Le renouvellement de l'échantillon-maître des enquêtes auprès des ménages et de l'échantillon de l'enquête Emploi de l'Insee

Insee Méthodes
N° 141 - Mai 2022

Insee
Mesurer pour comprendre

Auteurs

Ce document a été co-écrit par de nombreux auteurs ayant contribué au projet Nautile entre 2016 et 2019 : Martin Chevalier, Laurent Costa, Lionel Delta, Thomas Deroyon, Cyril Favre-Martinoz, Samuel Givois, Clément Guillo, Thomas Merly-Alpa, Nicolas Paliod, Pierre-Arnaud Pendoli, Thomas Sauvaget, Ludovic Vincent.

Sa rédaction a été coordonnée par Ludovic Vincent et Nicolas Paliod.

Résumé

Dans un contexte de développement de la collecte par Internet et par téléphone, la collecte des enquêtes auprès des ménages réalisée par l'Institut National de la Statistique et des Études Économiques (Insee) reste encore majoritairement menée en face en face. Afin de permettre la constitution d'un réseau d'enquêteurs stable et de limiter le déplacement des enquêteurs de l'Insee, le tirage des échantillons est restreint à des zones de collecte tirées aléatoirement et fixées pour une dizaine d'années. En 2019-2020, le renouvellement concomitant en France métropolitaine des zones de collecte utilisées pour les enquêtes en face-à-face, à l'exception de l'enquête Emploi en continu, et des zones de collecte utilisées pour l'enquête Emploi en continu a donné l'opportunité de mener plusieurs travaux méthodologiques, lancés dès 2017 et détaillés dans ce document. Les enjeux du renouvellement des zones de collecte font l'objet de la partie A.

D'une part, il est nécessaire de constituer des zones disposant d'un vivier de logements suffisamment important pour permettre d'enquêter des logements différents durant une dizaine d'année. D'autre part, ces zones doivent être suffisamment compactes pour éviter aux enquêteurs des déplacements importants. La partition du territoire en zones de collecte adaptées à ces objectifs est présentée en partie B.

Néanmoins, se restreindre à des zones de collecte pour l'interrogation des unités enquêtées implique d'enquêter des individus ou des logements ayant des profils moins variés qu'en interrogeant des unités réparties sur l'ensemble du territoire. Le tirage de zones de collecte génère ainsi un aléa d'estimation dans les enquêtes auprès des ménages et dans l'enquête Emploi en continu. Des études d'optimisation de variance ont été menées afin de maîtriser l'aléa engendré par le tirage de ces zones. Elles sont détaillées en partie C, au même titre que les paramètres ayant été retenus pour le tirage des zones de l'enquête Emploi et des enquêtes auprès des ménages. Les parties A, B et C approfondissent ainsi l'article de Sillard et al., 2020.

Enfin, l'ensemble des travaux nécessaires au renouvellement des zones de collecte ne saurait suffire pour enquêter des logements et des individus. Des questions relatives au calendrier de mise en collecte des zones de l'enquête Emploi en continu, à la sélection des logements interrogés au cours de cet enquête, à la méthode de tirage utilisée pour la sélection des échantillons pour les enquêtes auprès des ménages et à la gestion temporelle d'une base de sondage doivent être tranchées. Ces travaux de finalisation sont décrits dans la partie D.

Table des matières

Partie A

Contexte : le projet Nautile

Dans un contexte de développement de la collecte par Internet et par téléphone, la collecte des enquêtes auprès des ménages réalisée par l'Institut National de la Statistique et des Études Économiques (Insee) reste encore majoritairement menée en face en face. Afin de disposer d'un réseau d'enquêteurs stable et de limiter leurs déplacements, la logique de l'échantillon-maître (EM) est mise en oeuvre : il s'agit de réaliser un échantillon géographique de premier degré, représentatif de l'ensemble du territoire, permettant de concentrer la collecte de plusieurs enquêtes dans les mêmes zones géographiques, appelées unités primaires ; le tirage des échantillons de logements ou d'individus relatifs à ces enquêtes se fait alors au second degré au sein des zones sélectionnées. L'échantillon-maître de 2009 (EM2009), représentatif des données du Recensement de la Population de 1999 (RP99) et dont une description est donnée au point 1.1.2, s'étant dégradé de manière continue au cours du temps, son renouvellement est devenu indispensable (point 1.1.3).

En parallèle, afin de mesurer l'évolution structurelle et conjoncturelle du marché du travail, selon la définition du Bureau International du Travail (BIT), l'Insee produit trimestriellement et annuellement des indicateurs grâce à l'enquête Emploi en continu (EEC). Cette enquête doit permettre d'obtenir rapidement des résultats précis au niveau Nuts 2 [1], afin de répondre à la réglementation européenne. Pour ce faire, la solution adoptée est celle de l'échantillonnage « aréolaire », tirage d'aires géographiques dans lesquelles sont interrogées un sixième des logements, pendant six trimestres consécutifs. Au septième trimestre, un autre sixième des logements sont interrogés, et ce, jusqu'à épuisement de l'aire. La méthode est décrite plus précisément au point 1.2.1. Par construction, l'échantillon est construit pour une durée de 9 ans (6 * 6 trimestres) ; le dernier échantillon, introduit en 2009, arrive progressivement à expiration à partir de 2019 [2].

Le renouvellement des échantillons pose la question de la base de sondage. La construction du Fichier DEmographique des Logements et des Individus (Fidéli) est apparue comme une opportunité. Ce fichier, produit par l'Insee depuis 2012 à partir des fichiers fiscaux, vérifie tous les critères requis pour une base de sondage de qualité. Ces atouts sont décrits en section 2.1. La concomitance du renouvellement de ces échantillons et l'apparition d'une nouvelle source, dans un contexte de recherche de rationalisation des moyens d'enquête ont amené à réfléchir à la coordination de ces deux échantillons. La section 2.2 introduit les gains envisagés et les critères utilisés pour comparer les scénarios mis en place et décrits dans la suite du document.

Ces différents travaux ont été réalisé dans le cadre d'un projet regroupant de nombreux acteurs et ont requis des compétences variées : le projet Nautile, Nouvelle Application Utilisée pour le Tirage des Individus et des Logements dans les Enquêtes, qui a démarré en 2017 (cf. chapitre 3).

1. Les Nuts 2 forment un découpage du territoire de l'Union Européenne, à un niveau infranational. En France, les Nuts 2 correspondent aux anciennes régions administratives en vigueur jusqu'au 31 décembre 2015.

2. Après prolongation ponctuelle des aires introduites en premier en 2009-2010.

Chapitre 1

Échantillon-maître et enquête Emploi en continu : le renouvellement des échantillons

1.1 Le renouvellement de l'échantillon-maître

1.1.1 Le principe de l'échantillon-maître

Afin de mener les enquêtes auprès des Ménages, l'Insee a recours à son propre réseau d'enquêteurs. Cela étant, celui-ci n'est pas suffisant pour couvrir l'ensemble du territoire. Depuis les années 60, l'institut définit un échantillon-maître (EM), ensemble de zones géographiques au sein desquelles seront réalisées les enquêtes pendant une période.

La définition de ces zones sur lesquelles les enquêteurs travaillent doit répondre à certains critères. Tout d'abord, l'ensemble des zones sélectionnées doivent représenter du mieux possible le territoire. Idéalement, les unités primaires (UP) sont assez hétérogènes en intra, afin de limiter l'effet de grappe issu de leur sélection. Au sein de l'EM, elles doivent être en nombre suffisant, et suffisamment bien réparties sur le territoire. Ensuite, chaque zone doit être de taille réduite pour permettre à un enquêteur de la parcourir en un temps raisonnable. Enfin, elles doivent être assez denses en nombre de logements pour pouvoir être utilisées pendant une période assez longue, sans avoir à réinterroger un même logement [1].

1.1.2 L'échantillon-maître de 2009

Les enquêtes auprès des ménages sont tirées, pour la plupart [2], dans le recensement de la population. Aussi, l'Insee renouvelait classiquement son échantillon-maître à chaque recensement de la population. À partir de 2004, celui-ci a été rénové et le recensement de la population (RRP pour Recensement Rénové de la Population) est devenu annuel et rotatif.

1. La réinterrogation de ménages enquêtés dans d'autres enquêtes augmente la charge de collecte auprès des ménages et les risques de lassitude des enquêtés. Ce sujet fait l'objet du chapitre 4 de la partie D.

2. Certaines enquêtes sont tirées dans les sources fiscales ou dans d'autres sources (DADS...).

Le principe du recensement rénové de la population

On rappelle rapidement le principe du RRP : contrairement à ce qui a été réalisé jusqu'en 1999, le nouveau recensement ne se fait pas sur la totalité de la population française, mais concerne une partie des ménages français regroupés dans un groupe de rotation. Ces groupes sont au nombre de cinq, et sont construits de la façon suivante :

- dans les petites communes (moins de 10 000 habitants), chaque ville est rattachée à un groupe de rotation et est recensée exhaustivement tous les cinq ans, au moment de l'Enquête Annuelle de Recensement (EAR) associée à son groupe ;
- dans les grandes communes, chaque adresse est rattaché à un groupe différent, et une fraction de ses adresses est recensée chaque année, de telle sorte à disposer en fin de chaque cycle de cinq ans d'environ 40 % de la population municipale recensée.

Les contraintes appliquées à l'EM issu du RRP

Les unités primaires constituées en 2009 [3] pour définir un nouvel échantillon-maître doivent respecter de nombreuses contraintes (cf. Christine et Faivre, 2009) :

- contenir suffisamment de logements de chacun des groupes de rotation du RRP, afin d'être mobilisables dans chaque millésime de base de sondage ;
- être de taille suffisamment grande (en nombre de logements) afin d'éviter d'interroger deux fois un même logement sur une période de 5 ans ;
- être les moins étendues possibles, afin de limiter les déplacements des enquêteurs,

La constitution d'unités primaires cohérentes et mobilisables chaque année, pour le tirage des enquêtes dans l'EAR la plus récente, demandait donc que soit prise en compte dans leur construction l'affectation des petites communes à chaque groupe de rotation, de telle sorte à disposer d'au moins une commune (en fait, d'au moins 300 logements) dans chaque groupe de rotation et donc dans chaque millésime de base de sondage. Les unités primaires de l'échantillon-maître de 2009 sont plus communément appelées ZAE (Zones d'Action Enquêteurs).

Le plan de sondage de l'échantillon-maître 2009

Pour respecter les contraintes ainsi énoncées, les caractéristiques du plan de sondage de l'échantillon-maître de 2009 sont les suivantes :

- Le nombre total de ZAE à tirer est calculé en fonction de plusieurs paramètres : le taux de sondage moyen pour un échantillon d'enquête ménage et le nombre moyen d'interrogations réalisées pour une enquête par un enquêteur.
- Le tirage des ZAE est stratifié par région administrative (NUTS2). Le nombre total de ZAE à tirer est réparti entre les régions proportionnellement à leur taille en nombre de résidences principales.

3. Les unités primaires ou zones d'action enquêteurs de l'ancien échantillon-maître ont été constituées en 2007. Le tirage de l'ancien échantillon-maître a également été effectué en 2007, mais cet échantillon-maître n'est entré en vigueur qu'à partir de 2009. Par souci de clarté dans ce document, nous parlerons toujours des unités primaires ou ZAE de 2009 et de l'échantillon-maître de 2009.

— Les probabilités d'inclusion des ZAE sont calculées proportionnellement à leur taille en nombre de résidences principales.
— Dans chaque région, les ZAE sont sélectionnées selon un tirage aléatoire équilibré sur un certain nombre de variables : le nombre de résidences principales, le revenu fiscal, le nombre de résidences principales dans les différents types d'espace (rural, périurbain et urbain). A noter que l'équilibrage est réalisé par groupe de rotation, afin que l'échantillon de ZAE soit précis quel que soit le groupe de rotation mobilisé.

Par cette méthode, 3785 regroupements de communes ont été constituées dans ce cadre [4], parmi lesquelles 567 ont été sélectionnées pour la collecte [5].

1.1.3 Pourquoi renouveler l'échantillon-maître 2009 ?

Au bout de 10 ans, l'échantillon-maître sélectionné ne représente plus de manière satisfaisante l'ensemble du territoire, c'est-à-dire que la structure de population dans les ZAE de l'échantillon-maître a trop évolué pour que le tirage réalisé 10 ans auparavant permette toujours des estimations précises pour les enquêtes auprès des ménages [6]. Par ailleurs, continuer à utiliser le même échantillon-maître rend probable la réinterrogation d'un logement déjà enquêté dans le passé, et augmente ainsi le risque de lassitude de l'enquêté. C'est pourquoi, il paraît utile de refaire un échantillon-maître.

Si l'utilisation de la même méthode paraît l'idée la plus évidente, elle présente toutefois des inconvénients non négligeables. D'une part, parce que les unités primaires construites en 2009 ne respectent plus obligatoirement les critères initiaux. De plus, parce que le déséquilibre progressif des groupes de rotation du recensement engendre des difficultés quant à la constitution d'un échantillon-maître représentatif chaque année, tout en conservant des unités primaires de taille raisonnable. Enfin, l'utilisation de l'échantillon-maître de 2009 a montré certaines limites.

Il est à noter que le statut de chaque ville (grande commune ou petite commune) peut évoluer avec le temps : c'est ce que l'on appelle les franchissements de seuil, lorsque qu'une commune acquiert plus de 10 000 habitants par exemple. Ces évolutions complexifient la gestion de l'échantillon-maître, car elles conduisent à des ZAE hybrides comportant à la fois des grandes et des petites communes, qui sont ensuite traitées spécifiquement. Chaque année, il est donc nécessaire de mettre à jour la composition communale des ZAE, ce qui représente un travail important et qui amène à des écarts par rapport au cadre

4. Chaque arrondissement de Paris, Lyon et Marseille comptant comme une ZAE différente.

5. Dont 79 ZAE exhaustives retenues d'office (communes de plus de 40 000 résidences principales au RP 1999) et 488 ZAE sélectionnées aléatoirement. Une ZAE est en effet retenue d'office dans l'échantillon de 1er degré lorsque sa probabilité d'inclusion est égale à 1. Ce cas concerne les ZAE les plus grosses, les probabilités d'inclusion étant proportionnelles à la taille.

6. Cela provient du fait que l'équilibrage réalisé 10 ans auparavant permet de disposer d'un échantillon-maître précis sur des variables corrélées aux variables d'équilibrage représentant une structure de population vieille de 10 ans. Si cette dernière évolue avec le temps, la corrélation entre les variables actuelles et les variables d'équilibrage ne permet plus à l'équilibrage d'apporter un gain de précision pour les estimations dans les enquêtes auprès des ménages. La notion de tirage équilibré est détaillée en section 1.2 de la partie C.

théorique optimal du tirage auto-pondéré à deux degrés.

Ces éléments ont conduit à envisager d'autres sources pour le tirage des échantillons des enquêtes auprès des ménages.

1.2 Le renouvellement de l'échantillon de l'enquête Emploi en continu

1.2.1 Principe de l'échantillon de l'enquête Emploi en continu

Un échantillon pour 9 ans

L'échantillon de l'EEC n'utilise pas le principe de l'échantillon-maître, car l'unité échantillonnée n'est pas directement le logement mais le secteur, divisé en 6 grappes d'environ 20 logements. Chaque trimestre, tous les logements d'une grappe de chaque secteur sont interrogés (chaque grappe est enquêtée 6 trimestres consécutifs [7]) puis remplacés par ceux d'une autre grappe du même secteur [8]. La durée de vie de l'échantillon est donc, par construction, de 36 trimestres, soit 9 ans. Entré progressivement en service entre 2009 et 2010, il est impératif de renouveler l'échantillon de l'EEC pour mi 2019, soit dans un calendrier voisin de celui de l'EM.

Un échantillon aréolaire et rotatif

Pour répondre aux exigences de précision et de calendrier d'Eurostat, la collecte de l'enquête Emploi en continu doit être menée sur une période très restreinte (2 semaines et 2 jours) auprès d'un échantillon important (environ 92 000 logements enquêtés par trimestre). Pour répondre à ces besoins, l'échantillon a été construit selon deux caractéristiques principales : il est aréolaire et rotatif.

Aréolaire car il est issu d'une sélection de groupes de logements contigus, appelés "grappes". La concentration géographique des logements à enquêter facilite la collecte dans les délais restreints. Cependant, cette contrainte crée évidemment un "effet de grappe" car les ménages d'un même quartier ont très souvent des caractéristiques socio-économiques proches et donc la précision de l'enquête en pâtit par rapport à une enquête qui serait plus dispersée géographiquement.

Rotatif car les logements d'une grappe donnée sont enquêtés six trimestres consécutifs puis remplacés par ceux d'une grappe voisine. L'objectif est d'utiliser cette proximité socio-économique des logements proches pour remplacer les ménages sortant du panel par d'autres aux caractéristiques semblables. Cette méthode permet de gagner en précision lors de l'estimation des évolutions, tout en atténuant le phénomène d'attrition que peut générer la lourdeur d'une enquête en panel. Cet arbitrage entre précision longitudinale et limitation de l'attrition a conduit à fixer historiquement à six trimestres la durée de vie du panel entrant pour l'enquête Emploi en continu.

7. Par ailleurs, chaque trimestre, un sixième de l'échantillon est renouvelé.

8. L'unité enquêtée de l'enquête Emploi est le logement. Ainsi, lorsqu'un ménage quitte son logement, c'est le ménage qui s'installera à sa suite dans le logement qui continuera d'être enquêté.

1.2.2 L'échantillon de 2009 de l'enquête Emploi en continu

La taxe d'habitation comme base de sondage

Le tirage de l'enquête Emploi a été effectué dans la taxe d'habitation 2006 [9]. Pour constituer la base de sondage à partir de cette source exhaustive, les 25 millions de résidences principales de France métropolitaine ont été regroupées en grappes de 20 logements par tri sur le numéro de la parcelle cadastrale, les logements d'un même étage devant appartenir à une même grappe. Les 1,2 millions de grappes obtenues, triées par parcelles puis sections cadastrales ont alors été regroupées par 6 en 200 000 secteurs.

Tirage équilibré et stratifié, allocations proportionnelles

Au sein de chaque région, un tirage équilibré de secteurs est mis en place [10], l'allocation étant fixée proportionnellement au nombre de résidences principales. Les variables d'équilibrage retenues sont, par ordre croissant d'importance :

- la répartition des résidences principales par type d'espace selon le Zonage en Aires Urbaines (1999) : pôles urbains, couronnes périurbaines, communes multipolarisées, communes rurales ;
- la répartition des résidences principales par quintile de revenus ;
- 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.

3217 secteurs ont ainsi été sélectionnés [11]. L'échantillon de 2009 [12] a été introduit pour un tiers entre le T1 2009 et le T2 2010 (couvrant donc jusqu'au T4 2017 et T1 2019 respectivement, mais qui ont fait l'objet d'une prolongation temporaire grâce à un septième groupe ad hoc) et pour les deux tiers entre le T3 2010 et le T4 2011 [13] (couvrant donc jusqu'au T2 2019 et T3 2020 respectivement).

9. Menée depuis 1950, l'Enquête Emploi a connu de nombreuses évolutions jusqu'à aujourd'hui (modification du questionnaire, passage à une enquête en continu, panélisation de l'échantillon...). En 2006, le recours au recensement de la population, devenu une enquête rotative à partir de 2004, avait été abandonné au profit des fichiers de la taxe d'habitation.

10. Le principe du tirage équilibré est expliqué en section 1.2 de la partie C.

11. Parmi une base de sondage qui comptait alors 1 200 000 grappes de 21,1 résidences principales regroupées en 199 500 secteurs de 120 logements (voir Loonis, 2009).

12. L'ancien échantillon Emploi a été tiré en 2008 dans les fichiers de la taxe d'habitation 2006. Comme il est entré en vigueur à partir de 2009, nous le nommerons par la suite « échantillon Emploi de 2009 » par abus de langage et renverrons à 2009 pour toute date relative à cet échantillon.

13. L'intégration s'est faite en deux temps afin d'augmenter la taille d'échantillon globale de l'enquête Emploi grâce au premier tiers de secteurs. L'augmentation de 50 % de la taille d'échantillon permet une amélioration de la précision des estimations.

Chapitre 2

Les opportunités du renouvellement de ces échantillons

Alors que l'échantillon-maître et l'échantillon de l'enquête Emploi en continu arrivent en fin de vie et que leur renouvellement s'impose, l'apparition d'une nouvelle source, Fidéli (Fichier démographique des logements et des individus) a ouvert de nouvelles opportunités.

2.1 Une nouvelle source : Fidéli

Fidéli est un fichier d'individus et de logements issu des fichiers fiscaux, apurés, complétés par le répertoire des communautés et celui des résidences hôtelières, et enrichis d'informations de géolocalisation (coordonnées, zonage) et d'informations sur les revenus (issues de Filosofi).

Ce fichier, mis à jour chaque année, présente ainsi les propriétés d'une bonne base de sondage : l'exhaustivité, l'unicité (absence de doublons), et la fraîcheur des données.

L'utilisation de Fidéli permet de s'affranchir des groupes de rotation du RP, et ainsi de diminuer l'étendue des zones d'enquêtes, de diminuer le nombre de contraintes sur le tirage de l'échantillon-maître et d'améliorer la qualité de l'échantillon des zones sélectionnées (et donc des futures enquêtes).

En outre, le tirage dans Fidéli permet d'échantillonner des personnes et pas seulement des logements comme avec le recensement de la population. Un atout important de la source est qu'elle permet de repérer toutes les résidences connues au titre de la taxe d'habitation, et pas seulement les résidences principales comme dans le recensement. Ceci pourra permettre à l'avenir d'élargir le champ de certaines enquêtes afin de sélectionner des populations jusqu'alors inatteignables.

Enfin, certaines enquêtes étant déjà sélectionnées au sein des fichiers fiscaux, Fidéli offre la possibilité de tirer toutes les enquêtes dans une unique base, facilitant ainsi la disjonction et évitant la sélection d'un même logement pour plusieurs enquêtes, y compris l'EEC.

Par ailleurs, Fidéli possède de nombreuses informations caractérisant les logements et les individus, permettant ainsi l'élaboration de plans de sondage de qualité. Enfin, les nombreuses informations de repérage présentes dans la source permettent d'assurer une identification précise de l'unité de collecte sélectionnée. En particulier, un certain nombre de coordonnées (mails, numéros de téléphone) sont disponibles dans cette base de données, ce qui rend cette source particulièrement attractive pour l'Insee dans un contexte de diversification des modes de contacts et de collecte.

Le FIchier DÉmographique sur les Logements et les Individus (Fidéli)

L'objectif de Fidéli est de valoriser les informations prioritairement issues de l'administration fiscale sur l'impôt et les propriétés bâties afin de disposer d'une meilleure connaissance du parc de logement et de la démographique résidente. Il se situe dans la perspective d'une valorisation accrue des données administratives.

La constitution de Fidéli

Fidéli se présente comme un assemblage raisonné de données administratives, et est conçu pour répondre à des finalités en matière de statistiques démographiques. Il regroupe :

- des données d'origine fiscale : fichier de la taxe d'habitation, fichier des propriétés bâties, fichiers d'imposition des personnes et fichier des déclarations de revenus. Ces données, assemblées au moyen d'identifiants fiscaux sur les foyers et les locaux, constituent le cœur de Fidéli ;
- des données complémentaires contextuelles visant à enrichir les informations fiscales avec des variables permettant de mieux décrire l'environnement : les coordonnées des parcelles cadastrales sur lesquelles sont bâties les habitations, les communautés présentes dans le répertoire des communautés utilisé pour le recensement, des informations externes permettant de repérer les adresses de domiciliation administrative, le répertoire des logements sociaux du SDES (Service de la donnée et des études statistiques, au sein du ministère de la transition écologique et solidaire) qui permet de repérer les logements sociaux dans Fidéli ainsi que leurs occupants ;
- des informations sur les revenus et les montants de prestations sociales reçues par les ménages issues du processus Filosofi, disponibles dans un second temps.

Les objectifs de Fidéli

Fidéli vise en premier lieu à élaborer des fichiers de diffusion disponibles sur le Centre d'accès sécurisé aux données (CASD) ou mis à disposition des partenaires du service statistique public (SSP). Ceux-ci ont pour ambition :

- de fournir aux chercheurs en sciences sociales une base de données permettant de réaliser des travaux spécifiques, portant sur des populations relativement rares. Fidéli fournit également une description originale des migrations résidentielles sur le territoire puisque les données renseignent sur la situation avant et après la migration.

— de se substituer au fichier du parc de logement (Filocom) utilisé par le SDES, en fournissant des informations beaucoup plus nombreuses sur les personnes, notamment le suivi géographique des personnes ainsi que la connaissance de leurs ressources. Par rapport à Filocom, l'information produite par Fidéli est donc à la fois plus riche, annuelle au lieu de biennale et géolocalisée ;

L'autre objectif prioritaire de Fidéli est la constitution de la base de sondage pour les futures enquêtes ménages.

Une information complète

L'information sur le logement précise l'année de construction, le statut fiscal d'occupation (propriétaire, locataire,...), le fait qu'il s'agit d'une résidence fiscale principale ou secondaire, d'un logement social, ainsi que différentes caractéristiques physiques : surface, nombre de salles de séjour, chambres, cuisines, salles de bain, présence d'un ascenseur, d'un garage ou d'un box...

L'information sur les personnes permet de connaître leur sexe, ainsi que leurs dates et lieux de naissance (ce dernier faisant l'objet d'une codification dans le processus de production de Fidéli), ceci pour les personnes majeures. Sont également disponibles le statut matrimonial fiscal (célibataire, marié, pacsé, divorcé, veuve) ainsi que l'existence d'un changement d'état matrimonial au cours de l'année. Pour les personnes mineures, seule l'année de naissance est connue.

Une information localisée

Tout le bâti provenant des sources fiscales, qu'il soit résidentiel ou non, est disponible dans Fidéli. Lui sont associés :

— des éléments de repérage comme l'adresse au cadastre (code Rivoli, numéro de rue), un code repérant le bâtiment et l'entrée dans le bâtiment, et l'adresse postale courante utilisée par l'administration fiscale pour les correspondances ;
— des informations complémentaires comme la nature du propriétaire (particulier, société, HLM), pour les particuliers le fait qu'il s'agisse d'une maison ou d'un immeuble, le nombre d'étages voire le nom d'usage du bâtiment ;
— Les références cadastrales, ce qui permet un géoréférencement de l'information (X,Y) à la parcelle et par voie de conséquence un repérage des zonages infracommunaux : IRIS et quartiers de la politique de la ville.

Des volumes proches de ceux du recensement de la population

En matière de nombre de logements, les effectifs sont assez proches de ceux du recensement. Des travaux conduits en 2009 et toujours d'actualité permettaient d'y dénombrer 33 390 000 logements en comparaison des 32 952 000 logements présents dans le recensement (soit un écart de 1,3%). L'écart s'explique en partie par un excès de logements vacants dans Fidéli (3 353 000 contre 2 290 000 au recensement). Une fraction de cet excédent est dû à des mises à jour tardives, dans les fichiers fiscaux, des logements démolis.

Fidéli comptait 1,7 % de moins de résidences principales au sens de la taxe d'habitation (27 081 000) que le recensement (27 534 000). Mais certains logements ordinaires et considérés comme des résidences principales dans le recensement ne sont pas soumis à la taxe d'habitation, même si les personnes concernées sont présentes dans les fichiers fiscaux. On peut penser aux foyers logements pour étudiants ou personnes âgées qui ressemblent à des communautés mais n'en sont pas au sens du recensement

Au niveau national, la population de Fidéli excède de très peu celle du bilan démographique (0,3% en 2017). Les écarts de population restent faibles pour les régions et les départements, sauf pour les départements de Guadeloupe, Martinique, Corse, et surtout Guyane et Mayotte.

De fait, les différences de populations entre Fidéli et le recensement demeurent modestes si l'on s'intéresse à des agrégats géographiques de 2000 habitants ou plus. Ainsi, en ce qui concerne les communes de 2 000 habitants et plus, les écarts de population sont inférieurs à 2% pour 51% d'entre elles et supérieurs à 5% pour 13%. Ces écarts pourraient en outre être sensiblement réduits par un repérage fin des communautés et des domiciliations administratives dans les fichiers fiscaux. En revanche, pour les communes de moins de 2000 habitants , des écarts plus importants semblent à ce jour peu réductibles, une partie d'entre eux étant explicables par des décalages temporels entre les deux sources.

Fidéli représente une opportunité pour les DOM au même titre que pour la Métropole (cf. encadré sur le tirage des enquêtes auprès des ménages dans les DOM ci-après). Néanmoins, les problématiques différentes dans les DOM et en Métropole ont conduit à mener les travaux détaillés dans ce document uniquement sur le champ de la Métropole.

Le tirage des enquêtes auprès des ménages dans les DOM

Le projet Nautile est une opportunité pour les DOM de converger vers le processus national. Les DOM en intégrant l'application Nautile bénéficieraient d'un tirage standardisé des échantillons et d'un suivi intégré dans l'application de la disjonction des échantillons. Ce projet s'accompagne d'une autre opportunité/contrainte, celle de passer des EARs (base incomplète chaque année) à Fidéli (base a priori exhaustive).

Il est important de noter que les problématiques associées à ce projet sont différentes entre les DOM et la métropole. En métropole, le renouvellement de l'échantillon-maître et les impacts sur le réseau d'enquêteurs est un point central dans le projet. Dans les DOM, la problématique centrale est plutôt axée autour de la possibilité d'utiliser Fidéli comme base de sondage pour l'ensemble ou non des 4 ou 5 DOM. Il s'agit en effet de s'assurer que la base Fidéli dans les DOM dispose de tous les pré-requis d'une base de

sondage : exhaustivité, représentativité, absence de doublons, fraîcheur, présence d'information auxiliaire et informations pour le repérage des logements. Pour les DOM, les trois points de vigilance portent actuellement sur l'exhaustivité et la représentativité de la base, notamment en Guadeloupe et plus principalement en Guyane mais aussi sur les éventuelles difficultés de repérage associées au passage à Fidéli.

Pour que tout ou partie des échantillons domiens intègrent l'application Nautile (et donc un tirage dans Fidéli), il est nécessaire de vérifier que la base de sondage Fidéli mobilisée par cette nouvelle application Nautile soit de qualité suffisante dans les DOM. La notion de qualité suffisante peut être variable suivant le type d'échantillon à tirer. En effet, on n'exige pas le même degré de couverture de la base de sondage pour une enquête à représentativité nationale que pour une enquête à extension ayant vocation à fournir des statistiques fiables au niveau du DOM. Une des premières décisions a donc été (compte tenu des taux de couverture relativement corrects de Fidéli) de valider le tirage d'échantillons dans Fidéli pour toutes les enquêtes à représentativité nationale : la part des DOM dans l'échantillon national étant relativement faible (environ 2%), les légers défauts de couverture n'ont que peu d'incidence sur le calcul d'un agrégat national. Le passage à Fidéli pour les enquêtes nationales ou les enquêtes avec des faibles volumes comme l'enquête Loyers et Charges a nécessité la création dans chaque DOM historique d'un nouveau zonage d'enquête plus fin (appelé sous-SAE) que le zonage historique (basé sur des SAE : « Secteur Action Enquêteur »). Ces sous-SAE constituent un jeu d'unités primaires mobilisables lors d'un tirage de premier degré, permettant ainsi de réduire le périmètre géographique de l'enquête dans chaque DOM. La possibilité de mobiliser un zonage plus fin est intrinsèquement lié à l'exhaustivité de la base Fidéli et donc au volume disponible dans celle-ci. La construction d'un zonage d'enquête plus fin, de sous-SAE, n'aurait pas été envisageable à partir des volumes fournis par les enquêtes annuelles de recensement.

Même si le cas des enquêtes à représentativité nationale et de l'enquête Loyers et Charges sont traités, le cas des enquêtes à extension est toujours en suspens : des travaux complémentaires ou des expérimentations sont en cours sur les aspects déjà mentionnés précédemment : le repérage, l'exhaustivité et la représentativité de la base de sondage. Dans l'attente des conclusions, la base préférentielle pour le tirage des enquêtes à extension demeurent dans les 4 DOM historiques les Enquêtes Annuelles de Recensement. De ce fait, la génération des scans du recensement (appelés également BALS) et le maintien de la fonctionnalité de l'applicatif OCTOPUSSE qui permet les impressions des fiches adresses pour les enquêtes tirées dans le recensement demeurent indispensables pour la réalisation des enquêtes à extension dans les 4 DOM historiques.

Le cas du 5ème DOM, Mayotte, est très spécifique car non seulement aucune enquête n'a été réalisée à partir d'un tirage dans les sources fiscales mais aussi parce que la couverture de Mayotte par Fidéli est très faible : le taux de couverture de Fidéli 2017 en termes d'individus par rapport au recensement exhaustif de 2017 est de 85 % alors

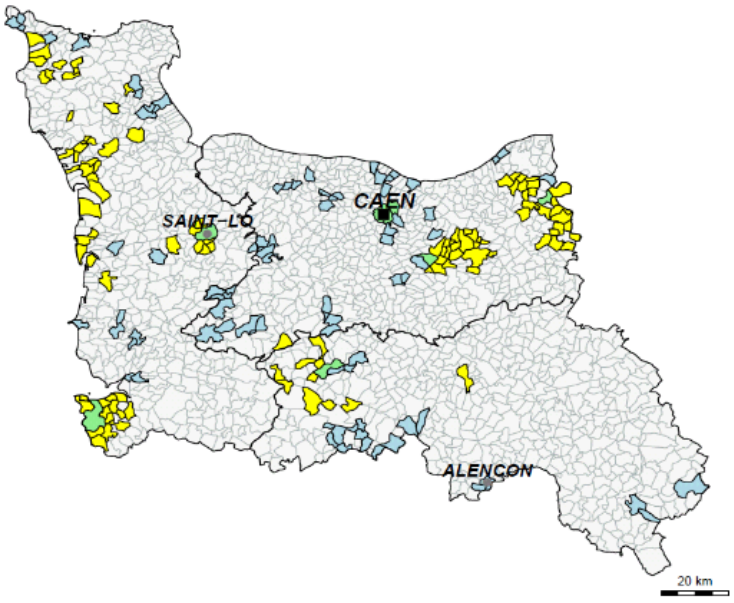
que le taux de couverture des résidences principales de Fidéli 2017 par rapport au recensement de 2017 est de l'ordre de 27 %. Ce taux de couverture très faible en nombre de résidences principales s'explique par la forte présence de bâti en tôle (appelé "bangas") qui représentent près d'un tiers du bâti à Mayotte et qui ne sont pas présents à la TH. Le tirage de résidences principales comme pour la Métropole ou comme pour les 4 DOM historiques n'est pas envisageable à Mayotte avec un taux de couverture aussi bas.

2.2 Les gains de la coordination des tirages d'échantillon

L'enjeux de la coordination : limiter le territoire couvert

Les renouvellements de l'échantillon-maître et de l'échantillon de l'enquête Emploi en continu suivant un calendrier rapproché et n'ayant que quelques trimestres d'écart, la coordination des tirages a été envisagée. Concrètement, la coordination des tirages se traduirait par l'implémentation d'une méthode permettant le tirage de secteurs de l'échantillon Emploi à proximité des unités primaires de l'échantillon-maître. Cela permettrait de limiter les déplacements des enquêteurs, d'éviter les secteurs isolés et de faciliter les remplacements en augmentant la charge sur une zone. Avant le renouvellement, la répartition des secteurs de l'EEC pouvait être dispersée sur l'ensemble d'une région, et éloignée des zones d'activité des enquêtes des autres enquêtes auprès des ménages, comme l'illustre la figure 2.1.

FIGURE 2.1 – Exemple de répartition de l'échantillon-maître et de l'échantillon Emploi dans l'ancienne région Basse-Normandie en l'absence de coordination



Note de lecture : les communes en jaune sont celles concernées par l'échantillon-maître de 2009 uniquement, les communes en bleu sont celles concernées par l'échantillon Emploi de 2009 uniquement, tandis que les communes en vert sont celles mobilisées pour les deux échantillons. Les communes grisées ne figurent dans aucun des deux échantillons.

En fait, la coordination des tirages de l'échantillon-maître et de l'échantillon Emploi relève des mêmes enjeux que le tirage de l'échantillon-maître pour les enquêtes auprès des ménages. L'objectif est de limiter à certaines zones géographiques représentant l'ensemble du territoire les déplacements des enquêteurs. Il s'agit ainsi de sélectionner un sur-échantillon-maître dans lequel seront tirées d'une part les enquêtes auprès des ménages, d'autre part les secteurs de l'EEC.

Une précision des indicateurs à respecter

Cependant, limiter le territoire à un échantillon de celui-ci accentue obligatoirement le risque d'effet de grappe, c'est-à-dire une perte de précision dû à un manque d'hétérogénéité au sein de l'échantillon retenu. Or, les enquêtes auprès des ménages et l'enquête Emploi doivent répondre à un certain nombre d'exigences réglementaires européennes de qualité. Le règlement IESS (Integrated European Social Statistics), adopté en 2019, fixe pour les prochaines années un cadre à respecter pour différentes enquêtes auprès des ménages, notamment pour l'enquête Emploi et d'autres enquêtes auprès des ménages (Statistiques sur les Revenus et les Conditions de Vie, Adult Education Survey, Budget de Famille...). Entre autres, le règlement impose des contraintes de précision aux instituts nationaux statistiques sur des indicateurs calculés à partir des enquêtes. Le taux de chômage national, le taux de chômage dans chaque région, la persistance dans la pauvreté monétaire au niveau national, le risque de pauvreté monétaire ou d'exclusion sociale au niveau régional, etc. sont autant d'indicateurs qui doivent être fournis par l'Insee avec une précision fixée par le règlement IESS. Les tirages de l'échantillon Emploi et de l'échantillon-maître doivent donc permettre l'estimation de ces indicateurs avec la précision requise par les normes européennes.

Des critères pour évaluer les scénarios de coordination

Les zones construites et les échantillons tirés doivent ainsi avoir une qualité statistique, être durables et permettre de faciliter la collecte des enquêteurs en limitant les déplacements. Pour évaluer la qualité des échantillons, nous retenons donc 6 critères :

1. la durée de vie de l'échantillon-maître : le réseau d'enquêteurs de l'Insee se constitue, au fil du temps, autour de ces zones. Un échantillon-maître est rentable budgétairement s'il dure suffisamment longtemps. La durée de 10 ans est communément retenue par l'Insee. Les tailles des UP doivent donc être suffisantes en termes de nombre de logements afin de pouvoir être exploitées sur cette durée pour les tirages des échantillons.
2. les précisions nationale et régionales de l'échantillon-maître, pour respecter les contraintes de précision imposées par le règlement IESS ;
3. la précision nationale et régionale de l'échantillon de l'enquête Emploi en continu pour respecter ces mêmes contraintes ;
4. l'étendue des unités primaires pour limiter les déplacements des enquêteurs ;
5. l'étendue des grappes et des secteurs Emploi et leur homogénéité en nombre de logements pour faciliter la collecte dans des délais très courts, et limiter la dispersion des poids de sondage ;
6. la proximité des secteurs Emploi et des unités primaires tirés pour améliorer la coordination des échantillons et ainsi faciliter le travail des enquêteurs.

Tous les scénarios de tirage présentés dans la suite du document de travail ont été étudiés à l'aune de ces différents critères. Les unités primaires, unités de coordination et secteurs présentés dans la partie B ont été construits afin de répondre à l'exigence de durabilité et de limitation de l'étendue.

Quelques critères spécifiques à l'enquête Emploi pour la construction des grappes et des secteurs (gestion des immeubles, des étages) sont également à respecter ; ils sont détaillés dans le point B 2.1.2.

Chapitre 3

Nautile : un projet coordonné

La définition des échantillons pour les enquêtes auprès des ménages de l'Insee a nécessité l'implication de nombreux acteurs.

D'une part, parce que cette opération s'organise autour de deux projets, l'un méthodologique, pour l'élaboration de la méthode de sélection de l'échantillon-maître des enquêtes auprès des ménages, et de l'échantillon de l'EEC, l'autre informatique pour la réalisation d'une application permettant de produire de manière sécurisée et selon une méthodologie rigoureuse les échantillons finalement enquêtés sur le terrain.

D'autre part, parce que la répartition des enquêteurs de l'Insee sur le terrain, et la précision des résultats des enquêtes dépendent des échantillons issus de ces méthodes. Aussi est-il primordial d'envisager des procédures de validation des échantillons produits à de nombreux niveaux (experts méthodologiques, acteurs locaux, responsables d'enquêtes...)

L'ensemble de ces travaux, débutés en 2016 et finalisés en 2020 par le premier tirage dans le nouvel échantillon-maître effectué grâce à l'application développée ont ainsi été menés au sein d'un projet : La Nouvelle Application Utilisée pour le Tirage des Individus et des Logements dans les Enquêtes, Nautile.

3.1 Deux projets pour une équipe élargie

Dès 2016, l'Insee s'est doté d'un chef de projet statistique pour le suivi des études méthodologiques et la spécification de l'application de tirage. La volonté de comparer différentes solutions aux enjeux méthodologiques pour assurer une méthode de qualité a conduit à impliquer plusieurs équipes d'experts méthodologiques, et ce, dès la constitution des unités primaires. Aussi, pour répondre à cette problématique, l'institut s'est appuyé sur ses compétences en sondage et en construction de données géolocalisées pour élaborer différentes méthodes pour la définition de ces zones, puis pour choisir la plus adaptée au contexte particulier de l'Insee.

La plupart des autres études présentées dans ce document (méthode de tirage d'unités primaires, élaboration des allocations, constitution de la base de sondage de l'enquête Emploi...) ont été menées au sein de la Division Sondage, puis validées par d'autres services ayant des compétences méthodologiques et métiers.

L'équipe s'est spontanément organisée selon certains principes Agile autour d'un binôme « chef de projet statistique / responsable de la section échantillonnage » et s'apparentant au binôme *product owner* / *scrum master*. Le chef de projet a priorisé les études méthodologiques en fonction des moyens disponibles (*product owner*) et de leur importance pour la production des différents échantillons, pendant que le responsable de la section échantillonnage a veillé à l'efficacité de l'équipe et à la bonne réalisation des études décidées [1].

Cette organisation construite sur des moyens de structure a permis de produire un maximum de scénarios de constitution des zones et de scénarios de tirage en des temps très courts, selon une priorisation assurée au quotidien, et validée par la hiérarchie de l'institut. La diversité des agents composant l'équipe (experts sondage, producteurs d'échantillons, data scientists) s'est présentée comme une aubaine pour l'exercice demandé et a entraîné, lors de brainstormings fréquents, la mise en commun de réflexions permettant de lever les obstacles rencontrés. Cependant, les travaux de production courante de la Division Sondage que l'équipe devait mener en parallèle, et le temps contraint pour le renouvellement de l'échantillon-maître et de celui de l'EEC n'ont pas permis de réaliser l'ensemble des études envisagées initialement. Cela a rendu la priorisation d'autant plus capitale. L'ensemble des travaux méthodologiques ont été réalisés entre mi-2016 et mi-2018. L'équipe au complet, ainsi décrite, a été mise à contribution entre octobre 2017 et avril 2018.

À partir de 2018, l'équipe s'est recentrée sur deux chefs de projet statistique pour la réalisation des spécifications, et une équipe de quatre développeurs pour la réalisation de l'application permettant le tirage des logements et des individus pour les enquêtes. Cette équipe a également suivi la méthode *scrum* pour réaliser le produit, selon les priorités données par le propriétaire de l'application. Cela a permis la production d'un premier échantillon dans l'application début 2020.

3.2 Une validation partagée

Si la réalisation des études méthodologique a reposé sur une équipe réduite, du Département des Méthodes Statistiques, la validation a, quant à elle, fait l'objet d'une large participation afin d'assurer à tous les niveaux la qualité des échantillons produits.

1. Dans la définition de la méthode *scrum* du guide *scrum*, le *product owner* est responsable de maximiser la valeur du produit résultant du travail de l'équipe *scrum*. Le *scrum master* est responsable de l'efficacité de l'équipe en lui permettant d'améliorer ses pratiques. La méthode *scrum* est un framework Agile qui aide les personnes, les équipes et les organisations à générer de la valeur grâce à des solutions adaptatives à des problèmes complexes. Voir `https://scrumguides.org`.

Ainsi, un groupe d'experts méthodologiques, composé notamment des producteurs des précédents échantillons-maîtres, a apporté une caution scientifique aux différentes solutions proposées. Les différentes méthodes établies ont été présentées à ce groupe, qui les a discutées afin de les confronter à leur expérience, de faire ressortir les qualités et les défauts de chacune, et ainsi d'émettre un avis sur celles apportant, à leurs yeux d'experts, le plus de certitude.

La qualité d'un échantillon se juge également à sa capacité d'être appréhendé sur le terrain par les enquêteurs, et utilisé avec justesse. Aussi, des représentants des agents en Directions Régionales ont été intégrés dans des groupes de travail, permettant de valider les options prises pour la constitution des zones, pour l'élaboration d'outils de contrôle, ou pour fixer les limites acceptables lors de la confrontation des échantillons au terrain (distance entre deux logements enquêtés, étendue d'une zone...). Plus largement, chaque Direction Régionale a été sollicitée à plusieurs étapes de ce travail (validation de la base d'unités primaires avant tirage de l'échantillon-maître, vérification des grappes de l'EEC, validation de la capacité à prendre en charge les échantillons finalement tirés), afin de garantir une caution pratique, en plus de la validation méthodologique.

Enfin, pour s'assurer que les solutions envisagées s'adaptent aux protocoles d'enquêtes, et que les résultats soient suffisamment précis pour l'exploitation des enquêtes, les producteurs des indicateurs résultant des enquêtes ménages, et particulièrement de l'enquête Emploi ont également été associés tout au long des études méthodologiques. Ils ont ainsi émis leur avis sur les choix pour la constitution des unités de tirage, pour les méthodes de sélection adoptées, et pour les allocations proposées.

À chaque étape de ces études méthodologiques, de nombreux avis ont été sollicités, et des validations demandées à des niveaux aussi bien pratiques que théoriques, afin d'assurer le bien fondé des choix, et la capacité à utiliser les échantillons résultant aussi bien par les enquêteurs que par les statisticiens. Ces nombreuses expertises ont permis au Comité de Direction de l'Insee de valider les échantillons finalement produits en juin 2018.

3.3 Calendrier

La fin annoncée de l'utilisation de l'échantillon de l'EEC et de l'échantillon-maître de 2009 a conduit à entamer les échanges sur les objectifs attendus pour leur renouvellement dès 2015. Mais c'est en 2016 que les premières études méthodologiques ont été entreprises, pour la constitution des unités primaires de l'échantillon-maître. La validation de cette base de sondage d'unités par les Directions Régionales est intervenue à l'été 2017. Cela a marqué la mise en place de l'équipe qui s'est saisie des études méthodologiques restantes et le lancement des réflexions pour la constitution des grappes pour l'EEC et la coordination des tirages. Cette équipe a mené ces opérations jusqu'au printemps 2018, ouvrant la voie à la validation des choix par les acteurs concernés (experts méthodologiques, enquêteurs, gestionnaires d'enquête...) pour une validation officielle en juin 2018,

permettant, dans la foulée, l'entrée sur le terrain de l'échantillon de l'EEC.

En parallèle, la liste des fonctionnalités attendue pour la future application, a été établie, et sa réalisation a commencé en octobre 2018, pour permettre une mise à disposition d'une première version de l'application au printemps 2020, avec le tirage de l'échantillon de l'enquête Camme (enquête mensuelle de conjoncture auprès des ménages) en mars 2020.

Partie B

Constitution des unités de tirage

En partie A, la description de la source Fidéli a montré les avantages que celle-ci présentait comparativement aux enquêtes annuelles de Recensement pour constituer une base de sondage des enquêtes auprès des ménages. Or l'utilisation des EAR avait imposé une construction particulière des zones de collecte, à partir des 5 groupes de rotation du RP. Le changement de base de sondage a permis de revoir cette partition du territoire (section 1.1), en apportant une meilleure réponse aux objectifs présentés en 1.2.1 : obtenir des UP géographiques constituées d'une ou de plusieurs communes, contenant un nombre minimum de logements et d'étendues les plus réduites possibles. Pour atteindre ces objectifs, la méthodologie retenue de constitution des UP, décrite dans les points 1.2.3 à 1.2.5, repose sur une application de l'algorithme du voyageur de commerce au sein des communes de chaque département. Elle a permis de réduire la taille des UP d'en moyenne 25% par rapport aux ZAE (section 1.3).

La méthode de tirage de l'enquête Emploi en continu implique la constitution d'une base de sondage de secteurs, comme ensemble de 6 ou 7 grappes, elles-même constituées d'une vingtaine de résidences principales (section 2.1). L'évolution du parc de logements et de la catégorie de ceux-ci impose également la reconstruction de cette base. Les sections 2.2 et 2.3 présentent les choix méthodologiques retenus pour la constitution respective des grappes et des secteurs afin de respecter plusieurs contraintes de collecte (20 résidences principales par grappe, étendue la plus réduite possible au sein de grappes et au sein de secteurs, étages indissociables...). L'utilisation de l'algorithme du voyageur de commerce entre des logements adjacents d'une même commune (ou d'un même Iris) permet de répondre à ces contraintes d'un point de vue pratique. Il en résulte, d'une part, une diminution de la dispersion du nombre de logements par grappe et, d'autre part, une baisse de l'étendue moyenne des grappes, comparativement aux grappes constituées pour l'échantillon Emploi de 2009 (section 2.4).

La coordination des deux tirages a également conduit à réfléchir à la création d'une nouvelle partition du territoire afin de sélectionner les zones dans lesquelles seraient tirées l'EEC et l'EM pour les autres enquêtes auprès des ménages (section 3.1). Pour la création de ces zones, appelées « unités de coordination » (UC), l'objectif reste le même : avoir des zones suffisamment grandes en nombre de logements pour garantir la précision des enquêtes et pour assurer la pérennité du système sur 10 ans, ces zones devant toutefois être suffisamment restreintes en superficie afin de limiter les déplacement des enquêteurs (section 3.2). Par ailleurs, ces zones ont été créées de telle sorte que chaque UP et chaque secteur appartiennent à une et une seule UC. Le chemin solution du problème du voyageur de commerce qui a permis de constituer les UP a été utilisé pour la construction de ces UC (section 3.3).

Chapitre 1

La constitution des nouvelles unités primaires

Pour déterminer un échantillon-maître, il est nécessaire, en premier lieu, de constituer la base de sondage, c'est-à-dire en l'occurrence de définir les unités primaires, partition géographique du territoire français. Les caractéristiques souhaitées des unités primaires, dites UP par la suite, sont les suivantes :

- les UP sont composées d'au moins 2 500 résidences principales [1], cette taille minimale devant permettre de ne pas réinterroger le même logement [2] plusieurs fois sur une période estimée à 10 ans ;
- les UP sont d'étendue minimale afin de limiter les déplacements des enquêteurs ;
- idéalement, les UP sont assez hétérogènes en intra, afin de limiter l'effet de grappe issu de leur sélection.

1.1 Pourquoi refaire les unités primaires ?

Les unités primaires de 2009, appelés Zone d'Action Enquêteur (ZAE) sont des groupes de communes constitués de telle sorte à minimiser leur étendue (Christine et Faivre, 2009). Cependant, d'autres contraintes entraient en jeu dans leur constitution, en particulier le fait de pouvoir disposer d'un nombre minimal de logements dans chaque groupe de rotation du recensement de la population.

Le relâchement de la contrainte liée aux groupes de rotation permet d'envisager de nouvelles zones répondant aux objectifs de taille en nombre de logements et de superficie plus limitée. Plusieurs briques ont alors pu être envisagées pour la construction des nouvelles unités primaires :

1. Par abus de langage, il arrive que le terme « logement » soit utilisé pour « résidence principale » dans la suite du document. La construction des différentes unités s'effectue bel et bien sur des critères liés au nombre de résidences principales, qu'il s'agisse des unités primaires, des secteurs emploi ou des unités de coordination. Le concept de logement est plus large que celui de résidence principale, puisqu'il inclut également les résidences secondaires et les logements vacants. Dans la suite du document, lorsque le terme de « logement » est utilisé pour désigner les résidences principales, les résidences secondaires et les logements vacants, cela est précisé.

2. L'objectif de ne pas réinterroger un même ménage est mis en oeuvre en pratique en évitant de tirer à nouveau le logement qui avait conduit à l'interrogation de ce ménage au cours d'une enquête passée. Des risques de réinterrogation existent donc pour les individus en cas de déménagement.

— La brique communale, dont l'intérêt est avant tout la continuité avec l'existant dans la gestion de l'échantillon-maître, en termes de processus, en termes statistiques et en termes de communication avec les différents acteurs, notamment les enquêteurs. Elle permet également de respecter les différents niveaux de diffusion demandés du fait de l'imbrication des niveaux administratifs. Cependant, la taille des communes varie ; les zones construites ne peuvent être homogènes en nombre de logements ;
— La brique carroyée, c'est-à-dire basée sur un découpage en rectangles constitués à partir des coordonnées cartographiques des logements et permettant de constituer des UP de taille homogène, au prix, dans certains cas d'une moindre cohérence avec les limites naturelles (rivières, etc.). À noter que d'autres pays tels que le Portugal l'étudient également (Schoenmakers et Santos, 2013)

Pour des raisons de simplicité des processus, la première solution a été retenue. Seule la réalisation de celle-ci est décrite dans la suite de ce paragraphe.

1.2 La méthode de construction des nouvelles UP

1.2.1 Les objectifs : éviter les réinterrogations et minimiser les étendues

On souhaite déterminer une méthode de construction des unités primaires, c'est-à-dire de regroupement de communes permettant de constituer un découpage du territoire en UP qui contiennent suffisamment de résidences principales pour éviter les réinterrogations durant la durée de vie du cycle de l'échantillon-maître Nautile. Chaque année, environ 125 000 ménages sont interrogés par l'Insee en face-à-face (hors EEC). Le réseau d'enquêteurs couvrant environ 500 zones, et l'échantillon-maître étant programmé pour une dizaine d'années, un seuil minimal de 2 500 résidences principales par unité primaire a été choisi [3]. À partir de là, la problématique est de minimiser l'étendue géographique de ces unités primaires afin de réduire les temps de déplacement des enquêteurs entre leur domicile et chaque logement de leur unité primaire d'affectation.

Deux éléments sont à préciser à ce stade :
— pour optimiser le regroupement des communes entre elles, il a semblé plus facile de considérer celles-ci comme des points et non comme des polygones. Les coordonnées associées aux communes sont le barycentre des coordonnées des adresses pondérées par le nombre de logements à l'adresse.
— afin de se rapprocher de l'indicateur d'étendue utile pour le travail d'enquête, la matrice de distance utilisée est la matrice de temps de trajet par la route entre toutes les communes d'un département. Cette distance permet ainsi de prendre en compte les obstacles naturels ainsi que la présence d'infrastructures spécifiques (autoroutes, ponts...) entre les communes [4]

3. 1 250 000 ménages sont enquêtés en 10 ans. En divisant par le nombre de zones - environ 500 - on obtient ce seuil de 2 500 résidences principales.

4. En revanche, on ne dispose pas de distances entre tous les logements, car cela demanderait des matrices de taille trop importante.

1.2.2 Trouver la partition optimale : un problème à réduire

Pour sélectionner le regroupement de communes en UP optimal, il faudrait tester l'ensemble des associations possibles. Or tester l'ensemble des partitions d'un ensemble à k éléments distincts n'est pas envisageable d'un point de vue computationel. Le cardinal de cet ensemble correspond au k-ième nombre de Bell. Or le 50^e nombre de Bell est déjà de l'ordre de 10^{47} ; c'est à dire qu'il est totalement impossible de résoudre de manière exhaustive ce problème dès lors que la zone considérée (la région, le département...) contient plus de quelques dizaines de communes. Une solution pourrait être de travailler à faire ces regroupements sur des zones géographiques plus petites : par exemple, en séparant les zones urbaines, péri-urbaines et rurales d'un département, ou au niveau du canton. Cependant, on constate que travailler à ces niveaux restreint énormément les degrés de liberté de la constitution des unités primaires, et aboutit à la constitution d'UP atypiques, de taille ou de forme non souhaitable.

1.2.3 Solution à l'algorithme du voyageur de commerce appliquée aux départements

Il est donc nécessaire d'utiliser une méthode approchée pour résoudre ce problème. Nous avons choisi, après exploration de différentes alternatives, d'utiliser des algorithmes solutions du problème du voyageur de commerce (voir, par exemple Applegate et al., 2003) pour construire un chemin que l'on pourra parcourir pour constituer les UP [5]. On rappelle que le problème du voyageur de commerce correspond à la situation d'un vendeur ambulant qui doit vendre sa marchandise dans k communes différentes, et qui se demande comment optimiser son temps de trajet pour passer dans toutes les villes. Ce problème est complexe à résoudre algorithmiquement, mais il existe de nombreuses solutions approchées, facilement mobilisables [6].

Par ailleurs, la construction des unités primaires a été mise en oeuvre au sein de chacun des départements français. D'autres niveaux auraient pu être considérés (national, régional), mais le niveau départemental présente deux avantages. D'une part, il permet de limiter les temps de calcul de l'algorithme de construction des UP. D'autre part, chaque unité primaire se trouve ainsi attribuée à une seule direction régionale de l'Insee, ce qui facilite l'affectation aux enquêteurs, le réseau d'enquêteurs étant géré au niveau régional, voire départemental pour l'Île de France.

1.2.4 Description de l'algorithme de découpage du territoire en UP

Ainsi, l'utilisation de cet algorithme à partir de la matrice de distances entre les communes d'un même département a permis d'ordonner ces communes en fonction d'une proximité géographique, et ainsi de dessiner un chemin. Nous créons alors un jeu d'unités primaires de la façon suivante : à partir du point de départ de l'algorithme, on parcourt les communes situées sur le chemin proposé jusqu'à respecter la taille minimale souhaitée

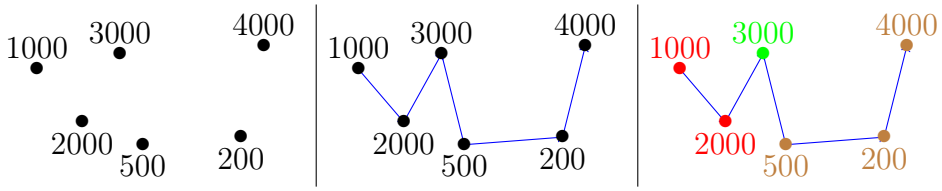
5. Pour tous les chemins calculés, on recourt à des chemins fermés, c'est-à-dire revenant au point de départ.

6. Ici, le package R `TSP` (Hahsler et Hornik, 2007) a été utilisé.

en termes de résidences principales (soit 2 500 résidences principales). On continue ensuite à parcourir le chemin pour construire les unités primaires suivantes.

Le schéma en figure 1.1 illustre la méthode de constitution : les six communes, représentées par des points en indiquant le nombre de résidences principales, sont reliées par un chemin solution du problème du voyageur de commerce, en bleu, puis séparées en suivant ce chemin en trois unités primaires respectant la contrainte de taille minimale de 2 500 résidences principales, de couleurs rouge, verte et marron.

FIGURE 1.1 – Méthode de découpage en Unités Primaires



La figure 1.2 montre un exemple de constitution des unités primaires pour le département du Morbihan. Le chemin solution du problème du voyageur de commerce est construit, puis découpé afin de construire des unités d'au moins 2 500 résidences principales les moins étendues possibles.

1.2.5 Critère pour le choix de la partition sélectionnée

La solution obtenue dépend du point de départ adopté. Afin de limiter cette dépendance, et ainsi d'obtenir une solution plus optimale, on effectue 1 000 réalisations [7] de l'algorithme du voyageur de commerce [8] par département [9], puis on choisit le jeu d'unités primaires du département qui minimise l'étendue moyenne [10], au sens de la distance de trajet nécessaire pour atteindre toutes les communes de l'unité primaire à partir de la plus grande commune de l'unité primaire [11], pondéré par la taille de ces communes : on utilise cet indicateur car, dans un échantillon d'enquête, une toute petite commune sera moins souvent présente qu'une grande commune de l'unité primaire, à proportion de leur population municipale. Cet indicateur est ainsi défini par la formule suivante pour une unité primaire *up* composée des communes *com*.

$$\sum_{com \in up} \frac{N_{com}}{\sum_{com \in up} N_{com}} D(com, princ) \tag{1.1}$$

7. Pour chaque réalisation départementale, on sélectionne aléatoirement la commune-point de départ et on met en oeuvre l'algorithme.

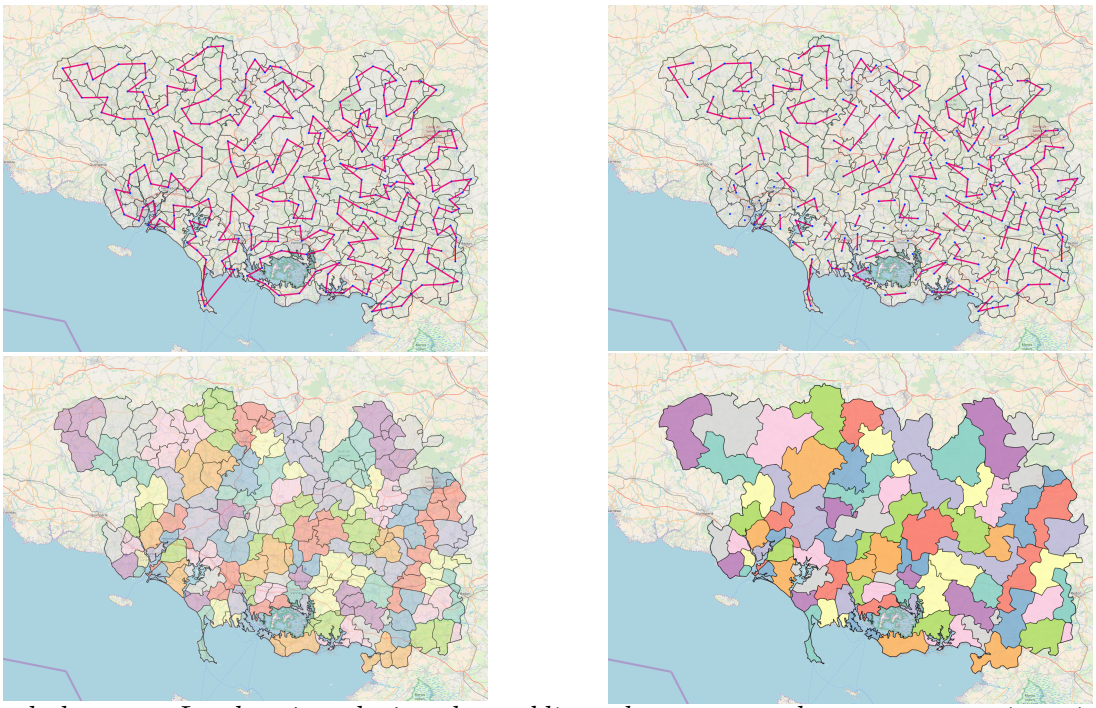
8. Notons que la construction des chemins au niveau départemental donne de meilleurs résultats qu'au niveau régional, car le fait de se limiter à 1 000 réalisations conduit à trouver plus facilement une solution de qualité en limitant le nombre de points de départ de l'univers.

9. Bien que les départements comptent tous moins de 1 000 communes, il ne suffit pas de tester chaque commune comme point de départ, car ce n'est pas la version exacte de la résolution du problème du voyageur de commerce qui a été utilisée. Avec l'algorithme de résolution retenu, on peut obtenir plusieurs chemins non optimaux associés à un même point de départ.

10. C'est-à-dire la moyenne de l'étendue des UP du département découlant d'un chemin solution du problème du voyageur de commerce. L'étendue d'une UP est donnée par la formule 1.1.

11. La plus grande commune de l'unité primaire est ici celle ayant le plus de résidences principales.

FIGURE 1.2 – Constitution des Unités Primaires dans le Morbihan (56)



Note de lecture : Le chemin solution du problème du voyageur de commerce présenté dans le cadre en haut à gauche est celui qui minimise l'étendue des unités primaires. Il passe par l'ensemble des communes du Morbihan. Dans le cadre en haut à droite, on parcourt le chemin en le découpant dès que la somme des résidences principales des communes regroupées dépasse 2 500 résidences principales. Cela permet de constituer des agrégats de communes qui forment des unités primaires dans le cadre en bas à gauche. Les unités primaires finalement retenues sont présentées dans le cadre en bas à droite.

La commune *princ* est la commune de l'UP *up* ayant le plus grand nombre de résidences principales. N_{com} décrit le nombre de résidences principales de la commune *com*. $D(com, princ)$ est la distance de trajet entre la commune *com* et la commune *princ*.

1.3 Description de la partition en unités primaires

Les 5 128 unités primaires ainsi obtenues (figure 1.3) forment une partition du territoire métropolitain. Elles ont alors été comparées aux ZAE de 2009. Comme attendu, la forte diminution de l'étendue moyenne des UP est confirmée pour chaque région, comme le montre le tableau 1.1.

Cette partition a été proposée pour validation aux Directions Régionales de l'Insee, afin que celles-ci puissent modifier le contour de certaines UP, en fonction de contraintes locales non modélisables dans les simulations (nouvel aménagement routier, bouchons fréquents...), tout en respectant la limite de 2500 résidences principales par UP.

