusion sera ce nombre divisé par 1000.

On sélectionne alors un seul échantillon auquel on attribue les probabilités simulées.

7.1.2. Sélection dans le cas 2

Si le plus grand des étages composé uniquement de logements non principaux a plus de $n-n_{ob}$, la procédure précédente ne fonctionne pas. Il faut faire en sorte que le critère soit vérifié. Pour cela, plusieurs solutions sont possibles :

1. On augmente la valeur de *n*, en la fixant à MAX- n_{ob}, où MAX est la taille du plus grand étage composé de logements non principaux. Si cette valeur est dans des bornes jugées acceptables, alors on applique la procédure du cas 1. Dans la pratique, les bornes jugées acceptables ont été fixées à n=20 environ[15].
2. La présence de grands étages avec uniquement des logements non principaux peut être le signe de la présence de communautés ou de résidences de vacances, qui sont hors champ de l'enquête. On peut alors procéder à un examen des logements de la grappe en plaçant hors champ le cas échéant certains étages. Si, calculé sur les seuls étages du champ, le critère de taille est vérifié, on se retrouve dans le cas 1.
3. Une dernière option est de diminuer la valeur de n_{ob}. Pour cela, on examine la liste des étages avec résidences principales et logements non principaux. Localement, on peut percevoir des erreurs de classement par exemple une résidence principale dans un étage avec 15 résidences secondaires. Dans ce cas, on prend la liberté de modifier le classement de la résidence principale, avant d'appliquer la procédure du cas 1.

Au total, la sélection des logements non principaux fait intervenir, sur peu de cas, des examens manuels de liste de logements. C'est pourquoi elle n'est appliquée qu'au fur et à mesure de l'entrée des sous-échantillons.

7.2. Réallocation, établissement des documents de collecte, disjonction

7.2.1. Réallocation

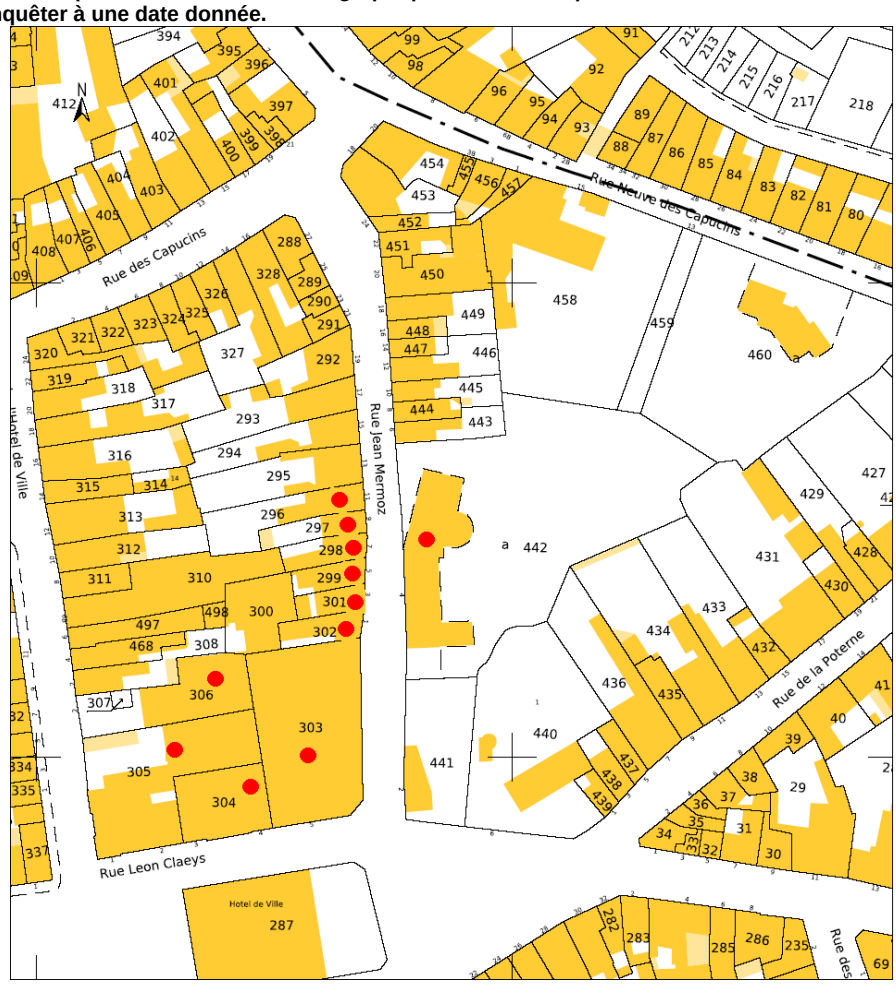
On a insisté sur l'attention portée aux questions de déplacements des enquêteurs. Il se peut malgré tout qu'in fine, dans un secteur échantillonné, il serait plus pertinent, d'un point de vue géographique, d'affecter certains logements d'une grappe à une autre grappe du même secteur. On a mis en place une procédure, appelée *réallocation*, pour peaufiner la répartition spatiale des logements dans les secteurs et les grappes. En toute rigueur statistique, cette procédure aurait dû intervenir avant l'attribution des numéros aux grappes, puisque cette attribution s'est effectuée en équilibrant les grappes. Toutefois, compte tenu du faible taux attendu[16] de logements réalloués et de tout l'intérêt de cette procédure sur l'établissement des documents de collecte, ces questions méthodologiques sont apparues mineures.