FIGURE 1.3 – Les 5 128 nouvelles Unités Primaires

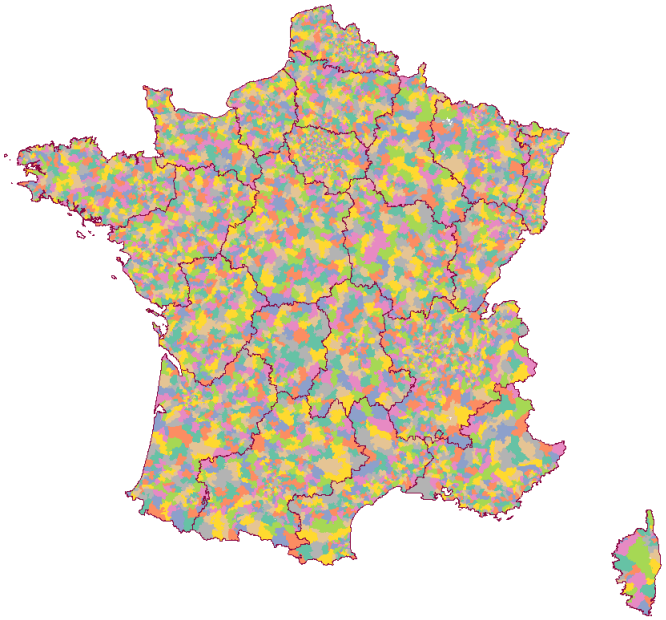


TABLE 1.1 – Étendues des nouvelles Unités Primaires

Code	DR	Nombre d'UP	Evolution de l'étendue
11	Île-de-France	222	-37,2 %
21	Champagne-Ardenne	213	-15,2 %
22	Picardie	199	-18,9 %
23	Haute-Normandie	231	-24,2 %
24	Centre	334	-25,5 %
25	Basse-Normandie	163	-16,5 %
26	Bourgogne	173	-15,0 %
31	Nord-Pas-de-Calais	297	-40,7 %
41	Lorraine	235	-28,5 %
42	Alsace	161	-32,5 %
43	Franche-Comté	132	-18,8 %
52	Pays de la Loire	310	-28,2 %
53	Bretagne	313	-36,1 %
54	Poitou-Charentes	205	-21,0 %
72	Aquitaine	313	-24,0 %
73	Midi-Pyrénées	244	-18,2 %
74	Limousin	78	-24,0 %
82	Rhône-Alpes	552	-33,7 %
83	Auvergne	149	-21,6 %
91	Languedoc-Roussillon	240	-39,4 %
93	PACA	272	-43,0 %
94	Corse	28	-24,8 %
France métropolitaine		**5 128**	**-25 %**

Au final, la France métropolitaine [12] a été partitionnée en **5 064 unités primaires**, constituant la base de sondage au tirage de l'échantillon-maître. Le contour de ces unités primaires est fixé en géographie communale 2017 [13].

12. 14 communes n'ont pas été intégrées à cette partition, car leur accès est difficile. Il s'agit d'îles non reliées au continent par la route.

13. L'algorithme du voyageur de commerce a permis de regrouper des communes en géographie 2016. Une fois les unités primaires construites, elles ont été adaptées aux évolutions de géographie 2017. Le contour des unités primaires est définitivement fixé en géographie 2017 et ne variera pas quelle que soit l'évolution future des contours des communes. Les logements neufs seront par la suite attribués à une unité primaire en fonction de leurs coordonnées cartographiques. Ainsi, la taille des unités primaires ne sera pas sujette aux évolutions futures de la géographie communale.

Chapitre 2

La constitution des secteurs de l'enquête Emploi en continu

Le tirage de l'échantillon Emploi nécessite de constituer une base de sondage de secteurs, qui sont des regroupements d'environ 120 logements proches géographiquement [1]. La logique de ces regroupements de logements est de travailler à partir de l'unité la plus élémentaire pour constituer des unités de plus en plus agrégées : on part d'une base de logements que l'on regroupe en briques au sein des adresses, puis en grappes à l'intérieur de zones définies et enfin, on assemble ces grappes en secteurs. Dans cette partie, nous présentons la constitution de ces différents regroupements dont le but est d'obtenir les unités de tirage (les secteurs) qui respectent au mieux les contraintes fixées par l'enquête.

Pour le tirage du nouvel échantillon de l'EEC en France métropolitaine, la base de sondage des secteurs constitués en 2009 ne pouvait être réutilisée du fait de la forte évolution du parc de logements. En outre, la taxe d'habitation, utilisée comme base de sondage pour le tirage en 2009, couvre un champ de logements plus petit que Fidéli et le concept de résidence principale de Fidéli est plus proche de celui défini par Eurostat dans le règlement IESS (cf. point 2.1.1) que la taxe d'habitation [2]. Il était ainsi nécessaire de construire une nouvelle base de sondage. C'est à partir de Fidéli que sont regroupés les logements en unités d'échantillonnage (les secteurs) dans les travaux présentés ci-après.

2.1 Base de logements et contraintes de construction

La collecte de l'enquête Emploi s'effectue chaque trimestre sur une courte période (16 jours). Une vingtaine de logements contigus sont interrogés pendant 6 trimestres consécutifs, puis remplacés par 20 logements géographiquement proches. La durée de vie de l'échantillon est fixée à 9 ans.

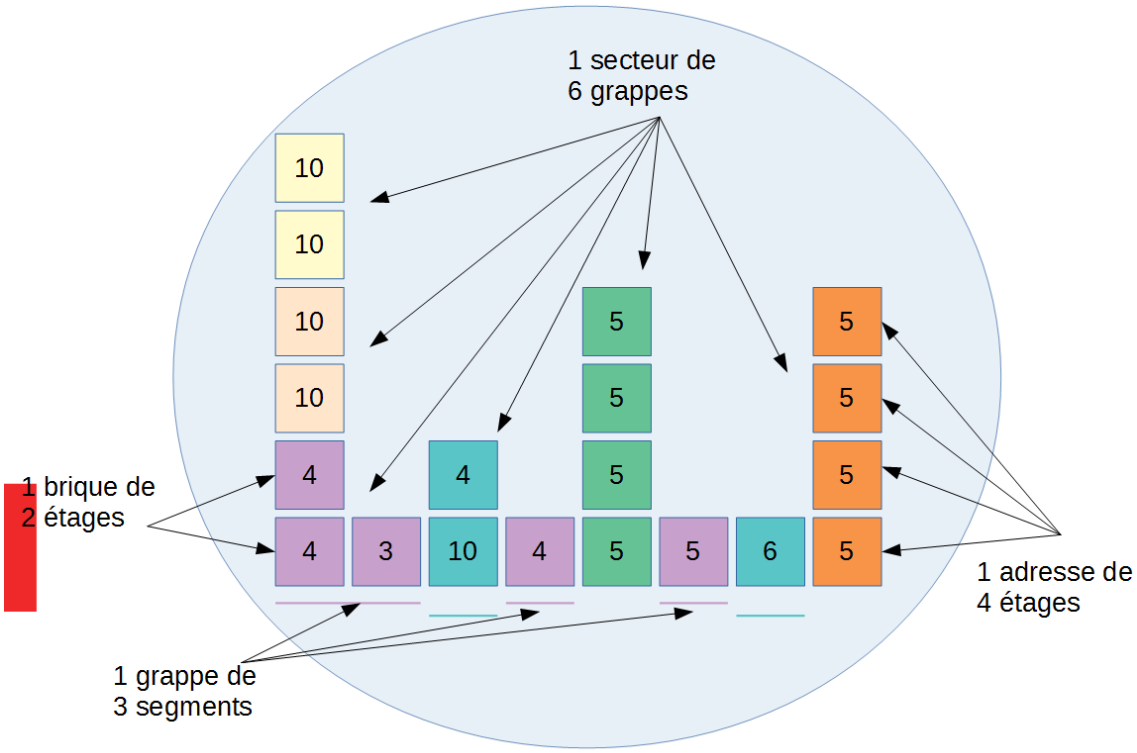
1. Voir le point 2.1.1 pour une définition plus précise des secteurs, également évoqués dans la section 1.2 de la partie A.

2. Fidéli permet de mieux identifier la résidence principale des individus quand ceux-ci sont localisés dans plusieurs résidences.

2.1.1 Notions utilisées pour la constitution des secteurs

Il est nécessaire, au préalable, de rappeler clairement le vocabulaire utilisé par la suite. La figure 2.1 illustre ces notions.

FIGURE 2.1 – Notions utilisées pour la définition des unités d'échantillonnage de l'EEC



Résidences principales

Le concept de résidences principales retenu pour le champ de l'enquête Emploi est celui de la source utilisée, Fidéli. Il s'agit de « résidences principales avec des occupants », notion proche de celle de « private household » définie par Eurostat dans le règlement IESS. Ce concept, plus large[3], diffère légèrement de celui de la taxe d'habitation.

Adresses

Chaque logement de la source Fidéli est géolocalisé et se voit ainsi attribuer les coordonnées (x,y) de son adresse. Une adresse correspond à l'ensemble des logements portant les mêmes coordonnées, c'est-à-dire possédant le même couple (x,y).

Étage

De même chaque logement possède un niveau. Tous les logements d'une adresse portant le même niveau constitue un étage. En pratique, une maison individuelle sera le

3. On dénombre 28 339 449 résidence principales en France métropolitaine, soit 170 000 de plus que dans la taxe d'habitation, du fait de la relocalisation de certains individus dans leur logement d'habitation.

seul logement portant ses coordonnées (x, y). Un immeuble verra ses logements avoir les mêmes (x, y), et tous les logements d'un même étage auront, de plus, le même niveau.

Grappes

La source est agrégée en regroupant les résidences principales par adresse et par étage. On cherche à constituer des grappes, ensemble d'environ 20 résidences principales géographiquement proches. La grappe est identifiée comme l'unité de collecte pour un trimestre et un enquêteur donné.

Briques

On regroupe alors les étages pour former, dans la mesure du possible, ces ensembles de 20 logements au sein des adresses. On définit alors la notion de brique comme ce regroupement d'étages au sein d'une même adresse, dans le but de constituer des grappes. Une brique est constituée de la manière suivante :

- Si le regroupement d'étages permet de constituer un ensemble de 20 logements environ, alors la brique formera à elle seule une grappe.
- Si le regroupement d'étages ne permet pas de constituer un tel ensemble (nombre de logements inférieur à 17 ou regroupement supplémentaire qui conduirait à plus de 24 logements [4]), alors la brique devra être associée à d'autres briques appartenant à d'autres adresses pour former une grappe.

Segments

On définit un segment comme étant un ensemble de briques contiguës, appartenant à une même grappe. Une grappe peut ainsi être constituée de plusieurs segments si les logements qui la composent ne sont pas tous contigus.

Secteurs

Les grappes sont alors réunies par 6 (ou 7 dans de rares cas) en secteur, afin de constituer la base de sondage de secteurs. Une fois 6 trimestres d'interrogation passés, une grappe d'un secteur est remplacée par une autre grappe du même secteur, constituant ainsi un pseudo-panel sur 36 trimestres, soit 9 ans.

2.1.2 Contraintes de construction des grappes

Du fait des spécificités de la collecte de l'enquête Emploi en continu, la construction des grappes doit répondre à de fortes contraintes :

- Les grappes doivent comporter entre 17 et 24 résidences principales, afin de limiter la variabilité de la charge d'enquête, une grappe devant être collectée par un enquêteur en deux semaines. Cette faible variation de la taille des grappes concourt aussi à la bonne précision des estimateurs ;
- Les grappes doivent être concentrées géographiquement pour faciliter la collecte, en limitant les déplacements de l'enquêteur, notamment au moment du repérage des logements à interroger ;

4. Dès lors qu'un regroupement supplémentaire conduirait à plus de 24 logements, il ne sera pas conservé. Les briques seront donc les regroupements d'étages de moins de 17 logements constitués avant ce regroupement supplémentaire.

— Les grappes doivent concerner le moins d'immeubles différents, pour limiter les problématiques liées à l'accès aux immeubles (digicodes, etc.) ;
— Un étage d'un immeuble doit être entièrement compris dans la même grappe Emploi, afin de simplifier la tâche des enquêteurs en ne demandant pas le repérage du logement à enquêter au sein d'un même étage, mais en interrogeant tous les logements d'un même étage. En effet, dès lors qu'il y a des déménagements et que les noms disponibles dans la base ne concordent pas avec ceux sur le terrain, déterminer quel appartement enquêter une fois arrivé à l'étage peut être compliqué.

Ces contraintes sont déterminantes dans la construction de la base de sondage.

2.2 L'algorithme de constitution des grappes

La méthode proposée ici vise à respecter au maximum les contraintes de constitution des grappes énoncées précédemment. Tous les logements d'un étage étant nécessairement inclus dans une unique grappe, l'étage sera l'unité élémentaire pour l'algorithme de constitution des grappes. En premier lieu, la base de logements initialement utilisée est transformée en une base d'étages, chaque étage correspondant à l'ensemble des logements appartenant à un même niveau d'une même adresse (cf. encadré sur la définition des étages). Ainsi, une maison individuelle formera un étage à elle toute seule ; un immeuble de 15 niveaux de 5 logements chacun constituera 15 étages de 5 logements chacun [5]. À partir de cette base, la méthode de regroupement des étages en grappes se décompose en deux étapes principales :

1. Une constitution des grappes au sein des grands immeubles, afin de limiter les accès à des adresses différentes, compliqués pour l'enquêteur ;
2. Une constitution des grappes entre les petits immeubles, une fois l'étape précédente effectuée, qui permet de gérer l'étendue géographique des grappes tout en respectant les contraintes sur le nombre de logements ;

Une étape finale de consolidation permet de gérer les cas limites des algorithmes précédents en lissant les grappes atypiques sur le reste des grappes constituées (cf. 2.2.3). Un bilan des grappes obtenues en termes d'étendue géographique (y compris par la route) et de nombre de logements par grappe est finalement présenté en section 2.4.

Quelques subtilités sur la définition d'étages

Les logements doivent être regroupés en étage afin de répondre aux contraintes de collecte (collecte de l'ensemble d'un étage pour un enquêteur). Cependant, la variable d'étage est basée sur des informations fiscales foncières qui peuvent être trompeuses. Par exemple, une copropriété de maisons indépendantes peut être associée à un unique point géographique, et conduire à construire un étage regroupant en fait les maisons. Une détection des "faux étages" est alors réalisée à partir des variables fiscales selon

5. Nous verrons, au paragraphe 2.2.1.2, que certaines corrections ont dû être apportées pour constituer cette base d'étages.

deux critères :

— si la variable d'étage est 0 (rez-de-chaussée) ou qu'elle n'est pas codée dans les fichiers fiscaux, et que le nombre d'appartements dépasse 50, alors on ne tient pas compte de cette variable d'étage, et chaque logement est alors considéré comme un étage indépendant ;

— si la variable d'étage est 0 (rez-de-chaussée) ou qu'elle n'est pas codée dans les fichiers fiscaux, et qu'un des logements est une maison, alors on ne tient pas compte de cette variable d'étage, et chaque logement est alors considéré comme un étage indépendant.

Une maison individuelle est donc considérée techniquement comme un étage d'un seul logement pour les besoins de construction de l'enquête. On dénombre, selon ces critères, **21 731 541 étages**. 5 228 030 d'entre eux sont considérés comme de "vrais" étages, c'est-à-dire appartenant à un immeuble, et contenant un ou plusieurs logements. Ils représentent 11 542 799 logements dans 1 448 793 immeubles différents.

Répartition du nombre de logements dans les "vrais" étages d'immeubles.

Nombre de logements par étage	Nombre d'étages
1	1 960 746
2	1 897 066
3	602 655
4	414 794
5	148 214
6 ou +	204 555
Total	**5 228 030**

Note de lecture : 414 794 étages ont 4 logements.

2.2.1 Première étape : création des grappes au sein des immeubles

L'objectif de cette étape est de limiter pour l'enquêteur le nombre d'entrées dans des immeubles différents, chacun ayant un coût d'accès propre, couvrant notamment l'obtention du digicode. Pour cela, l'idée est de créer un maximum de grappes au sein d'une même adresse en regroupant des étages pour former des ensembles de 20 logements. Les étages restants sont alors considérés comme des reliquats et seront utilisés par la suite, lors de la construction de grappes couvrant plusieurs immeubles.

2.2.1.1 Quel regroupement d'étages choisir ?

Différents regroupements d'étages possibles

Il peut être difficile, voire impossible, de construire des grappes de 20 logements au sein d'un même immeuble par regroupement d'étages. Cette contrainte (de collecte et de précision) peut être légèrement assouplie en acceptant des grappes de 17 à 24 résidences principales. Le nombre de regroupements possibles devient alors important.

Dans l'exemple du tableau 2.1, on peut ainsi regrouper les étages 0, 1, 2 et 3 dans une grappe, et garder l'étage 4 pour une autre grappe ; on peut également regrouper les

TABLE 2.1 – Jeu d'exemple – grappes au sein d'un immeuble

Etage	Nombre de logements
4e étage	10 logements
3e étage	5 logements
2e étage	5 logements
1er étage	6 logements
RDC	4 logements

étages 0, 1 et 4 dans une grappe, et 2 et 3 pour une autre grappe. Ces deux possibilités permettent d'obtenir une grappe de 20 logements, mais laisse un résidu de 10 logements à la sortie de l'algorithme. Une troisième possibilité est de regrouper les étages 1, 2 et 4 dans une grappe. On obtient alors une grappe de 21 logements, et deux résidus de 4 et 5 logements. De nombreuses autres solutions sont envisageables. Il faut donc définir un critère pour choisir la meilleure.

Limiter les reliquats, viser des ensembles de 20 logements

Le cas idéal est de réaliser des grappes internes de 20 logements et de n'avoir aucun reliquat. Le pire cas est de ne pas pouvoir regrouper d'étage et de considérer chaque étage comme un reliquat à traiter par la suite. Comme l'objectif est d'éviter à l'enquêteur d'avoir trop de grappes demandant d'entrer dans plusieurs immeubles, on cherchera tout d'abord à minimiser le nombre de reliquats restants à intégrer après construction des grappes au sein de l'immeuble. Ensuite, parmi les solutions qui minimisent ce nombre de reliquats, on choisira la solution permettant de constituer des grappes de taille la plus proche possible de 20 logements : la fonction de coût retenue est la somme quadratique des écarts de taille de grappe à 20.

Dans l'exemple précédent, parmi les trois découpages proposés, les deux premiers découpages permettent de limiter l'écart à 20 logements malgré la présence d'un reliquat (comme pour la troisième solution). Ces deux premiers découpages seront donc privilégiés.

Pour constituer les grappes au sein de tous les immeubles, on peut donc calculer le coût associé à chaque découpage possible de manière exhaustive.

2.2.1.2 Un problème trop lourd à traiter : optimisation de l'algorithme

Regrouper des paires d'étages plutôt que des étages

Pour des raisons de performance, cette solution fonctionne bien sur les immeubles de moins de huit étages. Par contre, elle n'est pas envisageable pour de plus grands immeubles. En effet, pour un immeuble de 20 étages, plus d'un million de combinaisons seraient à tester $\left(\sum_{k=1}^{20} \binom{20}{k} = 2^{20} - 1\right)$.

Pour les adresses de 9 étages ou plus, on va donc réduire le nombre de combinaisons possibles en commençant par constituer des paires d'étages, puis en explorant exhaustivement la création de grappes internes à l'immeuble à partir des paires ainsi créées. En effet,

si le nombre de partitions distinctes au sein d'un ensemble de 20 étages (par exemple) est immense, le nombre de paires d'étages distinctes est lui de 190 (i.e. $\binom{20}{2}$)), soit un nombre beaucoup plus raisonnable. Il est donc tout à fait possible de tester l'ensemble des paires d'étages possibles et de constituer la « meilleure » paire possible en minimisant le coût défini plus tôt.

Créer les *meilleures* paires d'étages possibles

Pour créer ces paires, on va chercher à nouveau à optimiser le regroupement. De façon schématique, pour un étage donné, « sa meilleure paire possible » peut s'entendre au moins en deux sens :

- soit un (ou plusieurs) rattachement(s) à un autre étage respecte(nt) les contraintes de tailles minimale et maximale et alors la meilleure paire est celle qui minimise la fonction de coût habituelle (la somme des écarts à 20 au carré) ;
- soit aucun rattachement avec un autre étage ne permet de constituer une paire qui respecte les contraintes de tailles minimale et maximale des grappes [6]. La meilleure paire est alors celle qui permet de se rapprocher le plus possible des 20 logements lorsqu'on ajoute un troisième étage à cette paire. Concrètement, à un étage donné, on associe un par un chaque étage, créant ainsi autant de paires. Puis pour chacune des paires, on ajoute le nombre de logements d'un autre étage (distinct), et on associe à ce triplet le coût tel que défini auparavant. La paire pour laquelle ce coût est le plus faible est alors constituée et « remise en jeu » en attendant d'être finalisée à l'étape suivante.

Extension du principe à des triplets et quadruplets d'étages

Si l'utilisation de paires est envisagée pour économiser les temps de calcul, elle diminue fortement le nombre de possibilités de regroupements d'étages en grappes et peut exclure la meilleure combinaison. Aussi, pour les immeubles entre 9 et 15 étages, on ne se contente pas d'examiner les paires, mais également les triplets et les quadruplets. En effet, pour un immeuble de 15 étages, il y a 105 paires, 455 triplets et 1365 quadruplets possibles. Il n'est donc pas très coûteux d'examiner toutes les paires, triplets et quadruplets possibles conduisant à des regroupements d'étages optimaux avant d'arriver à une dimension permettant une comparaison exhaustive de toutes les partitions possibles.

À la fin de cette phase, on a construit un premier ensemble de 160 000 grappes au sein des 18 millions d'immeubles du territoire métropolitain. Les étages restants au sein de chacun des immeubles forment quant à eux un reliquat, qu'on appelle « brique », qui va intégrer l'étape suivante de l'algorithme.

2.2.2 Deuxième étape : création de grappes entre les immeubles

Suite à l'étape précédente, on dispose d'une part de grappes construites au sein d'immeubles, d'autre part d'un ensemble de « briques » que l'on va chercher à regrouper pour construire des grappes sur plusieurs immeubles. Ces briques peuvent être de deux sortes :

- soit des maisons individuelles ;

6. Respectivement 17 et 24 logements.

— soit des reliquats de l'étape précédente, qui correspondent à des étages ou des groupes d'étages d'immeubles [7].

L'objectif est de combiner ces briques de sorte à construire des grappes d'une vingtaine de logements géographiquement compactes. Obtenir la solution la plus optimale à ce problème est très coûteuse d'un point de vue informatique, car il faudrait pouvoir tester tous les regroupements possibles. C'est pourquoi nous avons repris l'approche suivie pour l'échantillon de l'EEC 2009 : parcourir un court chemin entre les logements, facile à tracer, et le découper après chaque constitution de grappe.

2.2.2.1 Variante de la méthode suivie pour l'échantillon de 2009

Trier selon le cadastre : des effets indésirables, non contrôlables

En 2009, comme décrit dans Loonis, 2009, le tri selon les variables foncières dans la base de sondage avait permis de créer un chemin, le long duquel les grappes étaient construites : en pratique, on ajoute un immeuble à chaque étape du chemin, on incrémente en parallèle un compteur du nombre de logements de l'étage, et quand ce compteur atteint ou dépasse 20 logements, on finalise la grappe et on entame la suivante. Cette méthode permet de construire efficacement des grappes. En revanche, elle a deux inconvénients :

— On ne contrôle pas la taille maximale des grappes : deux étages de 15 logements seront regroupés ensemble et formeront une grappe de 30 logements ;
— L'ordre des références cadastrales ne correspond pas toujours à un cheminement géographique cohérent, notamment quand des parcelles ont été scindées ; on peut ainsi créer des grappes à partir de logements éloignés.

Utilisation d'une solution au problème du voyageur de commerce

Afin d'éviter ces écueils, l'idée est d'utiliser à nouveau une solution au problème du voyageur de commerce, comme pour la constitution des unités primaires (voir point 1.2.3). Pour rappel, ce problème consiste à chercher un chemin optimal pour un marchand itinérant qui doit parcourir plusieurs villes pour aller y vendre sa marchandise. Cet algorithme nécessite de limiter le nombre de briques sur lesquelles faire passer le chemin, pour des raisons de performance. Afin de déterminer des solutions locales cohérentes et de bénéficier de performances de calcul satisfaisantes, il est nécessaire de partitionner le territoire et de limiter les calculs de chemins à chaque zone ainsi constituée.

2.2.2.2 Découpage du territoire en zones

Si le découpage en zones est nécessaire pour permettre l'exécution du processus, il peut également permettre d'assurer une certaine concentration géographique, en fonction de la partition choisie. Ainsi, il apparaît judicieux de réaliser l'assemblage de logements en grappes dans un découpage de la France en petites zones, en contraignant l'inclusion de chaque grappe créée dans une et une seule zone. Concrètement, il s'agit de dessiner dans chaque zone créée un chemin, puis de constituer les grappes sur ce chemin.

7. Notamment tous les étages des immeubles lorsque ces derniers comprennent moins de 17 logements.

Il convient alors d'élaborer ce découpage de la France métropolitaine. En se plaçant au niveau communal, on peut envisager plusieurs scénarios :

— un découpage selon une « grille » : les logements de la commune sont affectés à des cases issues d'un découpage régulier en fonction des coordonnées cartographiques de leur parcelle cadastrale. Cette approche permet de limiter la distance à vol d'oiseau entre deux logements d'une même grappe, mais implique de gérer des cas « limite », où la grille ne se superpose pas très bien avec la géographie de la commune : cases avec quelques logements uniquement, communes avec une densité atypique, etc ;
— un découpage selon les sections cadastrales : si l'information est facilement mobilisable, de nombreuses sections cadastrales possèdent trop peu de logements pour permettre de constituer plusieurs grappes de 20 résidences principales. Il faudrait alors les regrouper entre elles, ce qui complexifierait grandement la partition du territoire ;
— un découpage selon les Iris (Ilots Regroupés pour l'Information Statistique) : ce découpage infracommunal construit par l'Insee en vue de la diffusion du recensement de 1999 (et retouchés depuis en 2008), regroupe 2 000 habitants environ par Iris. Il respecte l'homogénéité de l'habitat et ses limites s'appuient sur les grandes coupures du tissu urbain (voies principales, voies ferrées, cours d'eau, contours de communes...). Pour les communes plus petites (inférieures à 5000 habitants principalement), on assimile l'Iris à la commune. Cette partition du territoire permet d'obtenir dans la plupart des cas des zones de taille uniforme en nombre de logements et pas trop étendues géographiquement.

Le choix de l'Iris semble ainsi le plus simple à réaliser et le plus pertinent, d'autant plus qu'il gère en grande partie les problématiques d'obstacles infranchissables (rivière, etc.). On s'assure ainsi autant que possible qu'une grappe ne sera pas traversée par un fleuve et donc que la contiguïté des logements est très souvent assurée, ce qui est essentiel au bon déroulement de la collecte de l'enquête. Cependant, ce découpage présente deux limites :

— les étendues des Iris ne sont pas toutes semblables entre elles, certains Iris sont notamment très étendus ;
— certaines communes ou certains Iris sont trop petits en nombre de logements pour qu'on puisse constituer plusieurs grappes d'une vingtaine de logements.

Le premier cas est suffisamment rare pour être résolu manuellement en découpant les Iris considérés comme trop étendus. Le deuxième problème est plus fréquent (voir le tableau 2.2). La taille des grappes étant contrainte entre 17 et 24 logements, les zones, dites « à problème » c'est-à-dire qui ne comportent pas suffisamment de logements, sont les suivantes :

— Les zones très petites qui contiennent moins de 17 logements et ne peuvent contenir une grappe.
— Les zones contenant entre 25 et 33 logements, car elles ne peuvent contenir ni une (car trop grande) ni deux (car trop petite) grappes.
— Enfin, les zones contenant 49 et 50 logements sont à problème car elles ne peuvent contenir ni deux (car trop grande) ni trois (car trop petite) grappes.

À partir de 51 logements, les zones peuvent être regroupées en grappes respectant la contrainte de taille.

TABLE 2.2 – Dénombrement des zones dites « à problème ».

	Nb de zones concernées	Nb de logements concernés
Moins de 17 logements	677	6 760
Entre 25 et 33 logements	1 101	32 031
Entre 49 et 50 logements	276	13 660
Total	*2 054*	*52 451*

Ces trois cas de zones « à problème » se traitent de la manière suivante : chacune de ces zones est jointe à une zone dite « de renfort » qui est définie comme étant la plus proche [8] (et n'étant pas elle-même une zone « à problème »).

Afin de respecter une certaine unité géographique entre les différentes partitions créées pour les besoins de l'échantillonnage des enquêtes auprès des ménages, on regroupera les zones « à problème » aux zones de renfort de la même unité primaire de l'échantillon-maître [9]. Concrètement, chaque UP contient de nombreuses zones telles que définies ici (Iris ou communes de moins de 5000 habitants). Dans chaque UP contenant une zone dite « à problème », une matrice de distance géographique entre les coordonnées du barycentre de chaque zone est construite [10]. Cette matrice nous permet alors de déterminer la zone « de renfort » : il s'agit de la zone la plus proche par la distance de la zone à problème. Si plusieurs zones sont à problèmes dans l'UP, elles peuvent être reliées, de manière itérative, à la même zone de « renfort ».

Découpage des chemins en grappe Une fois la partition du territoire en zones effectuée, les chemins, solutions de l'algorithme du voyageur de commerce, peuvent être tracés entre toutes les briques d'une même zone. Intuitivement, il semble peu probable que ces chemins permettent d'obtenir des grappes entre 17 et 24 logements. Il est donc nécessaire d'autoriser des « sauts » pour la constitution des grappes. On entend ici par « saut » le fait de ne pas intégrer la brique suivante sur un chemin, car elle conduirait à une grappe de taille trop importante, mais à passer directement à la brique d'après, et donc de « sauter » la brique en question. Il est cependant nécessaire de s'assurer que ces « sauts » ne conduisent à avoir des grappes éclatées géographiquement. On impose ainsi plusieurs contraintes sur ces sauts :.

1. Si une grappe possède 20 logements ou plus, elle est considérée comme terminée ;
2. Si le fait de rattacher une brique à la grappe en cours de construction conduit à former une grappe de 25 logements ou plus, alors on autorise un saut ;
3. Si la grappe que l'on est en train de construire fait déjà au moins 17 logements, alors on limite le nombre de saut à un certain seuil (4 ou 5).

.

8. Au sens de la distance euclidienne avec les coordonnées cartographiques, X et Y.

9. Pour rappel, les unités primaires sont constituées comme des ensembles de communes, et donc, par extension, comme des ensemble d'Iris.

10. Les coordonnées du barycentre d'une zone sont calculées comme la moyenne des coordonnées des logements appartenant à la zone.

Considérons deux exemples d'illustration. Supposons que l'on ait une zone composée de 10 briques, et qu'un chemin obtenu par l'algorithme du voyageur de commerce les range dans l'ordre indiqué par le tableau 2.3.

TABLE 2.3 – Premier jeu d'exemple – constitution des grappes

Brique	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
Logements	8	10	8	2	4	1	6	8	2	12
Grappe	G1	G1	G2	G1	G2	G2	G2	G3	G2	G3

Dans ce cas, la construction d'une grappe commence avec les briques B1 et B2, ce qui donne déjà 18 logements. Si l'on rajoute la brique B3 à cette grappe en construction, on aboutit à 26 logements, ce qui dépasse la taille maximale acceptable pour une grappe. On décide alors de « sauter » cette brique, et d'ajouter plutôt la brique B4 à la grappe : on finalise ainsi la grappe G1 à 20 logements. On poursuit ensuite avec une nouvelle grappe, qui commence à la brique B3, première brique du chemin non rattachée à une grappe ; on ne considère pas la brique B4, déjà affectée, mais on passe directement à la B5, la B6 et la B7. À ce stade, la grappe en construction contient 19 logements ; rajouter la brique B8 conduirait à une grappe de 27 logements, ce qui n'est pas acceptable ; on la saute donc, puis on rattache la brique B9 ; la grappe G2 atteint 21 logements ; on passe à la grappe G3 qui réunit les briques B8 et B10 (20 logements).

Considérons maintenant une autre zone de 9 briques ordonnées comme indiqué dans le tableau 2.4.

TABLE 2.4 – Second jeu d'exemple – constitution des grappes

Brique	B1	B2	B3	B4	B5	B6	B7	B8	B9
Logements	8	10	7	7	7	10	10	8	2
Grappe	G1	G1	G2	G2	G2	G3	G3	G4	G4

Ici aussi, la grappe commence avec les briques B1 et B2, ce qui donne déjà 18 logements, et on décide de sauter la brique B3 (car on aurait une grappe de 25 logements). On considère alors la brique B4, qui n'est pas non plus acceptable, tout comme ne le sont pas les briques B5 et B6. On pourrait continuer jusqu'à la brique B9, mais afin d'éviter une étendue géographique trop importante, on décide de limiter le nombre de sauts possibles à 5 ; ainsi, la grappe G1 va s'arrêter avec les briques B1 et B2 et ne comporter que 18 logements, ce qui est acceptable car supérieur à 17 (seuil minimal pour la constitution d'une grappe).

Ainsi la méthode permet de construire des grappes comportant entre 17 et 24 logements ; cependant, il est possible que la dernière grappe de chaque chemin (et donc de chaque zone) soit trop petite, car il ne reste pas assez de briques non affectées pour atteindre 17 logements (comme c'est le cas pour la grappe G4 dans l'exemple du tableau 2.4). On se retrouve alors avec des reliquats, c'est-à-dire des grappes de 16 logements ou moins.

Limitation du nombre de reliquats Une solution pour limiter le nombre de reliquats est de jouer sur le chemin. En effet, cet algorithme est déterministe et la constitution des grappes dépend du chemin parcouru, lui même dépendant de la brique de départ utilisée. Pour obtenir le regroupement en grappes limitant les reliquats, on construit, dans chaque zone, plusieurs chemins en faisant varier les points de départs et les paramètres utilisés. Dans chaque zone, on peut alors construire plusieurs jeux de grappes, et choisir, s'il existe, celui qui ne contient pas de reliquat. S'il existe un reliquat, on choisira, dans la zone, le chemin qui donne les grappes les moins géographiquement dispersées en moyenne, c'est-à-dire qui minimise :

$$\frac{1}{n_{grappes \in zone}} \sum_{g \in zone} \frac{\sum_{i \in g} nblog_i \sqrt{(x_i - \bar{x_g})^2 + (y_i - \bar{y_g})^2}}{\sum_{i \in g} nblog_i} \tag{2.1}$$

en notant $n_{grappes \in zone}$ le nombre de grappes dans une zone pour un chemin donné, g les différentes grappes constituées dans cette zone et i les différentes briques constituant une grappe g. $nblog_i$ est le nombre de logements dans la brique i, x_i la coordonnée X de cette brique (respectivement y_i et Y). Les coordonnées $\bar{x_g}$ et $\bar{y_g}$ sont les coordonnées du barycentre de la grappe g, c'est-à-dire $\bar{x_g} = \frac{\sum_{i \in g} nblog_i \ x_i}{\sum_{i \in g} nblog_i}$ et $\bar{y_g} = \frac{\sum_{i \in g} nblog_i \ y_i}{\sum_{i \in g} nblog_i}$

Une fois les regroupements en grappes obtenus, une dernière étape de consolidation va permettre de répartir les reliquats résultants (grappes de moins de 17 logements) sur d'autres grappes proches.

2.2.3 Troisième étape : Consolidation et suppression des reliquats

Si la multiplication des chemins de constitution des grappes permet de limiter le nombre de reliquats, ce n'est pas suffisant pour gérer l'ensemble des cas de grappes trop petites. À l'issue des deux étapes de l'algorithme de constitution des grappes, environ 2% des grappes comptent moins de 17 logements et sont dites « en anomalie ». Plus précisément, 1% des grappes comptent strictement moins de 10 logements. Cette fréquence relativement élevée est préoccupante, dans la mesure où elle pose des difficultés sensibles en termes d'organisation de la collecte sur le terrain. Par ailleurs, d'un point de vue méthodologique l'homogénéité de la taille des grappes contribue à diminuer la variance d'un sondage en grappes.

Le principe de la consolidation L'objectif de la consolidation est ainsi de redistribuer le contenu de ces reliquats sur les grappes alentour, quitte à s'éloigner de la cible de 20 logements. Pour rappel, les grappes sont constituées de 1 ou plusieurs segments, chaque segment étant un ensemble de logements contigus sur le chemin et appartenant à une même grappe.

Le principe adopté est de déplacer de grappe en grappe les segments le long du chemin du voyageur de commerce de façon à supprimer les grappes reliquats issues des étapes précédentes de l'algorithme.

Plus précisément, la consolidation se déroule en deux étapes. Tout d'abord, les grappes reliquats de 8 logements ou moins sont démantelées et leurs briques sont associées aux segments qui les précèdent immédiatement sur le chemin solution au problème du voyageur de commerce. Les grappes de ces segments sont alors grossies, pouvant même dépasser le seuil de 24 résidences principales. La deuxième étape concerne les grappes reliquats de 9 à 16 résidences principales et les grappes affectées par l'éclatement des grappes reliquats de 8 logements ou moins.

En pratique, autour de chaque segment appartenant à une grappe reliquat, on définit une « fenêtre », c'est-à-dire l'ensemble des segments au voisinage du reliquat, qui pourront être affectées par cette consolidation.

Un segment est constitué d'une ou plusieurs briques. Au sein d'un segment, on peut isoler les briques « frontières » qui correspondent à celles qui, le long du chemin, jouxtent un segment n'appartenant pas à la même grappe. Dans une fenêtre, on teste alors l'impact du déplacement de chaque brique frontière dans le segment voisin [11] du chemin, sur la composition des grappes :

- si la composition des grappes de la zone se rapproche en moyenne de la cible de 20 logements [12], alors le déplacement de la brique est validé. On redéfinit alors les briques frontières de la fenêtre et on teste à nouveau le déplacement de chacune d'elle.
- si aucun déplacement de brique frontière dans la fenêtre ne permet d'améliorer la composition des grappes, alors l'algorithme s'arrête.

Considérons un exemple. Pour une zone comptant 154 logements individuels, l'enchaînement des algorithmes de construction des grappes présentés dans les parties précédentes conduit à la répartition des logements en 8 grappes ordonnées le long d'un chemin. Ces grappes comportent le nombre de logements suivant :

20 14 20 20 20 20 20 20

11. Ou les segments voisins de la brique, si celle-ci constitue à elle seule un segment.

12. La fonction de coût utilisée est la somme quadratique de l'écart entre le nombre de logements de chaque zone et la cible de 20 logements.

La grappe n° 2 compte seulement 14 logements et est donc en anomalie. Dans la mesure où elle comporte plus de 9 logements, l'objectif de l'algorithme est d'étendre cette grappe de façon à se rapprocher le plus possible de la cible de 20 logements tout en respectant l'ordre des briques sur le chemin. Les lignes qui suivent représentent les étapes successives de l'algorithme :

20	14	20	20	20	20	20	20
19	15	20	20	20	20	20	20
19	16	19	20	20	20	20	20
18	17	19	20	20	20	20	20
18	18	18	20	20	20	20	20
19	18	18	20	20	20	20	19
19	18	19	19	20	20	20	19

À noter que le chemin qui parcourt la zone étant circulaire (pas de point de départ ni de point d'arrivée), à l'itération 5 un déplacement de frontière intervient entre la « première » et la « dernière » grappe de la zone telle que représentée ici.

Le paramètre : la taille de la fenêtre L'utilisation de fenêtres autour des segments reliquats possède deux vertus :

- d'une part, elle limite considérablement le nombre de déplacements de briques frontières à tester, et ainsi le temps d'exécution de la consolidation ;
- d'autre part, elle limite les segments (et donc les grappes) affectés par la consolidation à ceux situés dans le voisinage immédiat des segments associés aux grappes en anomalie, et ne modifie pas la majorité des grappes issues des étapes 1 et 2 de l'algorithme.

La taille de ces fenêtres, symétriques autour des segments reliquats constitue à ce titre un paramètre important de la consolidation : plus la fenêtre sera large, plus le nombre de segments participants au lissage de la taille des grappes est important, et plus il sera efficace. Mais le nombre de grappes impactées par la consolidation et le temps de passage de cet algorithme seront également importants. Une fenêtre de taille 1 signifie que seuls le segment précédant et le segment suivant un reliquat seront impliqués dans la consolidation.

Il faut à présent déterminer la taille de cette fenêtre. Trois critères sont pris en compte (nombre de grappes de moins de 17 logements après consolidation, nombre de grappe de 19 à 21 logements, nombre de grappes modifiées du fait de la consolidation), et plusieurs tailles de fenêtres testées (de 1 à 6 segments de part et d'autre des segments reliquats des étapes 1 et 2 de l'algorithme).

À l'exception du cas où une fenêtre de largeur 1 est utilisée (c'est-à-dire un segment de part et d'autre du segment reliquat), la consolidation permet de réduire drastiquement

le nombre de grappes problématiques de 22 571 à environ 420 grappes de moins de 17 résidences principales. Les deux autres critères évoluent en parallèle : plus cette fenêtre est large, plus le pourcentage de grappes affectées par la consolidation augmente logiquement, et plus le nombre de grappes comptant de 19 à 21 logements augmente également. Le tableau 2.5 présente les valeurs des différents critères pour les différentes tailles de fenêtre étudiées. Au final, une largeur de 4 segments pour une fenêtre semble le meilleur compromis, permettant de conserver le nombre de grappes de 19 à 21 logements avant consolidation, tout en résorbant fortement les cas problématiques de grappe, et en limitant le nombre de grappes affectées par cette étape. À noter que les grappes de 1 à 9 logements sont entièrement résorbées, grâce à la première étape de l'algorithme de consolidation.

TABLE 2.5 – Caractéristiques des grappes en fonction de la taille de la fenêtre.

Taille de fenêtre	Nombre de grappes	Grappes de 17-24 logements		Grappes de 19-21 logements	
		Nombre	%	Nombre	%
Avant consolidation	1 394 522	1 364 824	97,87 %	1 127 201	80,83%
1	1 383 392	1 371 536	99,14 %	1 090 568	78,83%
2	1 383 392	1 375 781	99,45 %	1 096 903	79,29%
3	1 383 392	1 375 802	99,45 %	1 109 730	80,22%
4	1 383 392	1 375 810	99,45 %	1 116 863	80,73%
5	1 383 392	1 375 809	99,45 %	1 119 804	80,95%
6	1 383 392	1 375 808	99,45 %	1 120 700	81,01%

Note de lecture : Une taille de fenêtre égale à 4 permet d'obtenir 99,45 % de grappes ayant entre 17 et 24 résidences principales et 80,73 % de grappes ayant entre 19 et 21 résidences principales après consolidation.

Rappel des étapes de la constitution des grappes Au final, les étapes successives de la méthode ont permis :

- de regrouper les logements en étages pour les besoins de la collecte
- de regrouper les étages entre eux pour former un maximum de grappes internes aux immeubles
- de regrouper entre eux, au sein d'une même zone (Iris, commune ou regroupement d'Iris ou de communes), les logements des adresses quasiment contiguës, en limitant le nombre de grappes de moins de 17 logements, et en maximisant les grappes de 20 logements
- de gérer les grappes résiduelles en reventilant les logements sur les grappes voisines.

Mais les grappes ne sont pas l'unité d'échantillonnage de l'enquête Emploi : il faut à présent les regrouper par 6 ou 7 en secteurs.

2.3 Regroupement des grappes en secteurs

La base de sondage de l'EEC est une base de secteurs, qui correspondent à des ensembles de 6 ou 7 grappes géographiquement proches, dont chacune contient une vingtaine de logements. Chaque fois qu'une grappe sort de l'échantillon Emploi (après 6 trimestres

d'interrogation), c'est la grappe suivante *au sein du secteur* qui la remplace [13]. En raison de l'homogénéité spatiale des variables d'intérêt de l'EEC, ce mécanisme assure une forte corrélation entre les estimateurs construits à partir des grappes sortantes et ceux construits à partir des grappes entrantes, qui contribue à la bonne précision des estimateurs d'évolutions trimestrielles [14].

2.3.1 Les objectifs de la constitution des secteurs

Plusieurs considérations ont guidé la constitution des secteurs du nouvel échantillon de l'enquête Emploi en Continu :

1. Comme pour l'échantillon Emploi 2009, le choix a été fait de constituer d'abord une partition du territoire en grappes, puis de les regrouper en secteurs. Une stratégie alternative aurait consisté à constituer les secteurs comme des ensembles de 120 logements proches, puis de les diviser en grappes. Cependant, il est plus complexe de créer des ensembles de 20 logements au sein de secteurs de 120, si un tel découpage n'a pas été anticipé (compte-tenu des contraintes portant sur les étages).
2. Si la proximité des grappes au sein des secteurs est moins problématique que celle des logements au sein des grappes [15], elle est toutefois recherchée, afin de faire jouer l'homogénéité socio-démographiques des individus géographiquement proches, et pour faciliter l'association d'un secteur à un enquêteur sur le long terme.
3. Pour obtenir des poids les moins dispersés possibles, le nombre de grappes (et de logements) par secteur doit être stable. Ainsi, le processus de regroupement des secteurs cherche à minimiser le nombre de secteurs à 7 grappes. Ces secteurs sont constitués quand le nombre total de grappes à regrouper n'est pas un multiple de 6. Les grappes restantes sont alors réparties une à une sur d'autres secteurs. Pour un ensemble de grappes à regrouper, on obtiendra alors 1 à 5 secteurs de 7 grappes. L'algorithme de regroupement des grappes mis en œuvre vise à maintenir au strict minimum le nombre de secteurs à 7 grappes.

2.3.2 Le principe pour la constitution des secteurs

Pour répondre à ces critères, le principe suivi pour la constitution des secteurs est d'associer les grappes voisines en secteurs au sein d'un territoire, pas trop étendu pour assurer la proximité géographique des secteurs, mais suffisamment large pour limiter le nombre de secteurs à 7 grappes [16]. Si la logique est d'utiliser le chemin suivi et les

13. La détermination de l'ordre des grappes dans le secteur a fait l'objet d'investigations spécifiques, détaillées dans la section 1.3 dans la partie D.

14. Cela permet de limiter l'influence de la variance liée au tirage de grappes.

15. La collecte de deux grappes d'un même secteur ne s'effectuant pas les mêmes trimestres, l'enquêteur n'aura pas à optimiser ses déplacements entre ces deux grappes.

16. Il peut y avoir, potentiellement, cinq fois plus de secteurs à 7 grappes que de zones dans lesquelles sont regroupés les secteurs. Limiter le nombre de zones limitera donc le nombre de secteurs à 7 grappes.

zones adoptées (commune ou Iris, *cf. supra*) pour la création des grappes, le nombre de zones ainsi utilisées conduit à un nombre de secteurs à 7 grappes potentiellement très important. L'idée va donc être de joindre les chemins de deux zones voisines au sein d'un regroupement plus large, en limitant ainsi ces secteurs de moindre qualité statistique. Deux questions se posent : quel zonage plus large utiliser et comment regrouper deux chemins ?

Le zonage dans lequel constituer les secteurs La première question n'est pas la plus problématique puisque la seule contrainte est d'avoir un zonage permettant de respecter le plan de sondage défini plus tard. Ainsi, pour un tirage effectué sur l'ensemble du territoire métropolitain, sans contrainte de représentativité régionale, on pourrait ne définir qu'une seule aire, la métropole, et ainsi limiter le nombre de secteurs à 7 grappes à 5 au maximum. La proximité géographique des grappes d'un même secteur serait alors assurée, d'une part par chaque chemin utilisé pour les grappes, d'autre part par la jonction entre deux chemins. Cela étant, les résultats de l'EEC étant demandés au niveau Nuts2 (ancienne région), un découpage en 22 aires minimum semble inévitable. On va donc chercher le zonage le plus large respectant le plan de sondage dans chaque région (et minimisant ainsi les secteurs à 7 grappes) ; nous verrons dans le chapitre 3 que la coordination des tirages nous amène à choisir un regroupement d'unités primaires appelé unité de coordination. À ce stade, nous introduisons la notion de *section* comme étant ce zonage.

Jonctions des chemins Une section contient plusieurs zones, et donc plusieurs chemins de grappes. La question de la jonction des chemins est plus importante puisqu'elle va jouer sur la proximité des grappes d'un même secteur. L'objectif est d'arriver à joindre les zones voisines au niveau des grappes les plus proches pour constituer des secteurs peu dispersés. Pour cela, on adopte alors une démarche descendante : on partage le territoire de chaque section en deux, de telle sorte que chaque sous-section ainsi obtenue n'augmente pas le nombre de secteurs à 7 grappes, mais propose un territoire plus restreint. On procède alors de manière itérative sur chaque sous-section obtenue, jusqu'à ce qu'une nouvelle division entraîne l'augmentation du nombre de secteurs à 7 grappes, ou que la sous-section corresponde à une zone (Iris, Triris [17]). On obtient alors les sous-sections dans lesquelles les chemins sont joints, en minimisant l'impact de cette jonction sur la taille totale du chemin. Une consolidation nécessaire est alors effectuée pour prendre en compte certaines zones particulières, pour lesquelles l'algorithme adopté ne prend pas bien en compte l'alignement géographique.

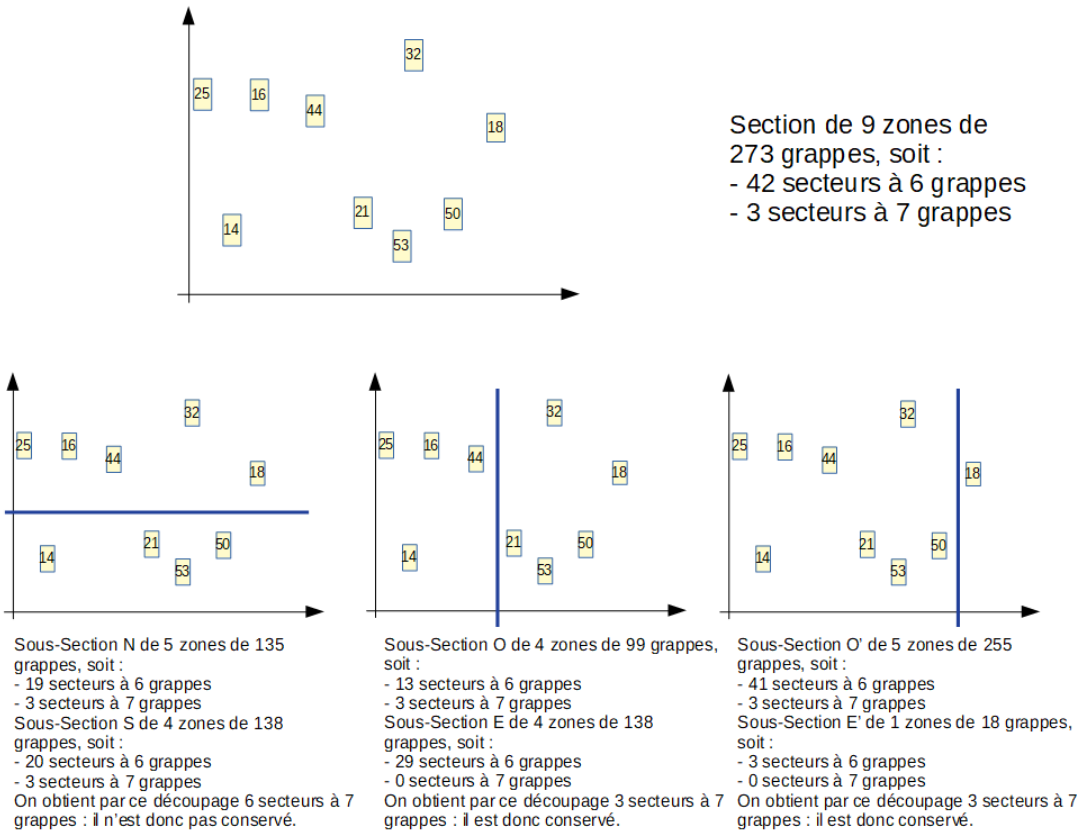
Création des sous-sections En pratique, on trie les zones [18] selon leur abscisse (coordonnée x). L'algorithme parcourt alors cet axe abscisse par abscisse, et à chaque

17. Un Triris est un regroupement d'Iris (en général 3 Iris). Le Triris a été créé en 1999 pour la diffusion de variables sensibles du recensement pour lesquelles l'Iris apparaît insuffisant pour garantir le secret statistique.

18. À cette étape, chaque zone est identifiée à un point, le barycentre des logements qui lui appartiennent.

nouvelle zone rencontrée, est calculé le nombre de secteurs à sept grappes nécessaires dans chaque sous-section si on divisait la section juste après cette zone. On effectue le même travail sur l'axe des ordonnées. On ne conserve alors que les partitions permettant de conserver le même nombre de secteurs à 7 grappes qu'avant division de la section [19]. Un exemple de ce parcours et de cette sélection est donné en figure 2.2. Entre ces partitions, on choisit alors celle qui maximise la distance entre le centre de chacune des deux sous-sections (est/ouest pour une séparation grâce à une abscisse, nord/sud pour une séparation grâce à une ordonnée). Ce choix est illustré par la figure 2.3. On réitère alors cet algorithme pour chacune des deux sous-sections ainsi créées. Il s'arrêtera si aucune division ne conserve le même nombre de secteurs à 7 grappes ou si la section correspond à une zone. La plupart des sous-sections ainsi obtenues ont un nombre de grappes multiple de 6. Chaque section est alors partitionnée en sous-section dans lesquelles sont joints les chemins.

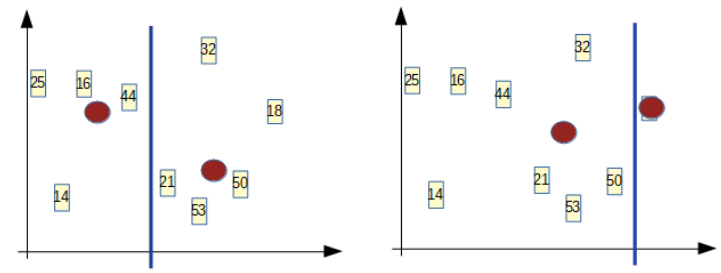
FIGURE 2.2 – Parcours d'une section pour identifier les sous-sections potentielles.



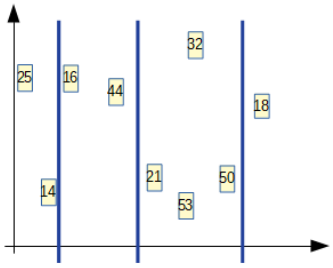
Note de lecture : Pour la section présentée dans la partie supérieure du graphique, 3 découpages possibles en 2 sous-sections sont présentés dans la partie inférieure. Seuls les 2 découpages les plus à droite sont susceptibles d'être conservés, car ils ne conduisent pas à augmenter le nombre de secteurs à 7 grappes dans la section.

19. Si une zone possède 6 × n grappes, on constituera n secteurs de 6 grappes dans cette zone, sans faire appel à une zone voisine. Cette zone ne participera donc pas à la constitution des sous-sections.

FIGURE 2.3 – Choix du partitionnement d'une section en deux sous-sections.



Le découpage de gauche est choisi, car la distance entre les centres des deux sous-sections est la plus grande.

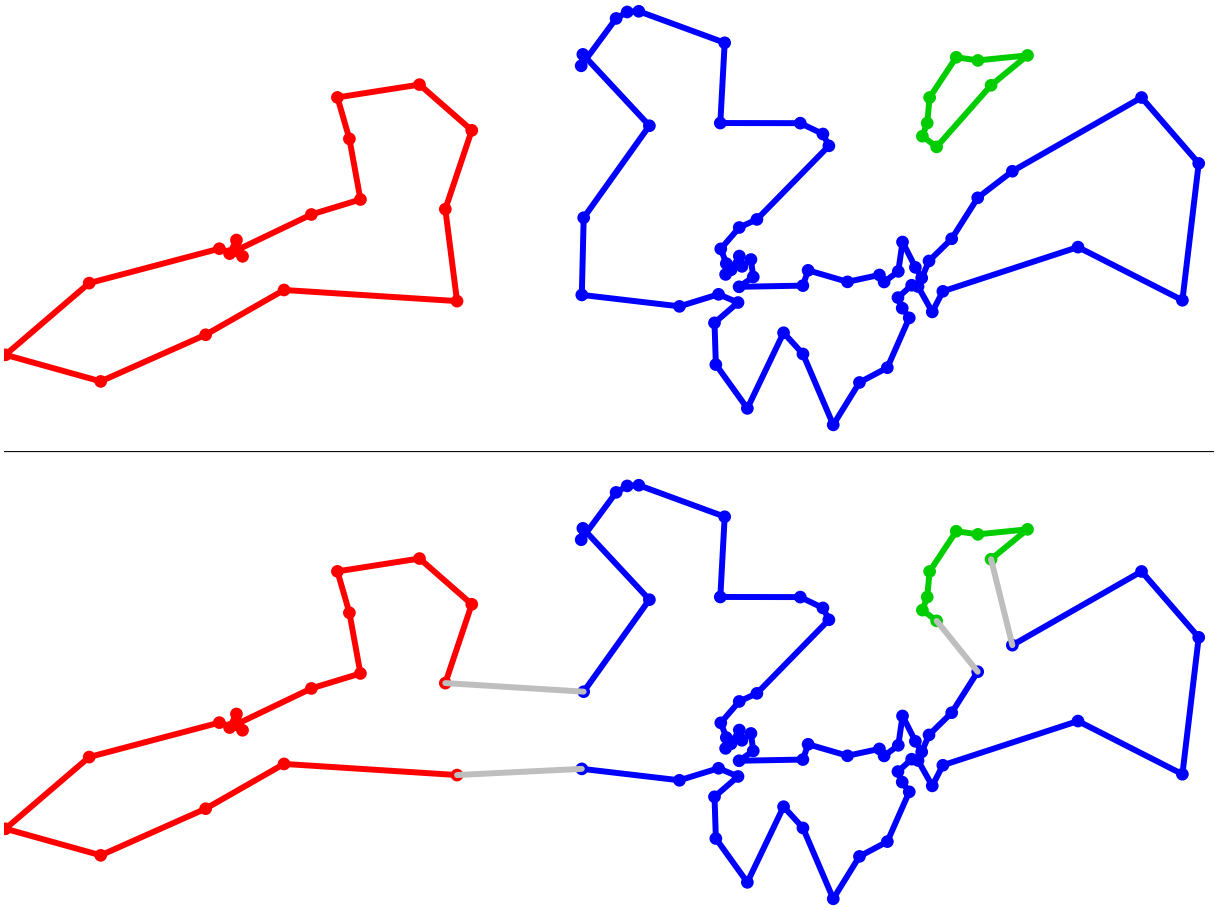


Le découpage final obtenu est le suivant : chaque sous-section ne peut plus être divisée sans augmenter le nombre de secteurs à sept grappes.

Note de lecture : Parmi les 2 partitions possibles de la section présentées dans la figure 2.2, celle en haut à gauche est retenue car elle maximise la distance entre les centres des deux sous-sections. Les 2 sous-sections ainsi constituées sont à nouveau divisées jusqu'à obtenir le découpage de la section en 4 sous-sections présenté dans la partie inférieure de la figure.

Jonction entre les chemins d'une sous-section Une fois les sous-sections construites, il reste à joindre les chemins des différentes zones, afin de disposer d'un unique chemin de grappes pour la sous-section. La jonction entre les chemins de deux zones implique nécessairement de rompre le chemin de grappes de chacune de ces deux zones. Pour une zone donnée, cette rupture intervient entre deux grappes adjacentes sur le chemin de la zone. Le chemin de la sous-section est alors construit de manière ascendante avec pour objectif de minimiser la longueur de ce chemin : toutes les jonctions entre deux zones de la sous-section sont testées et celle qui conduit au chemin de longueur minimale est retenue. Cet algorithme est ensuite réitéré jusqu'à ce qu'il ne reste qu'un chemin au sein de la sous-section. La figure 2.4 illustre le résultat de cet algorithme.

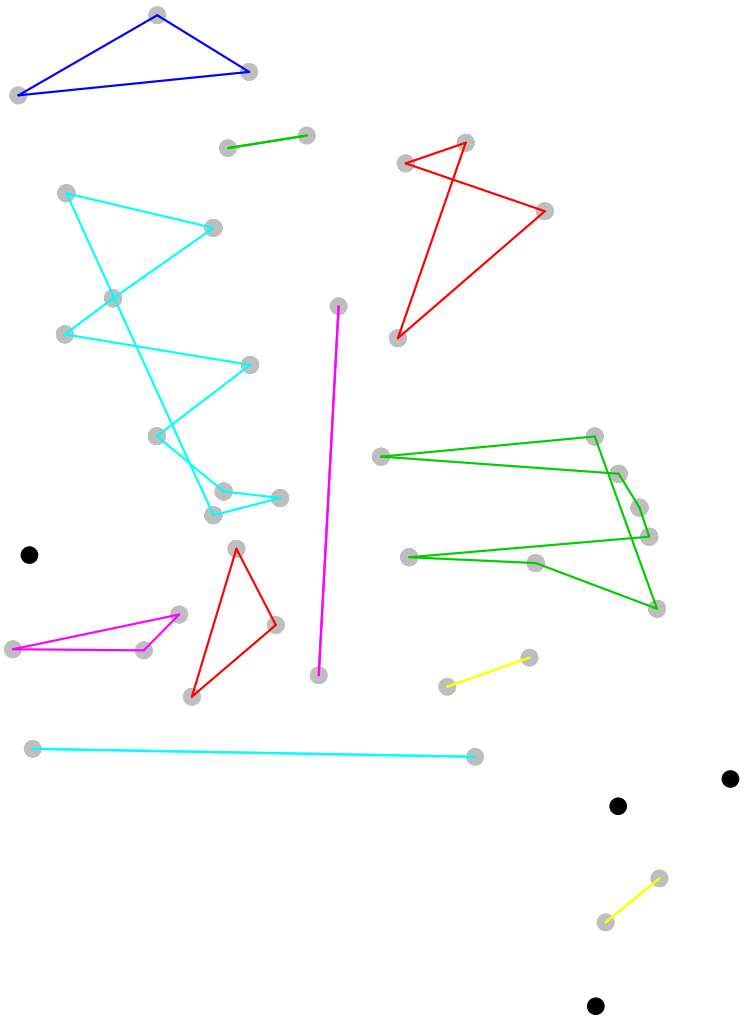
FIGURE 2.4 – Plöermel : Jonction des chemins des trois zones d'une *sous-section*



Note de lecture : Chaque point représente une grappe. Dans le premier graphique, chaque couleur représente le chemin de grappes dans une zone. Dans le second graphique, les segments gris représentent les points de jonction entre zones.

Consolidation Cela étant, dans de rares cas, le regroupement des grappes en secteurs au sein des sous-sections peut conduire à une distance élevée entre deux grappes d'un même secteur, mais de deux zones différentes. C'est pourquoi, après avoir obtenu tous les chemins dans une section, on a recours à une étape de consolidation qui consiste à regrouper deux sous-sections si la longueur du chemin joint est plus petite que la somme des longueurs des chemins passant dans chaque sous-section. La figure 2.5 illustre le découpage de l'UC (ou section) de Plöermel à l'issue de l'ensemble du processus du découpage de la section en sous-sections et de constitution des chemins de grappes dans les sous-sections.

FIGURE 2.5 – Plöermel : Illustration de la partition finale de la section en *sous-sections*

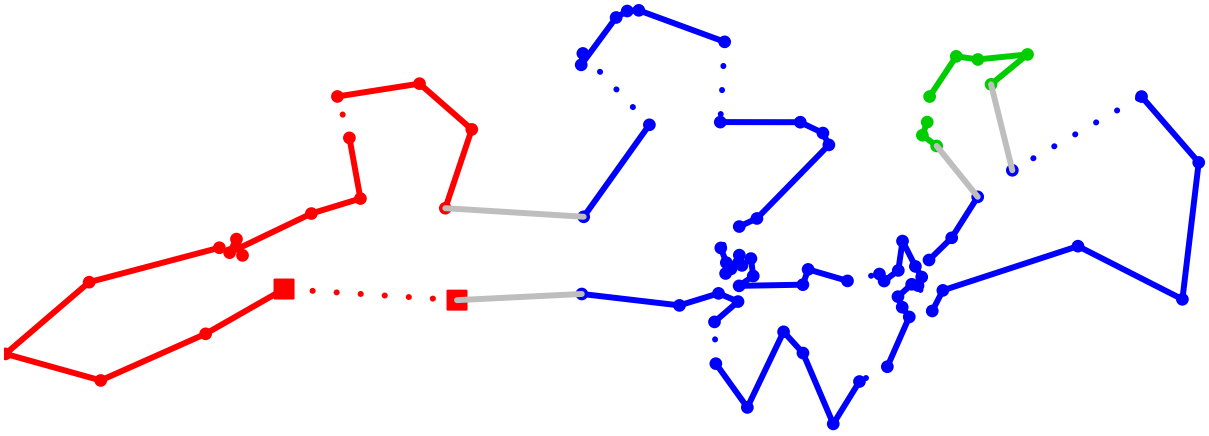


Note de lecture : Chaque point représente le barycentre d'une zone. Les traits de couleur relient les zones d'une même sous-section (l'ordre des zones le long des traits n'est pas signifiant). Les points noirs correspondent à des sous-sections composées d'une seule zone. La sous-section composée de 3 zones dans la figure 2.4 est représentée en rose dans cette figure.

Constitution des secteurs le long du chemin joint La constitution des secteurs à proprement parler s'effectue le long du chemin qui parcourt l'ensemble de la sous-section : à partir d'une grappe de départ judicieusement choisie, les secteurs sont constitués en regroupant 6 (ou plus rarement 7) grappes consécutives (figure 2.6). Plus précisément :

- On détermine, sur le chemin, une grappe de départ et une grappe d'arrivée : il s'agit des deux grappes consécutives les plus éloignées l'une de l'autre. Ce faisant, on garantit que les deux grappes consécutives les plus éloignées l'une de l'autre n'appartiennent pas au même secteur.
- On parcourt le chemin et on crée un secteur tous les 6 ou 7 grappes.
- Si le nombre de grappes dans la sous-section n'est pas divisible par 6, les secteurs à 7 grappes sont, par convention, constitués en début de chemin.

FIGURE 2.6 – Plöermel : Constitution des secteurs au sein d'une *sous-section* de trois zones



Note de lecture : Chaque point représente une grappe. Les traits pleins relient les grappes d'un même secteur, les traits en pointillés les grappes de secteurs différents. Les carrés représentent la grappe de départ et la grappe d'arrivée de l'algorithme de constitution des secteurs.

Effectué sur l'ensemble du territoire, la totalité de cet algorithme permet ainsi de créer des secteurs de 6 grappes dans la plupart des cas, limitant à 5 le nombre de secteurs à 7 grappes au niveau de chaque section. Par cette approche, les zones (Iris, Communes) dans lesquelles on peut constituer des secteurs de 6 grappes uniquement sont maintenues, ce qui permet d'assurer une proximité géographique, renforcée par l'utilisation des chemins solutions du problème du voyageur de commerce, tracés lors de la constitution des grappes (cf. 2.2.2).

2.4 Analyse des grappes obtenues

2.4.1 Elements chiffrés sur la base obtenue

Le fichier Fidéli contient 28 339 449 résidences principales. Après retrait des 26 communes présentes dans cette base mais inhabitées ou difficilement accessibles par le réseau des enquêteurs [20], cette nouvelle base contient 28 331 817 logements. Ont été également retirés de la base les logements ne contenant pas d'information de localisation (c'est-à-dire sans information sur leurs coordonnées cartographiques, X ou Y) soit 49 130 d'entre eux. La base finale de logements pour le nouvel EEC est donc constituée de **28 282 687 logements.**.

On dénombre **21 731 541 étages**, en considérant une maison individuelle comme un étage, regroupés en 17 952 304 adresses.

20. Il s'agit des communes de : Suzan, Île-d'Aix, Île-de-Brehat, Rochefourchat, Île-de-Batz, Île-de-Sein, Île-Molène, Ouessant, Ampriani, Beaumont-en-Verdunois, Bezonvaux, Cumières-le-Mort-Homme, Fleury-devant-Douaumont, Haumont-près-Samogneux, Louvemont-Côte-du-Poivre, Bangor, Groix, Hoedis, Île-d'Houat, Île-aux-Moines, Île-d'Arz, Locmaria, Le Palais, Sauzon, Île-d'Yeu.

1 395 866 grappes ont été créées à partir de cette base, regroupées en 231 966 secteurs dont 4 070 de 7 grappes.

2.4.2 Description et validation des grappes obtenues

Pour rappel, la constitution de la base de sondage devait prendre en compte certaines contraintes répondant à deux objectifs : faciliter la collecte pour les enquêteurs et permettre des estimations de qualité. Des vérifications ont été effectuées sur la base de sondage obtenue, afin de s'assurer de la réalisation de ces objectifs, notamment au regard de l'échantillon tiré en 2009.

Distance de parcours des grappes : analyse « par la route » En premier lieu, la dispersion géographique des grappes a été contrôlée. L'algorithme de constitution proposé permet déjà de limiter la dispersion géographique des grappes en s'appuyant sur un chemin optimisant la distance de parcours à vol d'oiseau entre les logements. Cependant, des topographies particulières peuvent être ignorées avec une telle méthode, dès lors que la distance à vol d'oiseau ne correspond pas à la réalité du terrain (rivière, montagne, autoroute...) [21].

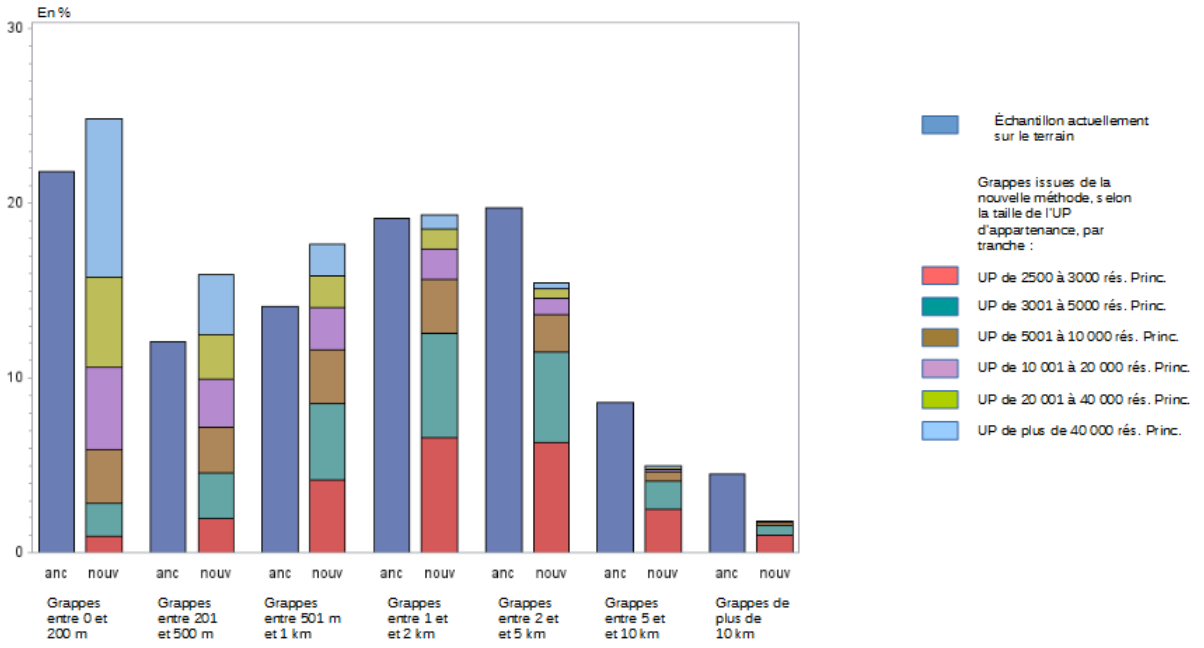
Pour obtenir une mesure « enquêteur » de la taille des grappes, les recherches se sont portées naturellement sur une distance par la route, comme, ce qui a été fait pour constituer les UP du futur échantillon-maître (cf. 1.2.1). La difficulté ici, réside dans le fait que le nombre de grappes est très important, empêchant l'utilisation de requêtes vers un outil commercial ou libre (type Google maps, OpenStreetView ou Géoportail). Nous avons donc opté pour l'outil *Métric*, développé à l'Insee. Cet outil permet notamment de calculer la distance en km et par la route entre deux points (deux couples de coordonnées cartographiques X, Y en Lambert 93) d'un même département. Métric calcule la distance à vol d'oiseau entre chaque point et son nœud routier le plus proche, et l'ajoute à la distance entre les deux nœuds routiers ainsi sélectionnés.

Si l'on voulait obtenir la distance de parcours optimal sur ces grappes, il faudrait tester, pour chaque grappe, toutes les permutations possibles entre les logements d'une grappe. Or, la volumétrie à traiter est bien trop importante pour envisager cette option. L'ordre des logements établi pour la constitution des grappes a donc été utilisé : la distance « par la route » d'une grappe est donc estimée comme la somme, sur l'ensemble des logements de la grappe, des distances données par Métric entre deux logements consécutif sur le chemin, en excluant les logements appartenant au même immeuble.

21. Ces risques sont toutefois limités par le choix de l'utilisation de l'Iris comme zone de constitution des grappes.

Les distances de parcours des grappes sont ainsi calculés [22], tant pour les nouvelles grappes que pour les anciennes, avec la même méthodologie (à savoir calculer une distance point à point, par grappe). Notons que des imprécisions demeurent par rapport aux distances de parcours réels, en particulier dans le choix du chemin de parcours, la position exacte des logements ou le calcul des nœuds routiers. Cela n'invalide pas pour autant le travail de comparaison, qui se fait dans les mêmes conditions pour les deux jeux de grappes : on suppose donc que les imprécisions se compensent. Les distributions de distances pour les grappes de 2009 d'une part et les grappes de 2019 d'autre part sont indiquées en figure 2.7.

FIGURE 2.7 – Comparaison des distances de parcours par grappe, entre les anciennes et nouvelles grappes.



Répartition des grappes en fonction de la distance « enquêteur » pour parcourir la grappe, pour l'échantillon actuellement sur le terrain et pour l'ensemble des grappes issu de la nouvelle méthode de constitution

Lecture : 22 % des grappes de l'échantillon actuel ont une distance de parcours de moins de 200m contre 25 % des grappes issues de la nouvelles méthode de constitution. Parmi ces 25 %, 9 % se situent dans des UP de plus de 40 000 résidence principales

Dans l'ensemble, les grappes de 2019 sont plus compactes que celles de l'échantillon de 2009 [23]. De façon plus précise, on constate que :

— La distance moyenne dans les grappes 2019 (1,5 km) est plus faible que dans l'échantillon de 2009 (2,3 km). C'est le cas pour toutes les anciennes régions.

22. Les temps de parcours des grappes par la route ont aussi été calculés, mais la faible dispersion des grappes conduit à présenter les distances de parcours plutôt que les temps de parcours.

23. On compare donc ici et dans les paragraphes suivants la base des grappes de 2019 et les grappes de l'échantillon de 2009. En effet, la base complète des grappes de 2009 n'est pas disponible.

La distance moyenne est particulièrement diminuée en Corse (elle passe de 6,5 km à 2,7 km).
— De même, sur l'ensemble des grappes, la distance médiane est réduite par rapport à l'échantillon de 2009 (0,6 km contre 1 km). La distance médiane est fortement diminuée en Corse (1,2 km dans les grappes 2019 contre 4,8 km dans l'échantillon de 2009).
— Dans les grappes 2019, 99 % des grappes ont une distance de parcours inférieure à 12,6 km. Dans l'échantillon de 2009, 99 % des grappes ont une distance inférieure à 17 km.
— Dans les grappes 2019, 27,8 % des grappes ont une seule adresse (distance à 0), contre 18,7 % dans l'échantillon de 2009. Ceci se vérifie dans toutes les régions.
— 1,6 % des grappes 2019 ont une distance de parcours de plus de 10 km, contre 4 % dans l'échantillon de 2009.

À noter que l'on travaille au niveau des anciennes régions, car elles correspondent à l'implantation des divisions gestionnaires de l'activité locale des enquêteurs et de la collecte de l'EEC.

Taille des grappes L'autre vérification apportée est relative à la taille des grappes, au sens nombre de résidences principales qu'elles contiennent. Pour rappel, l'objectif est de créer des grappes comprenant de 17 à 24 résidences principales, et préférentiellement 20 résidences principales. La distribution de la taille des nouvelles grappes est nettement plus concentrée. En particulier, les grappes comportant plus de 30 résidences principales sont désormais extrêmement marginales. Plus précisément :
— En moyenne, les grappes 2019 contiennent 20,2 résidences principales, contre 21,5 dans l'échantillon de 2009.
— Dans toutes les régions, 99 % des grappes 2019 ont moins de 24 résidences principales. Dans l'échantillon de 2009, 99 % des grappes ont moins de 33 résidences principales.
— De même, 90 % des grappes 2019 ont moins de 22 résidences principales, alors que dans l'échantillon de 2009, 90 % des grappes ont moins de 26 résidences principales.
— La part des nouvelles grappes comprenant de 17 à 24 résidences principales est de 97,9 % contre 73,5 % dans l'échantillon de 2009. En région PACA 99,3 % des grappes 2019 comprennent 17 à 24 résidences principales. Le minimum de cette proportion est atteint en Champagne-Ardenne où 94,9 % des nouvelles grappes comprennent 17 à 24 résidences principales.
— La part des grappes 2019 ayant plus de 30 résidences principales est très faible : de 0,01 % en moyenne, contre 3,5 % dans l'échantillon de 2009.

Ces résultats, que ce soit sur la distance de parcours des grappes ou sur le nombre de logements par grappe, sont toutefois à relativiser, car l'ajout de résidences non principales (voir section 2.1 de la partie D) dans ces grappes pénalisera nécessairement la qualité de la base. Cela étant, la base de sondage ainsi constituée semble répondre pleinement aux contraintes opérationnelles de l'enquête Emploi en continu.

Chapitre 3

Les unités de coordination

Les renouvellements de l'échantillon-maître (EM) pour les enquêtes auprès des ménages et de l'échantillon Emploi ont été réalisés à quelques trimestres d'écart. Rendue possible par la concomitance des tirages de l'EM et de l'échantillon de l'EEC, et pour répondre à des besoins de simplification de la gestion de la collecte présentés en section 2.2 de la partie A, la coordination spatiale des tirages des deux échantillons a été étudiée, dans le but que l'échantillon-maître et l'échantillon Emploi soient tirés dans des zones proches géographiquement. La coordination spatiale de deux échantillons peut être envisagée par plusieurs moyens, soit à travers des tirages prenant en compte une dimension géographique, soit grâce à plusieurs degrés de tirage, le premier permettant de déterminer la zone géographique utilisée pour les tirages suivants [1]. Un des objectifs que s'est fixé l'Insee au moment de redéfinir l'échantillon-maître et l'échantillon de l'EEC était de simplifier la méthode utilisée (par rapport à celle utilisée en 2009) pour la rendre plus intelligible sur le plan statistique [2]. Dans cette optique, l'utilisation de plusieurs degrés de tirage se présentait comme la meilleure solution, et seule celle-ci a été étudiée. Une fois choisie l'utilisation d'un degré de tirage supplémentaire, au même titre que pour la partition du territoire en unité primaire, la première étape consiste à définir les zones géographiques qui ont été retenues pour concentrer les secteurs Emploi et les unités primaires tirés dans chacun des deux échantillons, que l'on appelle unités de coordination.

La coordination spatiale diminue *a priori* la précision des échantillons puisque les tirages se font avec des contraintes supplémentaires. Un des critères de choix de la méthode de coordination spatiale était donc de préserver au maximum la précision de l'échantillon Emploi et de l'échantillon-maître, ce qui a été introduit dans les critères de choix pour les méthodes de tirage. D'abord, nous expliquons le choix de créer ces zones comme regroupement d'UP, plutôt que de tirer les secteurs directement dans les UP créées pour l'EM (chapitre 3.1). Puis, nous détaillons les enjeux autour du choix du seuil minimal de résidences principales (fixé à 10 000 en l'occurrence) par unité de coordination (chapitre 3.2). Enfin, nous présentons l'algorithme de constitution de ces unités de coordination (chapitre 3.3).

1. On retrouve ici le principe de l'échantillon-maître

2. En particulier, il s'agissait de s'affranchir de contraintes sur les tirages telles que celles générées par les groupes de recensement de la population qui impliquaient un calage sur marges annuel des ZAE.

3.1 Comment coordonner les tirages ?

Pour définir des zones où tirer les secteurs de l'enquête Emploi en Continu et les autres enquêtes auprès des ménages de manière coordonnée, deux solutions se présentent :

— Tirer les secteurs de l'EEC directement dans les unités primaires de l'échantillon-maître

— Créer des zones contenant les UP de l'EM d'une part, et où seront tirés d'autre part les secteurs de l'EEC.

3.1.1 Tirer l'échantillon de l'EEC au sein des UP de l'EM : une stratégie naturelle, mais peu exploitable en pratique

La solution la plus naturelle pour coordonner spatialement les deux échantillons est de tirer les secteurs de l'EEC au sein des unités primaires tirées pour l'échantillon-maître. Par construction des unités primaires dont l'étendue [3] se veut réduite, cette solution a l'avantage de garantir la proximité entre secteurs de l'EEC et des logements enquêtés pour les autres enquêtes auprès des ménages. Elle présente toutefois trois limites importantes : la précision de l'enquête Emploi, la précision des enquêtes auprès des ménages et la durabilité de l'échantillon-maître.

3.1.1.1 Un échantillon Emploi trop peu précis

Si l'on tire l'échantillon Emploi au sein des unités primaires de l'échantillon-maître, alors on effectue un tirage à deux degrés.

— D'abord, on tire les unités primaires qui vont constituer l'échantillon-maître. Ce premier degré de tirage permet d'atteindre, pour l'enquête Emploi, les secteurs Emploi des unités primaires de l'échantillon-maître.

— Dans un second temps, on tire l'échantillon de l'EEC parmi les secteurs des unités primaires faisant partie de l'échantillon-maître. Il s'agit du deuxième degré de tirage.

Le fait d'effectuer un tirage à plusieurs degrés pour tirer l'échantillon de l'EEC dégrade la précision de ce dernier, car la variance des indicateurs calculés à partir de l'échantillon Emploi intègre l'aléa lié à chacun des deux degrés de tirages. En particulier, plus le premier degré de tirage (le tirage des unités primaires) réduit le nombre de secteurs Emploi disponibles pour le deuxième degré de tirage (tirage de l'échantillon Emploi), moins les indicateurs produits grâce à l'enquête seront précis [4].

3.1.1.2 Une perte de précision pour les enquêtes ménage due à un recouvrement trop fort entre l'échantillon Emploi et l'échantillon-maître

Étant donné que l'Insee cherche à minimiser la charge d'enquête imposée aux ménages et ainsi à augmenter la qualité des réponses des enquêtés et diminuer la non-réponse,

3. L'étendue des UP est définie au point 1.2.5.

4. Ceci est assimilable à un effet de grappe, cf. Ardilly, 2006. Les secteurs d'une même unité primaire étant proches géographiquement, ils ont des caractéristiques similaires, ce qui réduit d'autant plus la précision de l'échantillon de secteurs au deuxième degré.

il n'est pas souhaitable d'interroger des ménages tirés pour l'enquête Emploi dans le cadre d'autres enquêtes. L'enquête Emploi est, en effet, assez chronophage pour les ménages qui sont interrogés 6 trimestres de suite. Le fait de tirer des secteurs Emploi dans les unités primaires de l'échantillon-maître diminue donc le nombre de logements pouvant être tirés pour les autres enquêtes auprès des ménages. Cette diminution du nombre de logements disponibles dans la base de sondage induit naturellement une perte de précision pour les enquêtes auprès des ménages. Plus le recouvrement entre l'échantillon-maître et l'échantillon Emploi est important, moins les logements échantillonnables pour les enquêtes auprès des ménages de l'Insee sont nombreux et moins les enquêtes auprès des ménages seront précises [5].

3.1.1.3 Un épuisement trop rapide de l'échantillon-maître

L'épuisement de l'échantillon-maître est une notion liée à des problématiques de terrain. Le critère de 2 500 résidences principales a été retenu pour la constitution des unités primaires, afin que les logements des unités primaires ne soient interrogés qu'une seule fois durant les 10 ans de l'échantillon-maître. Pour évaluer s'il est possible de tirer les secteurs Emploi dans l'échantillon-maître, il convient d'abord de vérifier si les secteurs Emploi tirés dans les unités primaires ne limitent pas trop le nombre de logements restants pour les enquêtes auprès des ménages. Des simulations de tirage ont donc été réalisées permettant de savoir à quel point les UP seraient épuisées rapidement lorsqu'on tire 3 216 secteurs dans 567 unités primaires [6]. Dans 50 % des simulations, ce sont plus de 9 %, plus de 21 % et plus de 28 % des unités primaires qui sont épuisées respectivement au cours de la 8^e, la 9^e et la 10^e année d'utilisation de l'échantillon-maître comme présenté dans la figure 3.2 du point 3.2.1. Dans de nombreux tirages, le fait de tirer des secteurs dans des unités primaires ne garantit donc pas la durabilité de l'échantillon-maître. Si trop de secteurs Emploi sont tirés dans une petite unité primaire de l'échantillon-maître, il restera alors trop peu de logements dans cette unité primaire pour qu'elle soit utilisable durant 10 ans sans réinterrogation.

La solution de coordination qui consiste à tirer l'échantillon Emploi au sein des unités primaires de 2500 résidences principales de l'échantillon-maître n'est donc pas satisfaisante, car elle dégrade la précision de l'ensemble des enquêtes et ne permet pas de garantir une durée d'utilisation de 10 ans de l'échantillon-maître. Elle n'a ainsi pas été retenue.

3.1.1.4 Pourquoi ne pas augmenter le nombre de résidences principales par unité primaire ?

Si la taille des UP semble la principale cause des inconvénients de la méthode, la question de construire des UP de plus grande taille se pose naturellement. On peut, par exemple, choisir un seuil qui garantit la présence de 2 500 résidences principales mobilisables par unité primaire pour les enquêtes auprès des ménages après tirage de

5. Concrètement, l'échantillon Emploi serait alors un degré de tirage supplémentaire entre le tirage des unités primaires et celui des logements. Les échantillons de logements pour les enquêtes auprès des ménages en interrogation en face à face seraient alors tirés après trois degrés de tirage. Le fait que le tirage de l'échantillon Emploi soit un tirage par grappes implique que les logements qui resteraient disponibles pour l'échantillonnage des enquêtes auprès des ménages, après tirage de l'échantillon Emploi, seraient également sélectionnés par grappes. Cela induirait une perte de précision.

6. Ce qui correspond aux nombres d'unités primaires et de secteurs dans les échantillons de 2009.

l'échantillon de l'EEC au sein de ces nouvelles unités primaires de l'échantillon-maître. Si cela apporte une solution aux problèmes de durabilité de l'échantillon-maître et de précision des enquêtes, l'étendue des UP ainsi construite augmente inévitablement ; elle sont alors plus difficiles à couvrir du point de vue de la collecte des enquêtes auprès des ménages. Si la coordination a pour objectif de rapprocher les secteurs de l'EEC des autres ménages collectés, la problématique de réduction des déplacements se pose principalement pour les autres enquêtes. Pour l'EEC, les logements enquêtés sont, par construction, regroupés dans des mêmes grappes de 20 logements environ, géographiquement proches, alors que pour les autres enquêtes, 20 logements sélectionnés peuvent être répartis sur toute l'unité primaire ; il est donc capital de réduire en priorité l'étendue de ces UP [7]. Ainsi, une telle solution aurait simplifié la collecte de l'enquête Emploi au détriment de la collecte des autres enquêtes auprès des ménages. Enfin, les travaux autour de la constitution des unités primaires étaient avancés et validés lorsque la question de la coordination des échantillons a émergé. La validation des contours des UP sollicitant un certain nombre d'acteurs en Directions Régionales de l'Insee, il était préférable de ne pas modifier le nombre minimal des résidences principales pour les unités primaires.

3.1.2 Les unités de coordination : une solution pour coordonner les tirages de l'échantillon-maître et de l'enquête Emploi

Le point 3.1.1 a permis de mettre en lumière que la durabilité de l'échantillon-maître et la précision de l'échantillon-maître et celle de l'échantillon Emploi n'étaient pas suffisantes lorsqu'on tire les secteurs Emploi au sein des unités primaires de 2 500 résidences principales. Le fait de restreindre la base de sondage des secteurs de l'EEC et de diminuer le nombre de logements disponibles pour les autres enquêtes rend cette solution difficilement acceptable. Il est donc nécessaire de trouver une solution limitant l'impact du tirage de l'EEC sur les autres enquêtes auprès des ménages, permettant de mobiliser une base de sondage de secteurs suffisamment conséquente, tout en réduisant autant que faire se peut les distances à parcourir pour les enquêteurs.

3.1.2.1 Concept d'unité de coordination

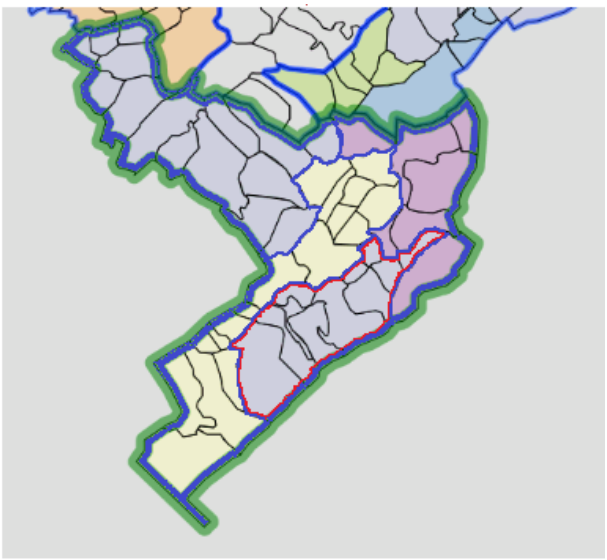
Afin de coordonner spatialement les tirages de l'EM et de l'EEC, on souhaite trouver une méthode permettant de tirer chacun des échantillons à proximité, sans pour autant que les secteurs de l'EEC soient totalement inclus dans les UP de l'EM [8]. Ainsi, on souhaite tirer les secteurs au voisinage des unités primaires de l'EM. La question revient donc à définir ce voisinage. La solution vient naturellement : plutôt que de tirer les secteurs Emploi uniquement dans les unités primaires de l'échantillon-maître, on va également rendre possible la sélection de secteurs Emploi dans les unités primaires voisines de celles de l'échantillon-maître. On élargit ainsi la base de sondage de secteurs Emploi, on limite le nombre de secteurs tirés dans l'EM, améliorant la précision de celui-ci et permettant de l'épuiser moins rapidement. Enfin, bien que la zone de sélection des secteurs emploi s'en trouve élargie, on limite le territoire couvert par les enquêteurs. Deux questions se posent

7. Augmenter la taille des UP signifie que les enquêteurs de l'Insee auraient plus de déplacements à effectuer pour effectuer le repérage et pour réaliser les enquêtes en face à face.

8. Tout particulièrement pour les petites unités primaires, plus rapidement épuisées.

alors : comment construire ce voisinage et comment définir la taille de ce voisinage. Pour construire le voisinage d'une UP, la partition du territoire a offert une maille naturelle : regrouper les UP pour constituer les unités de coordination (UC). L'idée est la suivante : les unités primaires sont regroupées entre unités primaires voisines, de manière à former de nouvelles zones. Afin de coordonner spatialement les tirages, un secteur Emploi ne peut être tiré que dans une unité de coordination dans laquelle une unité primaire est tirée et réciproquement. La figure 3.1 donne un exemple d'unité de coordination. Sur cet exemple, l'unité primaire aux contours rouges (de faible taille en nombre de logements) est réunie avec d'autres unités primaires voisines pour former une unité de coordination de taille suffisante, représentée avec des contours verts. Si cette unité primaire fait partie de l'échantillon-maître, sélectionner beaucoup de secteurs Emploi en son sein risque d'épuiser rapidement cette zone en termes de logements. Si le même nombre de secteurs Emploi sont sélectionnés dans l'unité de coordination associée à cette unité primaire, alors l'UP rouge durera plus longtemps, et les secteurs Emploi seront tirés dans une base de plus grosse taille. Bien évidemment, deux UP de l'EM pourraient appartenir à une même unité de coordination, mais nous verrons en section 1.2 présentant les algorithmes de tirage dans la partie C que l'EM doit être le plus spatialement réparti possible, limitant la possibilité d'avoir plusieurs unités primaires adjacentes tirées en même temps. Ce risque est donc limité.

FIGURE 3.1 – Découpage d'un territoire en unités primaires (contours bleus) et sélection de l'une d'elles (contour rouge), unité de coordination (contour vert) associée à l'unité primaire sélectionnée - Exemple dans le Doubs



3.1.2.2 Les unités de coordination : un bon compromis pour la collecte

Cette construction d'unités de coordination comme regroupement d'unités primaires a été retenue pour la coordination spatiale des tirages de l'échantillon-maître et de l'échantillon Emploi. Elle permet une meilleure précision des échantillons des enquêtes auprès des ménages, y compris EEC, que si les secteurs avaient été sélectionnés uniquement dans les unités primaires de l'EM, car elle augmente la taille de la base de sondage de l'EEC (grâce aux secteurs appartenant aux unité de coordination, mais pas à l'EM),

et permet une diminution moins importante du nombre de logements disponibles pour les autres enquêtes ménages (moins de secteurs étant tirés dans l'EM). De ce fait, elle permet également une durée de vie de l'EM (i.e. sans réinterrogation d'un logement pour deux enquêtes différentes) plus importante (une dizaine d'années). Enfin, si l'étendue du territoire à couvrir est plus importante pour l'EEC qu'en cas de tirage au sein des UP de l'échantillon-maître, elle reste limitée [9], et n'a pas d'impact sur le territoire à couvrir pour l'échantillon-maître. Ces éléments dépendent toutefois de la taille choisie (en nombre de résidences principales) pour les unités de coordination. Il convient donc de définir le seuil minimal de constitution d'une unité de coordination.

3.2 L'importance du choix du seuil de nombre de résidences principales

Comme les unités primaires sont formées au minimum de 2 500 résidences principales et que les unités de coordination sont des regroupements d'unités primaires, les seuils testés pour les unités de coordination sont des multiples de 2 500. Plusieurs seuils ont été testés : 5 000, 7 500, 10 000, 12 500 et 15 000 résidences principales.

3.2.1 Avoir un seuil élevé

Plus le seuil de constitution des unités de coordination est élevé, moins les unités primaires s'épuisent vite et moins l'échantillon-maître et l'échantillon Emploi se recouvrent. Ainsi, plus le seuil de constitution des unités de coordination est élevé, plus l'échantillon-maître dure longtemps et meilleure sera la précision des enquêtes auprès des ménages. Ceci s'illustre par les figures 3.2 et 3.3.

9. Pour rappel, en 2009, les secteurs de l'EEC ont été tirés sur l'ensemble du territoire.

FIGURE 3.2 – Épuisement de l'échantillon-maître dans le cas d'un tirage des secteurs au sein des unités de coordination

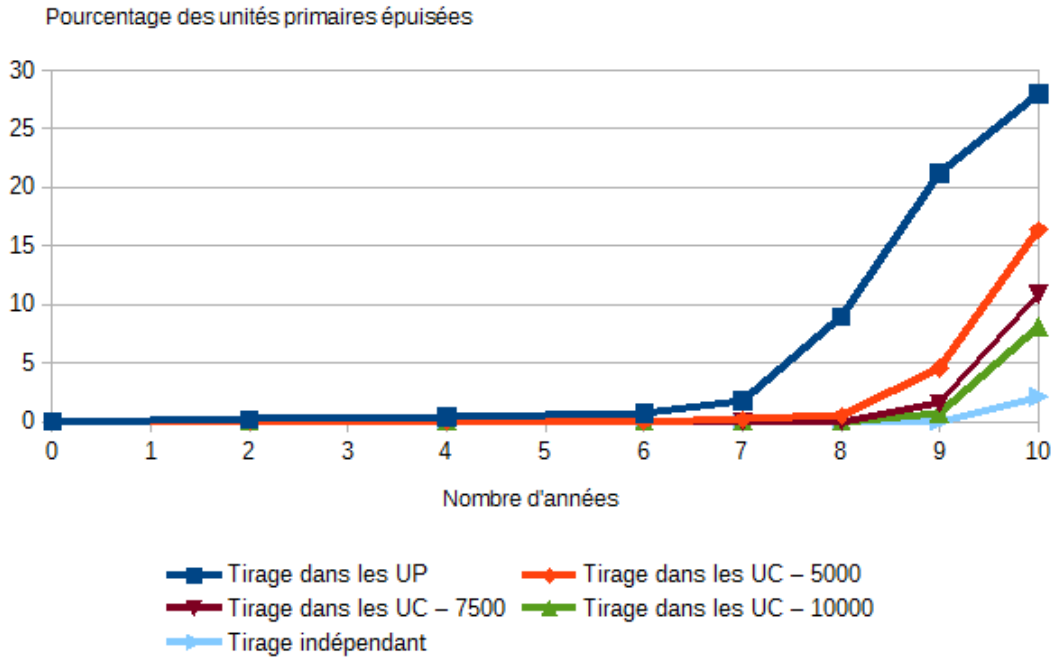
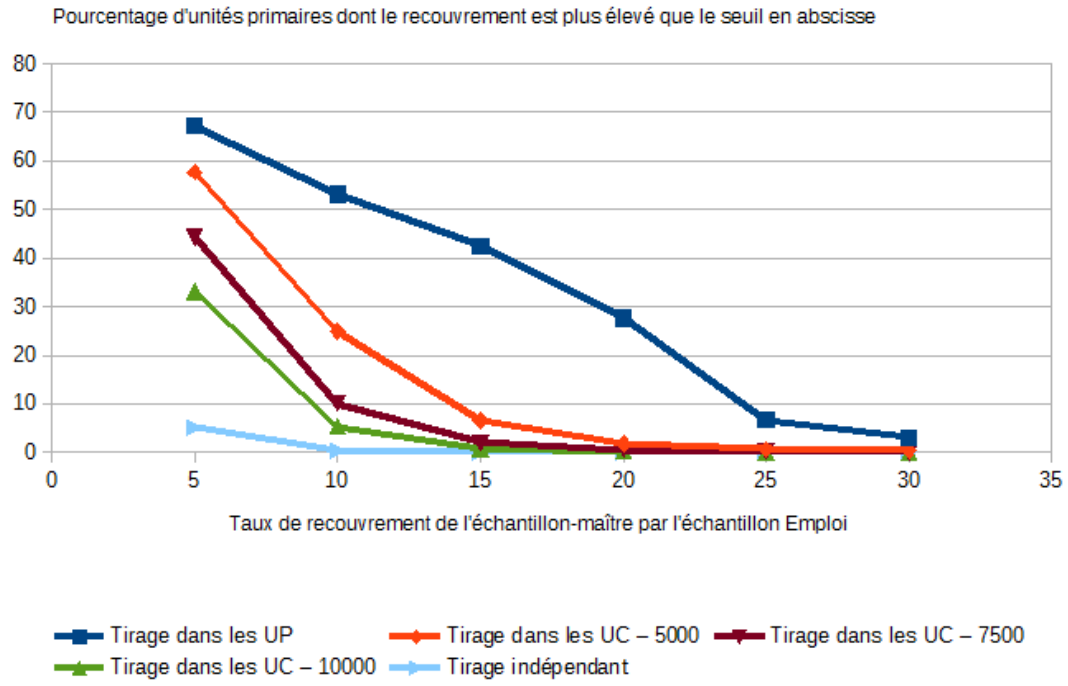


FIGURE 3.3 – Recouvrement de l'échantillon-maître et de l'échantillon Emploi dans le cas d'un tirage des secteurs au sein des unités de coordination



Elles présentent respectivement l'épuisement médian des unités primaires et le

recouvrement médian entre les échantillons. Ainsi, d'après la figure 3.2, dans 50 % des 14 000 simulations de tirage de 3 216 secteurs et de 567 UP avec des unités de coordination (UC) construites avec un seuil de 5 000 résidences principales, plus de 4,6 % des unités primaires sont épuisées au bout de 9 ans d'utilisation de l'échantillon-maître. D'après la figure 3.3, dans 50 % des 15 000 simulations de tirage de 3 216 secteurs et de 567 UP avec des unités de coordination construites avec un seuil de 7 500 résidences principales, plus de 10,1 % des unités primaires présentent un taux de recouvrement de 10 % ou plus [10]. Ces éléments amènent à opter pour un seuil suffisamment élevé, de 7 500 ou 10 000 résidences principales au minimum par UC.

3.2.2 Diminuer le seuil

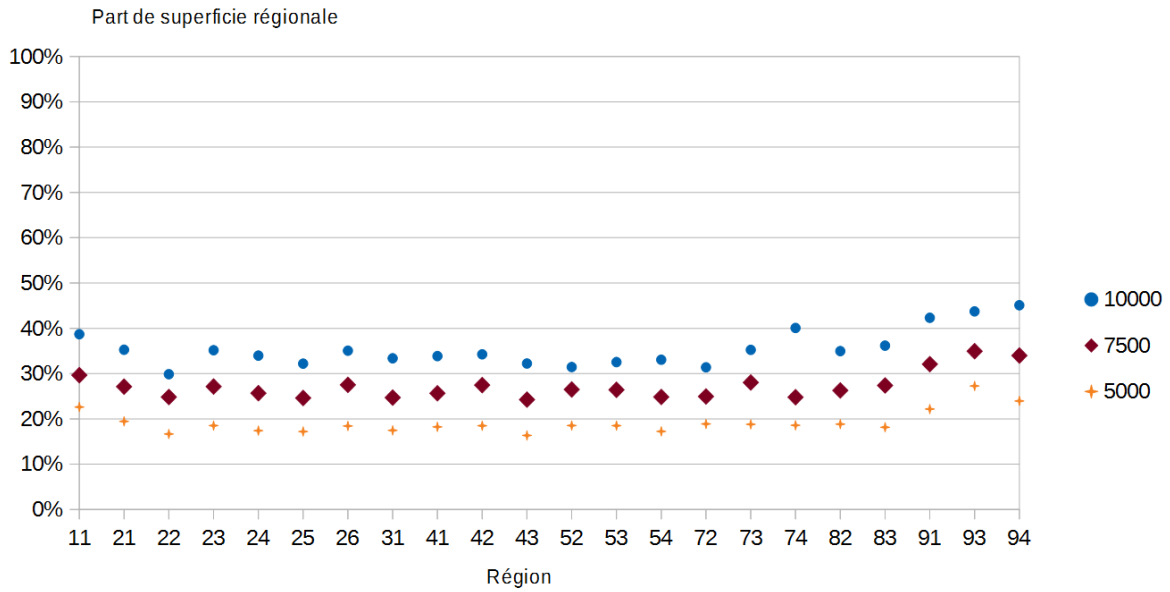
Cependant, l'objectif principal de la coordination est de diminuer les déplacements des enquêteurs en restreignant le territoire couvert. Ainsi, plus le seuil de nombre de résidences principales par unité de coordination est élevé, plus l'enquêteur est amené à effectuer de longs déplacements pour les repérages et la collecte en face à face auprès des ménages. On cherche donc à créer des unités de coordination qui partitionneront suffisamment finement le territoire. Sur la figure 3.4, on constate logiquement que plus le seuil est élevé plus la superficie de territoire couvert se rapproche de celle du département. Un seuil de 15 000 résidences principales conduit à n'avoir que trois unités de coordination ou moins dans 9 départements, un seuil de 12 500 dans 6 départements, ce qui apporte un gain limité de déplacement aux enquêteurs.

Il a donc été décidé de ne pas construire d'unités de coordination regroupant plus de 4 unités primaires [11], ce qui correspond à un seuil inférieur ou égal à 10 000 résidences principales minimum par UC. La figure 3.4 montre que le seuil de 10 000 résidences principales permet de se restreindre à des parts de territoire couvertes par les unités de coordination tirées variant entre 30 % et 45 % selon les régions dans 95 % des tirages. Ce seuil, bien que moins idéal que les seuils de 5 000 et 7 500 résidences principales, assure que, même dans un cas de tirage défavorable, les secteurs Emploi et les unités primaires sont concentrés dans moins de 50 % du territoire de chaque région.

10. Ces résultats ont été obtenus pour un scénario de tirage indirect des unités de coordination, avec tirage spatialement équilibré des secteurs avec une phase de vol par unité de coordination et un atterrissage régional. Ces paramètres de tirage sont expliqués dans les paragraphes 4.3.2.1 et 4.3.2.2 de la partie C.

11. Hormis certains cas exceptionnels en fin d'algorithme de constitution des UC.

FIGURE 3.4 – Part de la superficie couverte dans chaque région par les UC sélectionnées après tirage de 567 UP en fonction du seuil retenu pour la constitution des UC



Note de lecture : Représentation du 95e centile après 100 000 simulations dans le cas d'un tirage d'unités primaires stratifié par région avec calcul des allocations proportionnellement à la taille de la région. Lorsqu'on constitue des unités de coordination d'au moins 10 000 résidences principales, 95% des tirages aboutissent à la sélection d'unités de coordination couvrant moins de 39% du territoire d'Île-de-France.

3.2.3 Le choix du seuil de nombre de résidences principales

En résumé, le nombre minimal de résidences principales relève d'un arbitrage entre étendue des zones de collecte et qualité statistique de ces zones. Il découle des études menées que les seuils de 7 500 et 10 000 résidences principales sont les compromis les plus intéressants. Le seuil de 10 000 résidences principales a finalement été retenu sur le critère de précision de l'échantillon EEC, comme présenté dans le paragraphe 4.3.2.3 de la partie C.

3.3 Méthode de construction des unités de coordination

Le point 3.1.2 a introduit les unités de coordination qui sont des regroupements d'unités primaires voisines et qui contiennent au minimum 10 000 résidences principales. Les UC forment donc une partition de l'espace géographique, chaque commune appartenant à une seule UC. Dans cette partie, nous détaillons comment les unités primaires ont été associées pour former des unités de coordination.

De même que pour les unités primaires, les unités de coordination ont été construites par département. Une unité de coordination ne peut contenir des unités primaires et donc des communes que d'un seul département. *A fortiori*, une unité de coordination ne contient des communes que d'une seule région de gestion et d'une seule région administrative [12]. Ces choix ont pour objectif de simplifier l'organisation de la collecte pour les directions régionales. Par ailleurs, les grandes communes de plus de 40 000 habitants, qui chacune forme à elle seule une unité primaire, constituent également chacune une unité de coordination. Elles ne sont pas réunies avec d'autres unités primaires.

L'objectif de la construction des unités de coordination est de minimiser leur étendue géographique pour limiter les déplacements des enquêteurs, tout en s'assurant qu'elles contiennent un minimum de résidences principales (en l'occurrence 10 000). Les objectifs sont donc similaires à ceux de la constitution des unités primaires présentée dans le chapitre 1 de la partie B. Pour rappel, l'étendue d'une zone (UP/secteur/UC) est la somme des distances routières des communes la constituant à sa plus grande commune en nombre de résidences principales, pondérées par le taux de résidences principales de la commune parmi les résidences principales de la zone. Cet indicateur propose une mesure des déplacements entre la commune principale et les autres communes de la zone, en fonction de la probabilité d'avoir un logement tiré dans ces autres communes. Ainsi, plus une commune est grosse et éloignée de la commune principale, plus sa contribution à l'étendue sera importante.

Pour obtenir le meilleur regroupement d'UP en UC, la solution idéale aurait été de tester tous les découpages possibles d'unités primaires en unités de coordination d'au minimum 10 000 résidences principales. Or, comme pour la constitution des UP décrite au chapitre 1 de la partie B, même limitée au niveau des départements, cela pose des problèmes computationnels. Par exemple, dans le département du Nord, il y a 176 unités primaires constituées de communes de moins de 40 000 résidences principales : il faudrait tester 176 ! regroupements possibles. En réalité, cela ne s'avère pas nécessaire, car un premier ordonnancement des communes a déjà été proposé pour limiter le déplacer de l'une à l'autre, lors de la constitution des UP.

L'idée retenue consiste ainsi à exploiter le même chemin de communes que celui qui a permis le découpage des départements en unités primaires. Ce chemin, obtenu par un algorithme du voyageur de commerce décrit au point 1.2.4 et illustré par les figures 1.1 et 1.2, avait été sélectionné parmi 1 000 chemins du voyageur de commerce (obtenus par simulations) ayant chacun un point de départ différent. Le chemin retenu était celui qui minimisait l'étendue des unités primaires. Puisque l'algorithme de construction des unités primaires minimisait les distances entre communes par la route, cela implique

12. Il s'agit des régions administratives précédent la réforme territoriale de 2015. Cela correspond à la nomenclature NUTS 2.

que deux unités primaires successives sur le chemin sont soit voisines soit proches par la route. Ce chemin entre communes peut donc être vu comme un chemin entre unités primaires. Il est possible de regrouper les unités primaires se suivant sur ce chemin en unités de coordination. Les unités de coordination qui en découlent sont nécessairement peu étendues puisqu'elles regroupent des unités primaires voisines. Elles ne sont certes pas nécessairement optimales, et d'autres associations d'unités primaires auraient pu être testées, mais elles répondent aux contraintes de collecte : obtenir des unités de coordination de 10 000 résidences principales ou plus regroupant des unités primaires proches par la route et pouvant être couvertes facilement par les enquêteurs.

Il est toutefois possible d'améliorer à coût modique la solution proposée en parcourant ce chemin dans les deux sens et avec une unité primaire de départ aléatoire. Ceci permet de disposer simplement de plusieurs découpages en unités de coordination et de choisir celui qui minimisera la somme des étendues [13] des UC obtenues.

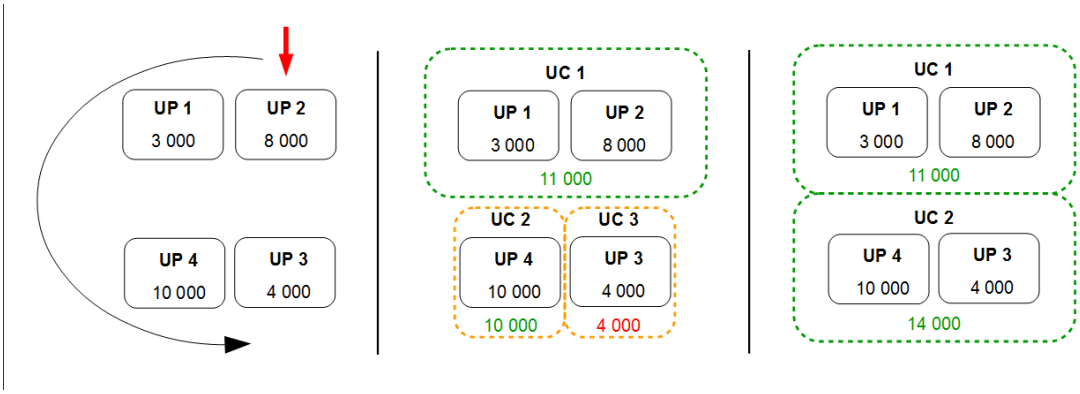
L'algorithme de constitution des unités de coordination d'un département pour un seuil de 10 000 résidences principales est le suivant :

1. Choix d'une unité primaire de départ et d'un sens de parcours du chemin
   - Parcours du chemin. À chaque fois qu'une unité de coordination a été constituée, les unités primaires suivantes sont parcourues jusqu'à ce que la somme des résidences principales dépasse 10 000. Ces unités primaires sont alors regroupées en une unité de coordination.
   - Fin du chemin. Lorsqu'on arrive à la dernière unité primaire, les unités primaires qui n'ont pas encore été affectées à une unité de coordination sont regroupées. Si la somme de leurs résidences principales dépasse 10 000, elles sont agrégées en unité de coordination. Sinon, elles sont ajoutées à la dernière unité de coordination constituée.
2. Calcul de l'étendue de chaque unité de coordination et calcul de la somme des étendues des unités de coordination sur le département.
3. Le découpage retenu est celui qui minimise la somme des étendues des unités de coordination sur le département.

L'algorithme est illustré dans la figure 3.5

13. L'étendue des unités de coordination est définie de la même manière que pour les unités primaires. Cet indicateur est défini par la formule suivante pour une unité de coordination *uc* composée des communes *com* ayant N_{com} résidences principales : $\sum_{com \in uc} \frac{N_{com}}{\sum_{com \in uc} N_{com}} D(com, princ)$. *princ* est la plus grande commune de l'unité de coordination en nombre de résidences principales. $D(com, princ)$ est la distance par la route de la commune *com* à la commune *princ*.

FIGURE 3.5 – Illustration de la première étape de l'algorithme de construction des unités de coordination



Note de lecture : Dans l'exemple, on part du chemin de constitution des unités primaires qui ordonne les 4 unités primaires de UP1 à UP4. La flèche rouge indique l'unité primaire de départ, l'UP2, et la flèche circulaire le sens de parcours du chemin. L'UP2 contient moins de 10 000 résidences principales. Elle est donc regroupée à l'UP suivante, l'UP1. Ces 2 UP réunies présentent 11 000 résidences principales, au-dessus du seuil de constitution de 10 000 résidences principales pour les unités de coordination. L'UP1 et l'UP2 forment donc la première unité de coordination. L'UP4 a suffisamment de résidences principales pour constituer elle-même une unité de coordination. Cependant, à la fin du chemin, la dernière UP, l'UP3, n'a que 4 000 résidences principales et ne peut être à elle seule une unité de coordination. Elle est donc regroupée avec la dernière UC constituée sur ce chemin. L'UP3 et l'UP4 forment ainsi l'unité de coordination UC2.

Enfin, une fois les unités de coordination constituées, de rares unités de coordination ont été modifiées à la main, par exemple lorsqu'une unité de coordination était de manière évidente trop étendue.

Si cette solution est tout à fait satisfaisante, d'autres solutions de même valeur existent probablement. Par exemple, il aurait été possible de regrouper en unités de coordination des unités primaires voisines sans qu'elles se suivent sur le chemin. Ou encore, plutôt que de choisir le découpage en unités de coordination sur le seul critère d'étendue géographique, il aurait été possible de retenir un indicateur mixte faisant intervenir l'étendue géographique mais aussi l'homogénéité des unités de coordination (ou leur hétérogénéité) sur des variables liées à l'emploi pour améliorer la précision du deuxième degré de tirage de l'échantillon Emploi. Cette solution aurait toutefois nécessité beaucoup d'investissement pour un résultat incertain, et n'a donc pas été étudiée. Ces questions ne sont cependant pas anecdotiques. En effet, comme nous le verrons au 4.3.2.3 de la partie C, la manière dont sont construites les unités de coordination a un impact sur la précision de l'enquête Emploi.

Au final, 1 646 unités de coordination ont été créées à partir des 5064 unités primaires. La répartition par région est donnée dans le tableau 3.1.

TABLE 3.1 – Nombre d'unités primaires et d'unités de coordination par ancienne région administrative et par région opérationnelle pour la collecte (i.e. région de gestion de chacun des établissements régionaux de l'Insee)

Code	Région	Nombre d'UP (adm)	Nombre d'UP (op)	Nombre d'UC (adm)	Nombre d'UC (op)
11	Île-de-France	455	222	236	141
21	Champagne-Ardennes	123	213	32	66
22	Picardie	199	199	56	56
23	Haute-Normandie	165	231	51	80
24	Centre	257	334	75	107
25	Basse-Normandie	163	163	44	44
26	Bourgogne	173	173	47	47
31	Nord-Pas-de-Calais	297	297	102	102
41	Lorraine	235	235	70	70
42	Alsace	161	161	49	49
43	Franche-Comté	132	132	35	35
52	Pays de la Loire	310	310	99	99
53	Bretagne	313	313	95	95
54	Poitou-Charentes	205	205	55	55
72	Aquitaine	313	313	96	96
73	Midi-Pyrénées	244	244	76	76
74	Limousin	78	78	19	19
82	Rhône-Alpes	552	552	176	176
83	Auvergne	149	149	42	42
91	Languedoc-Roussillon	240	240	74	74
93	PACA	272	272	109	109
94	Corse	28	28	8	8
France métropolitaine		**5 064**	**5 064**	**1 646**	**1 646**

La précision de l'échantillon-maître et de l'échantillon de l'EEC seront présentées dans les chapitres 4 et 5 de la partie C détaillant respectivement les paramètres du tirage des unités de coordination et les allocations de tirage. On peut néanmoins d'ores et déjà constater, dans le tableau 3.2, que le territoire couvert par les unités de coordination sélectionnées est bien inférieur à l'ensemble du territoire métropolitain : dans la grande majorité des régions, cela assure que les unités primaires et les secteurs Emploi sont tirés dans un espace qui couvre moins d'un tiers du territoire régional. C'est donc un gain substantiel par rapport au cas d'un tirage indépendant, dans lequel les secteurs Emploi peuvent être tirés sur tout le territoire, indépendamment de la zone dans laquelle les unités primaires sont tirées, ce qui justifie le recours à cette coordination.

TABLE 3.2 – Part de la superficie régionale administrative couverte par les unités de coordination tirées en production

Code	Région	Superficie
11	Île-de-France	23 %
21	Champagne-Ardennes	42 %
22	Picardie	32 %
23	Haute-Normandie	29 %
24	Centre	33 %
25	Basse-Normandie	31 %
26	Bourgogne	32 %
31	Nord-Pas-de-Calais	21 %
41	Lorraine	28 %
42	Alsace	29 %
43	Franche-Comté	31 %
52	Pays de la Loire	25 %
53	Bretagne	27 %
54	Poitou-Charentes	26 %
72	Aquitaine	23 %
73	Midi-Pyrénées	32 %
74	Limousin	58 %
82	Rhône-Alpes	24 %
83	Auvergne	28 %
91	Languedoc-Roussillon	21 %
93	PACA	25 %
94	Corse	40 %

Partie C

Le tirage des échantillons

Dans la partie B, nous avons présenté les méthodes employées pour constituer les bases de sondage de l'échantillon-maître (partition du territoire en unités primaires de 2500 logements minimum, constituées à partir de regroupements de communes) et de l'échantillon de l'enquête Emploi en continu (base de secteurs, ensemble de 6 grappes de 20 logements environ, géographiquement proches). Nous avons également introduit le concept d'unité de coordination, ensemble d'unités primaires de 10 000 logements minimum, utilisé à des fins de coordination des échantillons pour faciliter la collecte et la gestion des enquêteurs.

Dans cette partie, nous allons élaborer les plans de sondage conduisant à la sélection de l'échantillon-maître et de l'échantillon de l'EEC. Nous rappellerons au chapitre 1 les objectifs recherchés lors de la sélection d'un échantillon, et plus particulièrement dans le cadre présent, puis les méthodes de tirage équilibré et de tirage spatial. Dans le chapitre 2, nous présenterons les travaux ayant conduit à la méthode de tirage de l'EM, puis les premières réflexions autour de l'échantillon de l'EEC seront détaillées au chapitre 3. Le chapitre 4 décrit la mise en œuvre de la coordination des deux échantillons, et les impacts sur les méthodes de tirage, principalement sur celle de l'EEC. Enfin, la détermination des allocations sera établie au chapitre 5.

Chapitre 1

Objectifs d'un échantillon et méthodes de tirage usuelles

1.1 Objectifs recherchés pour l'échantillon-maître et l'échantillon de l'EEC

Lors de la sélection d'un échantillon, l'objectif de l'expert en sondage est de déterminer le plan de sondage qui permettra d'avoir l'échantillon le plus « représentatif » [1] possible de la population pour sa ou ses variables d'intérêt. Différents outils et différentes méthodes sont à sa disposition pour atteindre son but. L'objectif est d'autant plus facile à atteindre que les variables d'intérêt sont bien identifiées et que la base de sondage offre des variables fortement corrélées à celles-ci.

L'EEC vise à observer le marché du travail de manière structurelle et conjoncturelle. La qualité de son échantillon se mesurera donc par la précision des indicateurs liés à l'emploi. La source Fidéli proposant des informations sur ce thème, le cadre est propice à l'obtention d'un échantillon précis.

La problématique posée par l'échantillon-maître, par contre, est plus complexe. En effet, celui-ci est un premier degré de tirage pour un grand nombre d'enquêtes aux thématiques différentes. L'objectif va donc être de trouver des variables corrélées d'une part avec un grand nombre d'enquêtes, d'autre part avec certains indicateurs jugés comme les principaux.

Par ailleurs, ces deux échantillons sont appelés à être utilisés pendant 10 ans environ, ce qui demande de prendre en compte, si possible, des éléments permettant d'assurer sa qualité, malgré des évolutions différenciées entre les territoires. Pour cela, et du fait

1. Un échantillon représentatif pour une population est ici un échantillon tiré par un plan de sondage permettant d'estimer précisément des variables d'intérêt sur le champ de cette population.

de l'auto-corrélation spatiale entre indicateurs de territoires voisins, on cherchera à sélectionner un échantillon réparti sur toute la France métropolitaine.

La sélection des unités de coordination, quant à elle, doit être faite afin de coupler les deux objectifs. En effet, l'échantillon d'UC n'a pas à être représentatif d'une quelconque information en tant que tel, mais doit permettre de sélectionner en son sein un EM et un échantillon de l'EEC répondant aux objectifs énoncés précédemment.

À l'Insee, pour l'EM comme pour l'EEC – et pour d'autre échantillons –, on utilise communément le tirage équilibré. Les nouvelles techniques de sondage, avec l'apparition du tirage spatialement équilibré, offrent de nouvelles perspectives pour répondre à ces différents objectifs. La section suivante présente les méthodes de tirage équilibré et de tirage spatialement équilibré.

1.2 Rappels théoriques sur les méthodes de tirage équilibré et de tirage spatialement équilibré

1.2.1 Description de l'équilibrage et de l'équilibrage spatial

Le principe du tirage équilibré est d'être précis sur les variables d'intérêt corrélées aux variables d'équilibrage, car il vise à restituer exactement les informations disponibles de ces dernières dans la population. En effet, un échantillon S d'une population U équilibré sur la variable de contrôle X respecte la contrainte suivante :

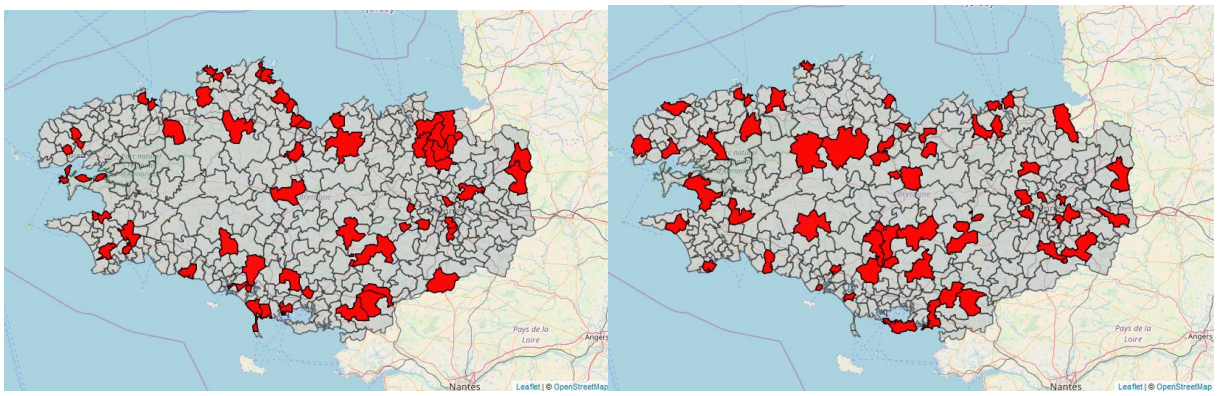
$$\sum_{i\in S}\frac{x_i}{\pi_i}=\sum_{i\in U}x_i \quad \text{soit} \quad \hat{\mathrm{t}}_{\mathrm{x}\pi}=\mathrm{t}_{\mathrm{x}}$$

Ainsi, l'estimateur Horvitz-Thompson $\hat{t}_{x\pi}$ du total de la variable de contrôle X coïncide exactement avec le vrai total t_x issu de la base de sondage : pour une variable qualitative, la proportion de chacune des modalités sera bien respectée par l'échantillon obtenu, et pour une variable quantitative, son total sera bien estimé par l'échantillon [2]. On peut alors dire que l'échantillon est « représentatif » de la population pour les variables X. L'information sur les variables d'intérêt est alors bien restituée grâce à la corrélation entre les variables d'intérêt et les variables d'équilibrage. C'est ce qui explique la capacité du plan de sondage équilibré à améliorer l'efficacité des estimateurs. La méthode du Cube développée par Deville et Tillé, 2004 permet la réalisation d'un échantillon équilibré en deux phases : la phase de vol (marche aléatoire dans l'espace des contraintes où, à chaque étape, l'algorithme arbitre sur l'appartenance à l'échantillon ou non d'une des unités de la base de sondage) et la phase d'atterrissage (permet de choisir un échantillon aussi équilibré que possible en relâchant les contraintes).
Le tirage spatialement équilibré est une « variante » du tirage équilibré dans laquelle on sélectionne peu d'unités proches géographiquement. Cette méthode permet de couvrir

2. En pratique, la variance des estimations sur les variables de contrôle n'est pas nulle en raison de la phase d'atterrissage de l'algorithme détaillé ci-après.

l'espace national (les unités sélectionnées sont alors mieux réparties que lors d'un tirage classique) et de limiter la corrélation spatiale (deux unités géographiquement proches partagent des caractéristiques socio-économiques proches) en plus des avantages cités précédemment pour un tirage équilibré.

FIGURE 1.1 – Exemples de tirage équilibré (à gauche) et tirage spatialement équilibré (à droite).



Cette façon de sélectionner est rendue possible par une méthode qui rend les probabilités d'inclusion double des unités voisines très faibles.

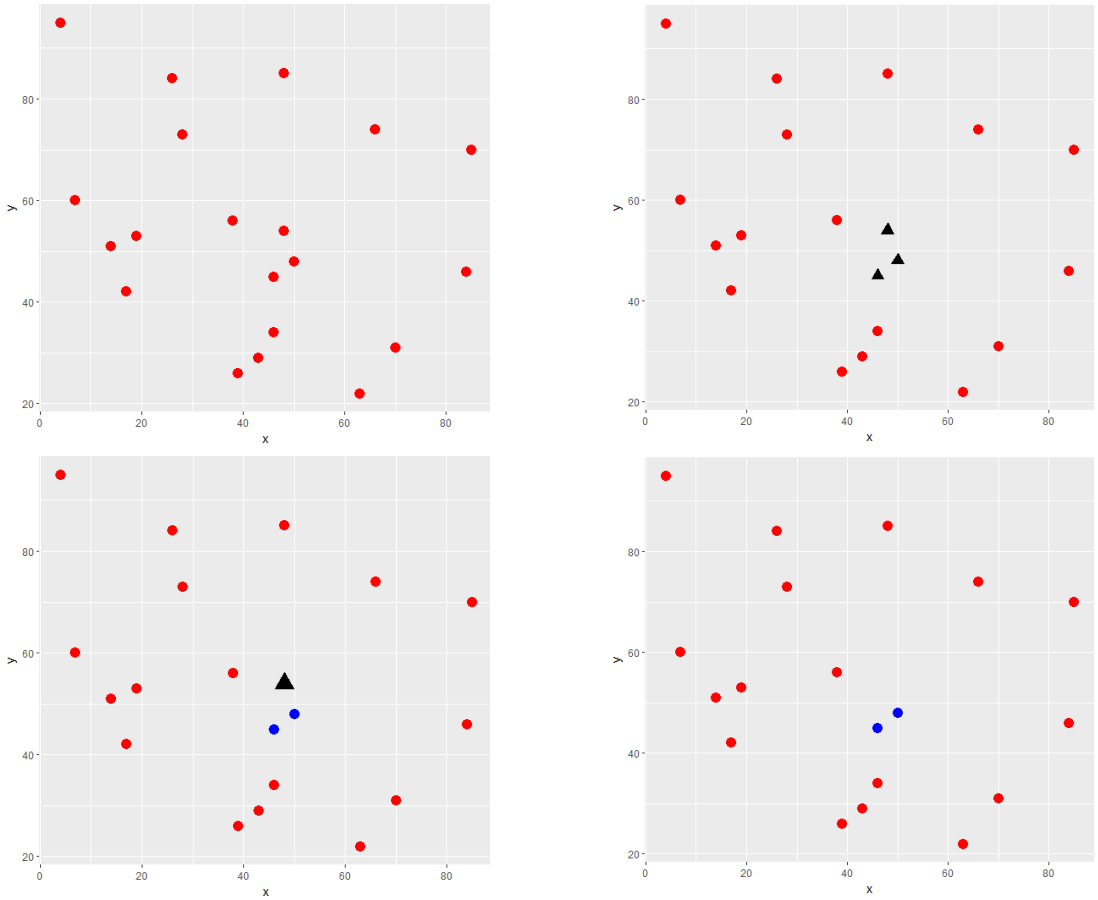
L'algorithme développé par GRAFSTRÖM et TILLÉ, 2013 est le suivant :

- On suppose que l'on équilibre notre échantillon sur p variables d'équilibrage.
- Tant qu'il reste au moins $p+1$ unités pour lesquelles aucune décision n'a été prise sur leur intégration ou non dans l'échantillon :
  - On retient aléatoirement un cluster de $p+1$ unités spatialement voisines pour lesquelles aucune décision n'a été prise sur leur intégration dans leur échantillon.
  - On effectue une phase de vol de la méthode du cube (DEVILLE et TILLÉ, 2004) sur ce cluster.
  - À l'issue de cette phase de vol, on a au moins statué sur la sélection d'une unité. Les autres unités du cluster voient leurs probabilités d'inclusion évoluer en conséquence.
- Lorsqu'il ne reste plus que p unités, la phase d'atterrissage de l'algorithme démarre (DEVILLE et TILLÉ, 2004).

La figure 1.2 montre l'application d'une itération de l'algorithme pour 20 unités de tirage et $p = 2$ variables d'équilibrage.

Dans le cas général, la dispersion de l'échantillon n'est pas nécessairement évidente au premier regard, surtout lorsque les probabilités de sélection sont inégales. Des outils permettent alors de mesurer cette dispersion. On peut par exemple tracer les polygones de Voronoï associés à cet échantillon. Un polygone de Voronoï relatif à une unité géographique contient l'ensemble des points de l'espace plus proches de l'unité considérée que de toute

FIGURE 1.2 – Une itération de l'agorithme de tirage spatialement équilibré.