7.2.2. Établissement des documents de collecte

La procédure de réallocation est une procédure par secteur. Elle peut être effectuée selon le rythme d'entrée des sous-échantillons, lissant la charge de travail des Directions Régionales. Le principe en est simple. Régulièrement sont envoyées aux Directions Régionales la compostion des secteurs devant entrer à une date donnée. Sur la base de la description des logements à partir des adresses ou des références cadastrales, sont examinées les possibilités de réallouer selon une meilleure logique géographique les logements entre les grappes d'un secteur. Les DR font remonter de manière centralisée le résultat du travail qui est intégré pour modifier l'échantillon. A cette occasion, les DR peuvent établir des plans, à partir du site du cadastre, qui seront fournis aux enquêteurs au moment de l'entrée de la grappe dans l'échantillon (voir graphique 12).

[15] On n'a rencontré pour le moment que des grappes à problème ayant de 11 à 40 logements non principaux.
[16] De l'ordre de 3 % des logements sont réalloués France Entière, et seulement 0.9% hors Île de France.

Graphique 12 : Exemple de documents cartographiques établis à partir du site du cadastre : localisation des logements à enquêter à une date donnée.



7.2.3. La disjonction

Afin de limiter la charge d'enquête des ménages, la disjonction des échantillons est l'opération qui consiste à s'assurer qu'un logement ne sera pas enquêté deux fois sur une période donnée. Dans les opérations standard de l'Insee cette disjonction est assurée par le tirage coordonné d'échantillons. Dans le cas de l'EEC, la disjonction doit s'opérer avec l'échantillon actuel mais aussi pour toutes les autres enquêtes. Compte tenu de la primauté de l'EEC, le problème de la disjonction est reporté sur la constitution des autres échantillons. On fournit pour cela la liste des logements échantillonnés, qui seront ôtés de l'Échantillon Maître[17] de l'Insee.

La disjonction avec l'échantillon actuel est une disjonction temporelle. La difficulté provient du fait que l'on ne peut pas apparier directement ancien et nouvel échantillon. Cependant, l'Enquête Revenus Fiscaux apporte une solution. En effet, cette dernière porte sur l'échantillon sortant de l'EEC. Pour répondre à ces objectifs, elle effectue un appariement avec les données fiscales. On peut ainsi obtenir l'identifiant pérenne des répondants à l'actuelle EEC. On n'a pas cependant ceux des logements qui pourront être interrogés d'ici la fin de vie de l'échantillon en 2011. On procède alors comme suit :

- On repère, par l'intermédiaire de l'enquête revenus fiscaux, les grappes du futur échantillon qui ont au moins 1 logement déjà enquêté au titre de l'actuel EEC.
- On fournit aux directions régionales les secteurs du nouvel échantillon contenant ces logements.
- A partir de la description physique des aires actuelles, dont elles disposent, les DR vérifient si on peut s'attendre d'ici la fin de vie de l'échantillon actuel à ce que d'autres logements du

[17] L'Echantillon Maître est la procédure de l'Insee qui permet d'obtenir des échantillons pour les enquêtes standard.

nouvel échantillon soient enquêtés au titre de l'actuel EEC, et quelles sont, le cas échéant, les grappes concernées.

- Pour ces grappes, on modifie le rang d'entrée pour faire en sorte qu'elles soient mobilisées le plus tard possible. On peut repousser ainsi jusqu'à 7,5 années le temps entre deux ré interrogations du même logement.
- On notera que cette disjonction affecte tout au plus quelques grappes.

8. Procédures de mise à jour de l'échantillon

Les procédures de mise à jour ont un triple objectif :

1. pour l'échantillon entrant à une date donnée, il s'agit de mettre à jour avec les dernières informations disponibles l'ensemble des variables permettant le repérage des logements sur le terrain : *nom, prénom, adresse, référence cadastrale…*
2. pour l'échantillon sur le terrain à une date donnée, il s'agit de disposer de l'information auxiliaire la plus fraîche possible pour procéder aux opérations de correction de la non-réponse,
3. pour l'échantillon constitué à une date donnée, il s'agit de l'enrichir par les logements nouveaux apparus depuis la dernière mise à jour.