Note de lecture : Dans l'exemple, 20 unités géographiques sont représentées sur un plan. Initialement, elles ont les mêmes probabilités d'inclusion (cadre en haut à gauche). Comme il y a 2 variables d'équilibrage, un cluster de 3 unités voisines est sélectionné aléatoirement, représenté par un triangle (cadre en haut à droite). L'algorithme statue alors sur une de ces 3 unités (celle représentée par un triangle), en l'occurrence en l'incluant dans l'échantillon. Les 2 unités restantes (ronds bleus) voient leur probabilité de sélection diminuer pour la suite de l'algorithme (cadre en bas à gauche). L'algorithme va ensuite être relancé sur les 19 unités restantes (cadre en bas à droite).

autre unité. La Figure 1.3 montre un exemple de tracé de polygones de Voronoi pour 20 points du plan.

Dans le cadre d'un tirage parfaitement réparti spatialement, lorsqu'on trace les polygones de Voronoï associés aux unités tirées, la somme des probabilités d'inclusion des unités de tirage à l'intérieur de chaque polygone est égale en espérance à 1. Intuitivement, cela signifie que l'on tire une unité par polygone, ce qui correspond à une répartition spatiale uniforme de l'échantillon. La figure 1.4 montre les polygones de Voronoï associés aux unités sélectionnées pour un échantillon. On peut alors déterminer l'écart à la répartition spatiale uniforme en calculant la dispersion de l'écart à 1 des sommes des probabilités d'inclusion de l'ensemble des unités participant au tirage au sein de chacun des polygones.

FIGURE 1.3 – Polygones de Voronoï.

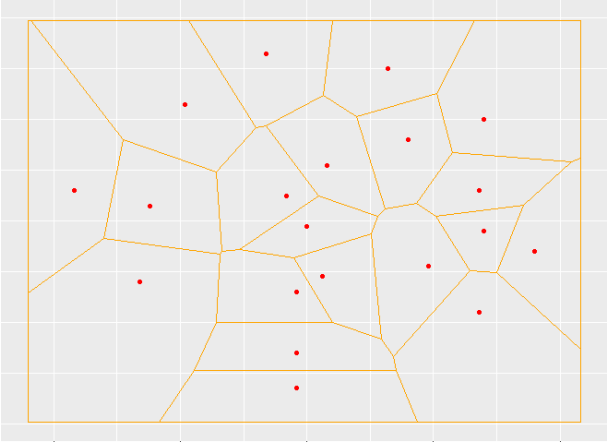
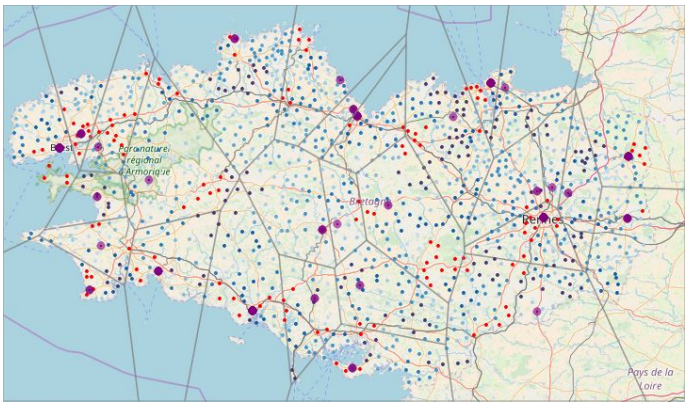


FIGURE 1.4 – Polygones de Voronoï des unités géographiques d'un échantillon en Bretagne.



Note de lecture : Les unités tirées et leurs polygones de Voronoi associés sont en violet.

1.2.2 Cadre théorique

On se place dans le cadre de l'article de GRAFSTRÖM et TILLÉ, 2013. On suppose que la variable d'intérêt de l'enquête y_i est régie par un modèle de la forme :

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \forall i \in U$$

où $\boldsymbol{x}_i$ est un vecteur de p variables auxiliaires connues pour l'unité i, $\boldsymbol{\beta} \in \mathbb{R}^p$ le vecteur des coefficients de régression, et les ϵ_i des variables aléatoires centrées de variance σ_i^2 et de covariance sous le modèle M :

$$\forall (i,j) \in U^2, \ i \neq j, \ cov_M(\epsilon_i, \epsilon_j) = \sigma_i \ \sigma_j \ \rho_{ij}$$

Conformément à l'idée selon laquelle les y_i sont spatialement corrélés, on suppose que $\rho_{ij} = \rho^{d(i,j)}$ avec $d(i,j)$ la distance géographique entre l'unité i et l'unité j et $\rho \in [0,1]$. Cette forme de ρ_{ij} traduit l'existence d'une corrélation spatiale positive pour la variable y qui décroit avec la distance : plus les unités sont géographiquement proches, plus la covariance entre les termes d'erreur est élevée. L'estimateur de Horvitz-Thompson du total

est donné par : $\hat{t}_{y\pi} = \sum_{i \in S} \frac{\boldsymbol{y}_i}{\pi_i}$. Grafström et Tillé, 2013 montrent que la variance de cet estimateur sous le plan de sondage (p) et suivant le modèle (M) ci-dessous est donnée par :

$$\mathbb{E}_p \mathbb{E}_M \{(\hat{t}_{y\pi} - t_y)^2\} = \mathbb{E}_p \left[\left\{ \left(\sum_{i \in S} \frac{\boldsymbol{x}_i}{\pi_i} - \sum_{i \in U} \boldsymbol{x}_i \right)^T \boldsymbol{\beta} \right\}^2 \right] + \sum_{i \in U} \sum_{j \in U} \sigma_i \sigma_j \rho_{ij} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \quad (1.1)$$

Si l'on suppose notre plan de sondage équilibré sur l'ensemble des variables auxiliaires alors on peut écrire :

$$\sum_{i \in S} \frac{\boldsymbol{x}_i}{\pi_i} = \sum_{i \in U} \boldsymbol{x}_i$$

Le premier terme de l'équation (1.1) devient nul, et la variance de l'estimateur d'Horvitz-Thompson se simplifie en :

$$\mathbb{E}_p \mathbb{E}_M \{(\hat{t}_{y\pi} - t_y)^2\} = \sum_{i \in U} \sum_{j \in U} \sigma_i \sigma_j \rho_{ij} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}$$

Pour limiter ce terme, il faut diminuer la probabilité d'inclusion double des unités i et j dont la distance $d(i, j)$ est faible (et dont le ρ_{ij} est élevé), pour abaisser la variance. Ceci est permis par l'algorithme d'équilibrage spatial précédemment détaillé.

Une autre méthode de sondage spatial : le pivot local

Méthode du pivot local

En présence d'autocorrélation spatiale positive, on espère mieux prendre en compte la variabilité des situations et améliorer *in fine* la qualité des estimations en répartissant les unités sélectionnées. Une façon d'obtenir cet étalement spatial est d'introduire un mécanisme de répulsion : des unités proches géographiquement doivent avoir peu de chances d'être présentes simultanément dans l'échantillon ; lorsqu'une unité est sélectionnée, les unités voisines doivent voir leur chance d'être sélectionnées réduite.

Grafström, Lundström et Schelin, 2012 ont proposé une méthode, le pivot local, permettant d'offrir un système de répulsion. Le procédé reprend l'approche introduite par Deville et Tillé, 1998 avec la méthode du pivot. L'idée du pivot est de mettre itérativement à jour les probabilités d'inclusion d'un couple d'unités. À chaque étape, une décision est prise (sélection ou non) pour au moins une unité, ce qui permet

de garantir la sélection en au plus N itérations.

La méthode du pivot propose une règle de décision. Elle nécessite aussi un critère d'ordre dans le choix des couples d'unité successifs. Le pivot local se base sur la proximité géographique. À chaque étape, une unité est choisie aléatoirement parmi les unités pour lesquelles aucune décision n'a été prise. Puis, l'unité qui lui est la plus proche est également choisie. Les probabilités d'inclusion sont mises à jour pour ces deux unités, et le procédé est ensuite réitéré jusqu'à la sélection de l'échantillon.

Le caractère répulsif de cette méthode vient de la règle de mise à jour des probabilités d'inclusion. En effet, sur les deux unités confrontées, l'une voit sa chance d'être sélectionnée réduite d'autant que celle de l'autre augmente. L'inclusion double des unités proches géographiquement est donc réduite. On peut ainsi espérer améliorer les estimations. Cependant, rien ne garantit qu'on compense dans ce cas le fait de ne pas équilibrer sur l'information auxiliaire.

Comparaison à l'équilibrage spatial

On observe bien un gain à étaler spatialement les unités sélectionnées. En introduisant un mécanisme de répartition spatiale dans le processus d'échantillonnage, on permet pour un certain nombre de variables de capter une part de l'hétérogénéité inobservée, liée à une autocorrélation spatiale positive. Hormis pour quelques variables auxiliaires servant à équilibrer le tirage, le cube local (algorithme du tirage spatialement équilibré) est presque systématiquement meilleur (plus précis) que le cube simple.

Un bon candidat peut également venir de la famille des méthodes du pivot local, à condition de choisir convenablement l'espace et la distance permettant l'étalement spatial. On a en effet observé que ces méthodes étaient très sensibles au choix de la distance, et qu'elles pouvaient rivaliser avec le cube local tout comme être nettement moins bonne que le cube simple. Une distance sociodémographique entre unités primaires, basée sur les ratios de variables centrés et réduits, paraît fonctionner correctement dans le cadre des simulations de tirage d'unités primaires et pourrait être envisagée pour le tirage de l'EM (voir Givois et Merly-Alpa, 2018). Cependant, contrairement au cube local, il n'est pas systématiquement meilleur que la méthode de référence, le cube classique strate par strate. De plus, le cube local se comporte légèrement mieux dans l'estimation des quantiles, et il permet une répartition plus homogène des unités sur le territoire.

Il serait nécessaire de renforcer cette étude en faisant varier les jeux de variables utilisés. Il n'est en effet pas à exclure que les résultats obtenus soient sensibles aux infor-

mations mobilisées, en particulier les méthodes du pivot local se basant sur des distances sociodémographiques, qui ne reposent que sur de l'information auxiliaire. Dès lors qu'une variable d'intérêt n'est pas corrélée à cette information auxiliaire, ces méthodes peuvent se traduire par une perte importante de précision. Le cube local semble offrir l'avantage d'intégrer une dimension géographique (coordonnées X et Y) pour répartir spatialement l'échantillon et capter une part de l'inobservable lorsque les variables d'intérêt sont peu corrélées aux variables auxiliaires utilisées pour l'équilibrage.

Autres méthodes

Enfin, d'autres méthodes existent : tessellation, tirage déterminantal... On peut se reporter au Manuel d'Analyse Spatiale (De Bellefon et al., 2018) à ce sujet. Ces méthodes n'ont pas été testées dans le contexte du tirage du prochain échantillon-maître mais pourraient avoir des propriétés intéressantes dans ce cadre.

Chapitre 2

Le tirage de l'EM

2.1 Pourquoi réaliser un tirage spatialement équilibré pour l'EM ?

L'objectif de l'échantillon-maître est de circonscrire le territoire à couvrir pour la collecte des différentes enquêtes auprès des ménages de l'Insee, et ainsi de limiter le nombre d'enquêteurs au sein du réseau, tout en garantissant des estimations de bonne qualité dans les enquêtes. Pour que cela soit rendu possible, l'échantillon de premier degré tiré doit être représentatif de l'ensemble du territoire, pour des indicateurs très divers. Si la méthode du tirage équilibré, utilisée notamment pour l'échantillon-maître en 2009, permet d'obtenir des résultats satisfaisants sur les indicateurs corrélés aux variables d'équilibrage [1], la dispersion de l'échantillon sur l'ensemble du territoire permet de mieux rendre compte de la diversité des territoires et des nombreuses variables d'intérêt non représentées par les variables d'équilibrage, mais ayant une autocorrélation spatiale positive. L'utilisation de la méthode de tirage spatialement équilibré garantit une meilleure précision pour l'ensemble des enquêtes ménages utilisant l'échantillon-maître [2], c'est pourquoi il a été décidé de favoriser cette approche.

Une fois la méthode choisie, il faut définir les paramètres de celle-ci, à savoir, dans le cas présent, les variables d'équilibrage, le niveau de la phase de vol et celui de la phase d'atterrissage, enfin les allocations.

2.2 Le choix des variables d'équilibrage

2.2.1 Choisir des variables socio-démographiques

Le choix des variables d'équilibrage est un enjeu crucial en ce qui concerne la qualité de l'échantillon d'unités primaires tirées. Si l'on considère que la variable d'intérêt y est régie par un modèle linéaire du type :

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \forall i \in U$$

1. Le nombre de logements par UP est notamment corrélé à la plupart des indicateurs issus des enquêtes ménages.

2. Une étude par simulations avait été réalisée par Favre-Martinoz et Merly-Alpa, 2016 et montrait un gain en précision pour des variables non nécessairement corrélées à celles intégrées au jeu d'équilibrage.

Alors l'objectif est de trouver un ensemble de p variables $\boldsymbol{x}$ suffisamment explicatives de la variable y. L'enjeu ici va être de trouver des variables d'équilibrage corrélées à un maximum d'indicateurs, sur une période de 10 ans.

Par ailleurs, la population (ainsi que l'échantillon) d'unités primaires est de petite taille. Le nombre de contraintes d'équilibrage qui peuvent être respectées (dépendant de la taille de l'échantillon sélectionné) peut ainsi rapidement être atteint, d'autant plus pour un cube spatialement équilibré. En effet, deux effets indésirables sont possibles :

— L'algorithme du cube peut ne pas arriver à répondre à toutes les contraintes d'équilibrage souhaitées, voire, potentiellement, se trouver en fin de phase de vol sur un échantillon équilibré sur aucune des variables d'équilibrage du jeu sélectionné du fait du trop grand nombre de contraintes.

— La phase de vol de la méthode de tirage spatialement équilibré est basée sur la sélection de p+1 proches individus (au sens des coordonnées X,Y avec p le nombre de variables d'équilibrage). Une phase de vol est appliquée sur ces p+1 individus et à l'issue de cette phase, on statue sur l'intégration ou non d'un individu dans l'échantillon. On recommence l'opération jusqu'à ce qu'il ne reste plus que p individus sur lesquels statuer, où commence alors la phase d'atterrissage. On voit que si le nombre de variables est trop élevé, les p individus restants s'étaleront sur une aire géographique trop vaste, et l'équilibrage spatial sera moins efficace. D'où la nécessité de choisir parcimonieusement le nombre de variables.

Il convient donc de trouver un ensemble de variables d'équilibrage ordonnées [3] qui permettront d'obtenir un échantillon-maître de qualité pour des variables d'intérêts issues d'enquêtes différentes.

Dans une première approche, deux types de jeux de variables d'équilibrage ont été considérés [4] :

— Le premier type de jeu est dit exhaustif, car il prend en compte l'ensemble des variables disponibles dans l'équilibrage. Ce jeu bien que très explicatif de notre variable intérêt y risque de surcharger inutilement le nombre de contraintes d'équilibrage dans la mesure où certaines variables explicatives sont corrélées entre elles.

— Pour éviter cette surcharge de contraintes, on peut définir un deuxième type de jeu dit "parcimonieux", où seulement quelques variables représentatives de groupes de variables préalablement définis seraient utilisées dans l'équilibrage. Ces groupes de variables sont sélectionnés par le biais d'une ACP effectuée sur l'ensemble des variables disponibles [5]. Les variables les plus contributives des premiers axes en sortie de cette ACP (analyse en composantes principales) sont alors retenues pour constituer le jeu de variables parcimonieux. Celui-ci réduit

3. L'ordre d'entrée des variables d'équilibrage dans l'algorithme est important. Les contraintes d'équilibrage sont abandonnées petit à petit, de la dernière à la première.

4. Pour ces deux types de jeu de variables, l'intégration des variables dans l'équilibrage a été testée sous forme de totaux et en structure. Il s'avère que les indicateurs permettant d'évaluer la qualité de l'équilibrage sont meilleurs lorsque les variables de totaux sont utilisées. On ne mentionne donc pas les scénarios incluant des variables en structure dans la suite du paragraphe.

5. Notons qu'il s'agit de variables quantitatives.

bien la dimension du problème, mais cette sélection de variables ne prend pas en compte toute l'information à disposition et risque d'être insuffisante pour expliquer la variable d'intérêt Y.

Le jeu de variables initiales [6], utilisé soit exhaustivement, soit parcimonieusement, est très étendu et comprend :

- Des variables démographiques issues du recensement de la population (âge et sexe) ;
- Des variables de catégorie socio-professionnelle ;
- Des variables relatives au niveau de diplôme et à la scolarisation ;
- Des variables relatives à l'activité de la population ;
- Des variables sur la composition des ménages ;
- Des variables sur le revenu, les salaires et les bénéfices industriels, agricoles, non commerciaux ainsi que le nombre d'imposés à l'ISF ;
- Des variables sur les caractéristiques du logement (surface, valeur locative, vacant, propriétaire, locataire) ;
- Des variables de zonage du territoire français, à savoir les tranches d'unité urbaine (TUU) et les aires urbaines (urbain, rural, couronne, etc.).

De ces variables, les résultats de scénarios issus de différentes combinaisons de variables d'équilibrage, soit exhaustive, soit parcimonieuses, sont comparés.

2.2.2 Comparaison des scénarios

Pour comparer les scénarios entre eux, on utilise à la fois les variables ayant servi à l'équilibrage mais aussi d'autres variables non utilisées à ce stade dont la qualité de l'estimation permettra de rendre compte de la qualité explicative des variables d'équilibrage et donc des gains potentiels en termes de précision dans les enquêtes futures.

L'indicateur privilégié est le coefficient de variation (CV) dont l'expression analytique est la suivante : $100\ \frac{\sigma(Y)}{T(Y)}$ avec $\sigma(Y)$ l'écart type de la variable Y considérée sous le plan de sondage utilisé et $T(Y)$ son total.

L'estimation de ces CV est réalisée en lançant un grand nombre de simulations, par la méthode de Monte Carlo. Plus l'estimation d'une variable est précise, plus le coefficient de variation est petit.

Certains éléments sont considérés comme primordiaux. Ainsi, le critère d'équilibrage sur le nombre de résidences principales doit être absolument respecté et donc les CV relatifs au nombre de résidences principales doivent être très faibles (proche de 0). En effet les probabilités d'inclusion simple des UP étant calculées proportionnellement au nombre de résidences principales dans l'UP (cf. 2.3.2), la contrainte d'équilibrage sur les probabilités d'inclusion est à la fois une contrainte de taille fixe [7] et une contrainte d'équilibrage sur le nombre de résidences principales.

Ce premier critère permet d'éliminer certaines combinaisons de variables. On ne

6. Ces variables sont aisément calculables au niveau communal, ce qui explique également pourquoi les unités primaires ont été définies comme des regroupements de communes.

7. La contrainte de taille fixe signifie que quel que soit l'échantillon tiré, il aura toujours la même taille. En pratique, selon les variables utilisées dans l'équilibrage, la taille fixe de l'échantillon n'est pas garantie. Or, pour permettre la collecte, l'échantillon-maître doit être adapté à la capacité du réseau d'enquêteurs de l'Insee. Le nombre d'unités primaires ne doit donc pas dépendre de la réussite de l'équilibrage.

conserve alors que les scénarios qui permettent d'obtenir un échantillon de la taille souhaitée.

Pour classer les différents scénarios restants, une première approche systématique est mise en œuvre : réaliser une ACP dont les individus sont les différents scénarios et dont les variables utilisées sont les coefficients de variation obtenus sur chacune des variables d'intérêt. Cela permet de repérer graphiquement et simplement quels jeux de variables d'équilibrage (parcimonieux ou exhaustifs) induisent une qualité d'équilibrage satisfaisante pour un maximum de variables d'intérêt. En particulier, les variables liées à la tranche d'unité urbaine se démarquent par une plus grande difficulté à obtenir des échantillons de qualité, selon l'un ou l'autre des types de jeux testés (exhaustifs ou parcimonieux). Pour discriminer les scénarios restants, une comparaison visuelle des dispersions des coefficients de variations par variable et par scénario est effectuée, afin de repérer les scénarios uniformément plus précis sur toutes les variables. Du fait de son caractère « manuel », cette méthode demande de se limiter à un faible nombre de scénarios à observer.

De ces comparaisons, on constate que la précision sur les variables telles que le nombre de résidences principales par tranche d'unité urbaine (dénommées par la suite « variables de TUU ») est rarement satisfaisante si on ne les utilise pas dans le jeu d'équilibrage. En conservant les variables de tranche d'unité urbaine comme variables d'équilibrage, en plus des variables parcimonieusement choisies, la précision est bien plus satisfaisante sur ces variables mais la précision calculée sur les autres variables se dégrade. En outre, la contrainte de taille fixe n'est pas toujours respectée [8].

En conclusion :

- Le jeu exhaustif ne respecte pas les contraintes de taille fixe ;
- Les scénarios parcimonieux :
  - Sont peu performants sur l'estimation des zonages urbain/rural lorsqu'on n'intègre pas les variables de TUU parmi les variables d'équilibrage ;
  - Ne respectent pas la contrainte de taille fixe dès lors qu'on intègre les variables de TUU, ce qui dégrade la précision sur d'autres variables.

Ces scénarios restent insatisfaisants. Aussi, une variante des scénarios parcimonieux a été testée : utiliser comme variables d'équilibrage les premiers axes d'une ACP effectuée sur les variables auxiliaires présentes dans la base de sondage.

2.2.3 Utilisation des axes d'une ACP comme variables d'équilibrage

L'idée est d'équilibrer le tirage sur les premiers axes d'une ACP exécutée sur l'ensemble des variables du jeu de données. Ces premiers axes peuvent être vus comme de nouvelles variables au niveau UP qui captent une forte part de l'information contenue dans notre base d'UP. Elles sont donc d'excellentes candidates pour la constitution des jeux de variables d'équilibrage d'un échantillon-maître, utile à de nombreuses enquêtes aux indicateurs différents. Par ailleurs, les axes d'ACP et leur capacité à bien résumer

8. D'autres travaux de la division Sondages de l'Insee sur des tirages spatialement équilibrés d'unités primaires mobilisent des variables de tranches d'unités urbaines ou des variables régionales. Plusieurs d'entre eux ont connu des problèmes de non-respect de la taille fixe. Sans avoir d'explication précise à fournir, cela conduit à supposer que l'algorithme de tirage équilibré utilisé lors de ces travaux a du mal à respecter la taille fixe de l'échantillon quand un certain nombre de variables ont une valeur nulle sur une majorité d'unités primaires.

l'information [9] permettent de répondre à l'objectif de parcimonie dans le choix des variables d'équilibrage dans le but de se prémunir contre d'éventuels effets indésirables liés à un trop grand nombre de variables d'équilibrage. Ces scénarios sont alors comparés avec les précédents selon les mêmes méthodes. L'utilisation des premiers axes issus d'une ACP en tant que variables explicatives semble être un bon compromis entre un nombre raisonnable de contraintes d'équilibrage et un bon pouvoir explicatif. Notamment, l'équilibrage sur les variables de tranche d'unité urbaine est correct (quand ces variables sont utilisées dans l'ACP). Cependant, si on ne garde que les axes, on ne garantit pas la taille fixe de l'échantillon, car la contrainte d'équilibrage des π_k n'est jamais totalement respectée. De même, d'autres variables socio-démographiques ou économiques telles le revenu total, la population, la catégorie socio-professionnelle ou le diplôme ne sont pas suffisamment bien estimées.

Une dernière approche est alors proposée pour garantir la bonne représentativité de l'échantillon sur ces variables, sans dégrader l'équilibrage sur les autres variables : il s'agit de mélanger les variables dont la contrainte d'équilibrage doit être respectée parfaitement (en particulier les π_k), et les principaux axes d'une ACP, c'est-à-dire les premiers axes qui permettent de capter 99 % de l'inertie de notre base d'unités primaires pour les variables qu'on souhaite estimer avec précision mais qui ne participent pas directement à l'équilibrage. Des simulations ont permis de tester différents scénarios en interchangeant les variables qui participaient à l'ACP et les variables qui participent directement à l'équilibrage. Elles ont également permis de comparer la qualité de l'équilibrage lorsque les axes issus de l'ACP étaient placés parmi les premières variables d'équilibrage ou, au contraire, parmi les dernières.

Cette approche permet ainsi d'obtenir un échantillon parfaitement représentatif des variables socio-démographiques mises dans le jeu de variables d'équilibrage, et d'être de bonne qualité pour les indicateurs corrélés aux variables intégrées à l'ACP dont les premiers axes sont utilisés comme variables d'équilibrage.

Au final, le jeu considéré est le suivant [10]. Toutes les variables sont relatives aux unités primaires :

- probabilités d'inclusion

- 14 premiers axes d'une ACP calculée sur les variables :
  - nombre de résidences principales par type de ménage (familles monoparentales, personnes seules, couples avec enfant(s), couples sans enfant)
  - somme de la surface des résidences principales
  - somme de la valeur locative des résidences principales

9. En particulier, car les variables auxiliaires disponibles sont des totaux, par nature bien corrélés aux totaux de populations et de résidences principales.

10. Notons d'une part qu'avec la coordination de l'EM et de l'EEC, ce jeu de variables se trouvera légèrement modifié (cf. section 4.3 de la partie C), et d'autre part que les modalités des variables utilisées présentent une modalité manquante pour éviter la colinéarité entre variables d'équilibrage.

  - nombre de femmes
  - somme des revenus par type de revenu (salaires, pensions de retraite, allocations chômage ou préretraite)
  - nombre de logements vacants
  - nombre de résidences principales qui sont des logements sociaux
  - nombre de résidences principales occupées par des ménages imposables sur la fortune
  - nombre d'individus scolarisés par tranche d'âge (2-5 ans, 6-10 ans, 11-14 ans, 15-17 ans, 18-24 ans, 25-29 ans, 30 ans ou plus)
  - nombre de résidences principales en milieu périurbain et nombre de résidences principales en milieu rural
  - nombre de résidences principales par tranche d'unité urbaine (tranches 1 à 8)

- population par tranche d'âge (15-29 ans, 30-44 ans, 45-59 ans, 60-74 ans, 75-89 ans, 90 ans et plus)

- somme des revenus des individus

- nombre de résidences principales occupées par leurs propriétaires

- nombre d'individus par PCS en 8 catégories (PCS 2 à 8)

- nombre de résidences principales en quartier prioritaire de la ville

- nombre d'individus de 15 ans ou plus, non scolarisés, ayant un niveau de diplôme donné (niveau certificat d'études primaires ou brevet des collèges, niveau CAP-BEP, niveau baccalauréat, diplôme du supérieur)

- nombre d'individus par statut d'activité et d'emploi (nombre de 15-64 ans actifs occupés, nombre de 15-64 ans au chômage, population en emploi salarié)

- somme des bénéfices par type de bénéfices (bénéfices industriels et commerciaux, bénéfices agricoles, bénéfices non commerciaux)

L'ordre de ces variables est celui effectivement utilisé pour l'équilibrage. Cet ordre a son importance puisque la phase d'atterrissage de l'algorithme retire les contraintes d'équilibrage au fur et à mesure en commençant par la dernière variable d'équlibrage.

2.3 Méthode de calcul des allocations et pondérations

L'utilisation du tirage spatialement équilibré demande de fixer encore trois paramètres : le niveau de la phase de vol, celui de la phase d'atterrissage, et les allocations.

2.3.1 Niveau d'équilibrage pour l'échantillon-maître

L'objectif de l'échantillon-maître est d'être représentatif aux niveaux de diffusion des indicateurs. La plupart des enquêtes auprès des ménages étant adossées au règlement européen IESS, le niveau recherché est le niveau Nuts2, correspondant aux anciennes régions administratives. C'est donc à ce niveau que l'équilibrage en phase de vol devra être effectué. Le tirage des unités primaires est donc un tirage spatialement équilibré stratifié par les anciennes régions administratives.

En ce qui concerne l'équilibrage en phase d'atterrissage, deux options se présentent :

— abandonner une à une les contraintes d'équilibrage dans chaque région
— une fois les phases de vol de chaque région terminées, effectuer une nouvelle phase de vol avec les unités primaires sur lesquelles l'algorithme n'a pas encore statué en les regroupant au niveau national, et en essayent d'équilibrer les résultats à ce même niveau national, puis abandonner une à une les contraintes au niveau national.

Le choix entre l'une ou l'autre des solutions dépend avant tout du nombre de cas restant à statuer après la phase de vol. Du fait du grand nombre de variables d'équilibrage au regard du nombre d'unités de tirage dans chaque région, il paraît plus opportun d'opter pour la deuxième solution. Cela permet de continuer la phase de vol (au niveau régional puis national) avec un nombre restreint de contraintes [11], et ainsi de ne les abandonner qu'ultérieurement si nécessaire.

2.3.2 Méthode de calcul d'une allocation au sein d'une région et pondération des unités primaires

Le dernier paramètre à définir est le jeu d'allocations, c'est-à-dire le nombre d'unités primaires à tirer dans chaque région, associé à la probabilité de sélection de chaque UP de la base de sondage. Ce paramètre est déterminé selon une triple contrainte :

— la probabilité de tirage d'un logement doit être identique pour tous les logements (une fois prise en compte le tirage des UP puis celui des logements au sein des UP tirées), c'est-à-dire qu'on souhaite effectuer un tirage équipondéré ;
— la charge de collecte doit être répartie au sein des UP tirées afin de permettre aux enquêteurs de réaliser l'enquête ;
— le nombre d'unités primaires tirées au sein d'une région doit permettre de garantir les contraintes de précision fixées par Eurostat.

D'abord, on montre l'influence de l'équipondération et des contraintes de collecte sur le calcul des probabilités d'inclusion des unités primaires sans UP exhaustive, c'est-à-dire sans UP dont la probabilité d'être incluse dans l'EM est égale à 1. Pour cela, on introduit le tirage auto-pondéré à deux degrés. Puis, on adapte ces calculs en cas de présence d'unités primaires exhaustives. Ensuite, la stratification régionale de l'algorithme d'équilibrage impliquant le calcul d'une allocation d'unités primaires par région, on

11. En effet, la première phase de vol se fait de façon indépendante pour chaque région sur le même jeu de variables d'équilibrage, soit un nombre de contraintes égal au nombre de variables d'équilibrage multiplié par le nombre de régions. La seconde phase de vol réduit la dimension des contraintes en s'abstrayant de la dimension régionale, excepté en ce qui concerne le respect des allocations.

montre comment le sondage équipondéré conduit à fixer les allocations régionales. Enfin, on évoque l'adaptation de ces pondérations et des allocations aux contraintes de précision régionale.

Principe du sondage auto-pondéré à deux degrés

L'objectif initial est l'équiprobabilité de sélection de chaque logement pour une enquête sans sur-représentation. Autrement dit, par le tirage de l'échantillon-maître, puis celui de l'échantillon de l'enquête concernée, on souhaite que tous les logements aient la même « chance » d'être tirés (π), soit :

$$\pi_l = \pi_{up}\pi_{l|up} = \pi = \frac{n}{N} \tag{2.1}$$

où n est le nombre de logements à enquêter, N le nombre de résidences principales sur le territoire, π_{up} la probabilité de l'unité primaire up d'être sélectionnée dans l'échantillon-maître, $\pi_{l|up}$ la probabilité d'un logement l de l'UP up d'être tiré dans l'échantillon conditionnellement à la présence de l'UP up dans l'échantillon-maître.

Si l'Insee utilise un échantillon-maître, c'est pour limiter le territoire à couvrir par son réseau d'enquêteurs en regroupant les logements enquêtés sur les zones de travail de ces agents. Aussi, chaque unité primaire peut être considérée comme la zone de travail d'un enquêteur [12]. Considérant qu'à temps plein, chaque enquêteur effectue le même nombre d'enquêtes, on sélectionnera le même nombre de logements c dans chaque UP. Pour un nombre d'enquêteurs donné, si on affecte un enquêteur par UP, on sélectionnera n_{up} logements dans chaque UP up avec :

$$c = n_{up} = \frac{n}{m} \tag{2.2}$$

où m est le nombre d'unités primaires tirées.

Au moment du tirage de deuxième degré de l'échantillon d'une enquête, la probabilité de sélection de chaque logement au sein d'une UP est la même et équivaut au nombre de logements tirés dans l'UP, rapporté au nombre de logements de l'UP :

$$\pi_{l|up} = \frac{n_{up}}{N_{up}} \tag{2.3}$$

où N_{up} décrit le nombre de logements de l'UP up. D'après (2.2) :

$$\pi_{l|up} = \frac{n}{mN_{up}}$$

12. Plus exactement, il va s'agir de la zone de travail pour une certaine quotité des enquêteurs, car on peut souhaiter avoir un enquêteur sur plusieurs zones, ou, à l'inverse, plusieurs enquêteurs sur une zone.

et, d'après (2.1),

$$\pi_{up} = \frac{\pi}{\pi_{l|up}} = \frac{\frac{n}{N}}{\frac{n}{mN_{up}}} = m\frac{N_{up}}{N}$$

En conclusion, pour que tous les logements du territoire aient la même probabilité d'être sélectionnés, et que le même nombre de logements soit tiré dans chaque UP, il faut que la probabilité de sélection d'une UP dans l'EM soit proportionnelle à sa taille en nombre de logements. Il s'agit de la méthode du sondage à deux degrés autopondéré.

L'adaptation du sondage auto-pondéré à la présence d'unités exhaustives

Ces probabilités ne peuvent cependant pas s'appliquer directement. En effet, il arrive, dans certains cas, que le nombre de logements d'une UP soit tel que sa probabilité de sélection soit supérieure à 1 [13]. En effet, si $N_{up} > \frac{N}{m}$ alors $\pi_{up} > 1$. On parle alors d'unités primaires exhaustives, car elles seront automatiquement sélectionnées dans l'EM. Cependant, nous montrons ci-après que le fait de tronquer à 1 la probabilité de sélection de ces UP exhaustives ne permet pas de conserver la même chance de sélectionner les logements sans relâcher la contrainte de tirer le même nombre de logements dans chaque UP au second degré. Ainsi, pour conserver cette équiprobabilité, il convient de tirer plus de logements dans ces UP exhaustives au second degré. Une fois ce point démontré, il s'agira de calculer les probabilités d'inclusion des UP exhaustives et celles des UP non exhaustives.

La probabilité de sélection d'un logement de ces UP est la suivante :

$$\pi_l^{exh} = \pi_{up}^{exh}\pi_{l|up}^{exh} = \pi_{l|up}^{exh} \tag{2.4}$$

En sélectionnant le même nombre de logements dans cette UP que dans les autres UP, on a :

$$\pi_{l|up}^{exh} = \frac{n_{up}}{N_{up}^{exh}} = \frac{n}{mN_{up}^{exh}}$$

La probabilité de sélection de ces logements vaudra :

$$\pi_l^{exh} = \frac{n}{mN_{up}^{exh}}$$

Pour rappel, pour les autres UP (non exhaustives et notées $nexh$),

$$\pi_l^{nexh} = \pi_{up}^{nexh}\pi_{l|up}^{nexh} = \frac{n}{N}$$

Or $N_{up}^{exh} > \frac{N}{m}$ ainsi $\pi_l^{nexh} > \pi_l^{exh}$. Il n'y a donc pas équiprobabilité (un logement d'une UP exhaustive a moins de chance d'être tiré qu'un logement d'une UP non exhaustive).

13. C'est le cas des grandes villes.

Pour assurer l'équiprobabilité, il faut sélectionner plus de logements dans les UP exhaustives que dans les UP non exhaustives de telle sorte que, pour toutes les UP : $\pi_l = \pi_l^{exh} = \pi_l^{nexh}$

Comme le nombre de logements sélectionnés dans une UP non exhaustive doit correspondre à la charge dévolue à un enquêteur, plusieurs enquêteurs devront être déployés sur une UP exhaustive (où plus de logements sont sélectionnés), et ce au détriment des UP non exhaustives, le contingent étant fixe : il est ainsi nécessaire de sélectionner moins d'UP non exhaustives pour réaliser ces différentes contraintes.

Trois éléments sont ainsi à déterminer pour permettre l'équiprobabilité des logements pour toutes les enquêtes :

— Combien d'UP non exhaustives m^{nexh} sélectionner pour une allocation initiale de m ?
— Combien de logements n_{up}^{exh} sélectionner pour une UP exhaustive, pour une enquête de taille n ?
— Combien de logements n_{up}^{nexh} sélectionner pour une UP non exhaustive, pour une enquête de taille n ?

Le calcul du nombre d'unités primaires non exhaustives se fait de manière itérative. D'abord, les unités primaires dont le nombre de logements $N_{up} > \frac{N}{m}$ sont identifiés. Elles sont au nombre de m_0^{exh} et contiennent N_0^{exh} logements. Puis, sont identifiées les unités primaires telles que $N_{up} > \frac{N - N_0^{exh}}{m - m_0^{exh}}$ et ainsi de suite.

Pour toutes les UP, du fait de l'équipondération souhaitée, on veut $\pi_l^{exh} = \pi_l^{nexh} = \frac{n}{N}$, d'après la formule (2.1). En combinant les formules (2.3) et (2.4), cela implique pour les logements dans les unités primaires exhaustives :

$$\pi_l^{exh} = \pi_{l|up}^{exh} = \frac{n_{up}^{exh}}{N_{up}^{exh}}$$
$$\text{et } \pi_l^{exh} = \frac{n}{N}$$

On obtient ainsi :

$$n_{up}^{exh} = n\frac{N_{up}^{exh}}{N} \tag{2.5}$$

Pour une enquête, le nombre de logements à tirer dans une UP exhaustive est proportionnel à la taille de l'UP exhaustive. Cela implique, en sommant (2.5), que le nombre de logements à tirer dans les unités primaires exhaustives est $n^{exh} = n\frac{N^{exh}}{N}$.

On déduit de cette allocation pour les UP exhaustives le nombre d'unités à enquêter dans les unités primaires non exhaustives $n^{nexh} = n - n^{exh} = n\frac{N^{nexh}}{N}$.

La charge constante sur l'ensemble des unités primaires non exhaustives tirées donne le nombre de logements à enquêter par unité primaire non exhaustive, d'après la formule (2.2) :

$$n_{up}^{nexh} = \frac{n^{nexh}}{m^{nexh}} = n\frac{N^{nexh}}{m^{nexh}N}. \tag{2.6}$$

Il reste à présent à déterminer les probabilités d'inclusion des UP non exhaustives π_{up}^{nexh}. En combinant (2.1), (2.3) et (2.6), on obtient :

$$\begin{aligned}
\pi_{up}^{nexh} &= \frac{\pi_{l}^{nexh}}{\pi_{l|up}^{nexh}} \\
&= \frac{n}{N}\frac{N_{up}^{nexh}}{n_{up}^{nexh}} \\
&= \frac{n}{N}\frac{m^{nexh}NN_{up}^{nexh}}{nN^{nexh}} \\
&= \frac{m^{nexh}N_{up}^{nexh}}{N^{nexh}}
\end{aligned}$$

Donc, avant l'introduction de contraintes régionales, on est en mesure d'identifier les UP exhaustives, on sait calculer le nombre d'UP non exhaustives à tirer, on dispose de probabilités d'inclusion pour les unités primaires exhaustives et pour les unités primaires non exhaustives.

L'adaptation du sondage auto-pondéré à un tirage spatialement équilibré stratifié par région

Les contraintes d'équlibrage sont définies à un niveau régional, afin de répondre aux contraintes de précision des règlements européens d'après le point 2.3.1. Or, rien ne garantit que la somme des probabilités des UP d'une région somme à un entier, ce qui est nécessaire pour définir une stratification régionale pour le tirage des UP. Pour chaque région i, il est ainsi nécessaire de calculer une allocation régionale qui arrondit :

$$\sum_{up \in i \cap exh} \pi_{up}^{exh} + \sum_{up \in i \cap nexh} \pi_{up}^{nexh}$$

Aux arrondis près, le nombre d'UP non exhaustives à sélectionner dans l'EM dans une région est proportionnel au nombre de résidences principales appartenant à ces UP non exhaustives.

L'adaptation du sondage auto-pondéré à des contraintes de précision régionales

Afin de faciliter le respect des précisions demandées par Eurostat, sans trop accroître la taille de l'EM, le nombre d'UP sélectionnées est défini au niveau de chaque ancienne région et non au niveau national. La coordination avec l'échantillon Emploi a conduit, comme évoqué plus loin au chapitre 5, à s'éloigner quelque peu de ces allocations arrondies issues d'un sondage équipondéré national, tout en essayant de rester le plus proche possible des allocations présentées ci-dessus. L'équiprobabilité de sélection des résidences principales est donc recherchée au niveau régional, et les formules présentées plus tôt peuvent être adaptées sans difficulté à une structure d'allocations m_i d'UP à tirer dans la région i.

Ainsi, la probabilité de sélection d'une UP up appartenant à la région i s'écrit :

$$\pi_{up}^{exh} = 1$$
$$\pi_{up}^{nexh} = \frac{m_i^{nexh} N_{up}^{nexh}}{N_i^{nexh}}$$

où m_i^{nexh} est le nombre d'unités primaires non exhaustives calculé itérativement, N_{up}^{nexh} le nombre de résidences principales de l'UP non exhaustive up, N_i^{nexh} le nombre total de résidences principales dans les UP non exhaustives de la région i.

Les allocations régionales m_i dont la sélection est détaillée au chapitre 5 sont les suivantes :

TABLE 2.1 – Allocations régionales de tirage de l'échantillon-maître

Région	Allocation
Île-de-France	85
Champagne-Ardenne	14
Picardie	19
Haute-Normandie	16
Centre	25
Basse-Normandie	16
Bourgogne	18
Nord-Pas-de-Calais	30
Lorraine	19
Alsace	14
Franche-Comté	13
Pays de la Loire	27
Bretagne	29
Poitou-Charentes	20
Aquitaine	27
Midi-Pyrénées	23
Limousin	10
Rhône-Alpes	51
Auvergne	15
Languedoc-Roussillon	21
PACA	44
Corse	5
Total	*541*

2.3.3 Méthode de calcul des pondérations

En pratique, la détermination des probabilités d'inclusion des UP se déroule en 4 étapes :

— Étape 1 : Définition du nombre initial d'UP par région (fixé en fonction de la précision souhaitée).
— Étape 2 : Calcul de la probabilité initiale de sélection de chaque UP
— Étape 3 : Détermination des UP dites exhaustives
— Étape 4 : Mise à jour des probabilités de sélection pour les UP non exhaustives.

Au cours de l'étape 4, certaines probabilités recalculées peuvent se retrouver supérieures à 1 ; les étapes 3 et 4 sont alors rejouées. Le poids d'une UP *up* sera donc, en notant i la région de l'UP, N_i le nombre de résidences principales dans cette région et N_{up}^{nexh} le nombre de résidences de principales dans l'UP si elle n'est pas exhaustive :

$$\begin{cases} w_{up}^{exh} & = 1 \text{ si } up \text{ est exhaustive} \\ w_{up}^{nexh} & = \frac{N_i^{nexh}}{m_i^{nexh} N_{up}^{nexh}} \text{ si } up \text{ n'est pas exhaustive} \end{cases}$$

Chapitre 3

Premiers arbitrages pour le tirage de l'EEC

3.1 Pourquoi réaliser un tirage spatialement équilibré pour l'échantillon de l'EEC ?

Pour rappel, l'objectif de l'EEC est d'obtenir des résultats précis au niveau Nuts2 [1] pour des indicateurs liés à l'emploi comme le taux d'emploi ou le taux de chômage. La méthode du tirage équilibré semble particulièrement bien répondre à cette problématique, la base de sondage issue de Fidéli possédant de nombreuses variables corrélées fortement à ces variables d'intérêt. Par ailleurs, la collecte de l'enquête doit être menée rapidement, ce qui a conduit à procéder à un échantillonnage par grappe, plus efficace en termes de repérage et de collecte. Pour contrer la perte de précision liée à l'homogénéité des logements d'une même grappe, il paraît opportun de disperser sur l'ensemble du territoire les grappes (et donc les secteurs) échantillonnées, et ainsi réduire l'autocorrélation spatiale observée. La méthode du tirage spatialement équilibré répond à ce double objectif.

Comme pour l'échantillon-maître, il faut à présent définir les paramètres de la méthode, et en premier lieu les variables d'équilibrage.

3.2 Choix des variables d'équilibrage

La recherche d'informations dans la base de sondage corrélées aux variables d'intérêt de l'enquête est simplifiée par l'unicité du thème abordé, l'emploi. Ainsi, le choix des variables d'équilibrage issues de Fidéli va consister à tester dans l'algorithme d'équilibrage, au moyen de nombreuses simulations, différents jeux de variables liées à l'emploi, et à choisir celui apportant la meilleure précision pour une variable d'intérêt fixée.

Le premier jeu de variables testé consiste à reconstruire, dans la base de sondage, les variables d'équilibrage [2] utilisées en 2009 pour la constitution de l'échantillon de l'EEC :

- la répartition des résidences principales par type d'espace selon le Zonage en Aires Urbaines (1999) : pôles urbains, couronnes périurbaines, communes multipolarisées, communes rurales ;
- la répartition des résidences principales par quintile de revenus ;
- le nombre de locataires ;

1. Critères définis dans le cadre du règlement européen IESS.
2. Ces variables sont listées par ordre croissant d'importance dans l'algorithme d'équilibrage.

— 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.

Ce jeu de données permettra d'étalonner les résultats des simulations par rapport à l'échantillon tiré en 2009.

De nombreuses variables de Fidéli peuvent être pertinentes dans l'élaboration du plan de sondage de l'EEC, comme le versement d'une allocation chômage ou d'un revenu d'activité à un individu de la base. A partir de ces données, des variables *proxy* sont construites pour chaque secteur de la base de sondage, se rapprochant au maximum du concept d'intérêt de l'enquête. Le nombre de chômeurs d'un secteur est ainsi approché par le nombre d'individus percevant ou non des allocations liées à la recherche d'emploi ; celui d'actifs occupés par le nombre d'individus percevant un salaire ou des revenus via une activité en tant qu'indépendant.

Les variables *proxy* sont particulièrement bien corrélées aux données recherchées par l'enquête (statut d'activité au sens du BIT, ...) et permettent de construire un échantillon équilibré de qualité. Malgré tout, les définitions des variables *proxy* et des variables d'intérêt de l'enquête sont différentes [3]. Par exemple, le fait d'approcher le statut de chômeur en fonction de la perception d'allocations liées à la recherche d'emploi diffère de la définition du BIT (Bureau international du travail) adoptée dans l'enquête Emploi, où un chômeur est défini comme :

— une personne de plus de 15 ans sans emploi (c'est-à-dire qui n'a pas travaillé au moins une heure durant la semaine de référence de l'enquête),
— disponible pour prendre un emploi dans les 15 jours,
— et qui a cherché activement un emploi dans le mois précédent (ou qui a trouvé un emploi qui commence dans moins de trois mois).

Un second jeu de variables d'équilibrage est élaboré avec l'aide des producteurs experts des indicateurs sur l'emploi, afin d'utiliser au mieux les nouvelles variables qui permettent de se rapprocher des concepts de l'enquête, tout en conservant des variables socio-démographiques pour prendre en compte les besoins de diffusion et de précision sur leurs différentes modalités [4] :

— Probabilités d'inclusion du secteur ;
— Nombre de personnes dans le secteur percevant des allocations chômage ou préretraite ;
— Nombre de personnes dans le secteur percevant des revenus d'activité ;
— Nombre de logements en quartier prioritaire de la politique de la Ville (QPV) dans le secteur ;
— Nombre de résidences principales par type de ménage dans le secteur [5] ;
— Nombre d'individus par croisement sexe/âge dans le secteur [6] ;

3. Ce qui justifie d'ailleurs l'existence de l'enquête...

4. Ces variables sont listées par ordre décroissant d'importance dans l'algorithme d'équilibrage.

5. Les types de ménage sont définis dans le fichier Fidéli en 6 catégories : homme seul, femme seule, couple sans enfant, couple avec enfant, famille monoparentale, autres.

6. 8 croisements sexe/âge ont été définis : mineurs, hommes de 18 à 25 ans, femmes de 18 à 25

— Somme des revenus disponibles du secteur.

Pour juger de la qualité de ces différents scénarios et les comparer, d'autres variables présentes dans le fichier Fidéli ont été utilisées comme variables d'intérêt, par exemple les totaux d'allocations chômages, de salaires ou de logements dans un secteur. Ces variables ne sont pas utilisées comme variables d'équilibrage car étant très corrélées aux autres variables d'équilibrage, elles n'apporteraient que peu d'informations, et augmenteraient le nombre de contraintes.

En comparant ces deux scénarios sur un tirage équilibré [7], on constate que celui apportant la meilleure précision est celui élaboré spécifiquement pour cette étude, c'est-à-dire le scénario avec le second jeu de variables d'équilibrage (cf. tableau 3.1). C'est donc ce scénario que l'on souhaite adopter pour le tirage de l'enquête Emploi en continu.

TABLE 3.1 – Coefficients de variation issus des scénarios de tirage pour la sélection des variables d'équilibrage

Variables	Scénario EEC 2009	Scénario nouvelles variables d'équilibrage
Nombre de personnes percevant une allocation chômage ou préretraite	0,53 %	0,04 %
Nombre de personnes percevant un revenu d'activité	0,31 %	0,03 %
Total des salaires	0,52 %	0,21 %
Total des allocations chômage ou préretraite	0,67 %	0,44 %
Nombre de logements	0,19 %	0,04 %

Note de lecture : Le tirage équilibré sur les variables de l'EEC 2009 permet d'estimer le nombre total de logements (y compris résidences secondaires et logements vacants) avec un coefficient de variation de 0,19% tandis que le scénario avec les nouvelles variables d'équilibrage permet cette estimation avec un coefficient de variation de 0,04%.

À ce stade, on pourrait déterminer les autres paramètres indispensables à un tirage équilibré, à savoir le niveau (région, France métropolitaine...) de la phase de vol et de la phase d'atterrissage de l'algorithme, et les allocations souhaitées. Pour un tirage équilibré au niveau régional, on serait amené à définir des allocations permettant d'obtenir dans chaque région la précision attendue par Eurostat dans le règlement IESS, et à équilibrer les phases de vol au niveau région, le nombre de variables d'équilibrage étant bien plus faible que le nombre de secteurs.

ans, hommes de 26 à 50 ans, femmes de 26 à 50 ans, hommes de 51 à 75 ans, femmes de 51 à 75 ans, personnes de plus de 75 ans. Ces croisements correspondent pratiquement à ceux demandés en diffusion par le règlement IESS.

7. Ces tirages sont équilibrés sans considération spatiales, afin de comparer uniquement les variables d'équilibrage. Les tirages sont donc ici indépendants du tirage de l'échantillon-maître.

Cependant, une autre contrainte s'impose à nous : la coordination des tirages de l'EM et de l'EEC). Pour rappel, le tirage direct des secteurs au sein des UP de l'EM n'étant pas souhaitable (cf. 3.1.1 de la partie B), nous avons constitué des unités de coordination comme regroupement de 1 à 4 UP. Nous souhaitons obtenir, après tirages, l'échantillon-maître et l'échantillon de l'EEC au sein des UC sélectionnées. Si cette contrainte existe également pour le tirage de l'EM, elle est moins prégnante, car les populations des UC et des UP sont comparables (1646 UC pour 5064 UP, des UP aux caractéristiques sociodémographiques proches au sein d'une même UC, un taux de sondage d'UP et d'UC proche en cas de tirage à deux degrés...). Autrement dit, les réflexions portant sur la définition des paramètres menant à l'EM (chapitre 2) peuvent être étendues à une sélection d'UC, si l'on souhaite passer d'abord par un premier degré de tirage, ce que nous étudierons au point 4.1.1. À l'inverse, la base de secteurs et la base d'UC sont très différentes (200 000 secteurs, des secteurs très différents au sein d'une même UC...). Ainsi, la coordination (et donc une forte diminution du nombre de secteurs dans la base de sondage limitée aux UC séletionnées) va obligatoirement influer sur le paramétrage du tirage des secteurs. Cela implique d'étudier le tirage des unités de coordination avant de poursuivre la réflexion sur le tirage des secteurs.

Le chapitre suivant décrit les différentes méthodes envisagées pour coordonner les tirages de l'EM et de l'EEC à travers la sélection d'UC.

Chapitre 4

Mise en œuvre de la coordination

Plusieurs plans de sondage ont été étudiés, soit par tirage d'unités de coordination puis d'unités primaires et de secteurs au sein de l'échantillon d'UC obtenu (tirage dit « direct »), soit par tirage de secteurs au sein des UC récupérées par tirage de l'échantillon-maître (tirage dit « indirect »). Ces plans de sondage sont présentés en section 4.1. Ces deux stratégies de tirage conduisent à des précisions différentes de l'échantillon-maître et de l'échantillon Emploi. Elles ont été comparées par le biais de simulations effectuées sur différents scénarios de tirages présentés en section 4.2, qui ont conduit à retenir la méthode de tirage indirect. Les pondérations associées aux unités de coordination dans le cas du tirage indirect sont également détaillées dans cette partie. Les différents paramètres du tirage de l'échantillon d'unités de coordination pouvant influer sur la précision de l'échantillon Emploi sont fixés dans la section 4.3.

4.1 Deux stratégies de tirage

L'objectif de la coordination est de sélectionner des UP et des secteurs de telle sorte que :

- l'échantillon-maître permette de tirer des logements représentatifs du territoire pour de nombreuses enquêtes et sur 10 ans ;
- l'échantillon de l'EEC permette de calculer des indicateurs sur l'emploi respectant a minima les contraintes d'Eurostat (réglement IESS) ;
- les logements à enquêter (pour l'EEC et les autres enquêtes) soient suffisamment proches, afin de limiter les déplacements des enquêteurs ;
- la coordination des deux échantillons n'oblige pas à interroger les mêmes logements pour deux enquêtes différentes.

Les unités de coordination ont été créées de telle sorte que ces deux derniers critères puissent être respectés (voir 3.1.2.1 de la partie B). L'objectif est à présent de déterminer la méthode de sélection de ces UC pour qu'EM et échantillon de l'EEC respectent les deux premiers objectifs.

Pour cela, deux possibilités s'offrent à nous :

- le tirage direct : on sélectionne en premier lieu des unités de coordination, puis on tire secteurs et UP au sein de cet échantillon d'UC.
- le tirage indirect : la sélection d'un échantillon-maître permet d'obtenir un échantillon d'unités de coordination dans lequel sont tirés les secteurs de l'EEC.

Il s'agit dans les deux cas de plans de sondage à plusieurs degrés. Dans le tirage direct, le premier degré consiste à sélectionner des zones géographiques (UC) au sein desquelles,

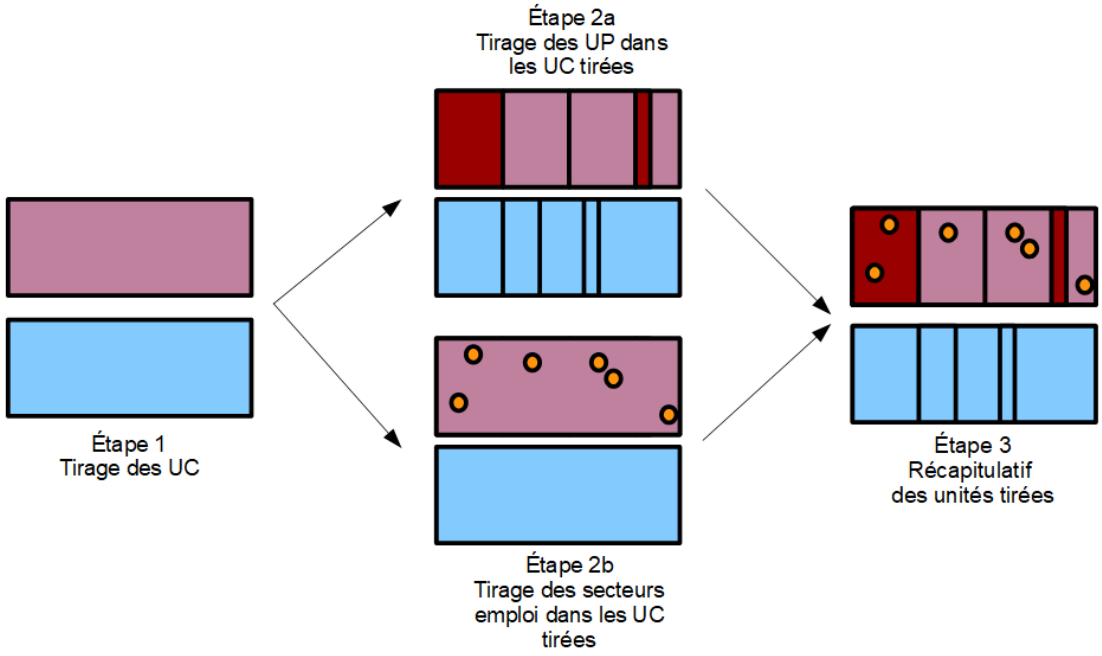
aux deuxièmes degrés, on sélectionne les secteurs de l'EEC d'une part et les UP de l'EM d'autre part. Dans le tirage indirect, le premier degré correspond à la sélection des UP et des UC, tandis que les secteurs de l'EEC sont tirés au deuxième degré.

4.1.1 Le tirage direct

Ce plan de sondage paraît le plus naturel, puisqu'il reprend la méthode de l'échantillon-maître : sélectionner un territoire, puis tirer un échantillon au sein de celui-ci.

La figure 4.1 illustre ce mode de tirage.

FIGURE 4.1 – Illustration du tirage direct



La première étape consiste à tirer un échantillon d'unités de coordination (rectangles). En l'occurrence, l'unité de coordination rose est sélectionnée ; la bleue ne l'est pas. On procède ensuite au tirage des unités primaires dans les unités de coordination tirées (étape 2a). Cette UC est divisée en 5 unités primaires, parmi lesquelles deux sont échantillonnées (en rouge) [1]. En parallèle (étape 2b), des secteurs Emploi sont tirés dans l'unité de coordination qui a été tirée à l'étape 1. Les secteurs Emploi ainsi sélectionnés sont les points orange. Ceux qui n'ont pas été tirés ne sont pas représentés. L'étape 3 montre l'ensemble des unités sélectionnées.

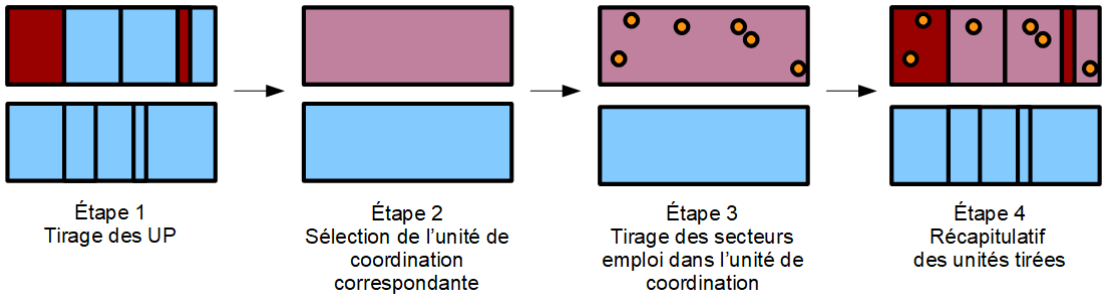
4.1.2 Le tirage indirect

Si une autre méthode est envisagée, c'est que le faible nombre d'unités primaire par unité de coordination semble contraindre fortement le tirage de l'EM, une fois l'échantillon d'UC sélectionné. On cherche donc une méthode permettant d'éviter cette contrainte.

1. Cette illustration a simplement une fonction didactique. En pratique le plan de sondage testé ne permet de tirer qu'une unité primaire par unité de coordination.

L'idée est de sélectionner l'EM indépendamment des UC, puis de « récupérer » des UC afin de tirer l'échantillon de l'EEC à proximité des UP [2]. La figure 4.2 illustre ce mode de tirage.

FIGURE 4.2 – Illustration du tirage indirect



Dans cet exemple, 10 unités primaires (petits rectangles) sont réparties en 2 unités de coordination (grands rectangles). L'étape 1 consiste à tirer un échantillon d'UP : ici, 2 UP, en rouge, sont sélectionnées (les 8 unités primaires qui n'ont pas été tirées sont en bleu). Comme les 2 unités primaires font partie de la même unité de coordination, seule une UC est conservée pour le tirage de l'échantillon Emploi (en rose dans l'étape 2). Dans l'étape 3, les secteurs Emploi (ronds orange) sont échantillonnés dans l'ensemble de cette unité de coordination. Les secteurs Emploi qui n'ont pas été tirés ne sont pas représentés. Enfin, l'étape 4 fait le bilan des unités tirées.

Dans chacune des stratégies, UP et secteurs sélectionnés sont situés à proximité. On note en particulier que les secteurs sont bien tirés dans l'ensemble des unités primaires de l'unité de coordination sélectionnée, et pas seulement dans les unités primaires tirées. Il est ainsi possible que certaines unités primaires de l'échantillon-maître (en rouge) ne contiennent pas de secteurs Emploi tirés dans l'échantillon Emploi et que d'autres unités primaires de l'échantillon-maître contiennent plusieurs secteurs de l'échantillon de l'EEC.

Il est nécessaire, à présent, de choisir la méthode à adopter.

4.2 Les termes du choix du tirage indirect

Le choix de la méthode repose sur l'objectif de variance minimale du sondage, tout en appliquant le principe de coordination spatiale.

4.2.1 L'inconvénient du tirage direct

Comme évoqué précédemment, si le tirage direct semble naturel, la sélection des UP dans un deuxième degré de tirage est particulièrement contraint, puisqu'il s'agit de tirer une UP par UC. Or chaque UC ne contenant que peu d'UP (4 au maximum), le

2. S'il avait pu être envisagé de construire les UC après tirage de l'EM, la probabilité de sélection des secteurs n'aurait pas pu être déterminée.

nombre de degrés de liberté dans l'algorithme d'équilibrage est trop faible pour que cette méthode puisse être appliquée au deuxième degré [3]. Il est donc impossible, par cette méthode de bénéficier de la diminution de variance que permet la méthode du tirage équilibré. Ce problème n'existe pas pour le tirage de l'EEC, la population de secteurs étant suffisamment nombreuse au sein de chaque UC pour qu'un tirage équilibré soit réalisé dans chacune des UC.

Pour le tirage indirect, l'échantillon-maître étant sélectionné en premier lieu, la méthode présentée au chapitre 2 est applicable. Ainsi, l'échantillon-maître est tiré par un algorithme de tirage spatialement équilibré statifié par ancienne région administrative et avec atterrissage national. Pour le tirage de l'échantillon de l'EEC, la population de tirage est la même dans les deux cas (ensemble des secteurs appartenant aux UC sélectionnées). En utilisant le tirage direct, on peut sélectionner les UC en fonction des variables de l'emploi, et ainsi, au deuxième degré, tirer l'échantillon de l'EEC équilibré sur ces mêmes variables. Pour rendre cela possible avec le tirage indirect, il faut que les UC « récupérées » lors du tirage de l'EM soient équilibrées sur ces mêmes variables de l'emploi. La méthode de partage des poids (Deville et Lavallée, 2006) permet d'aboutir à cet équilibrage des UC sélectionnées à partir de l'EM.

4.2.2 Calcul des pondérations des unités de coordination tirées

Pondération des unités de coordination dans le tirage direct

Dans le cas du tirage direct, les pondérations des unités de coordination se calculent simplement et selon le même principe que les pondérations des unités primaires présentées au point 2.3.2.

Pondération des unités de coordination dans le tirage indirect

Le calcul des pondérations des unités de coordination dans le cas du tirage indirect est plus complexe puisque les UC ne sont pas tirées directement. Le partage des poids est la méthode permettant de définir la pondération d'unités sélectionnées en tirant des unités d'une autre nature, en lien avec celles qui nous intéressent [4]. Dans le cadre des tirages de l'échantillon-maître et de l'échantillon Emploi, les unités primaires sélectionnées permettent d'accéder aux unités de coordination grâce aux liens entre ces deux types de zones.

La sélection des UC se fait par sondage indirect. Les UP sont d'abord échantillonnées, puis cela permet de déterminer indirectement un échantillon d'UC reliées aux UP de l'échantillon-maître. Naturellement, la pondération des unités de coordination tirées se calcule aussi indirectement à partir de la pondération des unités primaires *up* déterminées

3. Une solution possible aurait été d'effectuer le tirage équilibré parmi l'ensemble des UP appartenant aux UC sélectionnées, et non en en tirant une dans chaque UP, mais alors la coordination n'aurait pas de sens, puisque certaines UC se retrouveraient sans UP de l'EM.

4. L'utilisation de la méthode de partage des poids est courante à l'Insee dans le cadre de tirage d'individus quand il n'existe pas de base de sondage nationale de la population d'intérêt, dans le cadre d'un panel rotatif ou dans le cadre d'une enquête pour laquelle un échantillon est constitué de plusieurs sous-échantillons issus de bases de sondage différentes.

au point 2.3.2. La pondération de l'unité de coordination uc va être une somme pondérée des poids w_{up} des unités primaires up ayant permis de sélectionner uc. Pour les pondérations w_{up} des unités primaires, on retient l'inverse des probabilités d'inclusion des unités primaires $w_{up} = \frac{1}{\pi_{up}}$. Elles permettent de calculer un estimateur d'Horvitz-Thompson sans biais sur l'échantillon d'unités primaires.

La notion de lien est centrale pour le calcul des pondérations dans la théorie du partage des poids. Une unité de coordination uc est liée à une unité primaire up si l'échantillonnage de up permet de tirer uc dans l'échantillon d'unités de coordination. Dans le cas présent, il est très facile d'identifier un lien entre unités primaires et unités de coordination. Il existe un lien entre up et uc si up est incluse dans l'unité de coordination uc. Chaque UP est donc reliée à une seule unité de coordination. Une unité de coordination est par contre reliée à autant d'unités primaires que le nombre d'unités primaires qui la composent (cf. section 3.3 de la partie B détaillant la construction des UC).

La méthode généralisée du partage des poids (Deville et Lavallée, 2006) définit les liens pondérés $\tilde{\theta}_{uc,up}$ entre les UC et les UP tels que :

— lorsqu'il existe un lien entre uc et up, $\tilde{\theta}_{uc,up} \geq 0$
— lorsqu'il n'existe pas de lien entre uc et up, $\tilde{\theta}_{uc,up} = 0$
— la somme des liens pondérés de l'UC uc avec l'ensemble U_{up} des UP vaille 1 (y compris les UP ne figurant pas dans l'échantillon-maître) : $\sum_{up \in U_{up}} \tilde{\theta}_{uc,up} = 1$

Pour toutes les unités de coordination uc indirectement tirées, ces liens permettent de calculer des pondérations qui garantissent des estimations sans biais issus de l'échantillon d'UC :

$$w_{uc} = \sum_{up \in EM} \tilde{\theta}_{uc,up}\ w_{up} \tag{4.1}$$

Autrement dit, n'importe quelle moyenne pondérée des poids des UP respectant les 3 contraintes ci-dessus peut être utilisée pour calculer la pondération des UC tirées. Néanmoins, les liens pondérés ont une influence importante sur la variance des estimateurs. En effet, dans la formule (4.1), chaque unité primaire liée à uc et tirée dans l'EM contribue à hauteur de $\tilde{\theta}_{uc,up}\ w_{up}$ au poids associé à uc. Si cette contribution varie selon les unités primaires qui la composent, uc n'a pas le même poids selon l'unité primaire qui a été tirée et a permis de sélectionner uc dans l'échantillon d'unités de coordination. Une part de la variance des estimateurs sur les unités de coordination vient alors de la variation du poids de uc en fonction de l'échantillon-maître retenu. Cette source de variance peut être limitée par le choix de la valeur des liens $\tilde{\theta}_{uc,up}$.

Une solution intéressante est de choisir des liens pondérés tels que la contribution de chacune des unités primaires associée à une unité de coordination uc pour la pondération de uc soit identique lorsqu'elles sont tirées dans l'échantillon-maître. On retient donc :

$$\begin{aligned}\tilde{\theta}_{uc,up} &= \frac{\pi_{up}}{\sum_{up\in uc}\pi_{up}} \text{ si } up \in uc \\ &= 0 \text{ sinon}\end{aligned} \tag{4.2}$$

En combinant la formule du lien pondéré avec (4.1) et (4.2), et en notant $|E|$ le cardinal de l'ensemble E, on en déduit le poids des unités de coordination tirées :

$$\begin{aligned}w_{uc} &= \sum_{up\in EM} \tilde{\theta}_{uc,up}\ w_{up} = \sum_{up\in EM} \frac{\pi_{up}1_{up\in uc}}{\sum_{up\in uc}\pi_{up}}\frac{1}{\pi_{up}} \\ w_{uc} &= \frac{|up \in EM \cap uc|}{\sum_{up\in uc}\pi_{up}}\end{aligned} \tag{4.3}$$

Le poids utilisé pour les estimations sur les unités de coordination est donc le ratio entre le nombre d'unités primaires qui composent *uc* et qui figurent dans l'échantillon-maître, et la somme des probabilités d'inclusion dans l'EM des unités primaires qui composent *uc*. Ce poids présente l'avantage de ne pas varier selon l'unité primaire qui permet de tirer indirectement *uc*.

Néanmoins, le poids de l'UC *uc* varie en fonction du nombre d'unités primaires composant *uc* qui ont été tirées. Ce nombre est très souvent égal à 1 puisque les unités de coordinations regroupent des unités primaires voisines (cf. le paragraphe 3.1.2.1 de la partie B) et grâce à la propriété d'équilibrage spatiale qui diminue la probabilité d'inclusion double d'unités primaires voisines, comme évoqué en section 1.2. Toutefois, dans chaque simulation de tirage, plusieurs unités de coordination sont tirées par l'intermédiaire de plusieurs unités primaires, ce qui engendre une variabilité du poids pour une UC donnée.

Le poids des UC décrit par la formule (4.3) sera utilisé pour le calcul des pondérations des secteurs emploi décrites au chapitre 5.

4.2.3 Les variables Z du partage des poids pour l'équilibrage de l'EEC

L'équilibrage de l'échantillon Emploi ne peut se faire que sur la base de sondage dans laquelle il est tiré, c'est-à-dire sur les estimateurs calculés à partir de l'ensemble des secteurs figurant dans les UC tirées directement ou indirectement. Les estimateurs calculés à partir des UC tirées doivent donc être les plus précis possibles sur les indicateurs liés à l'emploi détaillés dans la section 3.2. Autrement dit, si le tirage direct ou indirect des UC présente une variance élevée sur les indicateurs de l'emploi, les estimateurs issus du tirage des secteurs incluront *a minima* la variance de tirage des UC.

L'équilibrage des UC dans le cas du tirage direct

Dans le cas du tirage direct des UC, l'équilibrage sur les indicateurs de l'emploi est facile à faire. Il suffit d'intégrer les indicateurs de l'emploi à la suite des variables d'équilibrage des UC, au même titre que les variables d'équilibrage des UP décrites au point 2.2.3 et calculées au niveau des UC. Les UC seront alors bien équilibrées sur les indicateurs de l'emploi et sur les indicateurs d'intérêt pour les autres enquêtes ménages.

L'équilibrage des UC dans le cas du tirage indirect

La situation est différente dans le cadre du tirage indirect. En effet, ce sont les UP qui sont équilibrées à l'issue du tirage et non les UC puisque l'échantillon d'UC est sélectionné indirectement. Intégrer des variables d'emploi calculées au niveau des UP parmi les variables d'équilibrage des UP ne garantirait donc pas la précision des estimateurs sur les variables d'emploi calculés à partir de l'échantillon d'UC.

Pour disposer d'un échantillon d'UC précis sur les estimateurs d'emploi, on procède donc différemment. On note s_{uc} l'échantillon d'unités de coordination, X une variable d'équilibrage, t_X son total, x_{up} sa valeur pour l'unité primaire up et x_{uc} sa valeur pour l'unité de coordination uc. Dans l'équilibrage des UP, on souhaite satisfaire la contrainte suivante sur les UC :

$$\begin{aligned} \sum_{uc \in s_{uc}} w_{uc}\, x_{uc} &= t_X \\ \sum_{uc \in s_{uc}} x_{uc} \sum_{up \in EM} \tilde{\theta}_{uc,up}\, w_{up} &= t_X \text{ en utilisant (4.1)} \\ \sum_{up \in EM} w_{up}\, (\tilde{\theta}_{uc,up}\, x_{uc}) &= t_X \end{aligned} \tag{4.4}$$

La double somme ci-dessus se simplifie puisque chaque up n'est reliée qu'à une uc. Ainsi, on peut définir une variable Z telle que $z_{up} = \tilde{\theta}_{uc,up}\, x_{uc}$ où uc est l'unité de coordination associée à l'unité primaire up. L'introduction de cette variable dite transformée Z parmi les variables d'équilibrage des unités primaires et calculée au niveau des unités primaires garantit donc l'équilibrage sur la variable X pour l'échantillon d'UC.

Cette propriété des variables transformées détaillée dans l'ouvrage Lavallée, 2007 permet ainsi d'équilibrer les unités de coordination sur les variables d'emploi bien que les unités de coordination ne soient pas tirées directement. On peut noter que ces variables n'ont pas de sens socio-économique pour les unités primaires. Cela ne garantit donc en rien la précision de l'échantillon-maître primaires sur les variables d'emploi. Seul l'échantillon d'UC bénéficie de cet équilibrage sur la variable transformée Z.

Ainsi, les variables utilisées pour l'équilibrage des UP sont dans l'ordre :
- la liste de variables retenues pour l'équilibrage des UP décrites dans le point 2.2.3 ;
- les variables transformées issues des variables sur lesquelles on souhaite équilibrer par la suite l'échantillon Emploi et décrites dans la section 3.2.

4.2.4 Le choix de la méthode indirecte

Les deux méthodes ont été comparées d'une part, pour la précision de l'échantillon-maître, à l'aide de différentes variables liées à plusieurs enquêtes au niveau des unités primaires, d'autre part, pour la précision de l'EEC, sur des variables relatives à l'emploi, calculées au niveau des unités de coordination. Cette comparaison suppose que la précision du tirage des secteurs au sein des UC découle directement de la qualité de l'équilibrage des unités de coordination au premier degré de tirage. On fait ainsi l'hypothèse que, lorsqu'on sélectionne les secteurs en ayant préalablement tiré des UC, la variance de l'ensemble des degrés de tirage est moindre si la variance associée au tirage des UC est plus faible. Si cette hypothèse paraît crédible, elle n'a pas été testée.

Dans les tableaux 4.1 et 4.2, la précision d'estimateurs sur des variables d'équilibrage obtenus respectivement sur les UP et sur les UC est présentée pour le tirage direct et pour le tirage indirect. La précision de ces estimateurs a été calculée par méthode de Monte Carlo sur 100 000 tirages d'échantillons de 567 UP pour chacune des deux méthodes, en utilisant des unités de coordination construites avec au moins 7 500 résidences principales. Les estimateurs calculés à partir des UP servent à évaluer la précision de la méthode de tirage pour l'EM, tandis que les estimateurs calculés à partir des UC permettent de connaître la précision de la base de sondage qui sera utilisée pour le tirage des secteurs emploi sur les indicateurs de l'emploi.

D'après le tableau 4.1, le tirage indirect aboutit à un échantillon-maître de bien meilleure qualité que le tirage direct pour l'EM. Sur les variables directement liées à l'emploi, et calculées au niveau des unités de coordination, la différence de précision issue de chacune des deux méthodes est très peu marquée (tableau 4.2) [5]. Pour d'autres variables, moins centrales et ne figurant pas dans ce tableau, le tirage direct est légèrement plus précis.

TABLE 4.1 – Coefficients de variation issus des scénarios de tirage direct et indirect pour des indicateurs estimés sur les unités primaires tirées

Scénario	Revenu total	Nombre de cadres	Nombre de 15-29 ans
Tirage direct	0,35 %	0,68 %	0,33 %
Tirage indirect	0,16 %	0,33 %	0,17 %

Note de lecture : Le tirage direct des UC conduit à un coefficient de variation de 0,68% pour la variable de nombre de cadres estimée sur les UP. Le tirage indirect d'UC conduit à un coefficient de variation de 0,33% pour la variable de nombre de cadres estimée sur les UP.

5. Notons que les variables utilisées ici sont intégrées dans le modèle d'équilibrage des deux méthodes. Les résultats obtenus permettent avant tout de mesurer la qualité des deux équilibrages, plutôt que de valider l'utilisation de ces échantillons pour des variables d'intérêt corrélées aux variables d'équilibrage.

TABLE 4.2 – Coefficients de variation issus des scénarios de tirage direct et indirect pour des indicateurs estimés sur les unités de coordination sélectionnées

Scénario	Nombre de 15 à 64 ans percevant une allocation chômage ou préretraite	Nombre de 15 à 64 ans percevant des revenus d'activité
Tirage direct	0,17 %	0,10 %
Tirage indirect	0,16 %	0,10 %

Note de lecture : Le tirage direct des UC conduit à un coefficient de variation de 0,17% pour la variable de nombre de 15-64 ans percevant une allocation chômage ou préretraite estimée sur les UC.

L'étude de ces deux scénarios de tirage aboutit à la conclusion que la méthode de tirage indirect présente plus de garanties méthodologiques, ainsi qu'une meilleure précision pour l'échantillon-maître et une précision similaire au tirage direct pour l'enquête Emploi. C'est donc cette méthode qui est retenue, pour laquelle il faut à présent déterminer les différents paramètres du plan de sondage.

4.3 La définition des paramètres du tirage indirect

La première étape de la méthode indirecte est le tirage de l'EM. Les paramètres du plan de sondage de l'EM ont déjà été présentés en sections 2.1 et 2.2. Pour rappel, l'EM est sélectionné par tirage spatialement équilibré au niveau Nuts2, suivi d'une nouvelle phase de vol avec les unités primaires sur lesquelles l'algorithme n'a pas encore statué en les regroupant au niveau national, et en équilibrant les résultats à ce même niveau national, puis en abandonnant une à une les variables d'équilibrage lors de la phase d'atterrissage. Les variables d'équilibrage sont les suivantes, déjà présentées au point 2.2.3, auxquelles s'ajoutent les variables indirectes de l'emploi découlant de la liste de variables déterminées en section 3.2 :

- probabilités d'inclusion des unités primaires

- 14 premiers axes d'une ACP calculée sur les variables calculées au niveau des unités primaires :
  - nombre de ménages par type de ménage (familles monoparentales, personnes seules, couples avec enfant(s), couples sans enfant)
  - somme de la surface des résidences principales
  - somme de la valeur locative des résidences principales
  - nombre de femmes
  - somme des revenus par type de revenu (salaires, pensions de retraite, allocations chômage ou préretraite)
  - nombre de logements vacants
  - nombre de résidences principales qui sont des logements sociaux
  - nombre de ménages imposables à l'impôt sur la fortune
  - nombre d'individus scolarisés par tranche d'âge (2-5 ans, 6-10 ans, 11-14 ans, 15-17 ans, 18-24 ans, 25-29 ans, 30 ans ou plus)

  - nombre de résidences principales en milieu périurbain et nombre de résidences principales en milieu rural
  - nombre de résidences principales par tranche d'unité urbaine (tranches 1 à 8)

- variables calculées au niveau des unités primaires :
  - population par tranche d'âge (15-29 ans, 30-44 ans, 45-59 ans, 60-74 ans, 75-89 ans, 90 ans et plus)
  - somme des revenus des individus
  - nombre de ménages propriétaires
  - nombre d'individus par PCS en 8 catégories (PCS 2 à 8)
  - nombre de résidences principales en quartier prioritaire de la ville
  - nombre d'individus de 15 ans ou plus, non scolarisés, ayant un niveau de diplôme donné (niveau certificat d'études primaires ou brevet des collèges, niveau CAP-BEP, niveau baccalauréat, diplômés du supérieur)
  - nombre d'individus par statut d'activité et d'emploi (nombre de 15-64 ans actifs occupés, nombre de 15-64 ans au chômage, population en emploi salarié
  - somme des bénéfices par type de bénéfices (bénéfices industriels et commerciaux, bénéfices agricoles, bénéfices non commerciaux)

- variables transformées issues de variables calculées par unité de coordination :
  - nombre d'individus percevant une allocation chômage ou préretraite
  - nombre d'individus percevant un revenu d'activité
  - nombre de résidences principales en quartier prioritaire de la ville
  - nombre de résidences principales par type de ménage (personnes seules, familles monoparentales, couples avec enfant(s), couples sans enfant)
  - nombre d'individus par tranche d'âge croisée au sexe [6]
  - somme des revenus disponibles

Les paramètres liés au tirage des UP sont figés par les travaux menés dans les sections 2.2, 2.3 et au point 4.2.3 en raison du choix de la méthode indirecte qui conduit à échantillonner d'abord des unités primaires. Il reste à présent à définir les paramètres du plan de sondage dont l'influence se mesure sur le tirage de l'échantillon de l'EEC au deuxième degré.

4.3.1 Liste des paramètres à fixer

Les différents paramètres à définir pour le plan de sondage sont les suivants :

- Niveau géographique auquel est réalisé l'atterrissage de l'algorithme d'équilibrage pour le tirage des secteurs : l'atterrissage peut se faire au niveau national (mise en commun des secteurs sur lesquels il reste à statuer au niveau national), de chaque région, ou de chaque UC ;
- Niveau géographique auquel est stratifié le tirage des secteurs : la sélection des secteurs peut être stratifiée par régions (Nuts2) ou par unités de coordination [7] ;

6. Hommes de 18-25 ans, femmes de 18-25 ans, hommes de 26-50 ans, femmes de 26-50 ans, hommes de 51-75 ans, femmes de 51-75 ans, individus de 76 ans ou plus.

7. L'option de réaliser le tirage parmi l'ensemble des secteurs des unités de coordination sélection-

— Taille des unités de coordination : au moment de la constitution des unités de coordination (voir le chapitre 3 de la partie B), le choix s'est porté sur des UC contenant un minimum de 7 500 ou 10 000 résidences principales ; il reste à statuer entre ces deux possibilités.

4.3.2 Comparaison des scénarios

Pour statuer entre les différentes valeurs de ces paramètres, il est nécessaire d'établir les critères sur lesquels effectuer ces choix, et la stratégie de comparaison de l'influence de ces paramètres sur la qualité du tirage. Les paramètres restant à définir n'influant que sur le tirage de deuxième degré des secteurs de l'EEC, les critères de comparaison retenus sont les suivants [8] :

— Précision nationale pour des variables corrélées aux variables d'intérêt de l'enquête Emploi
  — nombre d'individus percevant des allocations chômage ou préretraite (nommée également « nombre de chômeurs » par abus de langage) ;
  — nombre d'individus percevant des revenus d'activité (nommée également « nombre d'actifs » par abus de langage) ;
  — total des salaires ;
— Précision régionale pour ces mêmes variables.

Pour le calcul de ces précisions à partir des échantillons de secteurs, au minimum 10 000 simulations de tirage sont effectuées pour chaque scénario de tirage. Les résultats régionaux ne sont présentés que sur quelques régions pour simplifier la lecture des tableaux.

Les valeurs possibles pour ces différents paramètres, amènent à un nombre de scénarios difficilement comparables (jusqu'à 10 scénarios à allocations constantes). Aussi, la stratégie employée a été dans un premier temps d'étudier les paramètres un à un en figeant les autres paramètres. Cela a permis de sélectionner la valeur du paramètre d'intérêt pour laquelle les critères de comparaison était meilleure que pour les autres valeurs de ce même paramètre, toutes choses égales par ailleurs [9]. Dans un deuxième temps, le nombre de scénarios ayant drastiquement diminué, il a été possible de faire varier plusieurs paramètres en même temps.

4.3.2.1 Un premier paramètre : le niveau d'atterrissage

Deux niveaux géographiques de stratification sont possibles pour le tirage des secteurs : l'unité de coordination et la région (qui regroupe les secteurs des unités de co-

nées, sans stratification par régions, est écartée du fait de l'exigence régionale de précision d'Eurostat.

8. On peut noter que, à l'exception de la variable de salaires, il s'agit de variables d'équilibrage des secteurs Emploi, comme introduit en section 3.2. La variable de salaires est partiellement contenue dans la variable de revenus disponibles utilisée pour l'équilibrage, mais cette dernière intègre d'autres sources de bénéfices. En outre, elle est calculée après perception d'allocations sociales. Donc, bien que corrélée à la variable de revenus disponibles, la variable de salaires n'est pas une variable d'équilibrage. Sauf mention contraire, les résultats obtenus sur ces variables semblent se généraliser sur d'autres variables qui n'ont pas servi à l'équilibrage.

9. Par exemple, le paramètre de niveau d'atterrissage de l'équilibrage de l'EEC n'a été testé que pour une taille fixe d'UC.

ordination tirées). Il est à chaque fois possible de combiner ce paramètre de niveau de stratification avec le niveau d'atterrissage de l'algorithme d'équilibrage. Le choix du paramètre pour l'atterrissage est donc effectué à niveau de stratification donné. Dans un premier temps, on s'intéresse au paramètre d'atterrissage dans le cas d'un tirage de secteurs stratifié par UC. Dans un second temps, on étudie le paramètre d'atterrissage dans le cas d'un tirage de secteurs stratifié par régions.

Dans cette section, tous les scénarios présentés incluent un tirage de 580 UP en première étape et se basent sur des UC construites avec un seuil de 7 500 résidences principales.

Tirage des secteurs stratifié par unités de coordination

Trois niveaux d'atterrissage sont comparés pour un tirage équilibré de secteurs stratifié par unités de coordination :

— un atterrissage au niveau de l'unité de coordination ;
— un atterrissage au niveau de chaque région ;
— un atterrissage au niveau national.

L'atterrissage au niveau national est écarté en raison du temps élevé nécessaire à l'algorithme pour aboutir au tirage d'un échantillon de secteurs tirés [10]. L'atterrissage au niveau de l'UC s'avère de moins bonne qualité pour les indicateurs considérés, que ce soit pour les estimations nationales ou régionales (cf. tableaux 4.3 et 4.4). Donc, dans le cas où le tirage des secteurs est stratifié par unités de coordination, l'atterrissage se fera au niveau régional.

TABLE 4.3 – Coefficients de variation des estimateurs calculés sur l'échantillon Emploi au niveau national lors d'un tirage de secteurs stratifié par unités de coordination

Scénario	Nombre de chômeurs	Nombre d'actifs occupés
Atterrissage régional	0,26 %	0,18 %
Atterrissage dans les UC	0,28 %	0,19 %

Note de lecture : le coefficient de variation du nombre de chômeurs en France métropolitaine est de 0,26% dans le cas où le tirage des secteurs est stratifié par UC avec une phase d'atterrissage régional, en ayant au préalable tiré 580 UP et en utilisant des UC ayant au minimum 7 500 résidences principales.

10. Cette durée de l'algorithme est probablement due au nombre de contraintes non satisfaites à l'issue de la fin de la phase de vol dans les unités de coordination. En particulier, l'atterrissage au niveau national ne présente pas cette même problématique lorsque le tirage des secteurs est stratifié par régions.

TABLE 4.4 – Coefficients de variation des estimateurs calculés sur l'échantillon Emploi au niveau régional lors d'un tirage de secteurs stratifié par unités de coordination

Scénario	Nombre de chômeurs			Nombre d'actifs occupés		
	Île-de-France	Champagne-Ardenne	Provence-Alpes-Côte d'Azur	Haute-Normandie	Poitou-Charentes	Midi-Pyrénées
Atterrissage régional	0,68 %	2,40 %	1,06 %	1,90 %	2,46 %	1,47 %
Atterrissage par UC	0,79 %	2,43 %	1,18 %	1,95 %	2,46 %	1,54 %

Note de lecture : le coefficient de variation du nombre de chômeurs en Provence-Alpes-Côte d'Azur est de 1,06% dans le cas où le tirage des secteurs est stratifié par UC avec une phase d'atterrissage régional, en ayant au préalable tiré 580 UP et en utilisant des UC ayant au minimum 7 500 résidences principales.

Tirage des secteurs stratifié par régions

Deux niveaux d'atterrissage sont comparés pour un tirage équilibré de secteurs stratifié par régions :

— un atterrissage au niveau de chaque région ;
— un atterrissage au niveau national.

Le tableau 4.6 montre qu'un atterrissage régional donne logiquement de meilleurs résultats sur la précision des variables au niveau régional qu'un atterrissage national ; à l'inverse ces résultats sont légèrement moins bons quant à la précision des totaux métropolitains, mais l'écart est de moindre mesure (tableau 4.5). La précision demandée par Eurostat se situant notamment au niveau Nuts2, l'atterrissage au niveau de chaque région est préféré.

TABLE 4.5 – Coefficients de variation des estimateurs calculés sur l'échantillon Emploi au niveau national lors d'un tirage des secteurs stratifié régionalement

Scénario	Nombre de chômeurs	Nombre d'actifs occupés
Atterrissage national	0,26 %	0,21 %
Atterrissage régional	0,27 %	0,22 %

Note de lecture : le coefficient de variation du nombre de chômeurs en France métropolitaine est de 0,27% dans le cas où le tirage des secteurs est stratifié par régions avec une phase d'atterrissage régional, en ayant au préalable tiré 580 UP et en utilisant des UC ayant au minimum 7 500 résidences principales.

TABLE 4.6 – Coefficients de variation des estimateurs calculés sur l'échantillon Emploi au niveau régional lors d'un tirage des secteurs stratifié régionalement

Scénario	Nombre de chômeurs			Nombre d'actifs occupés		
	Île-de-France	Champagne-Ardenne	Bretagne	Haute-Normandie	Poitou-Charentes	Midi-Pyrénées
Atterrissage national	0,72 %	2,54 %	1,41 %	1,99 %	2,59 %	1,61 %
Atterrissage régional	0,72 %	2,43 %	1,38 %	2,00 %	2,51 %	1,59 %

Note de lecture : le coefficient de variation du nombre de chômeurs en Champagne-Ardenne est de 2,54% dans le cas où le tirage des secteurs est stratifié par régions avec une phase d'atterrissage national, en ayant au préalable tiré 580 UP et en utilisant des UC ayant au minimum 7 500 résidences principales.

Conclusion

Quel que soit le niveau géographique de stratification du tirage des secteurs, on observe que l'atterrissage au niveau régional est préférable. C'est donc cette option qui est retenue.

4.3.2.2 Un deuxième paramètre : le niveau géographique de stratification pour le tirage des secteurs

Une fois l'EM tiré et les UC sélectionnées, deux niveaux de stratification pour le tirage sont envisagés :

- on peut mettre en commun l'ensemble des secteurs des UC sélectionnées au sein d'une région et effectuer le tirage de l'échantillon de l'EEC région par région, parmi ces bases régionales (stratification régionale) ;
- on peut effectuer un tirage de secteurs dans chaque UC sélectionnée (stratification par UC).

Pour statuer sur ce paramètre, notons que la phase d'atterrissage est régionale dans tous les scénarios testés. Dans cette section, tous les scénarios présentés incluent un tirage de 580 UP en première étape et se basent sur des UC construites avec un seuil de 10 000 résidences principales.

Les résultats des tableaux 4.7 et 4.8 montrent que le tirage stratifié par unités de coordination est plus précis pour les variables les plus corrélées au taux de chômage national ou régional (indicateur phare de l'EEC) [11]. Aussi, le choix se porte sur un tirage équilibré stratifié par unités de coordination, avec atterrissage régional.

11. Ce résultat est à nuancer, car sur d'autres variables telles que le nombre de résidences principales en quartier prioritaire de la ville, le tirage stratifié par régions est plus précis. De plus, bien que non présentés dans ce document, des scénarios similaires à ceux des tableaux 4.7 et 4.8 avec des UC construites avec un seuil plus petit ont montré que la meilleure précision du tirage stratifié par UC sur les principaux indicateurs de l'emploi est moins marquée lorsque les UC sont plus petites. Toutefois, comme les UC qui seront finalement retenues sont construites avec un seuil de 10 000 résidences principales, ce sont bien les résultats à partir de cette taille d'unités de coordination qui sont à prendre en considération.

TABLE 4.7 – Coefficients de variation des estimateurs calculés sur l'échantillon Emploi au niveau national

Scénario	Nombre de chômeurs	Nombre d'actifs occupés	Totaux de salaires
Tirage par région	0,28 %	0,23 %	0,38 %
Tirage par UC	0,25 %	0,17 %	0,35 %

Note de lecture : le coefficient de variation du nombre de chômeurs en France métropolitaine est de 0,25% dans le cas d'un tirage de secteurs stratifié par UC, avec atterrissage régional, en ayant au préalable tiré 580 UP et utilise des unités de coordination avec un minimum de 10 000 résidences principales.

TABLE 4.8 – Coefficients de variation des estimateurs calculés sur l'échantillon Emploi au niveau régional

Scénario	Nombre de chômeurs			Nombre d'actifs occupés		
	Île-de-France	Bourgogne	Alsace	Île-de-France	Bourgogne	Alsace
Tirage par région	0,72 %	2,40 %	2,18 %	0,54 %	2,35 %	1,61 %
Tirage par UC	0,64 %	2,26 %	2,15 %	0,34 %	2,15 %	1,42 %

Note de lecture : le coefficient de variation du nombre de chômeurs en Bourgogne est de 2,26% dans le cas d'un tirage de secteurs stratifié par UC, avec atterrissage régional, en ayant au préalable tiré 580 UP et utilise des unités de coordination avec un minimum de 10 000 résidences principales.

4.3.2.3 Un dernier paramètre : la taille des UC

Nous avons vu en section 3.2 de la partie B que deux arguments s'opposaient pour définir la taille minimale adéquate des UC. Des UC de 10 000 résidences principales permettraient d'obtenir une meilleure précision pour les indicateurs issus de l'EEC (toutes choses égales par ailleurs), mais la coordination des deux échantillons (EEC et EM) sera moindre. En effet, les UC ayant au moins 10 000 résidences principales tirées indirectement couvrent une part plus importante du territoire que celles ayant 7 500 résidences principales. La dispersion géographique des secteurs tirés sera donc plus élevée dans des UC d'au minimum 10 000 résidences principales. Les enquêteurs mettraient ainsi plus de temps à enquêter l'ensemble des logements de l'échantillon de l'EEC, devant parcourir une plus grande distance. Une façon de contrer cette problématique est de diminuer le nombre de secteurs enquêtés (et donc la précision des indicateurs). Pour statuer sur la taille minimale des UC, plusieurs scénarios ont donc été comparés, en faisant varier, d'une part, les allocations de secteurs, d'autre part la taille minimale des UC.

Les tableaux 4.9 et 4.10 confirment en premier lieu le fait qu'à taille d'échantillon égale, un seuil de 10 000 résidences principales pour les UC donne toujours des indicateurs plus précis qu'un seuil de 7 500, et particulièrement pour des estimateurs régionaux. Pour confirmer l'influence de ce paramètre sur la précision, des simulations ont également été réalisées avec un seuil de 5000 résidences principales pour la constitution des UC. Les

scénarios présentés dans ces tableaux ont été réalisés avec un tirage de 580 unités primaires et un tirage de secteurs stratifié par unités de coordination avec atterrissage régional.

TABLE 4.9 – Coefficients de variation des estimateurs calculés sur l'échantillon Emploi au niveau national pour différents seuils de constitution d'unités de coordination

Taille seuil des UC	Nombre de chômeurs	Nombre d'actifs occupés	Totaux de salaires
5 000	0,27 %	0,18 %	0,37 %
7 500	0,26 %	0,18 %	0,35 %
10 000	0,25 %	0,17 %	0,35 %

Note de lecture : le coefficient de variation du nombre de chômeurs en France métropolitaine est de 0,25% lorsque le tirage des secteurs est stratifié par UC avec atterrissage régional, en utilisant des unités de coordination comptant au minimum 10 000 résidences principales et en ayant au préalable tiré 580 UP.

TABLE 4.10 – Coefficients de variation des estimateurs calculés sur l'échantillon Emploi au niveau régional pour différents seuils de constitution d'unités de coordination

Taille seuil des UC	Nombre de chômeurs			Nombre d'actifs occupés		
	Champagne -Ardenne	Picardie	Haute-Normandie	Champagne -Ardenne	Picardie	Haute-Normandie
5 000	2,61 %	2,46 %	2,51 %	2,60 %	2,12 %	2,05 %
7 500	2,40 %	2,44 %	2,33 %	2,37 %	2,06 %	1,90 %
10 000	2,30 %	2,08 %	2,20 %	2,24 %	1,88 %	1,79 %

Note de lecture : le coefficient de variation du nombre de chômeurs en Picardie est de 2,08% lorsque le tirage des secteurs est stratifié par UC avec atterrissage régional, en utilisant des unités de coordination comptant au minimum 10 000 résidences principales et en ayant au préalable tiré 580 UP.

Ainsi, il est préférable d'utiliser des unités de coordination comptant au minimum 10 000 résidences principales du point de vue de la précision. Mais cela n'est possible qu'en cas de diminution du nombre de secteurs pour ne pas augmenter la charge de collecte pour les enquêteurs. Il s'agit donc d'étudier l'impact de cette diminution sur la précision des indicateurs liés à l'emploi.

L'enseignement suivant est qu'une diminution de 10 % du nombre de secteurs tirés à taille d'UC constante permet d'obtenir, dans la plupart des régions, des indicateurs plus précis que lorsqu'on diminue l'étendue des UC tout en conservant l'allocation initiale. En effet, d'après le tableau 4.11, pour un tirage de 541 UP [12], le coefficient de variation du

12. Ce jeu d'UP se rapproche de celui retenu pour le tirage effectif de l'EM. Il comporte donc quelques surreprésentations régionales.

scénario avec diminution de la taille des UC excède de 0,05 points de pourcentage ou plus celui avec diminution du nombre de secteurs dans 6 régions sur 22 pour l'indicateur du nombre d'individus percevant une allocation chômage ou préretraite et dans 15 régions sur 22 pour l'indicateur du nombre d'individus percevant un revenu d'activité. Dans les autres régions, ces coefficients sont sensiblement identiques. Ces deux variables étant des variables d'équilibrage, ces résultats ont également été vérifiés sur le total des salaires, qui ne sert pas à l'équilibrage. On observe alors des résultats comparables à ceux pour le nombre d'actifs.

TABLE 4.11 – Coefficients de variation des estimateurs calculés sur l'échantillon Emploi au niveau régional pour différentes allocations de secteurs et tailles d'UC

	Nombre de chômeurs			Nombre d'actifs occupés		
	UC 7500 3216 sect	UC 10 000 2887 sect	UC 10 000 3216 sect	UC 7500 3216 sect	UC 10 000 2887 sect	UC 10 000 3216 sect
Île-de-France	0,73 %	0,73 %	0,70 %	0,37 %	0,38 %	0,36 %
Champagne-Ardenne	2,17 %	2,10 %	2,00 %	1,91 %	1,90 %	1,84 %
Picardie	2,26 %	2,04 %	1,96 %	1,84 %	1,74 %	1,68 %
Haute-Normandie	2,32 %	2,24 %	2,19 %	1,89 %	1,81 %	1,77 %
Centre	1,90 %	1,92 %	1,80 %	1,51 %	1,51 %	1,46 %
Basse-Normandie	2,25 %	2,34 %	2,19 %	2,26 %	2,13 %	2,07 %
Bourgogne	2,26 %	2,23 %	2,12 %	2,09 %	1,98 %	1,90 %
Nord-Pas-de-Calais	1,24 %	1,26 %	1,19 %	1,08 %	1,06 %	1,03 %
Lorraine	2,10 %	2,05 %	1,95 %	2,08 %	1,96 %	1,95 %
Alsace	2,32 %	2,33 %	2,24 %	1,56 %	1,53 %	1,51 %
Franche-Comté	2,29 %	2,36 %	2,27 %	2,12 %	2,06 %	2,03 %
Pays de la Loire	1,36 %	1,39 %	1,33 %	1,50 %	1,45 %	1,42 %
Bretagne	1,30 %	1,31 %	1,23 %	1,53 %	1,46 %	1,42 %
Poitou-Charentes	2,04 %	2,05 %	1,97 %	2,19 %	2,10 %	2,05 %
Aquitaine	1,59 %	1,61 %	1,57 %	1,43 %	1,38 %	1,36 %
Midi-Pyrénées	1,80 %	1,84 %	1,80 %	1,54 %	1,48 %	1,44 %
Limousin	2,31 %	2,07 %	1,92 %	2,56 %	2,40 %	2,36 %
Rhône-Alpes	1,26 %	1,29 %	1,24 %	0,81 %	0,80 %	0,77 %
Auvergne	2,45 %	2,43 %	2,30 %	2,39 %	2,23 %	2,18 %
Languedoc-Roussillon	1,83 %	1,83 %	1,77 %	1,88 %	1,81 %	1,79 %
Provence-Alpes-Côte d'Azur	1,18 %	1,20 %	1,14 %	0,97 %	0,97 %	0,96 %
Corse	5,65 %	4,53 %	4,45 %	3,43 %	3,22 %	3,15 %

Note de lecture : le coefficient de variation du nombre de chômeurs dans le Limousin est de 2,07% lorsque le tirage des secteurs est stratifié par unités de coordination avec atterrissage régional, en utilisant des unités de coordination comptant au minimum de 10 000 résidences principales, avec une réduction de 10% du nombre de secteurs (2887 secteurs) et en ayant au préalable tiré 541 UP.

Ainsi, une diminution du nombre de secteurs permet d'obtenir des résultats plus précis au niveau régional que l'abaissement du seuil pour la construction des unités de coordinations. Ces résultats plaident pour fixer le seuil minimal à 10 000 résidences principales pour les UC.

En résumé, le tirage de deuxième degré des secteurs est un tirage spatialement équilibré, stratifié par unités de coordination, avec atterrissage régional, les unités de coordination étant des regroupements d'unités primaires avec un seuil minimal de 10 000 résidences principales par UC.

Il reste à présent à définir les allocations des tirages, c'est-à-dire le nombre d'UP sélectionnées par région et le nombre de secteurs tirés.

Chapitre 5

Détermination des allocations finales

5.1 Objectif à atteindre, contraintes et paramètres ajustables pour la définition des allocations

La détermination des allocations résulte d'un équilibre entre précision souhaitée et coût acceptable. Ainsi, à nombre égal de logements tirés dans chaque UP au deuxième degré, plus le nombre d'unités primaires de l'EM augmente, plus les indicateurs obtenus à partir des enquêtes seront précis, mais plus la ressource enquêteur devra être conséquente. De même, un plus grand nombre de secteurs sélectionnés permettra d'obtenir de meilleurs résultats, mais aura un coût plus élevé.

Le règlement européen IESS définit une précision attendue pour les enquêtes auprès des ménages, enquête Emploi compris, au niveau national d'une part, et au niveau Nuts2 d'autre part. Les allocations doivent permettre d'atteindre cet objectif. Par ailleurs, la précision des échantillons-maître et emploi de 2009 donne une limite basse à ne pas franchir. En d'autres termes, une fois prises en compte toutes les étapes de tirage, on souhaite que la précision des enquêtes issues du nouvel échantillon-maître et du nouvel échantillon de l'enquête Emploi soit au moins aussi bonne qu'avec les anciens plans de sondage de l'EM 2009 et de l'EEC 2009.

L'utilisation accrue du multimode et les contraintes budgétaires plaident pour une diminution globale de la quotité « enquêteur », et donc des allocations des échantillons.

Étant donnés ces éléments, l'objectif se présente donc comme la définition d'un nombre minimal d'unités primaires et de secteurs au sein de chaque région, permettant d'atteindre la précision demandée par Eurostat, sans diminuer la qualité déjà existante avec les échantillons actuellement utilisés.

Le point 5.1.1 présente l'impact des allocations d'UP sur la précision des enquêtes tirées dans l'EM, et conclut sur une taille d'échantillon acceptable. Les mêmes éléments sont présentés pour l'EEC au point 5.1.2, en faisant varier d'une part le nombre d'UP, d'autre part le nombre de secteurs, ces deux éléments pouvant intervenir dans la précision de ce tirage à deux degrés. La section 5.2 donne les allocations finalement choisies, ainsi que les pondérations associées.

5.1.1 Impact de la taille de l'échantillon-maître sur la précision des enquêtes auprès des ménages

Dans cette partie, on s'intéresse à l'impact du nombre d'UP tirées sur la précision des échantillons sélectionnés au sein de l'échantillon-maître.

Pour déterminer le jeu d'allocations minimales recherché, on compare les variances obtenues sur plusieurs indicateurs, à partir d'échantillons sélectionnés dans l'EM de 2009 d'un côté et dans l'EM Nautile de l'autre. L'objectif est que le rapport des deux soit supérieur ou égal à 1 pour les variables d'intérêt Y et leur estimateur $\hat{Y}$:

$$\frac{V^{Octopusse}(\hat{Y})}{V^{Nautile}(\hat{Y})} \geq 1$$

La variance d'un tel tirage à deux degrés peut se décomposer en une partie liée au premier degré (tirage des unités primaires), et une partie liée au second degré (tirage des logements au sein des UP tirées au premier degré) du fait de l'indépendance du tirage des logements dans chaque unité primaire.

$$V^{Octopusse}(\hat{Y}) = V_1^O(\hat{Y}) + V_2^O(\hat{Y}) \text{ et } V^{Nautile}(\hat{Y}) = V_1^N(\hat{Y}) + V_2^N(\hat{Y})$$

Des résultats obtenus sur la variance de premier degré, bien que nécessaires, ne permettent pas de conclure directement sur la taille minimale de l'échantillon-maître. En effet, en fonction du nombre d'UP sélectionnées au premier degré, la base disponible pour le tirage d'une enquête au second degré est différente. Si ce nombre augmente, la base sera plus importante comme la précision des indicateurs calculés. Ainsi, l'allocation retenue joue également sur la variance du deuxième degré de tirage, et par conséquent sur la variance totale. Il est donc nécessaire d'estimer la variance totale des différentes variables d'intérêt étudiées (la diversité des enquêtes pour lesquelles l'EM est utilisé impose de ne pas se limiter à une seule variable).

Les simulations menées dans le cadre du projet n'intègrent que le premier degré de tirage (tirage des UP), et ne vont pas jusqu'au second degré (tirage des logements au sein des UP tirées), pour des raisons de temps de calcul ; ce deuxième degré est donc approché par des méthodes statistiques. Dans cette partie, on présente dans un premier temps les résultats obtenus au premier degré, puis la méthode permettant de revenir à la précision totale des enquêtes.

5.1.1.1 Précision obtenue au premier degré

Pour rappel, les UP sont sélectionnées selon un tirage spatialement équilibré au niveau des régions avec phase d'atterrissage au niveau national (voir point 2.3.1). Le jeu de variables d'équilibrage contient les variables d'intérêt principales, les premiers axes d'une ACP effectuée sur un grand nombre de variables corrélées aux différentes enquêtes ménages, et les variables indirectes utiles à l'équilibrage des UC pour le tirage coordonné de l'EEC (cf. section 4.3).

Le seul paramètre qui n'a pas encore été fixé est le nombre d'UP de l'échantillon-maître. La précision du nouvel EM dépend désormais uniquement du nombre d'UP tirées.

Les résultats suivants sont indépendants du nombre de secteurs Emploi [1] puisque le tirage d'UP s'effectue en amont du tirage de secteurs.

L'objectif est de mesurer l'impact d'une diminution du nombre d'UP tirées sur la précision de variables d'intérêt.

De nombreuses variables d'intérêt ont été sélectionnées dans l'étude des scénarios afin de couvrir toute la diversité des enquêtes qui seront tirées dans l'échantillon-maître. On se limite ici à trois variables pertinentes dans les enquêtes sociales : le revenu total disponible (issu de Fidéli), le nombre de cadres (issu du Recensement de la Population) et le nombre de 15-29 ans (issu du Recensement de la Population également), mais les résultats présentés sont généralisables à la plupart des autres variables.

Le tableau ci-dessous indique le ratio en base 100 entre le coefficient de variation de premier degré obtenu par simulations pour l'échantillon-maître Nautile et celui obtenu pour l'échantillon-maître de 2009 reproduit par simulations pour les trois variables d'intérêt étudiées, en fonction du nombre d'UP sélectionnées.

Les résultats présentés dans le tableau 5.1 montrent qu'une faible diminution du nombre d'UP tirées n'entraîne qu'une faible perte de précision au premier degré, quelle que soit la variable considérée. Par ailleurs, le gain lié à l'abandon des groupes de rotation du recensement de la population, à l'utilisation de Fidéli et des nouvelles variables d'équilibrage se confirme.

TABLE 5.1 – Rapport en base 100 entre les coefficients de variation des estimateurs calculés sur l'échantillon-maître au niveau national et la référence de l'échantillon-maître de 2009, en fonction du nombre d'unités primaires tirées

Nombre d'UP	Revenu total	Nombre de cadres	Nombre de 15-29 ans
580	17	23	15
535	19	26	17
494	20	29	19
463	22	30	20

Note de lecture : le coefficient de variation du nombre de cadres en France métropolitaine en tirant 535 unités primaires dans le plan de sondage Nautile est égal à 26% du coefficient de variation du nombre de cadres après un tirage avec le plan de sondage Octopusse avec 567 unités primaires.

5.1.1.2 Diminution du nombre d'unités primaires Nautile par rapport à Octopusse à qualité constante

Comparaison de la variance de premier et de second degrés entre Nautile et Octopusse

Comme dit précédemment, il n'était pas envisageable, pour des raisons de temps de calculs, de simuler des tirages à deux degrés pour comparer la variance totale entre

1. À noter que l'utilisation des variables indirectes introduites pour la coordination au point 4.2.3 conduit à dégrader légèrement la qualité de l'équilibrage de l'échantillon d'UP sur les autres variables, mais la perte est négligeable.

échantillons tirés à l'aide de la méthodologie de 2009 et échantillons tirés dans l'EM Nautile, selon des allocations d'UP différentes.

Par contre, il est possible d'exprimer approximativement le rapport des variances en fonction de l'allocation choisie. En notant n le nombre d'unités primaires, le rapport des deux variances que l'on souhaite estimé s'écrit :

$$R(n) = \frac{V^O(\hat{Y})}{V^N(\hat{Y})} = \frac{V_1^O(\hat{Y}) + V_2^O(\hat{Y})}{V_1^N(\hat{Y}) + V_2^N(\hat{Y})}$$

Or, pour les variances de second degré, si on suppose que les plans de sondage de tirage de logements respectent des propriétés d'invariance [2] et d'indépendance [3], on peut montrer, par la formule de Raj, 1966 [4] et au prix d'un certain nombre de simplifications, qu'elles sont égales, au facteur de ratio des tailles d'échantillon près, soit ici $\frac{567}{n}$ [5]. Ainsi :

$$V_2^O(\hat{Y}) = \frac{n}{567} V_2^N(\hat{Y})$$

Par ailleurs, les résultats obtenus précédemment permettent d'estimer le rapport des variances de premier degré pour plusieurs variables d'intérêt, entre l'EM de 2009 et un EM sélectionné avec la méthodologie présentée dans ce document de travail. En notant K ce rapport de variance, $V_1^O(\hat{Y}) = KV_1^N(\hat{Y})$, les simulations effectuées sur un grand nombre de variables montrent que le ratio K varie autour de 5 (cf. tableau 5.1 par exemple).

Pour obtenir le lien recherché entre rapport des variances et allocations, on souhaite obtenir une expression de la variance du premier degré sur la variance totale, pour l'échantillon-maître Nautile, notée X :

$$V_1^N(\hat{Y}) = XV^{Nautile}(\hat{Y})$$

Le plan de sondage utilisé au premier degré pour le tirage des unités primaires Nautile n'est pas adapté équitablement à toutes les variables d'intérêt, suivant que celles-ci sont plus ou moins corrélées aux variables d'équilibrage. Plus une variable sera corrélée aux variables d'équilibrage, plus la variance liée au premier degré sera réduite (et ainsi, la part de variance du premier degré sur la variance totale sera faible). Le coefficient X recherché dépend donc de la variable d'intérêt étudiée. Des études menées par Pascal Ardilly début 2017 (Ardilly, 2022) montrent que la part de variance au premier degré est de l'ordre de 1 % à 20 % selon les variables d'intérêt pour des tirages utilisant des variables d'équilibrage

2. L'invariance d'un plan de sondage à plusieurs degrés signifie que le plan de sondage au deuxième degré ne dépend pas du résultat du tirage de premier degré. Par exemple, cela signifie que le nombre de logements tirés dans une unité primaire ne dépend pas des autres unités primaires tirées. En pratique, cette hypothèse n'est pas totalement respectée dans les tirages de l'Insee. Ces derniers sont souvent autopondérés, ce qui signifie que le nombre de logements tirés dans une unité primaire dépend de l'estimation du nombre de logements en France métropolitaine à partir de l'ensemble des unités primaires tirées. Néanmoins, cette hypothèse n'est pas très éloignée de la réalité puisque cet estimateur varie peu d'un échantillon d'unités primaires à l'autre du fait de l'équilibrage sur les probabilités d'inclusion. Ces probabilités sont calculées proportionnellement au nombre de résidences principales des unités primaires en 2016.

3. L'indépendance au deuxième degré signifie que le tirage de logements est réalisé indépendamment entre les unités primaires. C'est effectivement le cas dans les tirages pour les enquêtes de l'Insee.

4. Le cadre d'application de cette formule est démontré dans Caron, Deville et Sautory, 1998.

5. 567 correspond au nombre d'UP pour l'échantillon-maître Octopusse de 2009. On peut noter ici une première approximation liée au fait que les unités primaires ne sont pas construites à l'identique pour les travaux sur l'échantillon-maître Octopusse et pour les travaux sur l'échantillon-maître Nautile.

analogues à celles mobilisées pour le tirage de l'échantillon-maître Nautile. Afin de couvrir toutes les variables d'intérêt des enquêtes, on considère ici plusieurs parts de variance au premier degré possibles [6] (1 %, 2 % et 10 %).

En reprenant ces différents éléments, on peut donc écrire que :

$$R(n) = KX + \frac{n}{567} * (1 - X)$$

Comme l'objectif est d'obtenir une précision au moins aussi bonne que précédemment ($R \geq 1$), on souhaite que :

$$n \geq 567 \, \frac{1 - KX}{1 - X}$$

Nombre d'unités primaires nécessaires pour conserver une aussi bonne qualité pour l'échantillon-maître Nautile que pour l'échantillon-maître Octopusse

On en déduit empiriquement le tableau suivant.

TABLE 5.2 – Nombre minimal d'UP à tirer pour avoir une meilleure variance dans Nautile que dans Octopusse

	$X = 1\%$	$X = 2\%$	$X = 5\%$	$X = 10\%$
$K = 2$	562	556	538	504
$K = 3$	556	544	507	441
$K = 4$	550	533	478	378
$K = 5$	545	521	448	315
$K = 6$	539	510	418	252
$K = 7$	533	498	388	189

Note de lecture : il faut au minimum tirer 538 unités primaires pour assurer une meilleure variance de tirage incluant les premier et deuxième degrés avec l'échantillon-maître Nautile qu'avec l'échantillon-maître Octopusse pour une variable dont la variance liée au tirage des UP Nautile vaut 5% de la variance totale du tirage d'unités primaires Nautile et de logements, et pour laquelle le tirage des unités primaires Nautile est 2 fois plus précis que celui de l'échantilon-maître Octopusse.

La première conclusion est une diminution dans tous les cas du nombre d'UP à sélectionner dans la solution entrevue. Si les allocations minimales sont très variées, notons qu'en réalité, les différentes études effectuées ont montré que plus une variable d'intérêt est bien expliquée par le plan de sondage de Nautile, plus le ratio K est élevé (gain important au premier degré par rapport à l'EM de 2009), et plus la variance totale dépendra du deuxième degré (part X faible). Si la variance de l'estimateur du total d'une variable n'est que peu réduite par le plan de sondage envisagé (par exemple, $K = 2$), alors, il est peu probable que la part de la variance de premier degré sur la variance totale soit faible (par exemple, X = 1 %). A l'inverse, une variable qui se trouverait bien mieux expliquée par le nouveau plan de sondage de l'EM ($K = 7$) devrait voir sa variance portée surtout par

6. Les situations où la part de variance au premier degré dépasse 10 % n'ont pas été étudiées, car elles sont plus favorables pour l'échantillon-maître Nautile et donc moins contraignantes que les cas présentés.

celle de deuxième degré ($X = 1$ % ou $X = 2$ %) Avec ces différents éléments, et afin de s'assurer une variance totale similaire dans le nouvel échantillon-maître et dans l'ancien échantillon-maître pour la totalité des variables, une diminution du nombre d'UP de 5 % environ par rapport aux 567 ZAE d'Octopusse semble envisageable.

5.1.2 Conséquence de la variation des allocations sur la précision de l'EEC

Pour rappel, le plan de sondage de l'EEC est un tirage à deux degrés, avec sélection d'un échantillon d'unités de coordination équilibré (voir point 4.1.2 et section 4.3) notamment sur le nombre de chômeurs au premier degré, et tirage équilibré de secteurs sur un jeu de variables liées à l'emploi au deuxième degré. Ainsi, si le nombre de secteurs tirés joue sur la précision des indicateurs de l'enquête, celle-ci dépend également du nombre d'UC, et donc indirectement de l'allocation de l'EM. L'objectif dans cette partie est de présenter l'impact sur la variance des indicateurs d'une modification du nombre d'UP (et donc d'UC) d'une part, et du nombre de secteurs d'autre part. Le but recherché est de respecter la précision demandée par Eurostat (nationale et au niveau Nuts2) pour l'enquête Emploi, sans dégrader la qualité des résultats issus de l'échantillon de 2009.

5.1.2.1 Cadre de référence et simulations

Pour mesurer l'impact des modifications des allocations sur la précision de l'échantillon Emploi, nous procédons à nouveau par simulations par la méthode Monte-Carlo. En particulier, nous comparons les résultats des scénarios de tirage testés pour l'EEC au scénario suivant :

- Méthode de tirage (tirage équilibré, variables d'équilibrage) utilisée en 2009 ;
- Allocation cible, définies par la division Emploi de l'Insee, de telle sorte à respecter les contraintes de précision régionales IESS en moyenne sur les trimestres tout en restant proche de la structure de l'allocation de l'échantillon utilisée en 2009.

Dans la suite du point 5.1.2, on appelle « scénario de référence » cette situation dans laquelle on aurait reproduit le plan de sondage de 2009 avec des allocations mises à jour pour respecter les contraintes de précision du règlement européen IESS. Les allocations régionales utilisées en 2009, et les allocations correspondant à cette mise à jour (allocation « cible ») sont présentées dans le tableau 5.3.

Pour faciliter la lecture des résultats de simulations, on définit pour chaque variable d'intérêt, une base 100 correspondant aux CV obtenus à partir des simulations du scénario de référence.

L'impact d'une diminution de la taille de chaque échantillon (EM et EEC) sur la précision des indicateurs a été étudié en deux étapes, en évaluant tout d'abord les conséquences d'une diminution du nombre d'UP tirées, puis à allocations régionales d'UP fixées, en concluant sur une réduction du nombre de secteurs dans l'échantillon. Comme dans le point 4.3.2, les variables étudiées sont :

TABLE 5.3 – Allocations régionales initiales et servant de base aux simulations de tirage.

Région	EEC 2009	Allocation cible
Île-de-France	572	565
Champagne-Ardenne	99	100
Picardie	90	89
Haute-Normandie	90	90
Centre	129	127
Basse-Normandie	86	85
Bourgogne	91	90
Nord-Pas-de-Calais	191	189
Lorraine	116	115
Alsace	92	91
Franche-Comté	79	80
Pays de la Loire	175	173
Bretagne	162	160
Poitou-Charentes	91	91
Aquitaine	161	159
Midi-Pyrénées	143	141
Limousin	77	82
Rhône-Alpes	303	299
Auvergne	73	76
Languedoc-Roussillon	130	128
PACA	251	248
Corse	16	38
Total	*3 217*	*3 216*

— le nombre d'individus percevant des allocations chômage ou préretraite appelée aussi nombre de chômeurs ;
— le nombre d'individus percevant des revenus d'activité appelée aussi nombre d'actifs ;
— le total des salaires.

Les deux premières variables sont des proxy des principales variables d'intérêt de l'enquête mais ce sont aussi des variables d'équilibrage. L'analyse ne peut donc se restreindre à ces variables. D'où l'analyse des résultats sur le total des salaires. Ce dernier sert de référence pour la précision d'autres variables qui seraient mesurées dans l'enquête.

Les résultats de ces scénarios ont été analysés aux niveaux national et régional.

5.1.2.2 Impact d'une diminution du nombre d'UP sur la précision de l'EEC

À nombre égal de secteurs tirés, la taille de l'échantillon-maître et sa répartition régionale modifie la précision des variables calculées à partir de l'échantillon de l'EEC. En effet, plus le nombre d'UP au premier degré est important, plus le nombre d'UC et la base de sondage disponible au second degré sont grands. Ainsi, cela permet de diminuer le nombre de secteurs tirés par UC, ce qui réduit l'effet de grappe et améliore la précision

des indicateurs. Cela étant, cet effet de grappe est plus ou moins marqué selon les régions. Aussi, afin de définir un nombre limite d'UP par région, permettant de respecter les seuils de précision demandés par Eurostat, sans dégrader la qualité actuelle de l'EEC, trois jeux d'allocations d'UP sont testés, pour une taille d'échantillon de 3216 secteurs au second degré répartis selon l'allocation cible du tableau 5.3 :

— les allocations proportionnelles : 567 unités primaires sont sélectionnées, réparties proportionnellement au nombre de logements dans chaque région ;
— les allocations optimisées : 567 unités primaires sont sélectionnées, en diminuant l'allocation dans les régions où la précision est suffisante dans le cas précédent, au profit d'autres régions dont la précision n'est pas satisfaisante ;
— les allocations optimisées réduites : 541 unités primaires sont sélectionnées, en poursuivant la diminution des allocations dans les régions dont la précision est jugée satisfaisante.

La répartition de ces allocations par région est présentée dans le tableau 5.4

TABLE 5.4 – Jeux d'allocations d'UP testés pour satisfaire les contraintes de précision sur l'EEC

Région	Allocations en nombre d'UP		
	Proportionnelle	Optimisée	Optimisée réduite
11 (Île-de-France)	*101*	*97*	*85*
21 (Champagne-Ardennes)	*11*	*14*	*14*
22 (Picardie)	*16*	*19*	*19*
23 (Haute-Normandie)	*16*	*16*	*16*
24 (Centre)	*23*	*25*	*25*
25 (Basse-Normandie)	*14*	*16*	*16*
26 (Bourgogne)	*15*	*18*	*18*
31 (Nord-Pas-de-Calais)	*34*	*33*	*30*
41 (Lorraine)	*21*	*21*	*19*
42 (Alsace)	*15*	*14*	*14*
43 (Franche-Comté)	*11*	*13*	*13*
52 (Pays de la Loire)	*31*	*27*	*27*
53 (Bretagne)	*29*	*29*	*29*
54 (Poitou-Charentes)	*17*	*20*	*20*
72 (Aquitaine)	*30*	*27*	*27*
73 (Midi-Pyrénées)	*24*	*23*	*23*
74 (Limousin)	*7*	*10*	*10*
82 (Rhône-Alpes)	*60*	*56*	*51*
83 (Auvergne)	*13*	*15*	*15*
91 (Languedoc-Roussillon)	*24*	*21*	*21*
93 (PACA)	*52*	*48*	*44*
94 (Corse)	*3*	*5*	*5*
Total	***567***	***567***	***541***

L'étude du tableau 5.5 montre que les précisions des indicateurs au niveau national sont dans tous les cas meilleures que celles du scénario de référence. Ici, les résultats sont présentés pour la variable de nombre d'actifs occupés.

Au niveau régional, l'utilisation d'une allocation proportionnelle donne des résultats très contrastés en fonction de la taille de la région. Ainsi, les allocations d'Île-de-

France, d'Alsace ou de Rhône-Alpes permettent d'améliorer sensiblement la précision des indicateurs par rapport au scénario de référence, contrairement à celles du Limousin ou de la Corse. La réallocation opérée dans le scénario d'allocations optimisées permet d'améliorer de façon très importante ces résultats, et de se rapprocher des coefficients de variation du scénario de référence (et donc de la précision requise par le règlement IESS) dans de nombreuses régions (Champagne-Ardennes, Picardie, Centre...), sans trop détériorer la situation dans les régions dans lesquelles on diminue le nombre d'UP sélectionnées. Ce gain vient du fait que l'effet de grappe est moins important dans les régions dans lesquelles le nombre d'UP tirées augmente, les secteurs Emploi étant tirés dans davantage d'UC ; ils sont donc mieux répartis et moins homogènes entre eux.

TABLE 5.5 – Précision régionale (CV en base 100 par rapport au scénario de référence permettant de respecter la précision fixée par le règlement IESS) du nombre d'actifs occupés estimés à partir des échantillons de secteurs pour les différents scénarios de jeux d'UP

Région	Allocations en nombre d'UP		
	Proportionnelle	Optimisée	Optimisée réduite
11 (Île-de-France)	54	54	56
21 (Champagne-Ardennes)	131	107	107
22 (Picardie)	117	105	104
23 (Haute-Normandie)	100	100	99
24 (Centre)	107	99	100
25 (Basse-Normandie)	113	104	104
26 (Bourgogne)	119	105	104
31 (Nord-Pas-de-Calais)	78	79	83
41 (Lorraine)	82	81	89
42 (Alsace)	44	47	47
43 (Franche-Comté)	132	118	118
52 (Pays de la Loire)	89	97	97
53 (Bretagne)	98	100	99
54 (Poitou-Charentes)	124	111	111
72 (Aquitaine)	88	95	95
73 (Midi-Pyrénées)	89	92	93
74 (Limousin)	180	138	137
82 (Rhône-Alpes)	73	78	81
83 (Auvergne)	123	113	110
91 (Languedoc-Roussillon)	88	98	99
93 (PACA)	66	72	79
94 (Corse)	151	117	117
France entière	**55**	**55**	**57**

Note de lecture : lorsque 567 UP sont réparties entre régions par l'allocation proportionnelle, le coefficient de variation de la variable de nombre d'actifs occupés en Franche-Comté de l'échantillon de 3216 secteurs vaut 132% de celui du scénario de référence qui reproduit les contraintes de précision européennes.

Notons toutefois que, dans certaines régions, le niveau de précision du scénario de référence n'est toujours pas atteint, même avec l'allocation optimisée. On pourrait continuer à augmenter le nombre d'UP tirées, mais le coût associé à la collecte pourrait être trop important [7].

7. Dans un cas extrême, le risque peut être de devoir sélectionner toutes les UP d'une région pour avoir une qualité aussi bonne que le tirage de référence, ce qui n'est pas envisageable.

Le troisième jeu d'allocations, dites optimisées réduites, permet de diminuer la taille de l'EM (et ainsi la surface à parcourir pour les enquêtes et la taille minimale nécessaire pour le réseau enquêteurs). Il donne également des résultats satisfaisants dans les régions impactées par une diminution de leur allocation, ce qui permet d'assouplir la contrainte sur la taille critique de l'EM.

C'est donc ce troisième jeu qui est utilisé pour la suite des simulations.

5.1.2.3 Impact de la modulation du nombre de secteurs par région sur la précision de l'EEC

Une fois les UC sélectionnées à travers le tirage des UP, le deuxième degré consiste à tirer les secteurs. Une baisse de l'allocation aura pour conséquence une diminution de la qualité de l'échantillon. Cela étant, l'impact n'est pas le même en fonction de la région concernée. Ainsi, comme pour les UP, deux scénarios sont testés :

— les 3216 secteurs (correspondant à la taille de l'échantillon de 2009) sont répartis d'après l'allocation cible qui permet d'atteindre la précision requise par IESS à plan de sondage équivalent à celui de 2009 (cf. tableau 5.3), allocation dite « initiale ».
— on applique une diminution de 10 % dans chaque région de l'allocation obtenue précédemment, sauf en Île-de-France où l'allocation initiale est conservée [8], allocation dite « réduite ».

Ces allocations sont présentées dans le tableau 5.6.

8. Ce choix résulte de la prise en compte de l'impact de la non-réponse sur les résultats de l'enquête en Île-de-France, dont l'évaluation est complexe.

TABLE 5.6 – Jeux d'allocations de secteurs testés pour satisfaire les contraintes de précision sur l'EEC

Région	Nombre de secteurs	
	Allocation initiale	Allocation réduite
11 (Île-de-France)	565	565
21 (Champagne-Ardennes)	100	90
22 (Picardie)	89	80
23 (Haute-Normandie)	90	81
24 (Centre)	127	114
25 (Basse-Normandie)	85	76
26 (Bourgogne)	90	81
31 (Nord-Pas-de-Calais)	189	170
41 (Lorraine)	115	103
42 (Alsace)	91	81
43 (Franche-Comté)	80	72
52 (Pays de la Loire)	173	155
53 (Bretagne)	160	144
54 (Poitou-Charentes)	91	81
72 (Aquitaine)	159	143
73 (Midi-Pyrénées)	141	126
74 (Limousin)	82	73
82 (Rhône-Alpes)	299	269
83 (Auvergne)	76	68
91 (Languedoc-Roussillon)	128	115
93 (PACA)	248	223
94 (Corse)	38	34
Total	***3216***	***2944***

On constate en premier lieu, dans le tableau 5.7, que malgré la diminution de près de 300 secteurs dans l'échantillon, la précision au niveau national reste bien meilleure que pour le scénario de référence. Au niveau régional, on constate également que l'impact d'une telle diminution est assez faible comparée par exemple à une diminution de quelques unités primaires (cf. tableau 5.5). Cela s'explique à nouveau par un fort effet de grappe lié au plan de sondage coordonné : tirer un secteur supplémentaire dans une UC, qui ne couvre qu'une part limitée du territoire n'apporte que peu en termes de précision de l'enquête. Cette diminution de 10 % dans l'ensemble des régions hors Île-de-France s'avère ainsi envisageable.

TABLE 5.7 – Précision régionale (CV en base 100 par rapport au scénario de référence permettant de respecter la précision fixée par le règlement IESS) du nombre d'actifs occupés estimés à partir des échantillons de secteurs pour les différents scénarios de jeux d'UP

Région	Allocations en nombre de secteurs	
	Initiale	Réduite
11 (Île-de-France)	56	56
21 (Champagne-Ardennes)	107	111
22 (Picardie)	104	109
23 (Haute-Normandie)	99	102
24 (Centre)	100	103
25 (Basse-Normandie)	104	105
26 (Bourgogne)	104	107
31 (Nord-Pas-de-Calais)	83	85
41 (Lorraine)	89	90
42 (Alsace)	47	48
43 (Franche-Comté)	118	119
52 (Pays de la Loire)	97	99
53 (Bretagne)	99	102
54 (Poitou-Charentes)	111	114
72 (Aquitaine)	95	96
73 (Midi-Pyrénées)	93	95
74 (Limousin)	137	140
82 (Rhône-Alpes)	81	85
83 (Auvergne)	110	114
91 (Languedoc-Roussillon)	99	100
93 (PACA)	79	80
94 (Corse)	117	120
France entière	**57**	**60**

Note de lecture : lorsque le tirage des secteurs consiste en un tirage de 541 UP par l'allocation optimisée, puis en la récupération des unités de coordination associées et le tirage de 3216 secteurs, le coefficient de variation de la variable de nombre d'actifs occupés en Franche-Comté vaut 118% de celui du scénario de référence qui reproduit les contraintes de précision européennes.