Les procédures de mise à jour sont annuelles, selon le rythme d'arrivée des fichiers de la taxe d'habitation. Les deux premières sont effectuées en même temps selon le même principe. Du fait de la disponibilité d'un identifiant pérenne des logements, il « suffit » d'apparier l'échantillon et le nouveau fichier TH[18] par l'identifiant pérenne en ne conservant que les variables nécessaires.

La sélection des logements nouveaux s'appuie sur l'identifiant pérenne des logements. On appellera « logements nouveaux » à la date *N+1*, tous les logements dont l'identifiant pérenne n'est pas connu sur la période N_0 *(=2006)* à *N*. On préfère le terme « nouveau » à « neuf » car une part de ces logements peut provenir de la réhabilitation. Le principe d'intégration de ces logements est relativement simple.

La constitution des grappes à l'intérieur d'une commune peut se résumer au tri des logements selon la section cadastrale et le numéro de parcelle. Dans les fichiers TH *N+1*, on peut ainsi :

- trier le fichier pour construire les grappes,
- identifier si un logement est nouveau ou non,
- repérer le premier logement non nouveau se situant avant un logement nouveau.
- Si ce premier logement est un logement échantillonné au titre de la sélection initiale dans la TH 2006 ou de la mise à jour des logements nouveaux jusqu'à N, alors le logement nouveau N+1 est échantillonné. Il prend l'ensemble des caractéristiques du logement auquel il a été associé, numéro de secteur, de grappe (et donc date d'entrée) et pondération.

La justification théorique relève du sondage indirect (P. Lavallée [10]). Les logements nouveaux sont échantillonnés par l'intermédiaire des logements existants. Comme à un logement nouveau n'est associé qu'un seul logement existant, l'application de la formule du partage des poids implique que la pondération se transmet sans modification. L'estimation finale sur les seuls logements nouveaux échantillonnés est sans biais sur la population des logements nouveaux.

On remarquera que cette méthode, outre sa relative simplicité théorique[19], permet de suivre la construction nouvelle en n'utilisant que les secteurs et grappes définies initialement et donc le réseau d'enquêteur existant.

On remarquera par ailleurs que tant qu'une grappe n'est pas entrante on l'enrichit par des nouveaux logements. Il se peut qu'au moment d'entrer le nombre de logements nouveaux qui lui sont attachés soient importants. Dans ce cas, il faudra prévoir une procédure d'échantillonnage qui est exactement

[18] On rappelle cependant que les fichiers TH ayant plus de 40 millions d'observations, une opération simple dans son principe peut ne pas être directe dans sa mise en œuvre.
[19] Ici aussi, il faut distinguer simplicité théorique et simplicité informatique…

la même que celle utilisée pour la sélection des logements non principaux. Les logements nouveaux obligés seront ceux se situant à un étage ayant au moins une résidence principale non nouvelle[20].

Bibliographie

[1] D. Place (2008), « Calcul de la précision des estimations longitudinales dans l'Enquête Emploi en Continu », in Méthodes de Sondage, P. Guilbert, D. Haziza, A. Ruiz-Gazen, Y. Tillé, Dunod 2008.
[2] « Rapport sur les statistiques d'estimation du chômage», Inspection Générale des Finances, Inspection Générale des Affaires Sociales.
[3] M. Christine (2002), « La construction de l'échantillon de la future l'Enquête Emploi en Continu à partir du Recensement 1999 », in Actes des Journées de Méthodologie Statistique, *Insee Méthodes*, n° 100, pp. 175-229.
[4] P. Ardilly (2006), «Les techniques de sondage », Paris : Éd. Technip, 2006.
[5] J-C. Deville et Y. Tillé (2004), Efficient balanced sampling : the cube method*, Biometrika*, **91**, 893-912.
[6] G. Chauvet 2006. De nouvelles macros SAS d'échantillonnage équilibré. Technical report, ENSAI, Rennes
[7] G. Chauvet et Y. Tillé. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21:53–61.
[8] G. Chauvet (2008), « Echantillonnage Equilibré Stratifié », Série des Documents de Travail du CREST, n° 2008-08.
[9] G. Chauvet (2009), note de travail Insee, n° : 28/DR35-SED
[10] P. Lavallée (2002). *Le Sondage Indirect, ou la Méthode généralisée du partage des poids.* Éditions de l'Université de Bruxelles, Brussels.

[20] Il s'agira la plupart du temps pour les logements nouveaux obligés de logement dont la vocation a été modifiée : cabinet de médecins vers logement d'habitation.