5.2 Tirage des échantillons

Après avoir étudié l'impact des différents jeux d'allocation sur les enquêtes auprès des ménages tirées dans l'EM d'une part et sur l'EEC d'autre part, il reste à concilier les deux afin de déterminer les allocations finales et pouvoir procéder au tirage effectif des échantillons.

5.2.1 Allocations retenues pour le tirage de l'échantillon-maître et de l'échantillon Emploi

Des études présentées précédemment à propos de l'impact du choix des allocations de l'EM et de l'EEC sur la précision des indicateurs issus des enquêtes auprès des ménages (enquête Emploi y compris), trois conclusions ressortent :

— une diminution jusqu'à 5 % du nombre d'UP tirées est acceptable pour les enquêtes ménages tirées dans l'EM ;

— une réallocation des UP par région, associée à une diminution globale de 5 % de la taille de l'échantillon-maître permet d'obtenir des résultats satisfaisant quant à la précision de l'EEC. Cette redistribution des allocations permet également d'améliorer la qualité des indicateurs issus des enquêtes auprès des ménages tirées dans l'EM dans les régions où la précision est moins bonne ;
— une diminution de 10 % du nombre de secteurs, hors Île-de-France, au second degré, permet d'obtenir des résultats *a priori* satisfaisants pour l'EEC.

Si une forte diminution de la taille des échantillons permet de limiter le coût de collecte, l'Insee doit avant tout garantir la qualité des estimations produites. Le jeu d'allocations optimisées réduites permet d'assurer cette qualité dans l'ensemble des régions métropolitaines. L'allocation correspondant à une diminution de 10 %, hors Île-de-France, du nombre de secteurs permet également d'obtenir un échantillon de qualité pour l'EEC. Le tableau 5.8 présente les allocations finalement retenues.

TABLE 5.8 – Allocations régionales de tirage de l'échantillon-maître

Région	Nombre d'UP	Nombre de secteurs
Île-de-France	85	565
Champagne-Ardenne	14	90
Picardie	19	80
Haute-Normandie	16	81
Centre	25	114
Basse-Normandie	16	76
Bourgogne	18	81
Nord-Pas-de-Calais	30	170
Lorraine	19	103
Alsace	14	81
Franche-Comté	13	72
Pays de la Loire	27	155
Bretagne	29	144
Poitou-Charentes	20	81
Aquitaine	27	143
Midi-Pyrénées	23	126
Limousin	10	73
Rhône-Alpes	51	269
Auvergne	15	68
Languedoc-Roussillon	21	115
PACA	44	223
Corse	5	34
Total	*541*	*2 944*

Les allocations régionales de secteurs ne peuvent toutefois pas être utilisées directement, car le plan de sondage est un tirage équilibré, stratifié au niveau des UC. Il reste donc, à partir de ce nombre de secteurs par région, à déterminer l'allocation dans chaque unité de coordination.

5.2.2 Détermination des allocations de secteur au niveau des UC

La ventilation de l'allocation n_{sect}^{reg} de secteurs de la région reg par unité de coordination est présentée ici. Cette ventilation dépend du nombre d'UC sélectionnées après tirage de l'EM. En effet, suite à ce tirage, si les UP sélectionnées de la région appartiennent à K UC, alors l'allocation régionale sera répartie parmi ces K UC. Le fait que les UC soient tirées indirectement induit que ce nombre d'UC est variable en fonction de l'échantillon d'UP tirées.

On cherche donc à définir ici l'allocation n_{sect}^{uc} de secteurs pour chaque UC uc issue du tirage des UP de l'EM dans la région reg. On souhaite ventiler cette allocation de telle sorte que chaque secteur de chaque UC ait la même probabilité d'être sélectionné dans une même région [9]. Ainsi, on souhaite que pour chaque secteur l, sa probabilité de sélection π_l soit : $\pi_l = \frac{n_{sect}^{reg}}{N_{sect}^{reg}}$ où N_{sect}^{reg} est le nombre total de secteurs dans la région reg.

En décomposant par chaque degré de tirage, la probabilité de sélection du secteur l de l'UC uc est :

$$\pi_l = \pi_{uc}\pi_{l/uc} = \frac{n_{sect}^{reg}}{N_{sect}^{reg}}$$

Du fait du tirage indirect des UC, leur probabilité de sélection π_{uc} n'est, en générale, pas connue. Par conséquent, la probabilité d'inclusion du secteur ne l'est pas non plus. On cherche donc à disposer de secteurs équipondérés, ce qui nous conduit à raisonner sur le poids w_l du secteur. La contrainte ci-dessus devient donc, en notant w_{uc} le poids de l'UC uc calculé au point 4.2.2 :

$$w_l = w_{uc}\frac{1}{\pi_{l/uc}} = C^{reg} \text{ avec } w_{uc} = \frac{|up \in EM \cap uc|}{\sum_{up \in uc}\pi_{up}}$$

Or, $\pi_{l/uc} = \frac{n_{sect}^{uc}}{N_{sect}^{uc}}$ où N_{sect}^{uc} est le nombre de secteurs dans l'UC uc. Donc, l'objectif est de déterminer l'allocation n_{sect}^{uc} de secteurs à tirer dans l'UC uc afin que le poids des secteurs w_l soit constant régionalement :

$$w_l = w_{uc}\frac{N_{sect}^{uc}}{n_{sect}^{uc}} = C^{reg}$$

Pour satisfaire cette contrainte, on va raisonner selon une logique similaire au sondage auto-pondéré présenté au point 2.3.2. Pour cela, il convient de distinguer les UP exhaustives [10] et les UP non exhaustives. Les UP exhaustives étant toujours sélectionnées, les UC auxquelles elles sont associées le seront également. On parle alors d'UC exhaustives. C'est le cas des très grandes communes. En pratique, les UC exhaustives ne sont constituées que d'une seule UP, elle-même exhaustive. Donc le poids de ces UC est égal à 1. On définit alors par $N_{sect}^{uc,exh}$ le nombre de secteurs appartenant à l'UC exhaustive uc de la région reg, et par $n_{sect}^{uc,exh}$ le nombre de secteurs à sélectionner au sein de cette UC exhaustive, de telle sorte que :

9. Du fait des surreprésentations régionales de l'EM et de l'échantillon Emploi, il est impossible que les probabilités d'être sélectionnés soient identiques pour 2 secteurs de 2 régions différentes.

10. Pour rappel, cela signifie que la probabilité de sélection initiale de l'UP est supérieure à 1 (cf. point 2.3.2).

$$n_{sect}^{uc,exh} = n_{sect}^{reg} \frac{N_{sect}^{uc,exh}}{N_{sect}^{reg}}$$

Le nombre de secteurs à tirer dans les UC exhaustives est alors fixé proportionnellement à la taille de l'UC (en nombre de secteurs).

Une fois toutes les UC exhaustives de la région ainsi traitées, il reste à allouer $n_{sect}^{reg,nonexh} = n_{sect}^{reg}$ - $n_{sect}^{reg,exh}$ secteurs au sein des UC non exhaustives de la région, sélectionnées avec le tirage de l'EM, où $n_{sect}^{reg,exh} = \sum_{i \in reg} n_{sect}^{i,exh}$. On cherche donc, pour chaque UC non exhaustive uc, le nombre de secteur $n_{sect}^{uc,nonexh}$ à tirer. D'après la logique du sondage auto-pondéré, on répartit ces secteurs de manière homogène dans les UC non exhaustives. Autrement dit, on divise l'allocation de secteurs à tirer dans les UC non exhaustives par le nombre d'UC non exhaustives de l'échantillon s_{uc}^{reg} :

$$n_{sect}^{uc,nonexh} = \frac{n_{sect}^{reg} - n_{sect}^{reg,exh}}{|UC_{nonexh \cap s_{uc}^{reg}}|}$$

Il reste à vérifier que les poids $w_l = w_{uc} \frac{N_{sect}^{uc}}{n_{sect}^{uc}}$ ainsi calculés sont identiques quel que soit le secteur l tiré dans la région. Ce poids se calcule aisément pour les UC exhaustives (composées uniquement d'une UP elle-même exhaustive, donc pour une UC uc exhaustive $w_{uc} = 1$). Pour les secteurs l d'une UC uc exhaustive :

$$w_l = w_{uc} \frac{N_{sect}^{uc}}{n_{sect}^{uc}} = \frac{N_{sect}^{uc} N_{sect}^{reg}}{n_{sect}^{reg} N_{sect}^{uc}} = \frac{N_{sect}^{reg}}{n_{sect}^{reg}}$$

Pour un secteur l tiré dans une UC non exhaustive, le poids w_l du secteur est :

$$w_l = w_{uc} \frac{N_{sect}^{uc}}{n_{sect}^{uc}} = \frac{|up \in EM \cap uc|}{\sum_{up \in uc} \pi_{up}} \frac{N_{sect}^{uc} |UC_{nonexh \cap s_{uc}^{reg}}|}{n_{sect}^{reg} - n_{sect}^{reg,exh}}$$

De prime abord, cette formule semble peu se simplifier. Néanmoins, les UC non exhaustives sont composées d'UP non exhaustives et ces dernières ont une probabilité d'inclusion proportionnelle à la taille de l'UP en nombre de résidences principales par rapport à la taille des UP non exhaustives de la région (cf. point 2.3.2). Or, les travaux sur la constitution des secteurs au chapitre 2 de la partie B ont conduit à des tailles de secteurs à peu près homogènes (120 résidences principales par secteur environ). Donc pour une UP non exhaustive up, $\pi_{up} \approx m^{reg,nonexh} \frac{N_{sect}^{up}}{N_{sect}^{reg,nonexh}}$ où $m^{reg,nonexh}$ est l'allocations d'unités primaires non exhaustives dans la région reg. Cela permet de déduire que $\sum_{up \in uc} \pi_{up} \approx m^{reg,nonexh} \frac{N_{sect}^{uc}}{N_{sect}^{reg,nonexh}}$. La formule précédente se simplifie donc en :

$$w_l \approx \frac{|up \in EM \cap uc| N_{sect}^{reg,nonexh}}{m^{reg,nonexh} N_{sect}^{uc}} \frac{N_{sect}^{uc} |UC_{nonexh \cap s_{uc}^{reg}}|}{n_{sect}^{reg} - n_{sect}^{reg,exh}}$$

$$w_l \approx \frac{|up \in EM \cap uc| N_{sect}^{reg,nonexh}}{m^{reg,nonexh}} \frac{|UC_{nonexh \cap s_{uc}^{reg}}|}{n_{sect}^{reg} - n_{sect}^{reg,exh}}$$

Or, $n_{sect}^{reg,exh} = n_{sect}^{reg} \frac{N_{sect}^{reg,exh}}{N_{sect}^{reg}}$ donc $n_{sect}^{reg} - n_{sect}^{reg,exh} = n_{sect}^{reg} \frac{N_{sect}^{reg,nonexh}}{N_{sect}^{reg}}$. Donc, la formule précédente se simplifie en :

$$w_l \approx \frac{|up \in EM \cap uc| N_{sect}^{reg,nonexh}}{m^{reg,nonexh}} \frac{N_{sect}^{reg} |UC_{nonexh \cap s_{uc}^{reg}}|}{n_{sect}^{reg} N_{sect}^{reg,nonexh}}$$
$$w_l \approx \frac{|up \in EM \cap uc|}{m^{reg,nonexh}} \frac{|UC_{nonexh \cap s_{uc}^{reg}}| N_{sect}^{reg}}{n_{sect}^{reg}}$$

Enfin, comme les unités primaires sont tirées par un algorithme spatialement équilibré et que les unités de coordination regroupent des unités primaires voisines, les risques de sélections conjointes d'UP proches géographiquement appartenant à la même UC sont limités. Empiriquement, on observe qu'une vingtaine d'UC sont tirées plusieurs fois lors d'un même tirage, d'après les simulations pour le tirage retenu de 541 UP. Donc $|UC_{nonexh \cap s_{uc}^{reg}}| \approx m^{reg,nonexh}$, ce qui permet de simplifier le poids des secteurs dans les UC non exhaustives :

$$w_l \approx |up \in EM \cap uc| \frac{N_{sect}^{reg}}{n_{sect}^{reg}}$$

Par conséquent, sauf quand une UC est tirée plusieurs fois, ce qui concerne environ 4% des UC tirées, les secteurs ont bien le même poids w_l ou presque, qu'ils soient tirés dans une UC exhaustive ou non exhaustive. Cela valide donc les allocations suivantes :

$$n_{sect}^{uc,exh} = n_{sect}^{reg} \frac{N_{sect}^{uc,exh}}{N_{sect}^{reg}}$$
$$n_{sect}^{uc,nonexh} = \frac{n_{sect}^{reg} - n_{sect}^{reg,exh}}{|UC_{nonexh \cap s_{uc}^{reg}}|}$$

Les différentes allocations obtenues de manière analytique n'ont pas de raison d'être entières. On arrondit alors chaque allocation des UC d'une région de manière à respecter le nombre de secteurs à tirer dans la région, tout en homogénéisant au maximum les différences de poids w_l des secteurs d'une unité de coordination à l'autre [11].

5.2.3 Pondérations des secteurs

La section précédente présente le calcul des allocations de secteurs au sein des UC. Ces allocations sont ensuite arrondies afin de tirer un nombre entier de secteurs par UC. Les pondérations des secteurs se déduisent simplement de ces allocations. Pour un secteur l tiré dans une UC uc, son poids w_l est

$$w_l = w_{uc} \frac{N_{sect}^{uc}}{n_{sect}^{uc}} = \frac{|up \in EM \cap uc|}{\sum_{up \in uc} \pi_{up}} \frac{N_{sect}^{uc}}{n_{sect}^{uc}}$$

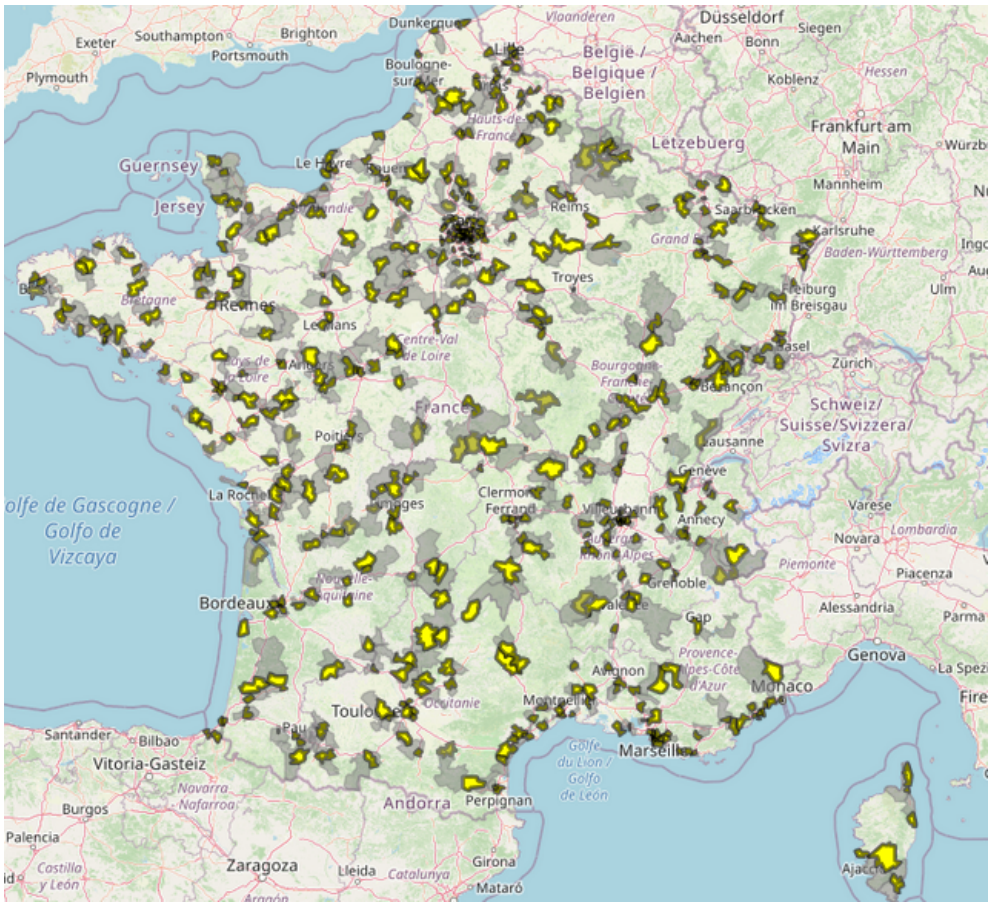
11. En pratique, cela implique notamment d'arrondir à l'entier supérieur les allocations dans les UC tirées plusieurs fois.

en notant N_{sect}^{uc} le nombre total de secteurs dans l'UC uc, n_{sect}^{uc} le nombre de secteurs à tirer dans l'UC uc, $\sum_{up \in uc} \pi_{up}$ la somme des probabilités d'inclusion des UP qui composent l'unité de coordination uc et $|up \in EM \cap uc|$ le nombre d'UP de l'UC uc qui sont tirées dans l'échantillon-maître.

5.2.4 Tirage des échantillons

Les allocations ainsi définies ont permis d'appliquer le plan de sondage coordonné de l'échantillon-maître et de l'échantillon de l'EEC. La figure 5.1 présente un exemple de tirage d'unités primaires (zones jaunes), tiré avec le même plan de sondage que l'échantillon-maître finalement retenu et les unités de coordination associées pour le tirage de l'enquête Emploi (zones jaunes et grises cumulées).

FIGURE 5.1 – Exemple de tirage d'unités primaires et d'unités de coordination avec le plan de sondage utilisé pour tirer l'échantillon-maître



Le tableau 5.9 détaille, pour quelques variables d'intérêt, les écarts relatifs entre les estimateurs sur l'échantillon de l'EEC tiré *in fine* obtenus par les pondérations du point 5.2.3, et les totaux calculés sur toute la population à partir de la base de sondage. Ces résultats sont très proches de ceux obtenus lors des simulations.

TABLE 5.9 – Écart relatif en valeur absolue entre les estimateurs dans le nouvel échantillon EEC et les vrais totaux des secteurs.

Variables	ER (%)
Nb de logements	0,064
Salaires	0,117
Nb de chômeurs	0,094
Allocation chômage	0,419
Nb d'actifs	0,095

Ces deux échantillons permettent à l'Insee de réaliser les différentes enquêtes ménages dont l'institut à la charge. La partie D décrit les réflexions suivantes menées pour le tirage des enquêtes au sein de l'échantillon-maître d'une part, et pour la ventilation de l'échantillon de l'EEC dans les différentes vagues de collecte d'autre part.

Partie D

Sélection finale des logements à enquêter : Finalisation des grappes, second degré des enquêtes ménages et marquage des échantillons

Les parties précédentes présentent la construction des échantillons de premier degré pour les enquêtes auprès des ménages (échantillon-maître et échantillon de l'enquête Emploi), de la construction des bases de sondage (partie B) à la sélection des unités géographiques - les UP et les secteurs (partie C). La coordination de ces deux échantillons, étudiée afin d'optimiser le travail de collecte, a posé des contraintes qu'il a fallu prendre en compte pour obtenir les résultats présentés dans la partie C.

Cela étant, ces échantillons ne sont que le premier degré d'un plan de sondage pour chaque enquête. En effet, pour l'enquête Emploi, les grappes de l'échantillon doivent encore être réparties afin de déterminer leur date d'entrée dans le processus de collecte. Cette répartition est présentée au chapitre 1.

Par ailleurs, une mise à jour de ces grappes est effectuée en deux temps, avec l'ajout initial des résidences secondaires et des logements vacants pour prendre en compte l'intégralité du champ des logements, puis l'intégration annuelle des modifications se rapportant aux logements de la grappe et aux logements alentours (construction de logement, modification du statut d'un logement de résidence principale à logement vacant...). Ces modifications pouvant générer un accroissement conséquent de la taille des grappes, une sélection supplémentaire des logements effectivement collectés est pratiquée. Cette finalisation de la composition de l'échantillon de l'EEC fait l'objet du chapitre 2.

Concernant le second degré des autres enquêtes auprès des ménages, l'étude comparative décrite au chapitre 3, entre différents plans de sondage, a amené à opter pour un tirage systématique.

Enfin, le dernier chapitre propose différentes options pour le rééquilibrage de la base de sondage, en fonction des évènements survenus sur celle-ci (tirage, rafraîchissement annuel de la base...) pour ne pas déformer la structure de cette dernière au fil du temps.

Chapitre 1

Détermination du calendrier de collecte pour les grappes et les secteurs

Le tirage de l'échantillon de l'EEC a permis de déterminer les secteurs enquêtés au cours des 9 ans à venir (voir chapitre 5 de la partie C). Si les secteurs définissent les unités de tirage de l'EEC, les logements entrent en collecte par grappes. Pour rappel, un secteur contient 6 grappes composées chacune d'une vingtaine de résidences principales [1]. Durant six trimestres consécutifs, une des grappes est enquêtée, puis est remplacée par la grappe suivante dans le secteur.

L'intérêt de cette méthode est de constituer un pseudo-panel, construit sur l'idée que les occupants de deux logements voisins ont des caractéristiques proches vis-à-vis de l'emploi, et peuvent donc se remplacer au bout d'une durée acceptable pour l'enquêté (établie ici à 6 trimestres), ce qui limite l'attrition tout en garantissant la qualité des estimations longitudinales. Au bout de 36 trimestres, soit 9 ans, toutes les grappes mises en collecte du secteur ont été interrogées.

Pour des raisons de collecte et de représentativité de l'échantillon utilisé pour des diffusions trimestrielles, il n'est pas envisageable de renouveler l'ensemble des grappes collectées un même trimestre, ni de poser des questions relatives à un même moment du trimestre pour l'ensemble des logements enquêtés. Il est ainsi nécessaire de répartir les secteurs enquêtés en définissant pour chacun le trimestre au cours duquel la première grappe du secteur entrera en collecte [2] et la semaine à laquelle font référence les questions posées aux enquêtés du secteur.

Les concepts de trimestre d'entrée et de semaine de référence sont présentés dans la section 1.1 avant de détailler en section 1.2 comment les secteurs sont chacun affectés à un trimestre d'entrée et une semaine de référence donnés. Enfin, si ces paramètres figent la

1. Un tirage est réalisé par sondage aléatoire simple pour sélectionner les 6 grappes qui seront collectées pour les secteurs contenant initialement 7 grappes.

2. Les logements des grappes mises en collecte étant enquêtés 6 trimestres consécutifs, cela fige par construction le trimestre d'entrée en collecte des 5 grappes suivantes dans le dispositif d'enquête.

date d'entrée en collecte de chaque secteur, il est nécessaire de déterminer l'ordre d'entrée en collecte de chaque grappe au sein du secteur pour fixer la date d'entrée en collecte de chaque logement, ce qui constitue la section 1.3

1.1 Pourquoi un trimestre d'entrée et une semaine de référence ?

1.1.1 Trimestre d'entrée

Il n'est pas souhaitable, ni pour la charge de collecte, ni pour la qualité des indicateurs, que les premières grappes entrant en collecte au sein de chaque secteur du nouvel échantillon soient collectées au cours du même trimestre.

En effet, en ce qui concerne la collecte, la première interrogation demande toujours une charge d'enquête plus importante que les interrogations suivantes. L'enquêteur doit repérer le logement, puis la collecte s'effectue en face-à-face, contrairement au 4 suivantes [3]. Si toutes les grappes entrent au cours du même trimestre en première interrogation, la charge de collecte risque d'être difficilement soutenable au premier et sixième trimestre, alors qu'elle sera bien plus faible au cours des quatre trimestres intermédiaires.

Par ailleurs, un enquêté aura un comportement de réponse potentiellement différent entre chaque interrogation. Faire entrer toutes les grappes au cours du même trimestre risque de générer un biais de *rang d'interrogation* difficile à appréhender. De plus, un renouvellement de l'ensemble des grappes au même trimestre présente un risque de rupture de série puisque, même si les habitants de grappes d'un même secteur sont supposés avoir des comportements similaires vis-à-vis de l'emploi, ils n'en sont pas moins des individus différents et n'ont donc pas une situation professionnelle identique.

Pour ces raisons, il convient de disperser les premières interrogations le plus possible sur des trimestres différents. Un secteur voyant une nouvelle grappe intégrer la collecte tous les 6 trimestres, on choisit naturellement de répartir les secteurs en 6 groupes, chaque groupe étant collecté, en première interrogation, un trimestre différent. Cette affectation fera l'objet de la section 1.2. Ainsi, chaque trimestre, $\frac{1}{6}$ des grappes entrent dans l'échantillon (les grappes du premier groupe de secteurs, en remplaçant d'autres grappes de ces mêmes secteurs, en sixième interrogation lors du trimestre précédent), $\frac{1}{6}$ sont en deuxième interrogation (les grappes du deuxième groupe de secteurs), $\frac{1}{6}$ en troisième interrogation (les grappes du troisième groupe de secteurs)... Le trimestre au cours duquel une grappe d'un secteur est en première interrogation est appelé **trimestre d'entrée** du secteur (ou de la grappe) dans l'échantillon.

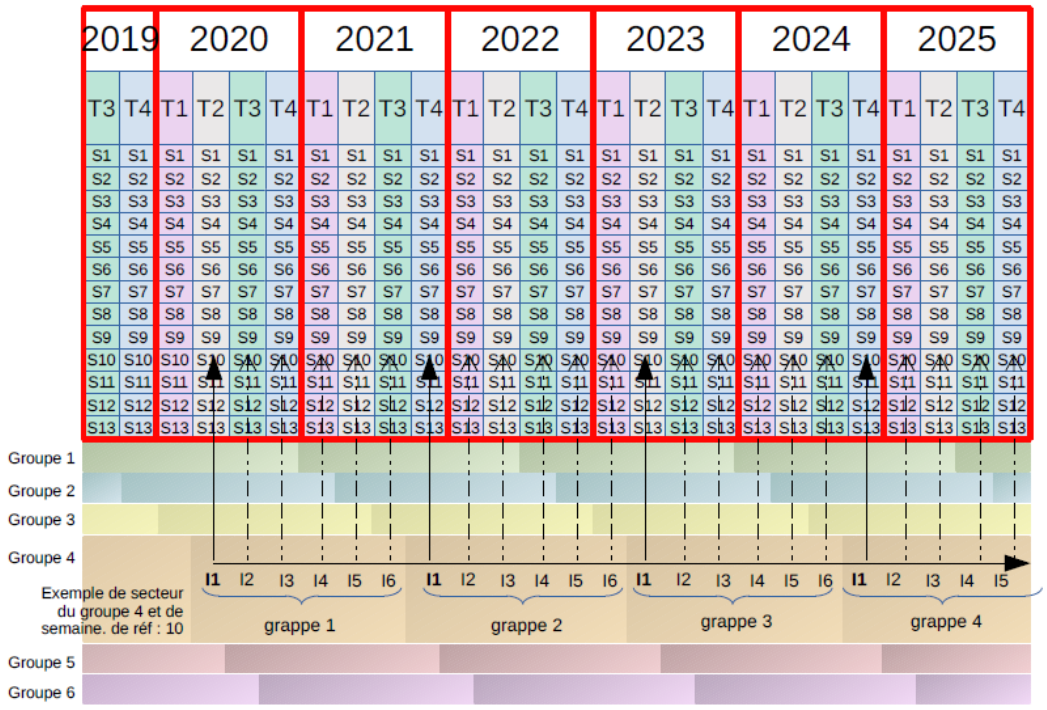
3. La sixième et dernière collecte, bien qu'elle ne nécessite plus de repérage, est également effectuée en face-à-face dans le protocole actuel. A compter de 2021, seule la première interrogation se déroulera en face à face. Toutes les interrogations ultérieures auront lieu par téléphone ou internet.

1.1.2 Semaine de référence

Pour atténuer les effets calendaires sur les estimations issues de l'enquête, la collecte se déroule sur l'ensemble des 52 semaines de l'année, soit 13 par trimestre. L'objectif est en effet que la collecte de l'EEC représente l'ensemble du trimestre et non pas seulement une semaine de chaque trimestre. La répartition de la collecte de l'EEC entre toutes les semaines de chaque trimestre permet également d'aplanir la charge des enquêteurs (voir l'encadré sur la collecte de l'EEC), chaque enquêteur pouvant avoir la responsabilité de plusieurs secteurs proches. Pour cela, en plus d'être alloués à un trimestre d'entrée, les secteurs se voient définir une **semaine de référence** dans le trimestre. Cette semaine de référence fige à la fois la semaine à propos de laquelle l'enquêteur posera des questions à l'enquêté et la période d'interrogation du logement.

La figure 1.1 explicite la répartition des grappes et des secteurs ainsi effectuée.

FIGURE 1.1 – Description du calendrier de collecte d'un secteur en fonction de son trimestre d'entrée et de sa semaine de référence.



La collecte de l'EEC

L'organisation de la collecte est structurée autour de la semaine de référence pour laquelle les personnes décrivent leur situation. Les réponses au questionnaire devront se rapporter à cette semaine-là.

Dans un premier temps, pendant la semaine de référence, l'enquêteur procède au repérage du logement à enquêter (il vérifie notamment que le logement existe toujours, qu'il est bien à usage d'habitation...) et prend contact avec les personnes à enquêter (envoi des lettres-avis, prise de rendez-vous). Dans un deuxième temps, l'enquêteur procède aux entretiens proprement dits. Il dispose pour cela de 2 semaines et 2 jours, ce qui limite le biais de mémoire de l'enquêté concernant la période d'interrogation. L'enquêteur a donc besoin d'environ 3 semaines pour interroger tous les logements d'une même grappe.

La première interrogation est habituellement la plus chronophage pour l'enquêteur parce qu'il doit d'une part repérer le logement, d'autre part ce premier entretien s'effectue en face-à-face, tout comme le sixième et dernier (en 2020). Les entretiens intermédiaires, quant à eux, se font en 2020 par téléphone.

Le renouvellement de l'échantillon de l'EEC démarre à compter du T3 2019 [4]. Sur l'exemple donné en figure 1.1, un secteur se voit attribuer :

— le groupe 4, ce qui signifie que ses grappes entreront en collecte tous les 6 trimestres respectivement aux 4^e, 10^e, 16^e, 22^e, 28^e et 34^e trimestres à compter du T3 2019 ;
— la 10^e semaine de référence : à chaque trimestre, la grappe est interrogée sur la situation vis-à-vis de l'emploi au cours de la 10^e semaine du trimestre.

La première grappe de ce secteur sera interrogée pour la première fois à propos de la 10^e semaine du 4^e trimestre de collecte, soit la première semaine de juin 2020 puis, lors des interrogations suivantes, au sujet des 10^e semaines de chaque trimestre. La deuxième grappe de ce secteur entrera en décembre 2021 et ainsi de suite jusqu'en juin 2029, date de la dernière interrogation de la dernière grappe du secteur.

Trimestre d'entrée et semaine de référence sont définis au niveau du secteur. Ils resteront les mêmes tout au long de l'enquête pour chacune de ses grappes. Cette double répartition doit être effectuée de telle sorte que, pour chaque semaine de référence, la taille de l'échantillon en collecte soit sensiblement la même au sein de chaque région, et qu'un enquêteur ne se retrouve pas confronté à une charge importante due à la collecte concomitante (donc à des semaines de référence voisines) de deux secteurs géographiquement proches. Ceci est particulièrement important quand il s'agit de l'entrée des grappes dans la collecte (voir encadré sur la collecte de l'EEC). Une fois les trimestres d'entrée et les semaines de référence définis pour chaque secteur, un ordre d'interrogation sera affecté à chacune de leurs 6 grappes. Cette affectation est décrite dans la section 1.3.

4. Ce renouvellement s'effectue par sixième à compter du T3 2019, c'est-à-dire qu'un sixième du nouvel échantillon entre en collecte à chaque trimestre entre le T3 2019 et le T4 2020, tandis qu'un sixième de l'ancien échantillon de l'EEC sort de la collecte à l'issue de chaque trimestre entre le T2 2019 et le T3 2020. Ainsi, du T3 2019 au T3 2020, des secteurs provenant de l'ancien échantillon EEC et des secteurs issus du nouvel échantillon de l'EEC cohabitent dans l'échantillon Emploi.

Il reste, à présent, à déterminer comment répartir par trimestre d'entrée et semaine de référence les 2944 secteurs sélectionnés et les grappes associées pour définir la date d'entrée en collecte de chaque logement de l'échantillon.

1.2 Affectation des trimestres d'entrée et semaine de référence aux secteurs

L'objectif recherché à travers la répartition des secteurs dans le temps est triple :
- lisser la charge d'un enquêteur, amené à enquêter les secteurs proches ;
- éviter une variabilité des estimations liées à une conjoncture ponctuelle locale (fermeture d'usine, travaux saisonniers...), c'est-à-dire éviter que des secteurs proches aient la même semaine de référence ;
- disperser les entrées dans l'échantillon tout au long des trimestres pour éviter des ruptures de série liées au renouvellement simultané de l'ensemble des grappes de l'échantillon.

Si l'on arrive à éloigner temporellement les collectes des secteurs géographiquement proches, on répond d'ores-et-déjà aux deux premiers objectifs. En effet, deux secteurs voisins sont susceptibles d'être affectés à un même enquêteur, et un évènement lié à l'emploi est souvent conséquence d'une spécificité territoriale.

C'est pourquoi, l'idée retenue est d'espacer sur un trimestre (soit 13 semaines) la collecte des secteurs d'une même UC, territoire au sein duquel a été tiré un échantillon de secteurs appelés à être enquêté par un voire deux enquêteurs. Plus précisément, on cherche à :
- étape 1 : déterminer l'ensemble des combinaisons trimestre d'entrée/semaine de référence *acceptables* pour les secteurs d'une UC (voir le point 1.2.1) ;
- étape 2 : puis tirer une de ces combinaisons (voir le point 1.2.2) ;
- étape 3 : enfin affecter chaque secteur de l'UC à l'un des éléments de la combinaison sélectionnée (voir le point 1.2.2).

On souhaite donc d'abord trouver une solution pour espacer les semaines de référence et les trimestres d'entrée des secteurs d'une même UC.

1.2.1 Les combinaisons *acceptables*

Il s'agit de déterminer ici l'ensemble des combinaisons dites *acceptables* pour chaque UC, c'est-à-dire l'ensemble des combinaisons des trimestres d'entrée et de semaines de référence des secteurs permettant d'espacer au mieux les semaines et trimestres affectés aux UC d'un secteur.

1.2.1.1 Critères à respecter pour les combinaisons trimestres d'entrée - semaines de référence au sein d'une UC

Une combinaison est considérée comme *acceptable* si elle répond à trois critères :

— Critère 1 : Dans la mesure du possible, espacer de 3 semaines les semaines de référence des secteurs d'une UC [5], et ainsi éviter que des secteurs voisins aient la même semaine de référence, améliorant de ce fait la précision des indicateurs, et facilitant le travail de l'enquêteur ;

— Critère 2 : A minima, espacer de 3 semaines l'entrée dans l'échantillon de plusieurs grappes appartenant à des secteurs d'une même UC, et ainsi éviter à un enquêteur une surcharge de collecte, liée à de nombreuses premières interrogations [6] ;

— Critère 3 : Affecter autant que possible à chaque secteur d'une même UC un trimestre d'entrée différent, et améliorer ainsi la précision des indicateurs.

Le premier critère fixe donc une contrainte sur les semaines de référence affectées aux secteurs de l'UC et le deuxième critère une contrainte sur la combinaison des semaines de référence et des trimestres d'entrée affectés aux secteurs de l'UC. Le troisième critère ne concerne que les trimestres d'entrée.

Le respect de ces trois critères au sein d'une UC va dépendre du nombre de secteurs dans l'UC. Plus une UC aura de secteurs, plus il sera difficile d'espacer les collectes de 3 semaines parmi les 13 semaines d'un trimestre, et de donner des trimestres d'entrée différents parmi les 6 possibles [7]. Le tableau 1.1 présente la répartition des UC en fonction du nombre de secteurs sélectionnés [8], en distinguant les UC exhaustives des UC non exhaustives.

5. On rappelle que tous les secteurs sont enquêtés tous les trimestres. C'est le nombre de fois où la grappe en collecte a été interrogée et la semaine de référence qui diffèrent d'un secteur à l'autre au cours d'un trimestre.

6. On rappelle que la première interrogation d'une grappe est plus longue que les interrogations suivantes et que la combinaison du trimestre d'entrée et de la semaine de référence déterminent la date de la première interrogation du secteur.

7. Pour rappel, l'affectation du trimestre d'entrée de la première grappe détermine les trimestres d'entrée des 5 autres grappes du secteur.

8. Les allocations du nombre de secteurs par UC ont été calculées à l'aide des formules présentées au point 5.2.2 de la partie C.

TABLE 1.1 – répartition des UC en fonction du nombre de secteurs sélectionnés, en distinguant UC exhaustives et UC non exhaustives

Nb secteurs tirés	UC exhaustives	UC non exhaustives
1	5	0
2	14	0
3	8	0
4	6	58
5	5	176
6	5	150
7	10	53
8	6	7
9	3	2
10	4	0
11	3	0
12	1	0
13	3	0
14	3	0
18	1	0
22	1	0
Total	*78*	*446*

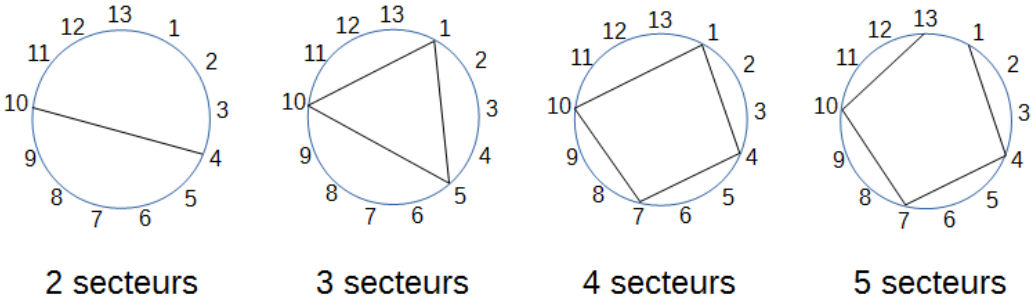
Note de lecture : 10 UC exhaustives comptent 7 secteurs enquêtés, tandis que 53 UC non exhaustives comptent 7 secteurs enquêtés.

Dans le paragraphe 1.2.1.2, on fait la liste des combinaisons de semaines de référence au sein d'une UC permettant de satisfaire le critère 1. Dans le paragraphe 1.2.1.3, le trimestre d'entrée sera croisé à ces combinaisons de semaines, afin d'aboutir à la liste des combinaisons trimestre d'entrée - semaine de référence permettant de satisfaire les trois critères présentés pour les UC. La liste de ces combinaisons sera ainsi établie pour chaque UC en fonction de son nombre de secteurs enquêtés.

1.2.1.2 Espacement des semaines de référence au sein d'une UC en fonction du nombre de secteurs

Dans cette section, on s'intéresse au critère 1. Pour respecter un espacement de 3 semaines pour les semaines de référence affectées aux secteurs d'une UC, il doit y avoir au maximum 4 secteurs dans l'UC. En effet, il faut tenir compte du fait que la première semaine d'un trimestre est consécutive à la dernière du trimestre précédent. En utilisant une représentation circulaire du trimestre, on constate que l'on peut répartir jusqu'à 4 secteurs d'une même UC sur un trimestre en respectant l'écart de 3 semaines entre deux secteurs, en tenant compte du chaînage des trimestres (cf. figure 1.2).

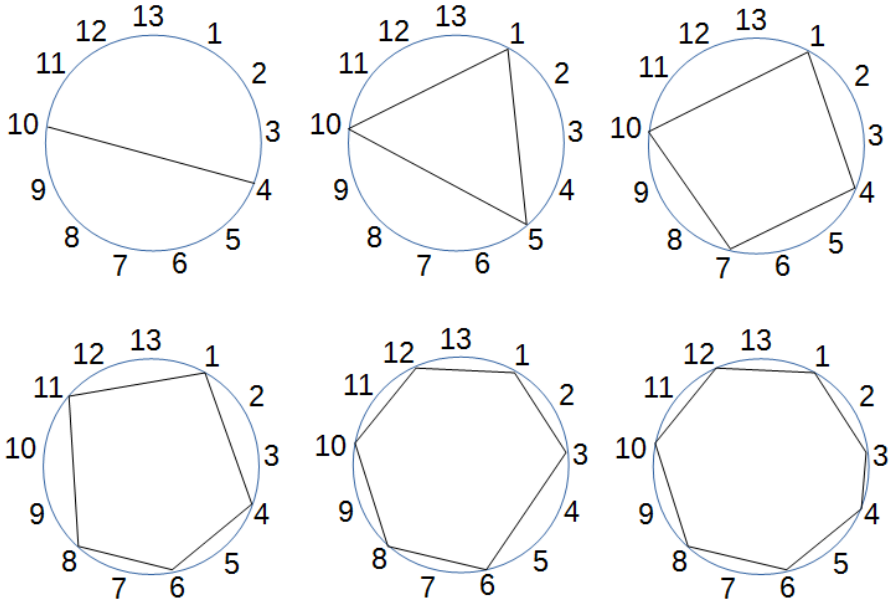
FIGURE 1.2 – Répartition de 2 à 5 secteurs avec un intervalle espéré de 3 semaines entre chacun.



Pour les UC ne contenant qu'un secteur, les 13 semaines peuvent être utilisées comme semaine de référence. Pour les UC contenant 2 et 3 secteurs, on les répartira au mieux afin d'optimiser l'espace entre chacun, soit 6 ou 7 semaines d'écart entre chaque secteur pour les UC contenant 2 secteurs et 4 ou 5 semaines d'écart pour celles contenant 3 secteurs.

Il n'est pas possible de respecter le critère 1 pour les UC contenant 5 secteurs ou plus. Pour les UC à 5, 6 ou 7 secteurs, on acceptera un écart moindre (1 ou 2 semaines) entre les semaines de référence des secteurs. La figure 1.3 présente des répartitions optimisées pour les UC contenant 2 à 7 secteurs.

FIGURE 1.3 – Répartition *optimisée* sur 13 semaines de 2 à 7 secteurs.



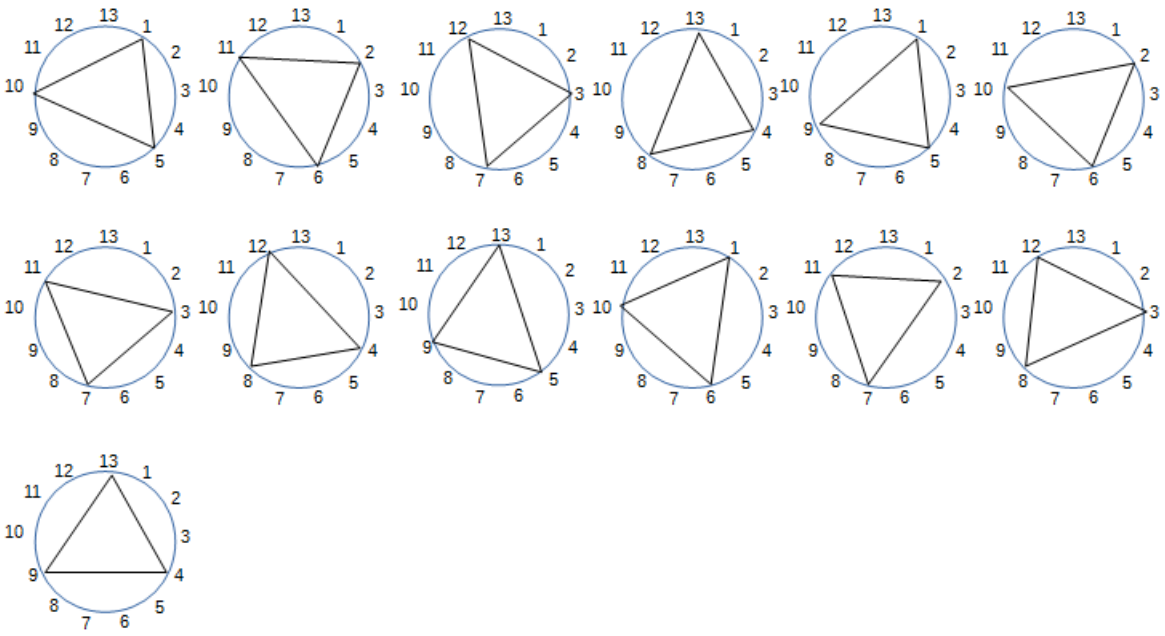
Au-delà de 7 secteurs dans une UC[9], un découpage de l'UC est réalisé, afin de

9. Le seuil a été fixé arbitrairement à 7 secteurs à la vue du tableau 1.1 ; on observe en effet bien

constituer des ensembles géographiques de 7 secteurs au plus, et ainsi de se retrouver dans un des cas précédents [10]. Les UC contenant 8 à 14 secteurs tirés sont scindées en deux grâce aux coordonnées cartographiques [11]. Dans le tableau 1.1, il apparaît que deux UC contiennent respectivement 18 et 22 secteurs, donc trop de secteurs pour être divisées en deux. Un découpage « à la main », regroupant intelligemment les secteurs en fonction de leur position est réalisé.

Au final, on dispose de regroupements de 1 à 7 secteurs correspondant à des UC entières ou à des partitions d'UC. Les combinaisons de semaines de référence acceptables sont établies exhaustivement pour chacun des 7 cas. La figure 1.4 montre, par exemple, l'ensemble des 13 combinaisons différentes permises pour une UC contenant 3 secteurs.

FIGURE 1.4 – Répartitions possibles des semaines de référence pour une UC à 3 secteurs sur 13 semaines.



On dispose ainsi de l'ensemble des combinaisons de semaines permettant de satisfaire au mieux le critère 1 au sein d'une UC pour un nombre de secteurs donnés.

moins d'UC contenant 8 secteurs tirés ou plus, que d'UC contenant 7 secteurs ou moins dans l'échantillon.

10. On présuppose, dans ces UC a priori plus peuplées, la présence d'au moins deux enquêteurs, rendant possible la levée des contraintes établies plus tôt.

11. On découpe horizontalement ou verticalement selon la dispersion géographique des groupes obtenus.

1.2.1.3 Répartition des trimestres d'entrée aux combinaisons de semaines de référence au sein d'une UC

Après avoir établi la liste des combinaisons de semaines de référence possibles au sein d'une UC, il convient de les croiser avec les trimestres d'entrée afin d'aboutir aux combinaisons trimestres d'entrée - semaines de référence satisfaisant l'ensemble des 3 critères.

Le troisième critère (affecter à chaque secteur d'une UC [12] un trimestre d'entrée différent) est possible dès lors que le regroupement étudié possède 6 secteurs ou moins. Ainsi, seules les UC composées de 7 secteurs ne peuvent pas respecter cette règle puisque chaque secteur doit nécessairement être affecté à un des 6 groupes de rotation. Dans ce dernier cas, il y aura donc deux secteurs enquêtés au même trimestre au sein d'une UC.

Si la répartition aléatoire des trimestres d'entrée aux secteurs d'une UC est *a priori* aisée, c'est à travers celle-ci que le critère 2 [13] sera respecté. L'affectation d'un trimestre d'entrée à un secteur ne sera possible que si aucun autre secteur de la même UC n'entre dans l'échantillon à moins de 3 semaines d'intervalle. Par exemple, si une grappe d'un secteur entre dans l'échantillon en 12^e semaine du troisième trimestre, une grappe d'un autre secteur de la même UC ne pourra être collectée pour la première fois en 1^e semaine du quatrième trimestre. Pour une UC, l'affectation des trimestres d'entrée à chacun de ses secteurs va donc dépendre de la combinaison de semaines appliquée.

Cas des UC contenant 4 secteurs ou moins

Pour les UC contenant 1 à 4 secteurs, la répartition des trimestres ne pose pas de question, puisque les combinaisons de semaines de référence considérées comme acceptables permettent déjà un écart de 3 semaines, quel que soit le trimestre d'entrée des secteurs.

Cas des UC de 5 ou 6 secteurs

Pour une UC à 5 ou 6 secteurs, les collectes de certains d'entre eux peuvent être espacées de deux semaines. L'affectation d'un trimestre différent pour chaque secteur évite pour la plupart des combinaisons que deux secteurs interrogés à 15 jours d'intervalle soient tous les deux en première interrogation. Le cas peut toutefois se produire si le premier secteur est affecté au trimestre consécutif à celui défini pour le dernier secteur, et si les semaines de référence sont espacées ainsi :

— semaine 1 pour le premier secteur et semaine 12 pour le dernier secteur
— semaine 2 pour le premier secteur et semaine 13 pour le dernier secteur.

12. Dans ce paragraphe, on entend par UC, soit une UC entière quand elle comporte au plus 7 secteurs, soit une partition d'UC construite dans la section précédente pour n'avoir que des groupes de 1 à 7 secteurs.

13. Pour rappel, le critère 2 est le suivant : espacer de 3 semaines au moins les entrées des secteurs d'une même UC.

Dans ces cas-là, une grappe du dernier secteur entrera en collecte deux semaines avant une grappe du premier secteur, ce qui ne convient pas.

Cas des UC de 7 secteurs

Les UC à 7 secteurs présentent deux difficultés. La première concerne l'affectation d'un même trimestre d'entrée à deux secteurs de l'UC. La deuxième rejoint la problématique décrite précédemment : l'affectation de deux trimestres consécutifs, à des secteurs portant la dernière et la première semaines de référence, et incompatible avec le respect des 3 semaines d'écart pour l'entrée en collecte.

Pour résoudre le premier problème, on choisit arbitrairement de répartir les 6 premiers secteurs (au sens de la combinaison) entre les différents trimestres, puis d'affecter un trimestre déjà utilisé au dernier secteur (portant donc la 12^e ou la 13^e semaine de référence, par construction des combinaisons possibles) [14].

C'est donc sur l'affectation du trimestre de référence au 7^e secteur de la combinaison que se porte l'enjeu de la satisfaction du critère 2. En effet, pour affecter correctement les trimestres d'entrée aux secteurs, il faudra :

— que le 6^e et le 7^e secteurs, dont les semaines de référence sont, par construction, espacées au maximum de 15 jours, n'aient pas le même trimestre d'entrée

— que le 7^e secteur ne soit pas affecté au trimestre précédant celui du 1^{er} secteur, puisque les semaines de référence sont, par construction, espacées au maximum de 15 jours.

A partir de cette contrainte et de la liste des combinaisons de semaines déterminées dans le paragraphe 1.2.1.2, on est en mesure de lister les combinaisons trimestres d'entrée - semaines de référence *acceptables* pour les UC à 7 secteurs.

1.2.1.4 Base de sondage des combinaisons trimestres d'entrée - semaines de référence pour chaque UC

Les paragraphes 1.2.1.2 et 1.2.1.3 ont permis de construire l'ensemble des combinaisons de couples (trimestres d'entrée, semaines de référence) acceptables.

Par exemple, pour une UC ayant 3 secteurs, 13 combinaisons de semaines de référence sont autorisées. Pour affecter un trimestre différent aux 3 secteurs de l'UC, on a, pour chaque combinaison, 6 possibilités d'affectation pour le premier secteur, 5 pour le deuxième et 4 pour le troisième. De ces 13 combinaisons de semaine, on génère alors $13 \times 6 \times 5 \times 4 = 1560$ combinaisons de couples (trimestres d'entrée, semaines de référence) possibles.

14. Si ce choix exclut certaines possibilités, il ne portera a priori pas conséquence sur les indicateurs, puisque l'affectation des trimestres aux 6 premiers secteurs est aléatoire.

On établit une liste de ces combinaisons de couples pour chacun des 7 cas définis plus haut, correspondant au nombre de secteurs par UC ou sous-ensemble de l'UC, en excluant du champ des possibles celles ne respectant pas le critère des 3 semaines d'écart pour l'entrée en collecte.

Cette liste constitue une base de sondage de combinaisons pour chaque UC au sein de laquelle sera tiré une combinaison par UC. Ainsi, il restera simplement à affecter aléatoirement à chaque secteur de l'UC un couple (trimestre d'entrée, semaine de référence) de la combinaison tirée pour disposer d'un trimestre d'entrée et d'une semaine de référence pour chacun des secteurs de l'échantillon de l'enquête Emploi en continu. Il reste à définir la méthode de tirage de ces combinaisons.

1.2.2 Tirage réjectif de la répartition des secteurs

Afin de lisser la charge des enquêteurs et estimer avec précision le nombre de chômeurs, chaque semaine de chaque trimestre, on souhaite obtenir une répartition équitable du nombre de secteurs pour chaque trimestre d'entrée et pour chaque semaine de référence. Pour ce faire, on définit des contraintes que l'échantillon des différentes combinaisons (trimestres d'entrée, semaines de référence) tirées pour chacune des UC va devoir respecter.

Ces contraintes sont les suivantes :

— Chaque échantillon restreint à une semaine de référence doit contenir un nombre de secteurs qui approche l'égale répartition des secteurs entre les 13 semaines à 5 % près ;
— Chaque échantillon restreint à un trimestre d'entrée doit contenir un nombre de secteurs qui approche l'égale répartition des secteurs entre les 6 trimestres à 1,5 % près ;
— Chaque échantillon restreint à une semaine de référence doit avoir un estimateur d'Horvitz-Thompson du total de nombre de chômeurs [15] avec un écart relatif de moins de 3,5 % avec le vrai total ;
— Chaque échantillon restreint à un trimestre d'entrée doit avoir un estimateur d'Horvitz-Thompson du total de nombre de chômeurs avec un écart relatif de moins de 2 % avec le vrai total.

On sélectionne une combinaison possible pour chaque UC à probabilités égales. L'échantillon obtenu définit ainsi, pour chaque UC, autant de trimestres d'entrée et semaines de référence qu'il y a de secteurs. Dans chaque UC, on affecte enfin aléatoirement à chaque secteur un couple (trimestre, semaine) de la combinaison retenue.

Après 2000 réitérations de cet algorithme, on obtient un échantillon satisfaisant les 4 contraintes précédentes grâce à un tirage réjectif (voir encadré).

15. Par « nombre de chômeurs » on entend ici le nombre de personnes percevant une allocation chômage ou préretraite.

Si le trimestre d'entrée et la semaine de référence de chaque secteur de l'échantillon sont à présent connus, il reste à déterminer l'ordre d'interrogation des différentes grappes de chaque secteur.

Le tirage réjectif

Un tirage réjectif est un algorithme dans lequel le tirage est réitéré jusqu'à ce que l'échantillon obtenu satisfasse les contraintes définies, en l'occurrence, les 4 critères fixés dans ce paragraphe. Le tirage réjectif présente le désavantage de déformer les probabilités d'inclusion par rapport à l'algorithme initial sur lequel il se base.

Néanmoins, dans ce cas précis, ce n'est pas problématique puisque toute combinaison de secteurs - semaines de référence - trimestres d'entrée qui décale les trimestres d'entrée d'un trimestre pour chacun des secteurs enquêtés ou les semaines de référence d'une semaine pour chacun des secteurs enquêtés continue à satisfaire les 4 critères précédents. Ces décalages peuvent aboutir à des combinaisons non acceptables mais cela ne remet pas en cause la symétrie des combinaisons.

Cette symétrie des combinaisons et leur tirage à probabilités égales garantit qu'un secteur a la même probabilité d'être affecté à toute semaine de référence ainsi qu'à tout trimestre d'entrée. Si les probabilités d'inclusion des combinaisons sont donc déformées par le tirage réjectif par rapport aux probabilités égales utilisées au départ, le secteur est bien affecté à probabilités égales à la semaine de référence et au trimestre d'entrée dans ce tirage réjectif.

1.3 Les rangs d'interrogation

Lors de l'entrée en collecte d'un secteur, les logements de l'une de ces grappes sont interrogés pendant six trimestres, puis c'est la grappe suivante qui entre en collecte au septième trimestre jusqu'au douzième, et ainsi de suite jusqu'à la dernière interrogation de la sixième grappe du secteur.

Attribuer le rang d'interrogation des grappes au sein d'un secteur correspond donc à définir quelle est la première grappe, la deuxième, etc. à entrer en collecte au sein du secteur. Toutefois, certains des secteurs sont composés non pas de 6 mais de 7 grappes. Il convient donc, en premier lieu, de sélectionner, par tirage aléatoire simple, 6 grappes parmi les 7 de ces secteurs. On peut alors attribuer le rang d'interrogation à chaque grappe de l'échantillon. La grappe non sélectionnée n'a pas de rang d'interrogation et ne sera pas collectée.

Aucun critère de collecte ou de précision sur l'échantillon ne semble justifier la mise en place d'une méthode particulière pour établir l'ordre de mise en collecte des grappes d'un secteur. Toutefois, ce nouvel échantillon de l'EEC fait suite à celui utilisé depuis 2009. Or, aucune disjonction entre ces deux échantillons n'a été appliquée, afin d'avoir la base de logements la plus complète possible, et ainsi la meilleure précision possible pour les indicateurs estimés. 7 833 logements sont ainsi présents dans les deux échantillons.

Pour éviter qu'un ménage récemment interrogé le soit de nouveau rapidement, on établit trois indicateurs précisant le nombre de logements interrogés récemment dans chaque grappe à partir desquels va être déterminé le rang d'interrogation recherché. Ainsi, pour chaque grappe des secteurs de l'échantillon de 2019, on décompte :

— **A** : le nombre de logements interrogés entre le T1 2015 et le T2 2016 ;
— **B** : le nombre de logements interrogés entre le T3 2016 et le T4 2017 ;
— **C** : le nombre de logements interrogés entre le T1 2018 et le T2 2019.

Dans chaque secteur, on trie alors les grappes par ordre décroissant des valeurs de C, puis B, puis A. Les premières grappes ainsi triées rentreront dans le nouvel échantillon le plus tard possible. A nombre de logements égal pour ces trois variables, un tirage aléatoire simple établit le rang d'interrogation [16].

Maintenant que le rang d'interrogation des grappes est établi, que la semaine de référence et le trimestre d'entrée des secteurs sont définis, la date exacte d'entrée en collecte de chaque logement du nouvel échantillon est désormais fixée.

Le chapitre suivant décrit la finalisation de la composition des grappes emploi, avec le rattachement de l'ensemble des logements autres que des résidences principales aux grappes constituées au chapitre 2 de la partie B, puis l'association des logements nouvellement construits à ces grappes et la sélection des logements effectivement mis en collecte.

16. En particulier, pour les secteurs dont aucun logement n'a été interrogé au cours des 6 dernières années, le rang d'interrogation des grappes n'est défini que par un tirage aléatoire simple.

Chapitre 2

Finalisation de la composition des grappes mises en collecte de l'échantillon EEC

Le chapitre 2 de la partie B décrit la construction des grappes de l'EEC à partir des résidences principales contenues dans Fidéli, puis le regroupement de ces grappes en secteurs. Ces secteurs sont alors échantillonnés en coordination avec le tirage de l'échantillon-maître utilisé pour les autres enquêtes ménages (voir les chapitres 4 et 5 de la partie C), puis, comme présenté dans le chapitre précédent, les logements sont méthodiquement répartis par date d'entrée en collecte, par l'intermédiaire des secteurs puis des grappes, pour permettre le respect des contraintes de collecte et de précision des indicateurs.

L'une des particularités de l'enquête Emploi en continu est que son échantillon est un panel tiré au départ pour une durée de 9 ans. Or la base de sondage et l'appartenance des logements au champ de l'enquête évoluent au cours du temps. En effet, au cours des 9 ans d'enquête, certains logements sont construits, d'autres détruits ; certaines résidences deviennent principales, d'autres secondaires ou vacantes, et réciproquement. Cette différence de statut des logements existe également entre l'information contenue dans la base de sondage et la réalité sur le terrain. Il est ainsi nécessaire d'interroger des logements qui, au moment de la constitution des secteurs étaient vacants ou des résidences secondaires, pour représenter l'entièreté du champ de l'enquête.

Le vieillissement des bases conduit à devoir ajouter les résidences secondaires et logements vacants de Fidéli 2016 aux grappes pour s'assurer de l'exhaustivité de l'enquête au moment de l'exploitation des données collectées. Ces logements, hors-champ au moment de la constitution de la base de sondage de secteurs, peuvent être devenus des résidences principales au moment de la collecte (section 2.1). Plus encore, l'évolution du bâti entraîne inexorablement un manque, sans le rafraîchissement de l'échantillon intégrant entre autres les logements nouvellement construits (section 2.2). C'est pourquoi, une telle mise à jour des grappes est pratiquée, conduisant à l'augmentation de la taille de certaines grappes

et entraînant par la suite un nouveau degré d'échantillonnage pour limiter la charge de collecte pour les enquêteurs. Ces éléments sont présentés en section 2.3.

2.1 Le rattachement des logements autres que des résidences principales

Pour atteindre le champ de l'enquête Emploi, on sélectionne des logements ordinaires dans lesquels vivent les individus. C'est pourquoi, les grappes ont été construites à partir des résidences principales de la source Fidéli. Cependant, l'imperfection des fichiers administratifs par rapport à la réalité du terrain ne permet pas d'assurer que certains logements autres que des résidences principales au sens fiscal [1] n'hébergent pas d'individus appartenant au champ de l'enquête [2]. Écarter ces logements entraînerait donc un défaut de couverture. D'ailleurs, les collectes passées de l'EEC, dont l'échantillon contenait des résidences supposées non principales, permettent de constater qu'en 2017-18, 34 % des résidences non principales d'après la base de sondage se sont avérées être principales lors de la collecte.

Le point 2.1.1 présente la base de résidences non principales issues de Fidéli 2016 qui devront être rattachées à des secteurs. Ce rattachement s'effectue sur un critère de distance entre les résidences non principales de Fidéli 2016 et les secteurs. La méthode naturelle pour rattacher les résidences non principales au secteur le plus proche s'avère dysfonctionnelle au point 2.1.2 car elle aboutit à un nombre très élevé de résidences non principales dans certains secteurs. D'autres méthodes visant à limiter le nombre de résidences non principales par secteur ont donc été envisagées. L'une d'elle, la méthode dite *cascade* a finalement été retenue et est présentée au point 2.1.3. Le choix du seuil de résidences non principales fixé pour le rattachement des résidences non principales aux secteurs fait l'objet du point 2.1.4. A ce stade, toutes les résidences non principales seront ainsi rattachées à des secteurs. Il ne restera donc plus qu'à rattacher ces logements à une grappe au sein de leur secteur dans le point 2.1.5.

2.1.1 Rattachement des résidences non principales à des secteurs

Au départ, on dispose des grappes constituées à partir des résidences principales connues dans le fichier Fidéli 2016 [3] ; elles comptent en très grande majorité entre 17 et 24 résidences principales. On rappelle que les résidences principales situées à un même étage appartiennent à une même grappe.

Pour rattacher les résidences non principales (RNP) aux grappes ainsi créées, on suit le même procédé, en regroupant par *étages* les 6 372 488 RNP de métropole [4], relevées

1. Les résidences non principales regroupent essentiellement les résidences secondaires et les logements vacants.

2. Plusieurs raisons peuvent être la cause de ces imperfections : un manque de fraîcheur de l'information, une erreur dans les fichiers administratifs ou de déclaration, un intérêt fiscal à déclarer une résidence secondaire comme sa résidence principale...

3. Voir la constitution des grappes en section 2.2 de la partie B.

4. Pour rappel de la section 2.1 de la partie B, un étage forme un ensemble de logements indivisible du point de vue de la collecte de l'enquête Emploi.

dans la source Fidéli 2016, et jugées exploitables [5]. Au total, le stock de logements de métropole, exploitables, principaux ou non, issus de Fidéli 2016, se répartit en 25 730 236 *étages* différents, que l'on peut répertorier en trois cas distincts :

— Cas 1 : les 20 475 696 *étages* ne comportant que des résidences principales (25 326 780 RP sont concernées)
— Cas 2 : les 3 998 695 *étages* ne comportant que des résidences non principales (4 527 591 RNP sont concernées)
— Cas 3 : les 1 255 845 *étages* composés à la fois de résidences principales et non principales (2 955 907 RP et 1 844 897 RNP sont concernées)

Le premier cas n'est pas concerné par le traitement présent, puisque aucune RNP n'est relevée dans ces *étages*. Le dernier cas est simple à traiter. En effet, chaque résidence principale d'un étage donné étant affectée à la même grappe (et au même secteur) pour des raisons de collecte, on impose aux résidences non principales de cet étage cette même grappe et ce même secteur. Il convient à présent de déterminer la méthode de rattachement des *étages* du cas 2, composés uniquement de RNP. Dans la suite de cette partie, on entendra par *étages* ou résidences non principales uniquement ceux relevant de ce deuxième cas. On rappelle que ces *étages* peuvent appartenir à des immeubles mais peuvent également être composés d'une unique maison individuelle.

Afin de limiter les temps de traitement, tout en assurant la proximité géographique des logements collectés par un même enquêteur, une même semaine, on choisit de procéder en deux temps en rattachant tout d'abord chaque étage à un secteur [6], puis au sein de celui-ci à une des 6 grappes.

2.1.2 Le rattachement naïf : un résultat insatisfaisant

Il s'agit donc d'abord de rattacher chaque *étage* de résidences non principales à un secteur, indépendamment du fait qu'il ait été tiré ou non. Cette première étape est la plus déterminante, puisqu'elle va réduire considérablement le nombre de grappes d'affectations possible pour chaque étage [7].

L'objectif est donc d'affecter chaque *étage* de RNP à son secteur *le plus proche*. Pour cela, il est donc nécessaire de définir une distance entre un *étage* et un secteur. Deux possibilités sont envisagées :

— le scénario dit « par barycentre » : à chaque *étage*, on attribue le secteur dont le barycentre des résidences principales est le plus proche de l'*étage* étudié ;
— le scénario du plus proche voisin : on affecte à l'*étage*, le secteur possédant la résidence principale la plus proche de l'*étage* étudié.

Le choix du scénario *barycentre*

Dans les deux scénarios, on restreint le champ des possibles à l'UC.

5. Quelques milliers de résidences non principales ne sont pas exploitables, car il manque certaines informations comme les coordonnées (x, y) ou le niveau dans le bâtiment.

6. Afin de suivre la même logique que pour la constitution des grappes dans la partie B, les étages de résidences non principales ont été nécessairement rattachés à un secteur de leur UC. Ainsi, il est uniquement nécessaire de rattacher les résidences non principales des UC dans lesquelles des secteurs ont été tirés. Au sein d'une UC tirée, le rattachement d'un étage de résidences non principales à un secteur sera indépendant du fait que ce secteur figure ou non dans l'échantillon EEC.

7. On travaille par étage, car on souhaite que tous les logements d'un même étage soient rattachés à la même grappe.

La distance retenue pour établir cette proximité est la distance euclidienne [8] entre les coordonnées cartographiques de l'*étage* étudié [9] et celles du barycentre d'un secteur, ou d'une résidence principale. Dans l'absolu, pour chaque *étage* de RNP d'une UC, il faudrait calculer la distance avec chacun des barycentres des secteurs de l'UC, et avec chaque *étage* de résidences principales de cette même UC. Ces temps de calcul seraient bien trop importants, pour une utilité toute relative dans la plupart des cas. Ainsi, on choisit de découper chaque UC en plusieurs cases à l'aide d'une grille (découpage à pas régulier, déterminé en fonction de l'étendue géographique des résidences principales de l'UC), puis de se restreindre, pour un *étage* de RNP, aux cases les plus proches.

A ce stade, le choix d'un des deux scénarios devrait dépendre a priori des différences de rattachements des RNP aux secteurs. Cependant, à l'issue de tests, il s'avère que le scénario du plus proche voisin est bien trop coûteux en temps de calcul. En effet, le scénario *barycentre* demande environ 50 heures pour traiter les étages de l'ensemble des 6 372 488 résidences non principales de métropole, alors que, pour une UC, le scénario du plus proche voisin demande environ 12 fois plus de temps [10]. Par ailleurs, la différence de résultats entre les deux scénarios est modérée, car les secteurs ont été construits dans l'objectif d'être aussi compacts que possible. Le barycentre d'un secteur est donc proche des résidences principales qui le composent. Le scénario *barycentre* est ainsi choisi pour rattacher chaque résidence non principale à un secteur.

L'inefficacité du rattachement de l'étage au secteur le plus proche

A partir d'ici, on entend à nouveau par résidences non principales, à la fois les RNP appartenant à un *étage* contenant uniquement des résidences non principales (cas 2), et celles situées au même étage que des résidences principales (cas 3).

Le rattachement, sans autre contrainte, des *étages* de RNP du cas 2 au barycentre du secteur le plus proche, entraîne, dans tous les départements, une forte disparité de la taille des secteurs en nombre de logements. Deux exemples sont proposés en figure 2.1 à travers la répartition des secteurs par nombre de résidences non principales qui leur sont rattachées :

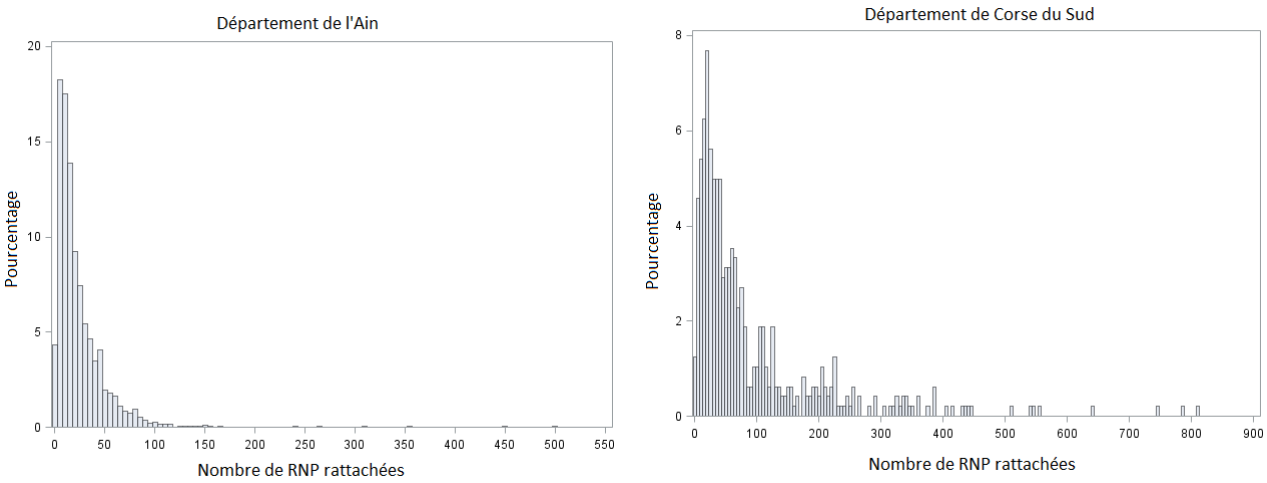
- l'Ain, qui compte 2142 secteurs dans 20 UC et environ 50 000 RNP, proche de la moyenne nationale de 24 résidences non principales par secteur ;
- la Corse-du-Sud, qui regroupe 481 secteurs sur 4 UC pour environ 45 000 RNP, atypique avec 94 résidences non principales par secteur.

8. Plusieurs autres méthodes ont été envisagées pour relier les *étages* au barycentre d'un secteur ou à la résidence principale la plus proche (arbres de décision utilisant les coordonnées cartographiques, partitions de Voronoi,...), mais elles n'ont pas abouti, soit par granularité trop élevée, soit par coût trop important en termes de performances.

9. Par construction, tous les logements d'un *étage* ont les mêmes coordonnées cartographiques.

10. Le temps de calcul dépend, par UC, du nombre de RNP à rattacher et du nombre de résidences principales.

FIGURE 2.1 – Histogramme des nombres de RNP rattachées par secteur dans les départements de l'Ain et de Corse du Sud, en l'absence de limite de nombre de résidences non principales



On observe, dans ces deux exemples, une disparité de la répartition des résidences non principales par secteur. Dans un département non atypique comme l'Ain, 21 secteurs se retrouvent sans RNP, alors que d'autres secteurs se voient rattacher jusqu'à 500 résidences non principales. Ce constat est encore plus marqué en Corse-du-Sud, département hautement touristique, possédant beaucoup de résidences secondaires : 4 secteurs sont dépourvus de RNP, et, à l'inverse, des secteurs s'en voient attribuer plus de 800.

Ces résultats montrent la nécessité d'une méthode pour limiter ces cas extrêmes, tant du point de vue de la collecte que de la dispersion des poids [11].

2.1.3 Méthode à seuil

Pour éviter l'écueil d'une disparité trop importante de la quantité de résidences non principales par secteurs, on ajoute une contrainte limitant ce nombre, aux dépens de la proximité géographique initialement recherchée [12]. On considère un paramètre *seuil* : il correspond au nombre maximum de résidences non principales que l'on souhaite rattacher à un barycentre de chaque secteur. Par définition, les RNP situées au même étage que des résidences principales ne peuvent être associées qu'au secteur de ces résidences principales. On ne pourra donc influer que sur le rattachement des *étages* ne possédant que des RNP. Cependant, le *seuil* à ne pas dépasser est bien relatif à l'ensemble des résidences non principales.

Pour modifier l'algorithme dit du barycentre, afin qu'il respecte la contrainte posée, on propose deux alternatives.

— Solution 1 – la méthode dite « grille »
— Solution 2 – la méthode dite « cascade »

11. La dispersion des poids provient du fait que, dans les grappes volumineuses en termes de nombre de logements, une nouvelle étape de tirage présentée en section 2.3 sera mise en oeuvre.

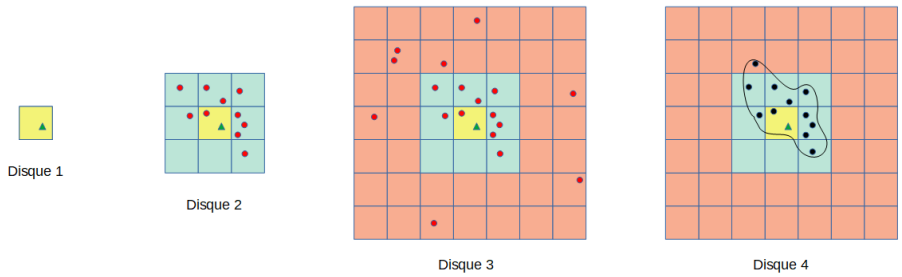
12. Pour rappel, un choix similaire avait déjà été effectué lors de la constitution des grappes, en autorisant des sauts entre les logements afin de se rapprocher d'une cible de 20 résidences principales par grappe (voir paragraphe 2.2.2.2 de la partie B).

La méthode grille

L'idée de cette méthode est d'appliquer à l'UC un découpage en grille permettant de créer un voisinage pour chaque *étage* composé uniquement de RNP, au sein duquel sera effectué le rattachement de l'*étage* à un barycentre, tant que le seuil n'est pas dépassé pour le secteur ainsi affecté. Plus précisément, on commence par appliquer la grille utilisée précédemment pour calculer les distances entre RNP et barycentre. On parcourt alors cette grille case par case. La suite de l'algorithme se déroule en deux étapes :

1. Étape 1 : Constitution des disques d'intérêt autour de l'*étage* étudié. Cette étape est illustrée par la figure 2.2.
   — Pour chaque case contenant au moins un *étage* de résidences non principales [13] (disque 1, en jaune dans la figure 2.2, l'étage étant le triangle vert), on liste les barycentres présents dans les 9 cases autour (la case de l'*étage* étudié et les 8 cases contiguës) ; s'il n'y en a pas, on étend la surface d'intérêt aux cases un cran plus éloignées, et ce jusqu'à ce qu'un barycentre, au moins, soit détecté comme illustré dans la figure 2.3.
   — Une fois le disque des barycentres obtenu (disque 2, en bleu), on étend à nouveau la surface d'intérêt de deux crans (disque 3, en rouge), afin de s'assurer de prendre en compte exhaustivement les résidences non principales dont la distance aux différents barycentres du disque 3 est inférieure à la distance barycentre – *étage* étudié dans le disque 2 [14]. Les barycentres correspondant forment le disque 4.
2. Étape 2 : Rattachement de l'*étage* à un secteur.
   — On sélectionne le barycentre le plus proche de l'étage, au sens de la distance euclidienne.
   — On calcule les distances entre le barycentre sélectionné et les *étages* de RNP, pour lesquelles ce barycentre figure à l'intérieur de leur disque 4.
   — Si les logements de l'*étage* étudié appartiennent aux *seuil* plus proches RNP du barycentre sélectionné [15], alors l'*étage* se voit attribuer le secteur du barycentre. Sinon, on passe au barycentre suivant (le deuxième plus proche).

FIGURE 2.2 – Étape 1 de l'algorithme de rattachement d'un étage par la méthode grille

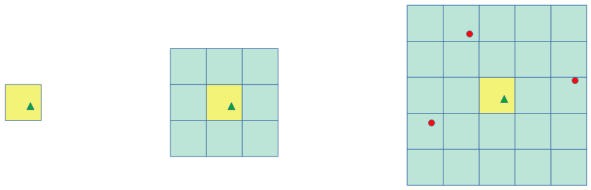


13. Pour rappel, il s'agit des *étages* sans résidences principales ; les *étages* avec résidences principales ont déjà été affectés.

14. En effet, la forme rectangulaire de la grille ne garantit pas de capter dans le disque 2 les barycentres les plus proches de l'*étage* étudié.

15. Les résidences non principales des *étages* contenant également des résidences principales sont également comptées ici.

FIGURE 2.3 – Élargissement du disque 2 dans l'étape 1 de l'algorithme de rattachement d'un étage par la méthode grille



A la fin de l'algorithme, certaines résidences non principales peuvent ne pas avoir été affectées. C'est le cas, par exemple, de maisons éloignées de toutes autres habitations et qui ne sont jamais parmi les *seuil* plus proches résidences principales des différents barycentres étudiés. Ces résidences principales non affectées sont une limite de l'algorithme, comme présenté plus loin dans les commentaires autour de la figure 2.5.

La méthode cascade

Cette méthode consiste à déverser les *étages* d'un secteur trop chargé sur le secteur le plus proche, jusqu'à respect de la valeur du *seuil*. L'algorithme est le suivant :

1. Étape 1 : On affecte chaque *étage* de résidences non principales au secteur dont le barycentre est le plus proche ;
2. Étape 2 : Après affectation de toutes les RNP, certains secteurs ne respectent plus la contrainte de seuil. Dans ces secteurs, pour chaque *étage* composé uniquement de RNP, on calcule la distance aux barycentres des secteurs ne saturant pas la contrainte de seuil ;
3. Étape 3 : On réaffecte alors le surplus de résidences non principales [16] du secteur étudié au secteur dont le barycentre est le plus proche. Tant qu'au moins un secteur ne respecte pas le seuil établi, on reproduit les étapes 2 et 3.

Contrairement à la méthode grille, cet algorithme ne laisse aucune résidence non principale sans affectation, sauf si le seuil fixé est trop contraignant (i.e s'il est inférieur au ratio $\frac{\text{nombre de RNP de l'UC}}{\text{nombre de secteurs de l'UC}}$ [17]).

Le choix de la méthode cascade

La comparaison des distances des RNP au barycentre de leur secteur d'affectation obtenu à l'issue de chaque algorithme va permettre de choisir la méthode finalement utilisée. Dans les deux cas, le seuil est fixé à 120 résidences non principales [18].

16. Plus exactement, les *étages* concernés par le surplus de RNP, afin de ne pas dissocier deux RNP d'un même étage.

17. En réalité, la valeur critique du seuil s'obtient de manière un peu plus complexe puisqu'un étage ne peut être subdivisé. La distribution de la taille des étages entre donc en compte dans l'établissement de cette valeur critique.

18. La détermination du seuil est présentée au point 2.1.4.

Afin d'illustrer ces différences entre les deux méthodes, considérons l'UC composée uniquement de la commune de Cagnes-sur-Mer dans les Alpes-Maritimes [19]. Il s'agit d'une commune en bord de mer s'étendant par des vallons dans les terres ; elle contient 197 secteurs et 6025 résidences non principales. Les moyennes des distances calculées entre chaque RNP et le barycentre de son secteur d'affectation sont assez proches :

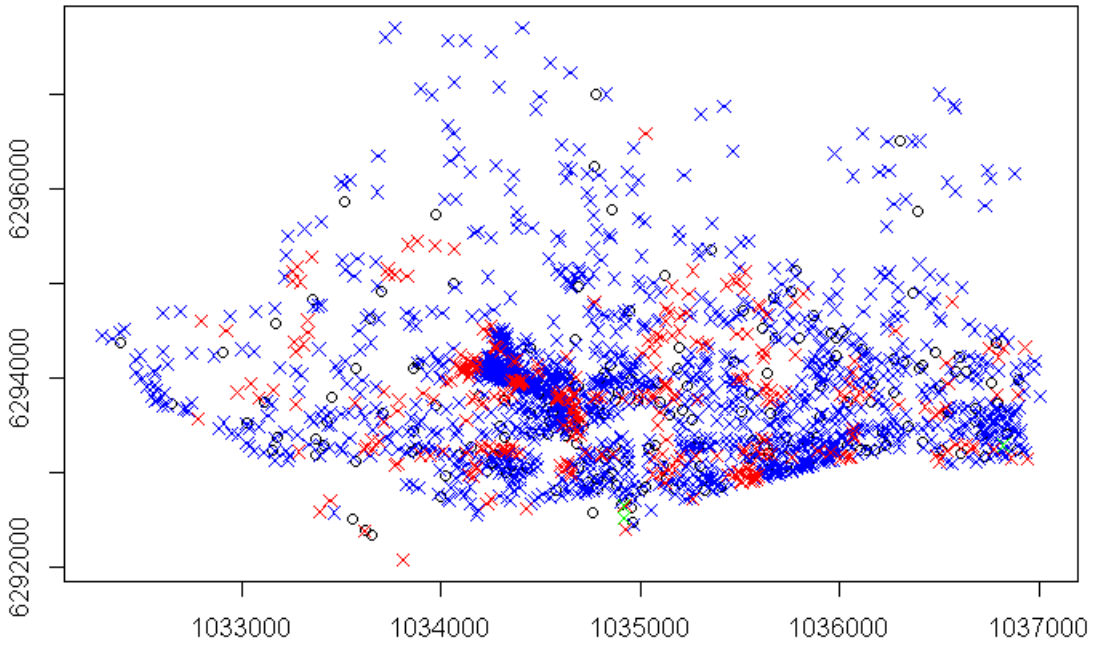
— 77 m pour la méthode *grille*
— 69 m pour la méthode *cascade*

Pour étudier la dispersion des distances obtenue par chaque méthode, on marque par des croix, sur une carte de Cagnes-sur-Mer (figure 2.4), chaque *étage* de résidences non principales :

— En bleu, l'*étage* est affecté au même secteur, quelle que soit la méthode retenue.
— En rouge, l'*étage* est affecté à deux secteurs différents. Le barycentre du secteur obtenu avec la méthode *grille* est plus éloigné de l'*étage* que celui obtenu par la méthode *cascade* . Par ailleurs, les *étages* non affectés par la méthode *grille* sont également repérés en rouge [20].
— En vert, l'*étage* est affecté à deux secteurs différents. Le barycentre du secteur obtenu avec la méthode *cascade* est plus éloigné de l'*étage* que celui obtenu par la méthode *grille*.

Les barycentres des différents secteurs sont repérés par des cercles.

FIGURE 2.4 – Différences dans le rattachement des résidences non principales entre la méthode *grille* et la méthode *cascade* pour Cagnes-sur-Mer



On constate en premier lieu que les méthodes conduisent souvent au même résultat

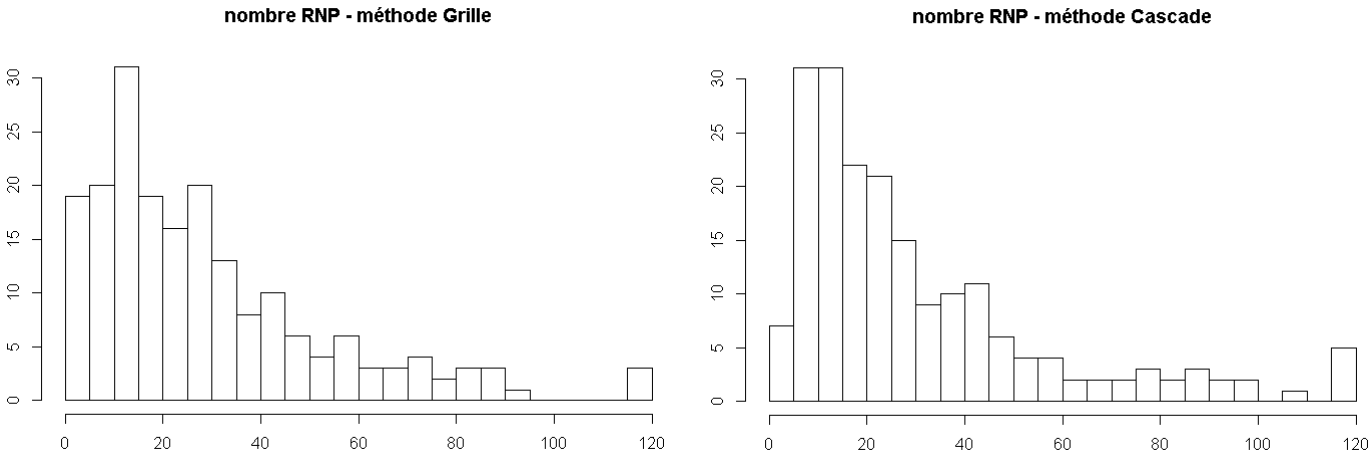
19. Ces deux méthodes ont été comparées pour d'autres UC avec des conclusions similaires à celles induites par l'UC de Cagnes-sur-Mer. Les comparaisons n'ont pas été étendues à l'ensemble des UC afin de limiter les temps de calcul.

20. On considère que si l'affectation de ces *étages* était forcée, elle conduirait à une distance plus importante que celle obtenue avec la méthode *cascade*.

(croix bleues). On observe toutefois de nombreuses croix rouges, confirmant que la méthode *cascade* attribue des secteurs au barycentre souvent moins éloigné de l'*étage* étudié que la méthode *grille*[21]. Enfin, la quasi-absence de croix vertes est certainement l'élément le plus concluant : pour seulement trois étages, la méthode *grille* donne un meilleur résultat que la méthode *cascade*.

Si la méthode *cascade* permet d'affecter les *étages* à des secteurs a priori plus proches, l'intégration du seuil avait pour objectif de diminuer la dispersion de la taille des secteurs en nombre de logements, aux dépens de leur étendue. L'histogramme 2.5 permet d'étudier cette dispersion pour chacune des méthodes.

FIGURE 2.5 – Histogramme des nombres de RNP rattachées par secteur au seuil 120 par la méthode *grille* et la méthode *cascade* pour Cagnes-sur-Mer



Les résultats observés sont assez proches. La méthode *grille* semble répartir un peu mieux les étages, notamment dans les secteurs obtenant peu de résidences non principales (0 à 5 résidences non principales). Ainsi, moins de secteurs se voient attribuer plus de 80 RNP. Il faut cependant relativiser ce deuxième constat, car 5 % des résidences non principales (éloignées, par construction, des barycentres des secteurs non saturés) ne sont pas affectées.

Au final, on choisit la méthode *cascade* : elle présente l'avantage d'affecter toutes les résidences non principales sans dégrader de manière rédhibitoire la compacité des secteurs. Par ailleurs, l'algorithme est beaucoup plus efficace, ce qui est non négligeable dans un contexte temporel contraint[22].

Un dernier élément est à définir : le nombre maximal de résidences non principales par secteur.

21. Parmi ces croix, 305 RNP correspondent à des *étages* non affectés par la méthode *grille*.

22. Petite UC (1 500 RNP) : environ 3 minutes pour la méthode *cascade*, et 4 minutes pour la méthode *grille* ; Grande UC (6 000 RNP) : environ 16 minutes pour la méthode *cascade*, et 40 minutes pour la méthode *grille*. Très grande UC (Nice avec 51 000 RNP) : environ 16 heures pour la méthode *cascade*, et plusieurs jours pour la méthode *grille*.

2.1.4 120 résidences non principales par secteur

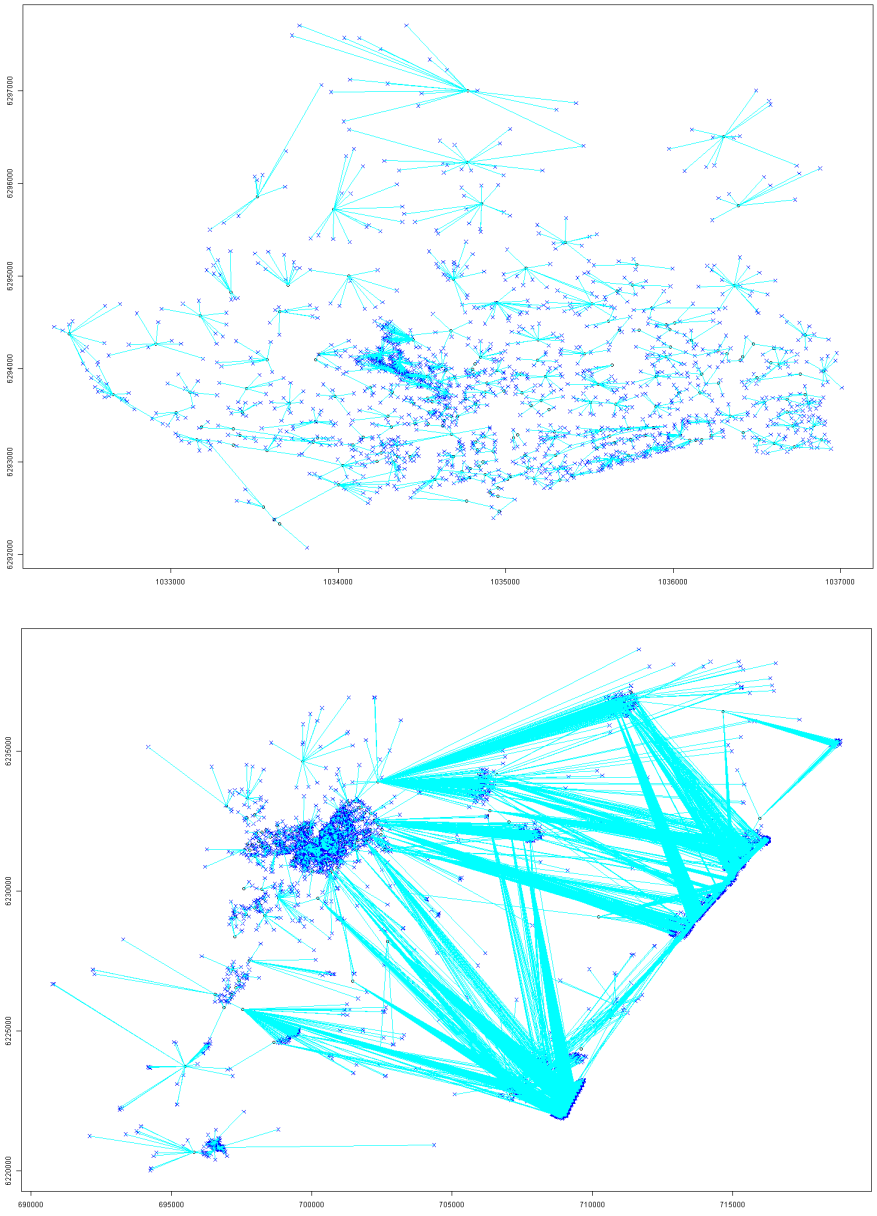
L'objectif recherché par la mise en place du seuil maximal de résidences non principales par secteur est d'éviter, après affectation, des secteurs trop déséquilibrés en nombre de logements. Dans l'absolu, on pourrait définir le seuil comme le rapport du nombre de RNP par le nombre de secteurs. Cependant, trois éléments rendent cela impossible. Tout d'abord, les RNP situées au même étage qu'une résidence principale n'entrent pas dans l'algorithme d'affectation [23] et peuvent déjà, dans certains cas, faire dépasser le seuil fixé. Ensuite, si le seuil concerne un nombre de logements, le rattachement se fait par *étages* et ne permet pas toujours une affectation suffisamment fine pour satisfaire un seuil trop bas. Enfin, abaisser le seuil conduit à augmenter les distances entre résidences non principales et barycentre du secteur d'affectation, ce qui peut nuire à la qualité de la collecte en raison de l'augmentation du temps de déplacement des enquêteurs.

On choisit de fixer le seuil à 120 RNP. Ce choix arbitraire correspond au nombre cible de résidences principales par secteur déterminé lors de leur construction [24] (cible de 20 résidences principales par grappe, 6 grappes par secteur). Si ce seuil permet d'obtenir des rattachements à des grappes proches pour la plupart des RNP dans la majorité des UC, la spécificité géographique de quelques communes de littoral ou de stations de ski demande de revoir sa valeur. En effet, lorsque des quartiers entiers ne sont pratiquement composés que de résidences secondaires, un tel seuil conduit à rattacher certaines à des secteurs situés dans d'autres quartiers.

Deux exemples sont présentés ci-dessous afin d'illustrer les situations dans lesquelles ce seuil de 120 RNP s'avère judicieux et celles dans lesquelles ce seuil est dysfonctionnel. Ces exemples représentent les rattachements obtenus pour un seuil de 120 RNP par secteur, dans l'UC de Narbonne et ses alentours, que l'on compare à ceux de Cagnes-sur-Mer. Chaque trait de la figure 2.6 correspond au rattachement d'un *étage* de résidences non principales (croix bleu) au barycentre de son secteur d'affectation (rond noir). Le rattachement des RNP apparaît aux secteurs de Cagnes-sur-Mer conduit à affecter des RNP à des secteurs proches, ce qui est pleinement satisfaisant. Au contraire, au voisinage de Narbonne, les RNP sont rattachées à des secteurs éloignés, ce qui est peu souhaitable.

23. Elles sont directement affectées au secteur des résidences principales situées à l'étage.
24. Ainsi, un secteur possédera, a priori, plus de résidences principales que de RNP.

FIGURE 2.6 – Rattachement des RNP par secteur au seuil 120 pour Cagnes-sur-Mer (en haut) et pour l'UC de Narbonne (en bas) par la méthode *cascade*



On observe aisément que ce seuil conduit à regrouper nombre de logements du littoral à des secteurs du centre-ville. Deux solutions sont envisagées :

— finalement, ne pas rattacher les résidences non principales trop éloignées du barycentre de leur secteur d'affectation ;
— relever le seuil afin de permettre de rattacher ces logements à des secteurs plus proches (mais possédant alors plus de 120 RNP).

Afin de ne pas avoir de défaut de couverture, on s'autorise, en définitive, à relever le seuil jusqu'à 400 résidences non principales, ce qui permet de résoudre tous les cas de ces UC atypiques.

Toutes les résidences non principales sont à présent affectées à un secteur ; tous les logements d'un étage sont reliés au même secteur. Il reste à présent à distribuer les *étages*

de RNP à l'une des 6 grappes du secteur qui leur a été attribué.

2.1.5 Rattachement à l'une des 6 grappes du secteur

Comme pour le rattachement au secteur, on affecte tous les logements (principaux ou non) d'un *étage* à une même grappe, soit, quand il y a présence d'une résidence principale à cet étage, à la grappe de celle-ci [25]. Pour attribuer une grappe aux *étages* composés uniquement de RNP, les mêmes arbitrages que précédemment se posent :

— soit on préfère une taille homogène des différentes grappes d'un secteur (permettant des poids de sondage homogènes [26], préférable pour la précision des estimateurs), et on distribue au mieux les *étages* entre les 6 grappes ;
— soit on privilégie la proximité géographique (facilitant la collecte), quitte à obtenir des grappes avec beaucoup de résidences non principales (par construction, jusqu'à 120 dans les secteurs non atypiques).

Deux algorithmes ont été implémentés, privilégiant l'un ou l'autre des critères. Le premier algorithme consiste à trier les *étages* composés uniquement de RNP par ordre décroissant de nombre de logements, puis à distribuer chaque étage, grappe après grappe. Cette solution a doublé l'étendue de chaque grappe (chemin de 1,8 km en moyenne entre les résidences principales, contre 3,7 km en moyenne en incorporant les RNP). Par ailleurs, les grappes sont, par construction, enchevêtrées les unes dans les autres. Pour la seconde méthode, les barycentres des 6 grappes sont déterminés, puis les *étages* de résidences non principales sont reliés au plus proche des 6 barycentres. Les grappes obtenues sont plus compactes (2,4 km en moyenne), et très peu enchevêtrées.

En définitive, l'arbitrage s'est porté en faveur des propriétés de compacité et de contiguïté des grappes, plutôt que de l'homogénéité des poids, et c'est le deuxième algorithme qui a été adopté.

Après ces différents traitements, chaque logement du champ de l'enquête Emploi, principal ou non, et issu de la source Fidéli 2016, est relié à une grappe, elle-même appartenant à un secteur. Cependant, le paysage urbain évolue chaque année, et des logements sont construits, d'autres détruits. Le paragraphe suivant décrit la méthode adoptée pour le rattachement des logements intégrant le champ de l'enquête lors du changement de millésime.

2.2 Rattachement des nouveaux logements

Chaque année, le parc de logements évolue, certains étant démolis, d'autre construits. Ces évolutions affectent la composition des grappes, car si la disparition de logements est *de facto* prise en compte, ignorer la création de logements reviendrait à exclure du champ

25. Pour rappel, lors de la construction initiale des grappes, toutes les résidences principales d'un même *étage* ont été rattachées à la même grappe (voir point 2.1.2 de la partie B).

26. L'argument de l'homogénéité des poids de sondage peut paraître surprenant, étant donné que jusqu'à présent, il n'a été question que de la mise en collecte de l'ensemble des logements d'une grappe sélectionnée. Ainsi, tous les logements des grappes mises en collecte semblent avoir le poids de leur grappe et donc un poids relativement homogène. En réalité, nous verrons en section 2.3 qu'un tirage d'étages est effectué dans les grappes comptant beaucoup de logements afin de limiter la charge de collecte des enquêteurs. Des grappes très volumineuses conduiront ainsi à une sélection de logements et donc à une dispersion des poids.

de l'enquête les ménages emménageant dans ces résidences, population probablement particulière face à l'emploi. Aussi, pour déterminer la méthode de rattachement des nouveaux logements aux grappes, on s'intéresse à l'évolution de la base entre 2016, millésime utilisé pour la création des grappes, et 2018, millésime le plus récent au moment de l'entrée en collecte de l'échantillon nouvellement tiré [27]. Entre ces deux éditions de Fidéli, 354 927 logements ont été démolis, et 996 652 nouveaux logements sont apparus, dont 68,8 % de résidences principales.

Le flux de logements concerné est assez faible par rapport à la population totale de logements (entre 2 et 3 % de logements créés, et 1 % de logements démolis). Par ailleurs, le rattachement de ces résidences aux grappes existantes peut difficilement s'effectuer sans modifier fortement le nombre de logements de la grappe ou étendre la surface de collecte [28]. Aussi, on décide d'opter pour un traitement simple consistant à rattacher chaque nouveau logement au secteur le plus proche (proximité du logement par rapport au barycentre), puis, au sein de ce secteur, à la grappe la plus proche.

En ne considérant plus que l'échantillon de l'EEC, on obtient 17 721 grappes dans 2944 secteurs, comprenant 555 994 logements répartis en 413 900 *étages*. On y compte 451 125 résidences principales, soit 81,13 % des logements. Les nouveaux logements sont au nombre de 14 677, soit 2,64 % des logements (répartis dans 9 355 étages , soit 2,26 %).

Les grappes ont en moyenne 25,11 logements dont 20,38 résidences principales. La distribution, observée dans le tableau 2.1 est assez resserrée, avec une médiane à 23 logements.

Table 2.1 – Distribution du nombre de logements par grappe après rattachement des nouveaux logements

	Min	P1	Q1	Q2	Q3	P99	Max
Nombre de logements	2	18	20	23	26	67	754

Les cas extrêmes correspondent à des grappes dans lesquelles de nombreux logements ont été détruits ou construits. Dans d'autres grappes, un nombre élevé de logements (jusqu'à 197 dans une grappe atypique à Toulouse) ont changé de statut générant un nombre de résidences principales très important. Ainsi, 95 grappes contiennent plus de 40 RP après cette mise à jour (soit 0,53 % des grappes). L'une d'elles contient 242 RP.

Ces extrêmes reflètent le fait que les grappes de l'échantillon, obtenues après ajout des résidences non principales et mises à jour annuelles, ne peuvent pas toutes être directement utilisées en collecte : un tirage de logements à enquêter au sein de ces grappes est nécessaire.

27. Une fois arrêtée, cette méthode sera appliquée à chaque nouveau millésime en partant du millésime précédent. Par exemple, à l'arrivée du millésime 2019, on repartira de la composition des grappes à l'issue de l'intégration des nouveaux logements du millésime 2018. On appliquera cette méthode afin d'intégrer dans ces grappes les nouveaux logements apparus dans le millésime 2019.

28. En effet, soit il s'agit de l'apparition d'une nouvelle maison, et la grappe affectée ne sera que peu modifiée, soit il s'agit de rattacher un immeuble, et dans ce cas, l'affectation d'un ou plusieurs *étages* aux différentes grappes entraînera obligatoirement une forte disparité en nombre de logements sur les grappes alentours.

2.3 Tirage de logements au sein des grappes de l'échantillon

Si les grappes ont été initialement construites pour contenir si possible 20 résidences principales, avec une tolérance entre 17 et 24, l'ajout des RNP (voir section 2.1) et la mise à jour annuelle des grappes (voir section 2.2) ont entraîné obligatoirement un éloignement de cette cible, conduisant à sélectionner au sein des grappes de l'échantillon, les logements effectivement collectés. Pour établir le plan de sondage mis en place, le service producteur a posé les contraintes devant permettre de réaliser la collecte tout en garantissant la qualité des résultats produits. Celles-ci sont présentées dans le point 2.3.1. Le plan de sondage alors défini est décrit dans le point 2.3.2. Il a abouti à un échantillon de logements mis en collecte dont les caractéristiques sont présentées en 2.3.3.1. La pondération des unités enquêtées établie en 2.3.3.2 permet de résumer les différents éléments ayant permis de produire cet échantillon complexe.

2.3.1 Les contraintes posées

La collecte de l'enquête Emploi en continu s'effectue dans des délais très contraints, obligeant de fixer des règles dans la définition de l'échantillon. C'est la raison principale de l'utilisation, dès 2009, d'un échantillon aréolaire, composé de grappes contenant un nombre restreint de logements obligatoirement proches et facilement repérables les uns des autres, afin de limiter toute déperdition de temps à l'enquêteur. Notamment, tous les logements d'un étage sont enquêtés durant la même période de 2 semaines et 2 jours. Si des premiers éléments cibles ont été définis au moment de l'élaboration de la base de sondage (voir 2.1.2 de la partie B), l'ajout des résidences non principales et la mise à jour annuelle des grappes tirées ne permettent pas de conserver en l'état les grappes entrant en collecte. Aussi, le service producteur des résultats de l'enquête, garant de la faisabilité de la collecte et de la qualité des indicateurs produits, a précisé les contraintes à appliquer pour sélectionner les logements enquêtés *in fine*.

L'objectif recherché est de ne pas détériorer les conditions actuelles de collecte, et ainsi de prendre pour cible les caractéristiques des échantillons utilisés entre 2011 et 2018. Les grappes mises en collecte durant cette période comptent 25,8 logements par grappe en moyenne, dont 4,4 résidences non principales, et moins de 10 % des grappes possèdent plus de 10 RNP.

Aussi, les contraintes fixées pour le tirage des logements à enquêter dans le cadre de l'EEC s'expriment ainsi :

- une cible de 20 résidences principales par grappe ;
- 4 à 6 résidences non principales par grappe en moyenne ;
- 95 % des grappes doivent compter entre 18 et 30 logements ;
- tous les logements d'un étage (ou aucun) doivent être sélectionnés simultanément dans l'échantillon.

Ces contraintes ont été déclinées de la manière suivante pour permettre la sélection des logements au sein des grappes :

- Si la grappe contient moins de 24 résidences principales, les sélectionner exhaustivement, sinon, en tirer 24 ;
- Si la grappe contient moins de 10 RNP, les sélectionner exhaustivement, sinon [29],

29. Il s'agit des seuils déjà utilisés depuis 2009 pour les résidences non principales.

  - entre 10 et 40, en sélectionner 10 ;
  - entre 41 et 100, en prendre le quart ;
  - plus de 100, se limiter à 25 RNP ;
- le tirage des logements à enquêter passe par un tirage d'étages [30], qui doit permettre de respecter les contraintes sur la distribution de taille des grappes tirées.

En particulier, les grappes ayant été construites afin de contenir autour de 20 résidences principales (cf. chapitre 2 de la partie B), le seuil de 10 résidences non principales enquêtées au maximum dans les grappes de taille inférieure 40 logements est fixé afin garantir que 95 % des grappes aient compter entre 18 et 30 logements.

2.3.2 Tirage stratifié d'étages

Les *étages* composant la base de sondage au sein d'une grappe peuvent être partagés en deux groupes :

- les *étages* contenant au moins une résidence principale
- les *étages* ne contenant que des RNP.

Pour sélectionner les logements envoyés en collecte, différentes méthodes de sondage auraient pu être adoptées (tirage Poissonnien ou tirage équilibré notamment). Le choix s'est porté sur un tirage aléatoire simple stratifié sur les deux groupes d'étages [31].

2.3.2.1 Première phase : sélection d'*étages* pour atteindre la cible

Les contraintes énoncées précédemment permettent de définir les probabilités d'inclusion des étages. On se place dans une grappe G de l'échantillon. On définit tout d'abord la probabilité de sélection π_A des *étages* contenant des résidences principales (strate A) qui se présente comme la contrainte prioritaire, d'où découlera celle des *étages* ne contenant que des résidences non principales (strate B), notée π_B. Dans une grappe entrant en collecte et possédant N_{RP} résidences principales, on pose :

$$\pi_A = \begin{cases} 1 \text{ si } N_{RP} \leq 24 \\ \\ \frac{24}{N_{RP}} \text{ sinon} \end{cases}$$

Ces *étages* peuvent également contenir des résidences non principales. En moyenne, la sélection de ces *étages* « rapporte »

$$n^A_{RNP} = \pi_A N^A_{RNP}$$

résidences non principales, où N^A_{RNP} est le nombre total de RNP dans les *étages* contenant également au moins une résidence principale.

30. Ce tirage d'étages correspond à un tirage de « grappes » de logements, le mot « grappe » se référant dans ce cas précis au sens usuel de ce mot en théorie des sondages et non aux grappes de l'enquête Emploi.

31. Le tirage poissonien ne permet pas de maîtriser la taille de l'échantillon (il peut y avoir, après tirage, des grappes finalement quasi vides, ou, à l'inverse contenant tous les logements de la grappe initiale). Quant au tirage équilibré, il rend improbable certaines combinaisons d'étages, ce qui peut s'avérer problématique compte-tenu de la faible taille des échantillons d'étages à sélectionner et du faible nombre d'étages participant au tirage au sein d'une grappe.

On peut alors déterminer le taux de sondage à appliquer à la strate B. Cette probabilité d'inclusion doit :

— respecter en moyenne la cible[32] n_{RNP} de nombre de RNP dans l'échantillon, définie à la fin du point 2.3.1 ;
— garantir qu'au minimum un *étage* de cette strate soit sélectionné, pour éviter tout biais (ce qui revient à dire qu'elle ne doit pas être inférieur à $\frac{1}{N^B_{RNP}}$ où N^B_{RNP} est le nombre total de RNP dans les *étages* sans résidence principale).

Ce qui se traduit ainsi :

$$\pi_B = \max\left(\frac{n_{RNP} - n^A_{RNP}}{N^B_{RNP}}, \frac{1}{N^B_{RNP}}\right)$$

Les allocations n_A et n_B de nombre d'*étages* à tirer de chacune des deux strates sont alors déterminées en sommant les probabilités d'inclusion de chaque *étage* de la strate (respectivement peuplées par N_A et N_B étages)[33]. Elles sont arrondies à l'entier inférieur. Pour la strate B, on impose, de plus, de tirer au moins un *étage* pour éviter tout biais. Ainsi :

$n_A = \lfloor \pi_A N_A \rfloor$ et $n_B = \max(\lfloor \pi_B N_B \rfloor, 1)$.

Un tirage aléatoire simple est alors effectué dans chaque strate (de chaque grappe) selon les allocations ici définies. Ce tirage peut toutefois conduire à un échantillon de plus de 40 logements dans certaines grappes, rendant la collecte compliquée. S'il n'est pas toujours possible de diminuer ce nombre (un étage contenant plus de 40 logements a été tiré), dans la plupart des cas, une seconde phase de tirage est mise en œuvre pour limiter à 40 logements la taille des grappes réellement mises en collecte.

2.3.2.2 Seconde phase : Diminution du nombre de logements interrogés

On se place dans une grappe où les n_E *étages* sélectionnés conduisent à un nombre de logements l supérieur à 40. On cherche à enlever x_E étages, contenant l' logements, de sorte que le nombre de logements $l - l'$ dans les $n_E - x_E$ *étages* restant dans l'échantillon soit inférieur à 40. Les *étages* sélectionnés initialement se répartissent :

— entre la strate A, dans laquelle n_A *étages* ont été sélectionnés parmi les N_A,
— et la strate B, dans laquelle n_B *étages* ont été sélectionnés parmi les N_B.

On cherche à déterminer les nombres x_A et x_B d'*étages* à enlever dans chacune des strates. Ceux-ci vont être déterminés par minimisation d'un objectif sous contraintes. Ces dernières sont :

— laisser au moins un *étage* de chaque strate dans l'échantillon : $0 \leq x_A < n_A$ et $0 \leq x_B < n_B$
— enlever suffisamment d'*étages* dans chaque strate pour que le nombre de logements restant passe sous la barre des 40 logements.

32. On entend par cible le fait de sélectionner 10 RNP quand la grappe compte entre 10 et 40 RNP, de sélectionner un quart des RNP quand la grappe en compte entre 41 et 100 et de tirer 25 RNP quand elle en compte plus de 100.

33. La formule de l'allocation n_A précisée ci-dessous ne semble pas garantir qu'au moins un étage de la strate A soit tiré dans chaque grappe concernée par ce tirage supplémentaire d'étages. Néanmoins, en pratique, pour les grappes entrant en collecte entre le T4 2019 et le T2 2020, au moins un étage de la strate A a systématiquement été mis en collecte pour chaque grappe.

Si la contrainte sur le nombre de logements est difficile à vérifier systématiquement, elle peut néanmoins se formuler en termes d'espérance : on cherche les valeurs de x_A et x_B qui, en moyenne, permettent d'enlever le surplus de logements dans la grappe. En notant e le surplus de logements dans la grappe à l'issue de la première phase ($e = l - 40$), m_A (respectivement m_B) le nombre moyen de logements dans les *étages* de la strate A (respectivement la strate B), la contrainte s'écrit alors :

$$x_A m_A + x_B m_B = e$$

.

Le plan de sondage adopté va chercher à respecter cette contrainte en minimisant la variance liée à ce tirage de seconde phase [34]. Conditionnellement à l'échantillon de première phase, cette variance est celle générée par un sondage aléatoire simple stratifié de $n_A - x_A$ *étages* dans la strate A (resp. $n_B - x_B$ dans la strate B) :

$$V = n_A^2 \left(1 - \frac{n_A - x_A}{n_A}\right) \frac{s_A^2}{n_A - x_A} + n_B^2 \left(1 - \frac{n_B - x_B}{n_B}\right) \frac{s_B^2}{n_B - x_B}$$

sous la contrainte $x_A m_A + x_B m_B = e$. En appliquant l'optimisation de Neyman [35], on obtient les allocations suivantes [36] :

$$\begin{cases} x_A = n_A - \frac{n_A}{\sqrt{m_A}} \frac{n_A m_A + n_B m_B - e}{n_A \sqrt{m_A} + n_B \sqrt{m_B}} \\ \\ x_B = n_B - \frac{n_B}{\sqrt{m_B}} \frac{n_A m_A + n_B m_B - e}{n_A \sqrt{m_A} + n_B \sqrt{m_B}} \end{cases}$$

On choisit d'arrondir ces allocations théoriques à l'entier immédiatement supérieur [37], si au moins un *étage* dans chaque strate est conservé :

$$x'_A = \begin{cases} 0 \text{ si } n_A = 1 \\ \\ min(\lceil x_A \rceil, n_A - 1) \text{ sinon} \end{cases}$$

$$x'_B = \begin{cases} 0 \text{ si } n_B = 1 \\ \\ min(\lceil x_B \rceil, n_B - 1) \text{ sinon} \end{cases}$$

Après suppression dans les deux strates du nombre d'étages (sélectionnés par tirages aléatoires simple) défini par les allocations ainsi déterminées, la taille de la grappe finalement envoyée en collecte, respecte, en moyenne, le seuil de 40 logements.

Les différents choix méthodologiques ont pu être mis en œuvre sur les premiers échantillons de l'EEC, et permettent d'évaluer leur pertinence. Des premiers résultats sont présentés dans le paragraphe suivant.

34. Un tirage réjectif a été étudié et même appliqué pour le troisième trimestre de 2019, mais n'a pas donné satisfaction, car il conduisait à privilégier la sélection des petits étages, rendant parfois impossible celle de grands *étages* (générant ainsi un biais de sélection). Par ailleurs, la variance d'un tel tirage n'étant pas connu, de lourdes simulations, trop coûteuses en temps, auraient été nécessaires.

35. Cette optimisation propose un jeu d'allocation minimisant une variance sous contrainte, voir Le Gleut, 2017.

36. Faute de savoir quelle variable Y retenir pour cette optimisation, les allocations x_A et x_B ont été calculées en supposant les dispersions s_A^2 et s_B^2 identiques pour les deux strates.

37. Cela conduit à retirer plus de logements, et permet de diminuer la probabilité que l'échantillon dans cette grappe contienne plus de 40 logements.

2.3.3 Résultats pour les tirages du T4 2019 au T2 2020

2.3.3.1 Nombre de logements des grappes obtenues

Au T3 2019, les premières grappes sélectionnées à partir de la méthode décrite dans ce document sont entrées en collecte [38]. Les logements entrants au dernier trimestre 2019 et aux deux premiers trimestres 2020 ont également été sélectionnés. Ces différents échantillons d'*étages* ont ainsi pu être analysés en regard de ceux sur le terrain de 2011 au T2 2019, sélectionnés par la méthode précédente. Le tableau 2.2 permet de comparer, selon chaque méthode [39], la dispersion des grappes en fonction du nombre de logements à collecter qu'elles contiennent.

TABLE 2.2 – Comparaison des nombres de logements des grappes en collecte de l'ancienne EEC et des tirages pour la nouvelle EEC Nautile

Nombre de logements par grappe (%)	Ancienne EEC	Tirage T4 2019 + T1 2020 + T2 2020
Moins de 4	0,1	0,1
Entre 5 et 9	0,1	0,1
Entre 10 et 14	0,2	0,4
Entre 15 et 19	3,7	5,4
Entre 20 et 24	43,7	62,1
Entre 25 et 26	13,7	10,2
Entre 27 et 29	15,7	11,5
Entre 30 et 34	16,3	8,3
Entre 35 et 39	5,6	1,4
Entre 40 et 44	0,8	0,4
Plus de 44	0,2	0,1
Total	100,0 [16 595 grappes]	100,0 [1 846 grappes]
Nombre moyen de RP	21,4	19,9
Nombre moyen de RNP	4,4	3,6
Pourcentile P90	33	30

Note de lecture : Dans l'ancien échantillon EEC, entre 20 et 24 logements sont enquêtés dans 43,7% des grappes mises en collecte. Entre 20 et 24 logements sont enquêtés dans 62,1% des grappes entrant en collecte entre le T4 2019 et le T2 2020. Ces dernières sont issues du nouvel échantillon.

Les résultats obtenus par la nouvelle méthode sont tout à fait satisfaisants, puisqu'elle permet d'obtenir des grappes en moyenne plus petites, et surtout plus nombreuses autour des 24 logements cibles (20 résidences principales et 4 RNP). Le nombre de grappes dépassant les 40 logements est, quant à lui, divisé par deux. Autre élément capital, l'étendue des grappes obtenues doit être suffisamment petite pour permettre à l'enquêteur de réaliser la collecte. Pour cela, on trace pour chaque grappe de chaque échantillon un chemin passant une et une seule fois par tous les logements, puis on mesure la taille du chemin. Le tableau suivant présente différentes caractéristiques de ces chemins pour chaque échantillon, en France métropolitaine, et dans quelques régions, plus ou moins denses.

38. Les grappes entrées en collecte au T3 2019 ayant fait l'objet d'une sélection d'étages par un tirage réjectif non conservé pour les autres trimestres, les propriétés de ce tirage ne sont pas détaillées dans les tableaux suivants.

39. Pour rappel, la méthode de tirage des *étages* au sein d'une grappe a évolué à partir du T4 2019, passant d'un tirage réjectif à un tirage aléatoire simple.

TABLE 2.3 – Comparaison avec Métric de la longueur des grappes de l'ancienne EEC et de la nouvelle EEC Nautile

Longueur (en km) [et nombre grappes]	France métropolitaine	Île-de-France	Aquitaine	Corse
moyenne T4 2018 à T2 2019	2,9 [1607]	0,8 [285]	4,2 [80]	10,2 [9]
moyenne T4 2019 Nautile	2,3 [618]	0,4 [119]	3,2 [30]	8,2 [7]
moyenne T1 2020 Nautile	2,4 [621]	0,8 [112]	3,3 [30]	5,6 [6]
moyenne T2 2020 Nautile	2,1 [607]	0,5 [116]	2,8 [28]	3,9 [8]
médiane T4 2018 à T2 2019	1,4 [1607]	0,8 [285]	2,4 [80]	0,4 [9]
médiane T4 2019 Nautile	0,9 [618]	0,0 [119]	1,3 [30]	3,4 [7]
médiane T1 2020 Nautile	1,0 [621]	0,2 [112]	1,9 [30]	0,1 [6]
médiane T2 2020 Nautile	0,8 [607]	0,2 [116]	1,5 [28]	2,0 [8]

Note de lecture : Pour les 80 grappes d'Aquitaine entrées en collecte entre le T4 2018 et le T2 2019 et issues de l'ancien échantillon EEC, le chemin pour parcourir l'ensemble des logements interrogés dans la grappe est en moyenne long de 4,2 km. Pour les 30 grappes d'Aquitaine entrées en collecte au T4 2019 et issues du nouvel échantillon EEC, ce chemin est en moyenne long de 3,2 km.

Là encore, les grappes nouvellement obtenues sont très satisfaisantes puisqu'elles permettent, en moyenne, et en médiane de diminuer la distance à parcourir pour passer par tous les logements. Ce résultat, constaté sur l'ensemble de la France métropolitaine, se retrouve quel que soit le profil de la région. S'il arrive ponctuellement, dans les zones urbaines, et pour certains trimestres, que les grappes les plus étendues soient issues de la méthode proposée ici [40], les échantillons des quatre derniers trimestres dans les zones rurales sont plus compactes.

Sur un aspect plus théorique, la dispersion des poids, à limiter pour améliorer la précision des indicateurs, est stable sur les 4 tirages, et comparable à celle des grappes sur le terrain entre le T3 2018 et le T2 2019 : le rapport de pourcentiles P99/P1 vaut 5,1 pour ces quatre trimestres, et entre 4,6 et 6,4 pour les tirages effectués dernièrement.

Notons enfin que les premiers chiffres produits à partir des échantillons tirés grâce à cette nouvelle méthodologie (grappes tirées pour le T3 2019) ont été publiés le 14 novembre 2019 [41]. En particulier, aucun problème n'a été constaté, tant au niveau de la collecte terrain, que sur les travaux méthodologiques pour la production des indicateurs (calage et redressement notamment).

Le dernier paragraphe résume les différentes étapes de sélection de l'échantillon à travers l'établissement des pondérations.

2.3.3.2 Pondérations

L'échantillon de l'EEC envoyé en collecte est le résultat de nombreuses étapes devant satisfaire des contraintes de collecte et de qualité statistique. Toutes ces étapes ont été construites afin de permettre une estimation sans biais des variables de populations.

40. Cela peut également provenir de la construction initiale de la grappe, qui autorisait de légères discontinuités dans la contiguïté des étages, pour approcher au maximum de la cible de 20 résidences principales. Ces « sauts » se sont avérés utiles particulièrement en zone urbaine où les *étages* contiennent plus de logements qu'en zone rurale.

41. Voir `https://www.insee.fr/fr/statistiques/4247277`

Pour ces estimations, il est nécessaire d'extrapoler les résultats obtenus sur les différents échantillons, à partir d'une variable de poids associé à chaque logement [42], produit des différents éléments suivants, caractérisant chaque étape de construction des grappes :

- du poids des unités de coordination obtenu par application de la méthode généralisée du partage des poids (cf. 4.2.2 de la partie C) ;
- de l'inverse de la probabilité de sélection utilisée pour le tirage des secteurs conditionnellement aux unités de coordination (cf. 5.2 de la partie C) ;
- d'un facteur égal au nombre de grappes du secteur divisé par 6 . Ce facteur permet de tenir compte du fait que, dans les secteurs contenant 7 grappes, seules 6 grappes, sélectionnées aléatoirement par un sondage aléatoire simple, sont mises en collecte (cf. 1.3 de la partie D) ;
- d'un facteur égal à 6, tenant compte de l'affectation du secteur à l'un des six trimestres d'entrée (cf. 1.2 de la partie D) ;
- d'un facteur égal à 6, tenant compte de la sélection de la grappe interrogée dans le secteur (cf. 1.3 de la partie D) ;
- du poids de sélection des *étages* dans les grappes en première phase, après mise à jour annuelle de l'échantillon (valant 1 pour les grappes respectant les contraintes de taille), cf. 2.3.2.1 de la partie D ;
- du poids issu de la restriction du nombre d'*étages* dans les grappes ayant encore plus de 40 logements (1 sinon), cf. 2.3.2.2 de la partie D.

Au final, le poids w_{lgt} d'un logement s'écrit :

$$w_{lgt} = 6 w_{uc} \frac{N^{uc}_{sect}}{n^{uc}_{sect}} n^{sect}_{g} \frac{N^{j}_{etage}}{n^{j}_{etage} - x'^{j}_{etage}} \tag{2.1}$$

où w_{uc} est le poids de l'unité de coordination du logement, N^{uc}_{sect} le nombre total de secteurs dans l'unité de coordination, n^{uc}_{sect} le nombre de secteurs tirés dans l'unité de coordination, n^{sect}_{g} le nombre de grappes dans le secteur, N^{j}_{etage} le nombre total d'étages dans la strate $j = A$ ou B de la grappe et $n^{j}_{etage} - x'^{j}_{etage}$ le nombre d'étages finalement collectés dans cette strate (pouvant être égal à N^{j}_{etage}) [43].

La complexité du tirage de l'EEC est avant tout liée aux problématiques de collecte, devant être réalisée dans un délai contraint. Les autres enquêtes auprès des ménages, sélectionnées dans l'échantillon-maître, ne rencontrent habituellement pas ces difficultés. Aussi, les méthodes de sélection au deuxième degré posent, a priori, moins de questions. Cela étant, plusieurs plans de sondages peuvent être envisagés. Le chapitre suivant propose de mesurer l'intérêt de privilégier un tirage systématique à un tirage équilibré, pourtant considéré comme préférable habituellement.

42. Le poids des logements est égal au poids de leur *étage*, puisque l'on interroge tous les logements d'un *étage* tiré (principe du tirage par grappe).

43. Le terme $\frac{N^{j}_{etage}}{n^{j}_{etage} - x'^{j}_{etage}}$ regroupe les probabilités d'inclusion pour les tirages des paragraphes 2.3.2.1 et 2.3.2.2. Ce terme vaut 1 dès que la grappe contient 24 résidences principales ou moins et 10 résidences non principales ou moins.

Chapitre 3

La méthode d'échantillonnage pour le second degré des enquêtes auprès des ménages

L'échantillon-maître issu d'un tirage d'UP spatialement équilibré a été constitué afin de répondre qualitativement à la plupart des enquêtes auprès des ménages (cf. chapitre 2 de la partie C). L'utilisation d'un EM est liée à la ressource enquêteur, et ne peut être spécifique à une enquête au détriment d'une autre. Les caractéristiques du plan de sondage, lié à une enquête particulière sont donc intégrées au second degré de tirage, qui détermine les logements ou individus effectivement collectés. Plusieurs plans de sondage peuvent être imaginés pour améliorer la précision de l'enquête en question. Le premier paragraphe de cette partie précise la raison qui fait hésiter entre tirage systématique et tirage équilibré. Le deuxième paragraphe présente la méthodologie utilisée pour évaluer comparativement ces deux méthodes de tirage. Il ressort de cette évaluation qu'il est préférable d'utiliser au second degré de tirage un sondage systématique, en choisissant les variables de tri les plus pertinentes selon l'enquête considérée.

Échantillonner des logements ou des individus : le dilemme du tirage de second degré

Le logement est l'unité historique d'échantillonnage à l'Insee. Une des nouveautés apportées par le répertoire Fidéli, désormais utilisé à l'Insee comme la base de sondage principale pour les enquêtes métropolitaines, est de permettre également l'échantillonnage d'individus (voir section 2.1 de la partie A). Une fois les unités primaires sélectionnées se pose la question de l'unité de tirage au second degré. Le choix de l'unité de tirage dépend essentiellement de quatre facteurs : l'unité d'observation, le mode de collecte, le protocole de collecte et les domaines de diffusion de l'enquête.

Lorsque le plan de sondage inclut deux degrés de tirage, le mode de collecte est presque toujours le face-à-face. Pour les enquêtes en face-à-face, dans le cas d'un tirage de logements, le protocole de collecte consiste toujours à enquêter les individus qui occupent le logement tiré au moment de la collecte. En cas de déménagement récent, ceux-ci peuvent ne pas être les mêmes que les derniers occupants connus dans le logement

dans les bases administratives. Dans les immeubles, cela pose alors des difficultés de repérage aux enquêteurs qui doivent identifier le logement échantillonné sans connaître le nom des nouveaux occupants. Dans le cas d'un tirage d'individus, il s'agit d'interroger l'individu échantillonné, quand bien même ce dernier aurait déménagé. L'enquêteur doit ainsi trouver la nouvelle adresse de l'individu. Lorsque ce dernier a déménagé loin de son ancien logement, soit un autre enquêteur récupère la responsabilité d'interroger l'individu, soit cet individu n'est pas interrogé car la réorganisation des unités affectées aux enquêteurs est trop coûteuse. Généralement, la complexité de l'organisation liée aux déménagements implique la non-enquête des individus ayant déménagé à une adresse trop éloignée de celle connue dans la base de sondage. Le tirage de logements présente ainsi un avantage majeur en face-à-face : malgré les difficultés de repérage que peuvent engendrer les déménagements, les logements à enquêter sont ceux échantillonnés et ne changent pas d'adresse. Il est donc certain que les logements attribués pour la collecte à un enquêteur devront être interrogés par cet enquêteur, ce qui est bien plus simple en termes d'organisation pour la collecte. Par conséquent, dans les collectes en face-à-face, le logement est l'unité de collecte la plus naturelle.

Néanmoins, dans un certain nombre d'enquêtes, l'objectif est de collecter des informations sur les individus et non sur les ménages ou les logements. Pour éviter un effet grappe au sein du ménage et pour alléger la charge de collecte sur le ménage, il est préférable d'interroger un seul individu dans le logement. Dans le cas d'un tirage d'individus, l'individu à enquêter est celui échantillonné. On peut donc réaliser le tirage afin d'optimiser la dispersion des pondérations des individus tirés, ce qui permet de diminuer la variance d'estimation. Au contraire, dans le cas d'un tirage de logements, lorsqu'on cherche à interroger un unique individu dans le logement, celui-ci est sélectionné aléatoirement parmi les occupants du logement au moment de l'enquête (tirage « kish »). Cela se traduit par un degré de tirage supplémentaire et donc par une augmentation de la variance. Ainsi, pour diminuer la variance dans une enquête qui interroge un seul individu par logement et qui diffuse sur des variables agrégées sur les individus, l'échantillonnage d'individus au second degré est idéal. Par contre, il est préférable d'échantillonner des logements lorsque des informations sur le logement ou le ménage sont recueillies dans l'objectif d'une diffusion sur les logements ou les ménages.

Enfin, lorsqu'une enquête s'intéresse à une population rare, un profil donné d'individus est ciblé. Il n'est pas envisageable d'échantillonner des logements pour ensuite interroger une personne sélectionnée aléatoirement dans ce logement. En effet, cette dernière peut ne pas appartenir à la population d'intérêt de l'enquête. Dans les enquêtes dans lesquelles la principale population d'intérêt est minoritaire dans la population (par exemple, les immigrés dans l'enquête Trajectoires et Origines, les personnes en situation de handicap ou de perte d'autonomie dans l'enquête Autonomie), la seule unité d'échantillonnage envisageable est donc l'individu.

Pour résumer, le choix de l'unité d'échantillonnage au second degré est évident dans deux situations. Si l'enquête diffuse sur les logements ou les ménages, dans le cas d'une collecte en face-à-face, le tirage de logements s'impose. C'est par exemple le cas de l'enquête Statistiques sur les Revenus et les Conditions de Vie de l'Insee ou de l'enquête Logement. Si l'enquête diffuse sur une population rare en interrogeant une personne

par logement, alors seul le tirage d'individus est viable, comme dans le cas de l'enquête Trajectoires et Origines ou de l'enquête Autonomie. Mais, dans le cas où une enquête vise à diffuser des indicateurs sur les individus en ayant interrogé un individu par logement et sans cibler une population rare, il est possible de tirer directement des individus ou bien de tirer des logements dans la base de sondage puis de réaliser un tirage kish une fois les individus occupant le logement recensés dans l'enquête. La divergence entre l'enquête Formation tout au Long de la Vie qui échantillonne des logements et l'enquête Piaac sur l'évaluation des compétences des adultes pour laquelle sont échantillonnés des individus illustre bien ce dilemme.

Ces enjeux sont amenés à évoluer dans les prochaines années pour les tirages à deux degrés. Le développement des enquêtes multimodes pourra conduire à sélectionner des unités primaires pour des enquêtes qui ne se déroulent pas exclusivement en face-à-face. Par exemple, il sera possible de réaliser des tirages à deux degrés pour concentrer la répartition géographique des individus ou des logements non-répondants à un mode de collecte auto-administré et qui seront ensuite enquêtés en face-à-face. Le choix d'échantillonner des logements ou des individus dépendra alors également du protocole de collecte retenu, puisqu'il est plus complexe d'identifier les occupants actuels d'un logement dans un mode auto-administré que les anciens occupants, au contraire du face-à-face.

Lorsque d'autres modes de collecte (Internet, papier ou téléphone) sont utilisés, il n'est pas nécessaire de procéder à un tirage de zones de collecte, puisqu'il n'y a aucun déplacement d'enquêteurs. On peut ainsi tirer les individus ou les logements sur l'ensemble de territoire. On parle alors de tirage à 1 degré. Pour contacter les unités tirées, on dispose de coordonnées postales, mail et téléphoniques des individus. Il est ainsi plus facile de suivre des individus que des logements, puisqu'en cas de déménagement, on ne dispose pas des coordonnées des nouveaux occupants des logements. Échantillonner des individus présente alors l'avantage de pouvoir interroger des individus dont les numéros de portable et adresses mail sont stables, ce qui facilite le suivi des déménagements lorsque la collecte se déroule par Internet ou téléphone. Le tirage de logements pose plus de questions : doit-on interroger les occupants actuels du logement, dont on ne dispose pas des coordonnées s'ils ont emménagé à une date ultérieure de la date à laquelle se réfère la base de sondage ? Doit-on interroger le ménage qui vivait dans le logement au moment du tirage ? Le cas échéant, comment prendre en compte correctement le fait que les ménages ne sont pas des entités stables dans le temps et que leurs contours varient au gré des événements personnels de la vie des individus ? Ces questions, qui se posent pour des tirages à 1 degré, sont très différentes des enjeux des tirages à 2 degrés. Le choix d'un plan de sondage pour une enquête multimode combine généralement les enjeux des tirages à 1 degré et ceux des tirages à 2 degrés.

3.1 Tirage systématique ou tirage (spatialement) équilibré

L'Insee utilise depuis un certain nombre d'années, au second degré, pour les enquêtes auprès des ménages, le tirage systématique (voir encadré). Il présente l'avantage

d'être simple d'implémentation, rapide d'exécution, et prend en entrée une information a priori corrélée aux indicateurs de l'enquête.

D'autres méthodes, plus complexes, mais présentant des résultats habituellement plus précis, sont classiquement utilisées quand on dispose d'une base de sondage riche en information auxiliaire. C'est notamment le cas du tirage équilibré et du tirage spatialement équilibré (cf. section 1.2 de la partie C). Si ces tirages ont apporté des résultats très satisfaisants pour le tirage d'enquêtes à un degré (et également pour le tirage du premier degré menant à l'EM), ils sont d'autant plus efficaces que les nombres d'unités à tirer dans la population sont grands par rapport au nombre de variables d'équilibrage. Or, dans le cadre des enquêtes ménages, le second degré de tirage ne concerne que quelques dizaines de logements par UP, légèrement plus dans les UP exhaustives (cf. point 2.3.2 de la partie C). L'implémentation de telles méthodes est par ailleurs plus complexe, aussi est-il pertinent d'estimer le gain apporté par celles-ci (tirage équilibré et spatialement équilibré) par rapport à un tirage systématique, particulièrement dans ces UP exhaustives.

Le tirage systématique

Le tirage systématique à probabilités égales est un algorithme de tirage couramment utilisé pour le tirage d'enquêtes auprès des ménages. Cet algorithme utilise une base de sondage préalablement triée sur des variables auxiliaires. Il permet une stratification implicite sur ces variables auxiliaires, c'est-à-dire que l'échantillon obtenu a une structure proche de celle de la base de sondage pour les variables utilisées pour le tri.

Concrètement, l'algorithme consiste à définir un aléa entre 0 et 1, à calculer les probabilités d'inclusion cumulées des unités de la base de sondage, puis à sélectionner les unités pour lesquelles les probabilités cumulées encadrent l'aléa incrémenté d'un nombre entier. Un exemple est donné en figure 3.1.

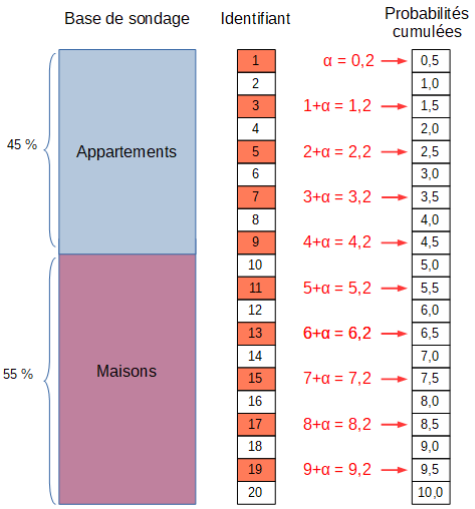
Dans cet exemple, 20 logements composent la base de sondage et 10 logements sont tirés à probabilités égales $(0,5)$. 11 sont des maisons et 9 sont des appartements. La base de sondage est triée par type de logement, préalablement au tirage systématique. L'aléa du tirage systématique est fixé à $0,2$. L'unité 1 est donc sélectionnée car $0,2$ se situe entre 0 et $0,5$. On incrémente ensuite l'aléa de 1, ce qui implique que l'unité 3 est sélectionnée car $1,2$ se situe entre 1 et $1,5$, et ainsi de suite. Au final, les unités 1, 3, 5, 7, 9, 11, 13, 15, 17 et 19 sont ainsi sélectionnées.

L'échantillon est donc composé de 5 appartements et de 5 maisons. Tout aléa inférieur ou égal à $0,5$ aurait conduit à ce résultat, tandis que tout aléa supérieur à $0,5$ aurait conduit à un tirage de 4 appartements et de 6 maisons. Le tirage systématique à probabilités égales dans une base de sondage triée par des variables qualitatives permet ainsi de tirer un nombre fixe d'unités par modalité à une unité près, se rapprochant ainsi fortement d'un tirage aléatoire simple stratifié avec allocations proportionnelles, aux arrondis près.

Le tirage systématique à probabilités égales avec une base de sondage triée par des variables qualitatives garantit ainsi le nombre d'unités tirées à une unité près pour chaque modalité de la première variable de tri. Le nombre d'unités tirées pour chaque modalité de la deuxième variable de tri est garanti au nombre de modalités de la première variable près. Le nombre d'unités tirées pour chaque modalité de la troisième variable de tri est assuré au produit des nombres de modalités de la première et de la deuxième variables près, et ainsi de suite.

Ainsi, dès l'utilisation de quelques variables de tri qualitatives ou l'utilisation d'une variable de tri continue, une variable de tri supplémentaire n'a que peu d'incidence sur la précision du tirage.

FIGURE 3.1 – Exemple de tirage systématique



3.2 Le cadre choisi pour comparer les méthodes

On souhaite évaluer différentes méthodes de tirage et les comparer entre elles. Il s'agit de préciser la stratégie adoptée pour comparer les différentes méthodes, et les paramètres des différentes méthodes que sont la base de sondage, les allocations et les variables auxiliaires utilisées.

3.2.1 Stratégie de comparaison : méthode de Monte-Carlo

Comme précédemment, on choisit pour cela le procédé de Monte-Carlo, qui consiste à comparer les erreurs quadratiques moyennes estimées pour chaque méthode à partir d'un grand nombre M de tirages.

Pour un estimateur $\hat{t}_{y_{(m)}}$ du total t_y (ou de la moyenne) d'une variable Y, calculé lors de la m-ième simulation, on estime le biais relatif (BR) Monte-Carlo en pourcentage

par :

$$\hat{BR}_{MC}(\hat{t_y}) = \frac{1}{M}\sum_{m=1}^{M}(\hat{t}_{y_{(m)}} - t_y) \text{ x } \frac{100}{t_y}$$

Le coefficient de variation (CV) en pourcentage est lui estimé par :

$$\hat{CV}_{MC}(\hat{t_y}) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{t}_{y_{(m)}} - \frac{1}{M}\sum_{s=1}^{M}\hat{t}_{y_{(s)}})^2} \text{ x } \frac{100}{t_y}$$

Enfin l'erreur quadratique moyenne (EQM) est estimée par :

$$\hat{EQM}_{MC}(\hat{t_y}) = \frac{1}{M}\sum_{m=1}^{M}(\hat{t}_{y_{(m)}} - t_y)^2$$

et en notant RMSE (pour Root-mean-square error) la racine carrée de EQM, on a alors en termes relatifs et en pourcentage :

$$\hat{CV(RMSE)}_{MC}(\hat{t_y}) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{t}_{y_{(m)}} - t_y)^2} \text{ x } \frac{100}{t_y}$$

On fixe, ici, le nombre de simulations à 100 000 pour toutes les méthodes testées à l'exception du tirage spatialement équilibré, porté à 20 000 uniquement compte tenu de la durée nécessaire pour générer des échantillons.

3.2.2 Paramètres de tirage : base de sondage et allocations

Pour mettre en œuvre les différents tirages, on choisit de travailler sur une unité primaire exhaustive. C'est en effet au sein de ces unités primaires, dans lesquelles le nombre de logements tirés est plus important, que la différence entre les méthodes sera a priori la plus marquée. La ville de Poitiers est un bon candidat pour effectuer ces simulations (voir encadré). On dispose ainsi de 45 473 résidences principales servant de base de sondage pour les multiples tirages attendus.

Après avoir retenu la commune de Poitiers comme cas type, la taille des échantillons est fixée à $n = 50$, allocation proche de la moyenne pour les enquêtes déjà tirées dans cette ville.

Dès lors quel que soit le plan de sondage retenu, le taux de sondage est de l'ordre de 1 pour 1000 ($\frac{50}{45473}$ très exactement) correspondant également à la probabilité d'inclusion simple π_k de chaque logement étant donné que nous ne considérons que des sondages à probabilités égales.

Pourquoi le choix de Poitiers ?

Disposant d'une base de sondage comprenant $N = 45\,473$ unités secondaires au sein de l'unité primaire exhaustive que constitue la commune de Poitiers, nous cherchons à tirer des échantillons de n logements. L'objectif de ces tirages est de modéliser différentes procédures d'échantillonnage pouvant être mises en place dans le cadre de certaines enquêtes de l'Insee auprès des ménages.

Le choix de Poitiers comme d'étude s'explique par le fait que les tirages et plus particulièrement ceux opérés par la méthode du cube local s'avéraient chronophages : générer un échantillon par le cube local nécessitait sur Poitiers une quinzaine de secondes selon les paramètres utilisés. Dès lors que l'on souhaite réaliser un très grand nombre de tirages, cette durée unitaire prend une grande importance. La taille n de l'échantillon ainsi que le nombre et la nature des variables d'équilibrage mises en jeu pouvaient avoir une incidence sur cette durée mais pas de manière significative. C'est plus spécifiquement la taille de la base de sondage dans laquelle l'échantillon est tiré qui joue ici un rôle crucial. En effet, la commune de Nantes qui est également une unité primaire exhaustive avait dans un premier temps été envisagée. Or le temps de tirage relevé pour Nantes lors de la phase de test était de l'ordre d'une minute soit 4 fois plus qu'à Poitiers alors qu'elle est regroupe environ deux fois plus de logements appartenant au champ considéré. D'autres tests ont mis en évidence que le temps de tirage selon la méthode du cube local tendait à croître avec le carré de la taille de la base de sondage.

Afin de limiter le temps nécessaire pour effectuer les simulations, il semblait donc préférable de travailler sur le cas d'une commune de taille moins importante mais néanmoins suffisante pour constituer une unité primaire exhaustive.

3.2.3 Choix des variables auxiliaires

Les différentes méthodes de tirage étudiées dans la suite de cette partie mobilisent de l'information auxiliaire pour améliorer la précision des échantillons tirés. Préalablement à l'analyse comparative des différentes méthodes de tirage, des travaux ont été menés dans le but de sélectionner l'information auxiliaire qui semble la plus pertinente pour les différentes méthodes.

De nombreuses informations sont disponibles dans les bases fiscales pouvant servir d'information auxiliaire (nombre de personnes par logement, sexe, âge, revenus, caractéristiques du logement, revenus sociaux...). Différentes combinaisons de variables sont testées pour choisir les variables auxiliaires pertinentes dans l'implémentation des méthodes de tirage. En ce qui concerne le tri systématique, il n'y a guère d'intérêt à tester des combinaisons de plus de 2 ou 3 variables auxiliaires (cf. encadré sur le tirage systéma-

tique). Pour les méthodes du tirage équilibré et spatialement équilibré, les combinaisons étudiées peuvent être multiples, allant jusqu'à utiliser l'ensemble des variables disponibles.

Après avoir testé différentes combinaisons de variables d'équilibrage [1], le principal enseignement qui ressort de l'analyse des coefficients de variation (CV) estimés à partir des simulations effectuées est qu'inclure l'ensemble des variables auxiliaires ne permet pas d'améliorer significativement la précision des indicateurs.

En revanche, ne pas inclure le nombre de personnes ou les revenus dégrade sensiblement cette précision. En conséquence, seules la taille des ménages et celle sur leurs revenus sont conservées comme variables d'équilibrage.

De même, après avoir analysé différentes combinaisons de variables de tri, on ne retient que ces deux mêmes informations auxiliaires pour ordonner la base en vue du tirage systématique.

Les paragraphes suivants présentent donc les résultats issus de la comparaison par méthode de Monte-Carlo sur 100 000 tirages (ou 20 000 dans le cadre du tirage spatialement équilibré) de 50 logements dans l'unité primaire de Poitiers des scénarios suivants [2] :

- tirage systématique sur base triée par revenu fiscal puis par nombre de personnes par logement (scénario systématique, utilisé comme base de référence) ;
- tirage équilibré sur le revenu fiscal puis sur le nombre de personnes par logement (scénario cube 1) ;
- tirage spatialement équilibré sur le revenu fiscal et le nombre de personnes par logement (scénario cube local 1) ;
- tirage équilibré sur la distribution de revenus [3] et sur le nombre de personnes par logement (scénario cube 2) ;
- tirage spatialement équilibré sur la distribution de revenus et le nombre de personnes par logement (scénario cube local 2).

Ces scénarios sont ensuite comparés suivant selon un certain nombre de variables d'intérêt corrélées à des indicateurs mesurés dans des enquêtes auprès des ménages :

- le total d'individus de Poitiers [4] ;
- le taux de pauvreté dans l'unité primaire ;
- le total de la variable de niveau de vie des ménages composant les résidences

1. Là encore, les méthodes de Monte-Carlo sont utilisées pour comparer les différentes combinaisons de plan de sondage. Les résultats sont présentés en annexe 2.

2. Tous les scénarios de tirage équilibrés et spatialement équilibrés utilisent comme première variable d'équilibrage les probabilités d'inclusion, même quand cette variable n'est pas explicitement mentionnée.

3. La population est divisée en quartile de revenu, le premier correspondant à tous les ménages au revenu nul ou non renseigné.

4. Notons que, comme il s'agit de tirages de résidences principales, le nombre de résidences principales est parfaitement estimé pour chaque tirage tandis que le nombre d'individus estimé présente une variance non nulle.

principales poitevines ;
— le total de la variable de revenus fiscaux des ménages vivant dans les résidences principales de Poitiers avant abattements ;
— le total d'aides aux logements perçues à Poitiers ;
— le total de minima sociaux perçus à Poitiers.

La comparaison de ces scénarios est réalisée en premier lieu sur l'ensemble des échantillons tirés, puis avec prise en compte d'une non-réponse totale pour une partie des échantillons. En effet s'il apparaît naturel dans un premier temps de se concentrer sur les propriétés de l'échantillon au moment du tirage, la collecte des enquêtes ménages fait l'objet de non-réponse qui influe sur les propriétés des estimateurs. Il apparaît donc intéressant d'étudier la performance des algorithmes de tirage en intégrant de la non-réponse dans la chaîne de simulations.

3.2.4 Comparaison des méthodes avant non-réponse

Dans un premier temps, on compare les scénarios présentés en 3.2.3 uniquement sur la performance liée aux algorithmes de tirage. Les résultats des rapports de CV sont présentés dans les tableaux 3.1 et 3.2, respectivement avant et après un calage sur marges effectué sur les variables d'équilibrage [5]. Le tirage systématique est utilisé comme base dans les résultats présentés.

TABLE 3.1 – Comparaison de la précision de différentes méthodes de tirage en l'absence de non-réponse et de calage

Estimateurs	CV systematique	CV cube 1 (base systématique)	CV cube 2 (base systématique)	CV cube local 1 (base systématique)	CV cube local 2 (base systématique)
Population	*1.87*	1.08	1.21	1.24	1.45
Taux de pauvreté	*31.10*	1.08	1.03	1.03	0.98
Niveau de vie	*9.00*	0.83	0.91	0.80	0.88
Revenus fiscaux	*33.97*	0.95	0.97	0.89	0.89
Aides logement	*20.61*	1.08	0.98	1.02	0.97
Minima sociaux	*37.71*	1.04	1.00	1.00	0.97

Note de lecture : Le CV empirique du total de niveau de vie est de $9,00$ *lorsqu'on utilise un tirage systématique, en l'absence de non-réponse et sans utilisation du calage sur marges. Le CV empirique du total de niveau de vie dans le scénario cube 2 est de 0,91 fois celui du CV pour le tirage systématique.*

Avant calage (tableau 3.1), les méthodes sont assez proches, la précision obtenue par un scénario sur certaines variables se faisant aux dépens d'autres indicateurs. Par exemple, le tirage systématique est plus précis que les tirages équilibrés pour mesurer la pauvreté et la population et moins précis sur les niveaux de vie et de revenus moyens.

5. L'utilisation des variables d'équilibrage pour le calage est une pratique particulièrement utile dès lors que la structure de l'échantillon est déformé par la collecte, ce qui sera le cas dans les points 3.2.5 et 3.2.6.

TABLE 3.2 – Comparaison de la précision de différentes méthodes de tirage en l'absence de non-réponse et en présence d'un calage

Estimateurs	CV systematique	CV cube 1 (base systématique)	CV cube 2 (base systématique)	CV cube local 1 (base systématique)	CV cube local 2 (base systématique)
Population	*0.00*	1.00	1.00	1.00	1.00
Taux de pauvreté	*26.47*	1.27	1.25	1.21	1.18
Niveau de vie	*4.92*	1.29	1.15	1.20	1.11
Revenus fiscaux	*2.21*	0.69	0.85	0.68	0.92
Aides logement	*18.89*	1.19	1.12	1.13	1.11
Minima sociaux	*35.57*	1.11	1.09	1.07	1.06

Note de lecture : Le CV empirique du total de niveau de vie est de $4,92$ *lorsqu'on utilise un tirage systématique, dans le cas d'utilisation d'un calage sur marges en l'absence de non-réponse. Le CV empirique du total de niveau de vie dans le scénario cube 2 avec calage sur marges est de 1,15 fois celui du CV pour le tirage systématique avec calage sur marges.*

Dès lors qu'on applique le calage sur marges (tableau 3.2), le tirage systématique aboutit à des estimations nettement plus précises (hormis sur les niveaux de revenus) que les tirages spatialement équilibrés [6]. Ce constat est renforcé lorsqu'on compare les résultats du tirage systématique et ceux du tirage équilibré [7].

Ce résultat semble assez surprenant dans la mesure où les méthodes équilibrées visent des totaux (et par extension des moyennes) approximativement exacts sur les variables d'équilibrage et devraient permettre d'être plus précis sur les variables qui leur sont corrélées. Or, ces simulations révéleraient qu'avant même que la non-réponse ne vienne déformer la structure des échantillons, un tirage équilibré ne serait pas nécessairement plus intéressant qu'un tirage systématique compte tenu notamment de la nature de l'information auxiliaire disponible (variable de revenus notamment avec une distribution asymétrique) et de la faible taille des échantillons (50 logements). Une explication pourrait être que le faible nombre auxiliaires utilisées rend le tirage systématique performant. Cette première conclusion ne saurait néanmoins suffire à guider le choix de la méthode de tirage dans la mesure où dans la pratique, la plupart des enquêtes sont confrontées à des phénomènes de non-réponse.

6. Les résultats sur la variable de revenus fiscaux sont difficilement interprétables car il s'agit d'une variable de calage qui aurait donc dû être parfaitement estimée. Même si le CV de cette variable est faible après calage, il n'est pas nul, sans que nous ayons trouvé d'explication à ce sujet. Par contre, dans le point 3.2.6, le CV de cette variable est bien nul en présence de non-réponse et d'utilisation du calage sur marges, comme attendu.

7. Notons, par ailleurs, que l'équilibrage sur distribution de revenus donne, après calage, de meilleurs résultats que celui sur le niveau de revenu fiscal, dans tous les scénarios utilisant la méthode cube.

3.2.5 Influence de la non-réponse sur la précision

Chaque enquête est confrontée à deux types de non-réponse dégradant la qualité des indicateurs produits (par introduction d'un biais et diminution de la précision) :

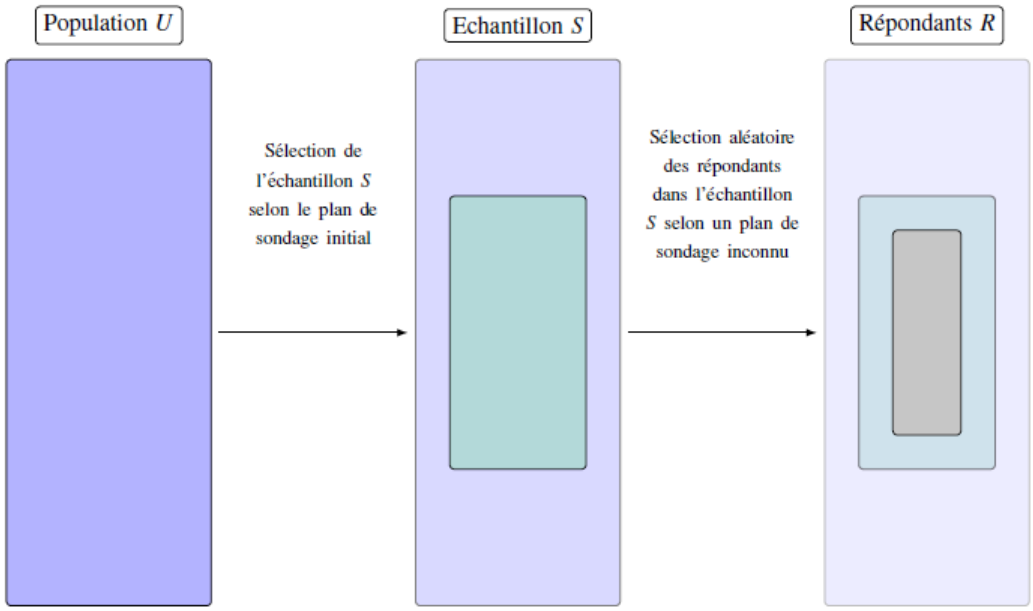
— la non-réponse partielle, correspondant à l'absence de valeur pour une variable de l'enquête ;

— la non-réponse totale, correspondant à une absence de valeur collectée pour l'ensemble du questionnaire [8].

Seule la non-réponse totale est ici prise en compte pour comparer les différentes méthodes de tirage.

La non-réponse peut être décrite comme un phénomène aléatoire : chaque unité de l'échantillon a une certaine probabilité de répondre, ρ_i. Ainsi, la sélection des répondants dans l'échantillon peut être vue comme une phase additionnelle du plan de sondage (voir figure 3.2). Les répondants sont de fait sélectionnés dans la population totale en deux étapes : la sélection de l'échantillon enquêté dans la population, suivant un plan de sondage connu et maîtrisé ; puis la sélection des répondants dans l'échantillon, suivant un plan de sondage inconnu.

FIGURE 3.2 – La non-réponse comme phase additionnelle du plan de sondage



8. Cela peut être dû au refus de l'enquêté, à un problème technique, à l'impossibilité d'enquêter l'individu ciblé...

Deux méthodes ont été implémentées pour modéliser le plan de sondage « inconnu » et ainsi introduire de la non-réponse dans les simulations :

— la première a consisté à stratifier l'échantillon en fonction d'informations contenues dans la base de sondage, connues comme étant liées au comportement de réponse pour des enquêtes passées, puis à attribuer des probabilités de réponse qui diffèrent d'une strate à l'autre, fixées en fonction des résultats observés au cours des enquêtes passées [9] ;

— la seconde option pour modéliser la non-réponse a consisté en un modèle non-paramétrique, au contraire de la première option, qui affecte à chaque unité échantillonnée le comportement de réponse du logement de Poitiers dont elle est le plus proche géographiquement et qui a été interrogé dans une enquête passée [10]. Ainsi, si le logement avait répondu, l'unité de l'échantillon simulé est considérée comme répondante, sinon, elle est considérée comme non répondante.

Une fois la non-réponse simulée, différentes méthodes existent pour corriger cette non-réponse, et ainsi limiter le biais et atténuer la perte de précision qu'elle induit. L'une d'elles étant le calage sur marges, on ajoute aux différentes variables définissant les marges, celles expliquant le comportement de réponse [11].

Les deux modèles de simulation de non-réponse appliqués ont mené à des conclusions similaires. Aussi, seuls les résultats de la première méthode sont présentés dans le paragraphe suivant. Dans le scénario de modélisation paramétrique de la non-réponse, l'information auxiliaire sur le type de logement a permis de sur-représenter, en moyenne, les logements individuels parmi les répondants.

3.2.6 L'effet de la non-réponse sur les différents plans de sondage

Comme précédemment, un grand nombre de simulations ont été effectuées pour chacun des 5 plans de sondage retenus (cf. 3.2.3), et à chacune d'elles, une étape de sélection de répondants a été ajoutée. Le calage sur marges a également été complété pour corriger cette non réponse construite.

La taille des échantillons obtenus n'est désormais plus fixe et répond à un processus aléatoire. Le tableau 3.3 permet de comparer la précision obtenue en fonction des plans de sondage qui diffèrent selon la méthode de tirage de la première phase [12].

9. Le choix d'inclure ou non chacune de ces unités se fait indépendamment des autres et correspond à la réalisation d'une variable aléatoire X_i qui suit une loi de Bernoulli de paramètre ρ_i.

10. Les réponses aux enquêtes « Condition de vie et sécurité 2012 à 2017 » ont été utilisées. Ces enquêtes présentent une taille d'échantillon proche de celle utilisée pour les simulations.

11. Dans notre cas, le type de logement (individuel/collectif), qui a été utilisé pour stratifier l'échantillon dans la première modélisation de la non-réponse, est naturellement utilisé.

12. Sont comparés, à présent, les coefficients de variation relatifs à l'erreur quadratique moyenne pour tenir compte du biais introduit par la non-réponse.

TABLE 3.3 – Comparaison des CV(RMSE) des différentes méthodes de tirage avec non-réponse par type de logement en présence de calage

Estimateurs	CV(RMSE) cube 1 (base systématique)	CV(RMSE) cube 2 (base systématique)	CV(RMSE) cube local 1 (base systématique)	CV(RMSE) cube local 2 (base systématique)
Population	1.00	1.00	1.00	1.00
Taux de pauvreté	1.09	1.08	1.07	1.06
Niveau de vie	1.14	1.06	1.07	1.05
Revenus fiscaux	1.00	1.00	1.00	1.00
Aides logement	1.06	1.03	1.05	1.03
Minima sociaux	1.07	1.05	1.04	1.03

Note de lecture : Le CV empirique calculé à partir de l'écart quadratique moyen du total de niveau de vie dans le scénario cube 2 est de 1,06 fois celui obtenu par tirage systématique, lorsqu'on intègre une étape de non-réponse avec correction par calage sur marges.

La présence de non-réponse dégrade, bien évidemment, la précision de l'ensemble des méthodes de tirage pour les différents estimateurs considérés. Après calage, les CV obtenus par la méthode de tirage systématique sont plus faibles que ceux des tirages équilibrés [13].

Les simulations réalisées en appliquant différentes méthodes d'échantillonnage tendent à montrer qu'un tirage systématique des unités secondaires au sein des UP exhaustives est plus efficace que des tirages équilibrés, que la non-réponse soit prise en compte ou pas.

Toutefois, c'est seulement une fois le calage sur marges appliqué, que l'avantage comparatif du tirage systématique est clairement identifié. Sans calage, les différentes méthodes équilibrées affichent des niveaux de précision généralement comparables à ceux du tirage systématique voire meilleurs sur certains estimateurs.

Tout se passe comme si les échantillons obtenus à partir d'un tirage systématique avaient des combinaisons d'unités plus favorables à la repondération. Une analyse de la structure des différents échantillons, présentée en annexe 3, a été menée, afin de comprendre ce qui assure au tirage systématique des propriétés plus favorables. Le calcul des probabilités d'inclusion double suggère que l'algorithme du cube tend à générer des associations privilégiées de logements, principalement entre ménages à revenus très élevés et ménages sans revenus, des sortes d'effets "granulaires" que l'on ne maîtrise pas bien à ce jour et qui s'avèrent manifestement pénalisants pour les petites tailles d'échantillon, comme celles des échantillons de second degré.

13. A nouveau, le tirage spatialement équilibré est un peu plus précis après calage (à l'exception de l'estimateur du total de personnes pauvres) que le tirage équilibré.

On utilisera donc des tirages systématiques pour les tirages de logements au second degré. Les variables de tri changeront d'une enquête à l'autre, en fonction de leur pertinence vis-à-vis du sujet de l'enquête.

Si les plans de sondage au second degré présentés ici semblent indépendants les uns des autres, la base de sondage utilisée est la même pour la plupart des enquêtes, et le comportement de réponse d'un individu à une enquête peut dépendre de son appartenance à l'échantillon d'une autre enquête de l'Insee. Aussi, l'Insee procède-t-il au marquage de ces échantillons, afin de ne pas sélectionner un même logement pour plusieurs enquêtes. Cette problématique pose différentes questions sur la représentativité de la base, les échantillons surreprésentant fréquemment certaines catégories de population, et la base évoluant annuellement. Le chapitre suivant propose une introduction aux difficultés que pose cette opération de marquage, et aux premières réponses apportées.

Chapitre 4

Le marquage

Afin de ne pas réinterroger à court terme un même ménage pour des enquêtes différentes, l'Insee a recours à une disjonction des échantillons en gardant en mémoire les unités déjà sélectionnées : on appelle cette opération le marquage. S'il peut concerner tout type d'unité, il est jusqu'à présent réservé au tirage des logements, car c'est l'unité d'échantillonnage habituellement utilisée, et la plus facile à suivre dans le temps dans la base de sondage [1]. Les raisons de ce marquage, présentées dans le premier paragraphe, concernent autant l'Institut, pour la qualité de ses indicateurs [2], que l'enquêté pour la charge liée à l'obligation de répondre. Cette opération de marquage n'est toutefois pas sans conséquence puisqu'elle implique une modification de la composition de la base de tirage à l'issue du tirage d'un nouvel échantillon. Le deuxième paragraphe décrit les traitements réalisés par l'Insee pour conserver une structure identique de la base de sondage au fil des tirages dans un millésime de Fidéli, puis au moment du changement annuel de millésime. Enfin la pratique du marquage pose plusieurs questions, notamment au sujet de l'interdépendance du marquage de logements et du marquage d'individus (ces deux types d'unités pouvant faire l'objet de tirages dans les bases issues de Fidéli) puisque, si les marquages de ces deux types d'unités interagissent, les bases de tirage de logements et d'individus deviennent interdépendantes.

4.1 Les raisons et les limites du marquage

Les enquêtes auprès des ménages de l'Insee ont vocation, pour la plupart, à représenter l'ensemble de la population résidant en France. Aussi, pour chacune de ces enquêtes, la base de sondage devrait contenir tous les logements/individus correspondant à ce champ. Cela étant, tirer continuellement dans la même base de sondage rend possible la sélection d'une même unité pour deux enquêtes différentes. Cette probabilité augmente avec l'utilisation d'un échantillon-maître, puisque la population éligible pour un tirage d'échantillon est limitée à celle appartenant aux unités primaires de l'EM. En considérant deux échantillons de 20 000 logements dont les tirages sont indépendants et réalisés avec

1. Depuis la rédaction de ce document et la refonte des échantillons-maître et emploi, des marquages d'individus ont été mis en oeuvre à millésime fixé de base de sondage. Cependant, la question de la conservation des marquages d'individus issus des millésimes passés reste à instruire lors du passage à un nouveau millésime pour la base de sondage. Cet enjeu est détaillé dans le point 4.3.2.

2. En cas d'interrogation pour diverses enquêtes, l'enquêté peut se lasser. Les réinterrogations entraînent donc un risque de diminution des taux de réponse aux enquêtes.

remise. un logement a une probabilité d'être sélectionné pour les deux enquêtes égale à $\frac{1}{2\ 000\ 000}$ si le tirage est réalisé sur tout le territoire, i.e. selon un sondage à un degré [3]. En revanche, dans le cas d'un tirage réalisé au sein des 541 UP de l'EM, i.e. selon un sondage à deux degrés, l'allocation par UP est de 40 logements environ. Alors pour un logement situé dans une UP de 2 500 logements, la probabilité d'être tiré deux fois est d'environ $\frac{1}{4000}$ [4]. Si l'évènement semble rare, dès le deuxième tirage (de 40 unités parmi 2500), il y a une chance sur deux qu'au moins une unité soit sélectionnée deux fois [5]. En considérant l'ensemble des UP de l'EM, la probabilité qu'aucun logement ne figure dans les deux échantillons est très faible. Ainsi, quand le nombre de tirages augmente, il est hautement improbable de ne pas tirer à nouveau des logements déjà sélectionnés pour des enquêtes passées. Le risque prépondérant est une non-réponse totale ou partielle, voire une réponse sciemment erronée, consécutive à la lassitude de l'enquêté. L'image de l'Insee, le moral de l'enquêté, la qualité des indicateurs sont alors pénalisés.

Pour éviter cela, l'application Nautile opérant les tirages d'échantillons garde en mémoire les unités enquêtées grâce à un procédé appelé le marquage. Concrètement, il s'agit de considérer les unités sélectionnées pour un échantillon, et de les enlever de la base de tirage qui servira aux prochains tirages. Cette opération, a priori simple, présente plusieurs imperfections ou défauts.

4.1.1 Le marquage du logement, l'interrogation du ménage

La première imperfection de cette méthode réside dans le fait que la plupart des enquêtes de l'Insee sont tirées au niveau du logement afin d'atteindre le ménage qui y réside. C'est donc le logement qui est marqué. Or, la charge de réponse incombe aux individus y habitant, qui, en déménageant, peuvent se trouver à nouveau sélectionnés pour une autre enquête si leur nouveau logement est tiré dans un nouvel échantillon. L'objectif recherché n'est alors pas atteint. Il faut toutefois nuancer ce risque, car seuls 11 % d'individus (France hors Mayotte) déménagent chaque année de leur logement (Levy et Dzikowski, 2017) et seuls 20 % des ménages (France métropolitaine) déménagent en 4 ans (Delance et Vignolles, 2017).

4.1.2 Une réduction accélérée de la base de sondage et une réinterrogation quasi-inévitable

La constitution des unités primaires a répondu à la contrainte d'obtenir des zones comportant suffisamment d'unités pour être viables 5 à 10 ans, sans devoir réinterroger un logement déjà enquêté. En effet, en marquant chaque échantillon, on restreint petit à

3. Le territoire France entière comporte approximativement 28 000 000 logements résidences principales. Donc pour 2 tirages de 20 000 logements, un logement se retrouve dans les 2 tirages avec une probabilité $\frac{1}{2\ 000\ 000} = \left(\frac{20\ 000}{28\ 000\ 000}\right)^2$.

4. $\frac{1}{4000} = \left(\frac{40}{2500}\right)^2$

5. La probabilité d'une unité de figurer dans les deux échantillons est $\left(\frac{40}{2500}\right)^2$. L'événement qu'au moins une unité soit tirée deux fois est le complémentaire de l'événement dans lequel aucune unité n'est tirée deux fois. Si on fait l'approximation que le tirage des logements est indépendant d'un logement à l'autre, alors la probabilité qu'aucune des 2500 unités ne soit tirée dans les deux échantillons est $[1 - (\frac{40}{2500})^2]^{2500} = 0,53$.

petit la base de sondage disponible pour le tirage d'un échantillon. Au bout d'un certain nombre de tirages, tous les logements de la base initiale ont été sélectionnés, et il devient nécessaire de démarquer certaines unités si l'échantillon-maître n'est pas renouvelé. En reprenant l'exemple développé plus haut (UP de 2 500 résidences principales, tirages de 40 logements par UP), avec marquage des échantillons la base de sondage serait totalement marquée dans certaines UP à l'issue de 63 tirages. Afin de garantir un stock d'unités disponibles dans la base de sondage, il pourra être nécessaire de démarquer certains logements échantillonnés. Ainsi, une unité déjà enquêtée pourrait être tirée à nouveau ultérieurement en cas de démarquage d'anciens échantillons.

Dans ce dernier cas, l'usage vise à s'assurer qu'un même ménage ne peut pas être interrogé à moins de 5 ans d'intervalle et on suppose qu'il est acceptable pour un individu d'être réinterrogé de nouveau 5 ans après avoir été sélectionné pour une première enquête [6].

Cette restriction progressive de la base de sondage laisse craindre une diminution de la précision des enquêtes, tirage après tirage [7]. En réalité, en considérant chaque tirage comme aléatoire, on peut voir cette succession d'échantillons comme un unique tirage sans remise. L'impact du marquage sur la qualité des indicateurs est ainsi négligeable.

4.1.3 Un marquage plus étendu que les seuls échantillons tirés

Le troisième défaut (et certainement le plus pénalisant) de l'opération de marquage simple, consistant à n'enlever de la base de tirage que les unités des échantillons précédemment tirés, provient de la modification de la structure de la base consécutive au marquage. En effet, pour que la base garde les mêmes caractéristiques avant et après le tirage, il faut que l'échantillon possède lui aussi cette même structure. Or, si, en espérance, cela est vérifié pour des tirages aléatoires simples dans l'ensemble de la base, la sur-représentation de populations d'intérêt, voire la restriction de champ, pour des enquêtes modifient obligatoirement la composition de la base résiduelle à l'issue du marquage. La conservation de la structure de la base de sondage initiale dans la base résiduelle doit obligatoirement passer soit par un marquage partiel des unités interrogées, avec le risque de tirer de nouveau des unités déjà marquées, soit par le marquage complémentaire de logements non enquêtés, impliquant un épuisement plus rapide de la base.

Cette deuxième option, celle d'un marquage complémentaire de logements non enquêtés, est mise en oeuvre par l'Insee. Elle est utilisée également au moment du changement de millésime de la base de sondage pour éviter la surreprésentation des logements neufs. Le paragraphe suivant présente la méthode choisie pour marquer les différentes

6. Cette durée de 5 années est héritée des enquêtes annuelles de recensement qui étaient utilisées jusqu'en 2019 comme base de sondage. Un logement peut être recensé au maximum une fois tous les 5 ans, ce qui implique qu'un logement ne pouvait figurer dans la base de sondage qu'une fois tous les 5 ans jusqu'en 2019. Ainsi, dès lors qu'on était vigilant à ne pas tirer deux fois le même logement pour deux enquêtes différentes la même année, il était certain de l'interroger une fois au maximum tous les 5 ans. Par ailleurs, l'utilisation de Fidéli comme base de sondage implique une charge d'enquête plus faible pour les ménages puisque les logements interrogés dans les enquêtes auprès des ménages de l'Insee et dans les enquêtes annuelles de recensement ne sont plus nécessairement les mêmes.

7. Un nouvel échantillon est tiré dans une base plus petite que le précédent échantillon.

unités dans la base (marquage de l'échantillon à enquêter, marquage d'un échantillon complémentaire, marquage de logements neufs).

4.2 La part d'unités à marquer

On distingue trois cas de marquage d'unités dans la base :
— le marquage des unités tirées pour un échantillon
— le marquage d'unités pour rééquilibrer la base suite au tirage d'un échantillon
— le marquage de logements neufs, au moment du changement de millésime.

4.2.1 Le marquage des unités tirées

Lorsqu'un échantillon est sélectionné, toutes les unités tirées sont habituellement marquées dans la base de sondage afin de n'être pas sélectionnées de nouveau lors les tirages futurs. Dans l'absolu, ce marquage concerne toutes les enquêtes, même celles sélectionnées dans d'autres bases de sondage [8]. Certaines exceptions dérogent à cette règle.

— Il n'est pas toujours aisé de faire le lien entre diverses bases de sondage. Aussi, l'opération peut être abandonnée si elle risque de mener à des résultats non souhaités (marquage par erreur, ou oubli de marquage)
— L'unité de tirage n'est pas toujours le logement. L'Insee réalise aussi de plus en plus de tirages d'individus. Or les logements et les individus échantillonnés peuvent être dépendants. Aussi, il peut être complexe de gérer un marquage croisé de la base des logements et de la base des individus (cf. point 4.3.3).
— Certaines enquêtes sur-représentent des populations marginales. Dans de tel cas, le déséquilibre de la base à l'issue du marquage risquerait d'être trop important. Les unités tirées ne sont alors pas marquées.
— Enfin, des unités, pourtant interrogées, peuvent échapper à cette règle du marquage, comme pour les tests d'enquête dont les échantillons répondent parfois à des logiques de zonage qui n'est pas établi aléatoirement.

4.2.2 Le marquage pour rééquilibrer la base

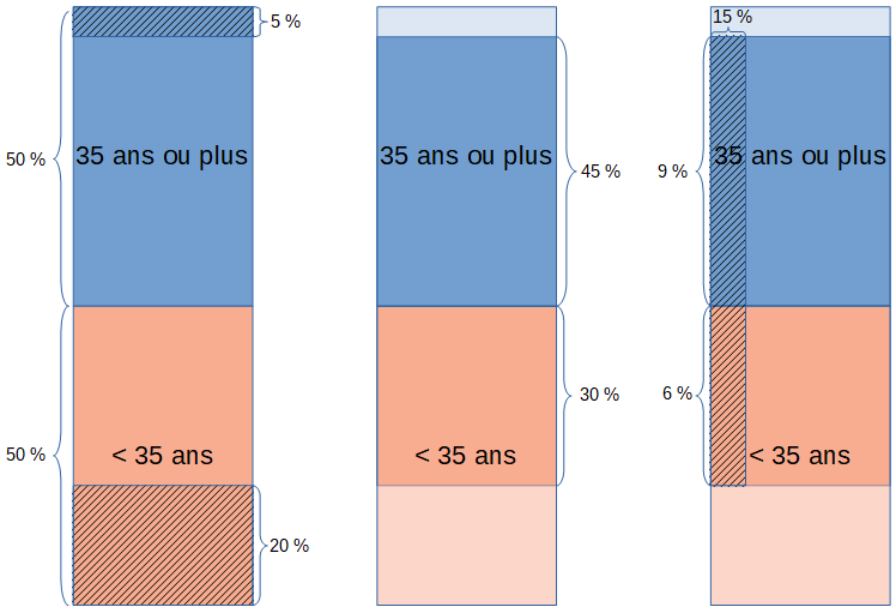
Comme évoqué précédemment, la plupart des enquêtes auprès des ménages de l'Insee procèdent à la sur-représentation de certaines populations d'intérêt. Si l'on marque les logements d'un échantillon n'ayant pas la même structure que la base de sondage, le complémentaire restant dans la base, pour la prochaine enquête, n'aura plus la même structure que l'ensemble de la population.

Par exemple, pour une enquête, on souhaite interroger 40% des moins de 35 ans, et 10% des 35 ans ou plus, chaque strate représentant la moitié de la population. Marquer

8. En particulier, les logements des secteurs sélectionnés pour l'enquête Emploi en continu ont été marqués dès la première utilisation de l'application de tirage dans la base de sondage, réduite au nouvel échantillon-maître ; d'autres enquêtes, échantillonnées dernièrement, ont également été marquées dans la nouvelle base, afin d'éviter de réinterroger les unités sélectionnées.

l'échantillon tiré détériorerait la base résultante. En effet, il resterait 60% des moins de 35 ans et 90 % des 35 ans ou plus. Les 35 ans ou plus représenteraient alors 60% de la base non marquée au lieu de 50% dans la base initiale. L'échantillon de l'enquête suivante, pour lequel on souhaite un taux de sondage de 15 % de la population générale, sans sur-représentation, comporterait, en moyenne, trop de personnes de plus de 35 ans par rapport à la population réelle. La figure 4.1 illustre ce phénomène.

FIGURE 4.1 – Déformation de la base de sondage en cas de marquage du seul échantillon tiré



Note de lecture : La base de sondage initiale est constituée de 2 strates réparties équitablement. Un premier échantillon est tiré, sur-représentant les moins de 35 ans. L'échantillon est composé de 5% de la base initiale issus de la strate 35 ans ou plus et de 20% de la base initiale issus de la strate moins de 35 ans (à gauche). Suite à ce tirage, l'échantillon est marqué, c'est-à-dire qu'on conserve pour les tirages ultérieurs la base de sondage initiale privée de l'échantillon tiré. Ne restent alors que 75% de la base initiale (45% issus de la strate 35 ans ou plus et 30% issus de la strate moins de 35 ans) mobilisables pour les tirages ultérieurs (au centre). Le tirage suivant, à probabilités égales, respecte la structure de cette base intégrant le marquage et comporte donc plus d'individus de 35 ans ou plus (à droite).

Pour éviter cette déformation de la base de sondage, on procède à son rééquilibrage, en marquant un échantillon complémentaire [9] dans la base. D'un point de vue théorique, cela revient à faire un tirage en deux phases :

— la première phase, en population générale et sans sur-représentation, permet de tirer l'échantillon à marquer ;
— la deuxième phase permet de tirer, au sein de l'échantillon de première phase, l'échantillon à enquêter en prenant en compte la sur-représentation.

9. C'est-à-dire, en ne marquant pas uniquement les unités interrogées.

Habituellement, un plan de sondage à deux phases est utilisé quand on n'arrive pas à atteindre la population cible, souvent par manque d'information dans la base de sondage. Ainsi, on sélectionne, dans la population générale, un échantillon pour lequel on obtient une information auxiliaire (par enquête filtre, par exemple), puis on tire l'échantillon final de deuxième phase, auprès duquel est effectuée l'enquête complète. Pour échantillonner des populations spécifiques, la première phase implique de tirer un échantillon assez conséquent, afin d'être certain d'atteindre les allocations ciblées en deuxième phase. Dans le contexte du marquage, il est possible de limiter la taille de l'allocation de première phase au plus juste, car l'information auxiliaire permettant de stratifier la population est déjà connue.

Pour un tirage dans s strates où a_i et N_i sont respectivement l'allocation de deuxième phase et le nombre d'unités de la strate i, le taux de sondage théorique *minimal* tx de première phase permettant d'assurer le tirage de deuxième phase dans chaque strate est :

$$tx = max_j(\frac{a_j}{N_j})$$

Ainsi, pour chaque strate i, l'allocation a'_i de première phase sera :

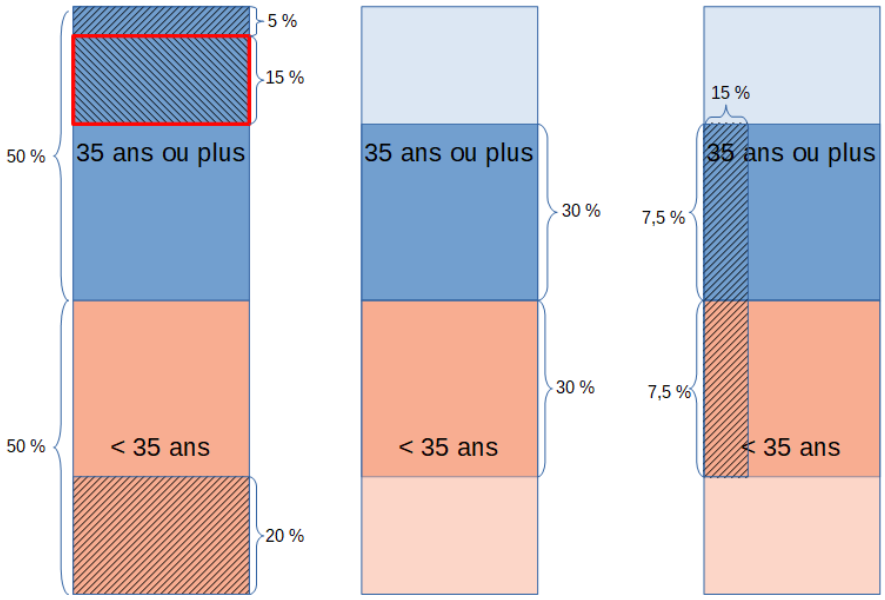
$$a'_i = \lceil N_i * max_j(\frac{a_j}{N_j}) \rceil$$

Pour cette strate i, le marquage affectera a'_i logements au lieu de a_i unités [10]. Quand l'enquête est menée au sein de l'EM, la population générale considérée est celle de la base de tirage, soit celle de l'EM. Par construction, il est censé avoir la même structure que la population totale.

La figure 4.2 illustre cette méthode en reprenant les allocations de l'exemple précédent, mais en appliquant une première phase de tirage entre les deux enquêtes.

10. En pratique, on tire un échantillon de deuxième phase en respectant les allocations a_i, puis, au sein de chaque strate i, on sélectionne $a'_i - a_i = \lceil N_i * max_j(\frac{a_j}{N_j}) \rceil - a_i$ unités.

FIGURE 4.2 – Absence de déformation de la base de sondage en cas de marquage de l'échantillon avec rééquilibrage



Note de lecture : La base de sondage initiale est constituée de 2 strates réparties équitablement. Un premier échantillon est tiré, sur-représentant les moins de 35 ans. L'échantillon est composé de 5% de la base initiale issus de la strate 35 ans ou plus et de 20% de la base initiale issus de la strate moins de 35 ans. Dans la logique du tirage en 2 phases, un échantillon complémentaire est tiré (cadre rouge) dans la strate 35 ans ou plus afin que l'échantillon complet ait le même taux de sondage dans chaque strate (à gauche). Suite à ce tirage en deux phases, l'échantillon et son complémentaire sont marqués, c'est-à-dire qu'on conserve pour les tirages ultérieurs la base de sondage initiale privée de l'échantillon tiré. Ne restent alors que 60% de la base initiale (30% issus de la strate 35 ans ou plus et 30% issus de la strate moins de 35 ans) mobilisables pour les tirages ultérieurs (au centre). Le tirage suivant, à probabilités égales, respecte la structure de cette base intégrant le marquage qui est la même que celle de la base de sondage initiale. L'échantillon contient donc autant d'unités de chaque strate (à droite).

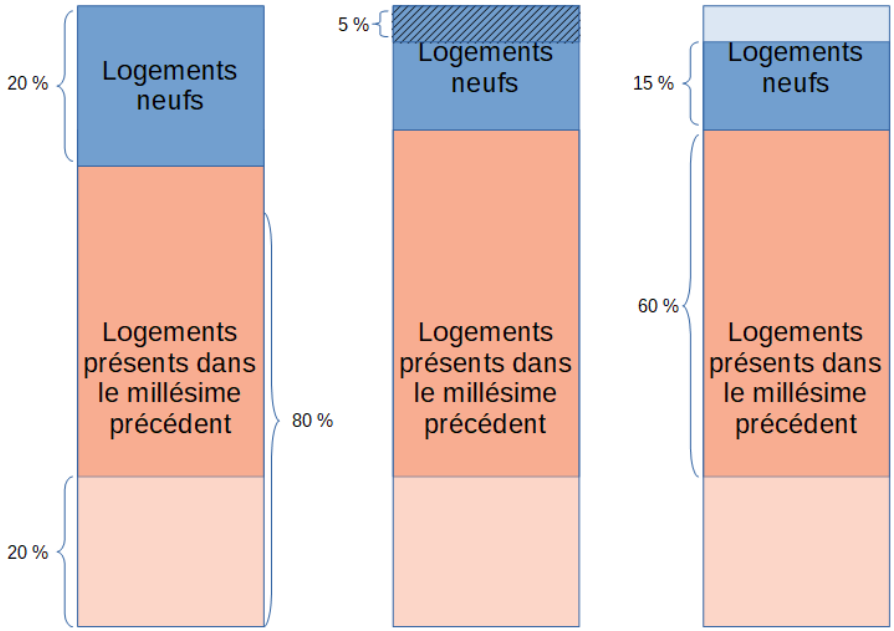
On constate que cette procédure accélère l'épuisement de la base, mais permet de conserver, après chaque tirage, une population respectant la structure initiale. La sur-représentation de populations très particulières (donc rares et peu nombreuses) empêche cependant le recours à cette méthode, car elle risquerait d'amener à marquer pratiquement toute la population, pour une interrogation concernant très peu d'individus. Ainsi, la probabilité de tirer à nouveau ces unités, ainsi que la déformation de la population restante, induite par la sur-représentation effectuée, sont négligeables au regard du marquage nécessaire pour rééquilibrer la base, et donc de son épuisement [11].

11. Par exemple, l'enquête Histoire de Vie et Patrimoine sur-représente fortement les hauts patrimoines. Rééquilibrer pleinement la base conduirait à tirer un échantillon complémentaire d'au moins 100 000 logements dans l'échantillon-maître, ce qui épuiserait la base de sondage rapidement. Le choix est donc fait de marquer l'échantillon enquêté sans tirer de complémentaire, au risque de déformer légèrement la base pour de futurs tirages.

4.2.3 Le marquage des logements neufs

La déformation de la base peut avoir lieu également lors de la mise à jour de la base de sondage. En effet, entre deux millésimes de Fidéli, certains logements disparaissent, d'autres apparaissent. Si la part d'unités interrogées parmi les disparitions est, du fait de la sélection aléatoire effectuée lors des tirages antérieurs, la même que dans la population restante, tous les nouveaux logements n'ont au contraire jamais été interrogés. Ils seraient donc sur-représentés dans les échantillons à venir si aucun traitement n'était effectué. C'est pourquoi, un procédé similaire à celui présenté précédemment est entrepris. La figure 4.3 modélise la mise à jour de la base de sondage. Un échantillon est ainsi tiré parmi les logements neufs avec un taux de sondage égal au taux de logements marqués parmi les logements qui sont présents à la fois dans l'ancien et dans le nouveau millésime.

FIGURE 4.3 – Rééquilibrage de la base de sondage lors du passage au nouveau millésime



Note de lecture : La base de sondage initiale est constituée de 2 strates. Les logements déjà présents dans l'ancien millésime et dont le quart est marqué correspondent à 80 % de la base. Les logements neufs forment 20 % de la base (à gauche). Pour rééquilibrer la base, un quart des logements neufs est tiré (au centre). Ces logements neufs sont ensuite marqués dans la base de sondage. La base mobilisable pour les tirages dans le nouveau millésime correspond à 75 % de la base complète (60 % issus de l'ancien millésime et 15 % issus des logements neufs). La structure de cette base est la même que la base initiale (à droite).

Si le marquage, ainsi pratiqué, assure de ne pas réinterroger un logement déjà sélectionné, et évite de détériorer, petit à petit, la structure de la base, il ne garantit pas la non-réinterrogation d'un individu s'il est tiré dans un échantillon d'individus, et ce, même lorsque son logement a déjà été enquêté.

4.3 La complexité du marquage d'individus dans les sources fiscales

Cette méthode de marquage résout la plupart des difficultés rencontrées, car l'Insee, jusqu'à maintenant, sélectionne, en majorité, des échantillons de logements. Le logement est une unité dont les caractéristiques sont factuelles, qui n'évolue que peu dans le temps, et dont l'apparition et la disparition de la base de sondage sont maîtrisées.
Cependant, de plus en plus, l'Insee est amené à sélectionner des échantillons d'individus [12]. Le marquage d'individus a alors été envisagé. Cela pose toutefois de nouvelles questions non résolues à l'heure actuelle.

4.3.1 L'individu, une unité plus difficile à suivre

Pour que le marquage soit efficace, il est indispensable de pouvoir suivre dans le temps les unités concernées. Dans le cas contraire, on risque de réinterroger des unités déjà collectées, et d'en exclure d'autres, à tort, de la base de tirage. Si les logements sont faciles à suivre (car ils ne "bougent" pas et font l'objet de déclarations lors de leur construction ou destruction), il n'en est pas de même pour les individus. À chaque changement de millésime, une partie des individus du nouveau fichier ne sont pas retrouvés par appariement dans la base précédente. Cependant, ce phénomène concerne principalement les individus mineurs pour lesquels les informations connues dans Fidéli sont moins riches. En se limitant aux individus majeurs, seuls 4,6 % ne sont pas chaînés entre deux millésimes [13]. Dans ce cadre, il est envisageable de marquer ces unités, sachant que la grande majorité des enquêtes se concentrent principalement sur un champ d'individus majeurs ou ayant 16 ans ou plus au moment de l'enquête.

4.3.2 Naissance, décès, sortie et entrée dans le champ

Si la *création* et la *suppression* d'individus liées à leurs naissance et décès se rapprochent des concepts de création et de suppression de logement, elles peuvent être également le fait d'autres phénomènes, comme des séjours longs à l'étranger. Aussi, le marquage à mettre en œuvre pour ces individus est beaucoup plus complexe à appréhender. Faut-il conserver le marquage d'un individu quittant le champ momentanément ? Faut-il mettre en œuvre un processus identique au marquage des nouveaux logements ? Par ailleurs, comment gérer le fait que les individus sont mobiles et que les flux des individus entre des UP de l'EM et hors des UP de l'EM peuvent ne pas être aléatoires [14] ?

12. Notamment du fait du développement de la collecte ayant recours à d'autres modes d'interrogation que le face-à-face (multimode) qui conduit à privilégier, pour le contact et la collecte, des coordonnées reliées à la personne plutôt qu'au logement.

13. Un tiers de ces individus majeurs non chaînés ont entre 18 et 21 ans dans le nouveau millésime.

14. Par exemple, les individus entrant dans une UP avec de nombreux étudiants vont être plus jeunes que ceux en sortant, dont certains seront marqués ; ces personnes sortantes auront également une structure particulière, différentes de la population demeurant dans l'UP.

4.3.3 Une base de logements reliée à une base d'individus

L'utilisation de Fidéli comme unique base de sondage, d'individus et de logements, permet d'envisager une coordination des marquages entre individus et logements. En effet, quelle que soit l'unité marquée, ce sont bien les individus que l'on souhaite décharger d'une nouvelle enquête. Si marquer tous les individus d'une résidence principale tirée pour une enquête ne semble pas produire d'effet indésirable, marquer un logement dès qu'un de ses résidents est échantillonné est plus complexe. Un tel traitement conduit rapidement à modifier la structure de la base de logement, car tous les logements n'ont pas la même probabilité d'être retirés de la base : à probabilité de sélection égale des individus, un logement contenant i individus aura i fois plus de chance d'être marqué qu'un logement n'abritant qu'une seule personne. Or ces logements et leurs occupants possèdent des caractéristiques bien particulières (âges et revenus des habitants différents, surfaces et nombre de pièces des logements différents...). Un tel marquage de logements déforme la structure de la base restante.

Pour pallier à ce défaut de marquage, une solution moins satisfaisante pour la charge d'enquête des ménages, mais de meilleure qualité statistique a été considérée : ne marquer un logement que si l'individu que l'on considère comme l'individu de référence est sélectionné. Deux propositions ont été étudiées :

- Utiliser le référent fiscal du logement comme individu de référence ;
- Sélectionner un individu du logement aléatoirement, indépendamment du tirage de l'échantillon en lui-même, et le considérer comme l'individu de référence.

Que l'une ou l'autre de ces propositions soit mise en oeuvre, la déformation de la base est toujours moindre que si le logement est marqué quel que soit l'individu qui est tiré dans le logement. La deuxième proposition est néanmoins préférable, le référent fiscal utilisé dans la proposition alternative présentant des caractéristiques moyennes qui peuvent, sur le long terme modifier légèrement l'équilibre initial [15]. Malgré cela, sélectionner un individu aléatoirement est également la solution qui garantit le moins que la personne répondant habituellement aux enquêtes avec tirage de logements ne soit pas échantillonnée ultérieurement [16].

Enfin, plus généralement, le marquage d'un logement à partir d'un ou plusieurs individus complexifie grandement le changement de millésime de la base de sondage : faut-il conserver ou supprimer le marquage des logements dont l'individu de référence a changé entre deux millésimes ?

Aussi, actuellement, les tirages d'individus ne conduisent pas à des marquages de logements et réciproquement [17].

15. On y constate une sur-représentation des hommes notamment.

16. L'individu du ménage qui répond aux enquêtes pour lesquelles on procède à un tirage de logements est plus probablement le référent fiscal puisque c'est l'individu du logement auquel la lettre-avis de l'enquête est alors adressée.

17. Il existe une exception : afin de ne pas détériorer les taux de collecte de l'enquête Emploi en continu, l'échantillon EEC est retiré de la base de tirage pour les échantillons d'individus.

Bibliographie

Applegate, Bixby, Chvátal et Cook (2003). « Implementing the Dantzig-Fulkerson-Johnson algorithm for large traveling salesman problems ». In : *Mathematical Programming* B 97, p. 91–153.

Ardilly (2006). *Les techniques de sondage*. Technip.

— (2022). « Projet Nautile : arbitrage sur la nature des unités primaires ». In : *Note interne de l'Insee* N2022_7084_DG75-L110.

Caron, Deville et Sautory (1998). « Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel Poulpe ». In : *Série des Documents de Travail Méthodologie Statistique de l'Insee* 9806.

Christine et Faivre (2009). « OCTOPUSSE : un système d'Échantillon-Maître pour le tirage des échantillons dans la dernière Enquête Annuelle de Recensement ». In : *Journées de Méthodologie Statistique 2009*.

De Bellefon, Audric, Durieux, Loonis, Le Gleut, Bouayad-Agha, Floch, Marcon, Puech, Le Saout, Le Gallo, Védrine, Genebes, Renaud, Sémécurbe, Favre-Martinoz, Fontaine, Lardeux-Schutz et Merly-Alpa (2018). « Manuel d'analyse spatiale ». In : *Insee Méthodes* 131.

Delance et Vignolles (2017). « Ça déménage ? La mobilité résidentielle et ses déterminants ». In : *Les conditions de logement en France*, p. 55–73.

Deville et Lavallée (2006). « Sondage indirect : Les fondements de la méthode généralisée du partage des poids ». In : *Techniques d'enquête* 32.2, p. 185–196.

Deville et Tillé (1998). « Unequal probability sampling without replacement through a splitting method ». In : *Biometrika* 85.1, p. 89–101.

— (2004). « Efficient balanced sampling : the cube method ». In : *Biometrika* 91.4, p. 893–912.

Favre-Martinoz et Merly-Alpa (2016). « Utilisation des méthodes d'échantillonnage spatialement équilibré pour le tirage des unités primaires des enquêtes ménages de l'Insee ». In : 9^{eme} *Colloque Francophone sur les Sondages*.

Givois et Merly-Alpa (2018). « Échantillonnage spatial via des distances socio-économiques : comparaison de méthodes pour le tirage d'un Échantillon-Maître ». In : *Journées de Méthodologie Statistique 2018*.

Grafström, Lundström et Schelin (2012). « Spatially balanced sampling through the pivotal method ». In : *Biometrics* 68.2, p. 514–520.

Grafström et Tillé (2013). « Doubly balanced spatial sampling with spreading and restitution of auxiliary totals ». In : *Environmetrics* 24.2, p. 120–131.

Hahsler et Hornik (2007). *TSP - Traveling Salesperson Problem - R package*. URL : `https://github.com/mhahsler/TSP`.

Lavallée (2007). *Indirect Sampling*. Springer.

Le Gleut (2017). « Stratification et calcul d'allocations dans les enquêtes auprès des entreprises ». In : *Fiches méthodologiques de l'Insee*. URL : https://www.insee.fr/fr/information/2838097%22.

Levy et Dzikowski (2017). « En 2014, un quart de la population qui déménage change de département ». In : *Insee Première* 1654.

Loonis (2009). « La construction du nouvel échantillon de l'Enquête Emploi en Continu à partir des fichiers de la Taxe d'Habitation ». In : *Journées de Méthodologie Statistique 2009*.

Raj (1966). « Some remarks on a simple procedure of sampling without replacement ». In : *Journal of the American Statistical Association* 61, p. 391–396.

Schoenmakers et Santos (2013). « A Geografia de referência do Ficheiro Nacional de Alojamentos ». In : *Jornadas de Classificação e Análise de Dados 2013*.

Sillard, Faivre, Paliod et Vincent (2020). « Pour les enquêtes auprès des ménages, l'Insee rénove ses échantillons ». In : *Courrier des statistiques* N4, p. 80–101.

Annexes

Annexe 1 - Étendue géographique des grappes

TABLE 1 – Distribution des étendues géographiques des anciennes et nouvelles grappes EEC (en km par la route) : troisième quartile, médiane et premier quartile ; part des anciennes et nouvelles grappes dont la totalité des logements sont à la même adresse

	Échantillon actuel				Nouvelles grappes			
Région	Q3	Médiane	Q1	Grappes à une seule adresse	Q3	Médiane	Q1	Grappes à une seule adresse
Alsace	1,8	1,0	0,3	15,4 %	1,2	0,6	0,1	24,4 %
Aquitaine	4,9	1,9	0,5	12,4 %	2,7	1,1	0,3	17,9 %
Auvergne	3,8	1,7	0,6	8,2 %	2,8	1,2	0,4	13,9 %
Basse-Normandie	5,0	1,7	0,7	9,5 %	3	1,2	0,4	14,7 %
Bourgogne	3,0	1,4	0,4	13,5 %	2,1	1	0,3	17,5 %
Bretagne	4,4	1,8	0,6	9,7 %	2,6	1,2	0,4	14,0 %
Centre	3,7	1,5	0,6	9,4 %	2,1	1	0,3	15,9 %
Champagne-Ardenne	2,1	1,0	0,3	16,7 %	1,5	0,7	0,2	20,8 %
Corse	9,1	4,8	0,2	18,4 %	3,3	1,2	0	28,5 %
Franche-Comté	2,7	1,4	0,6	10,0 %	1,9	1	0,3	16,8 %
Haute-Normandie	3,7	1,3	0,3	16,0 %	1,9	0,8	0,1	22,9 %
Ile-de-France	0,8	0,2	0,0	43,4 %	0,4	0	0	59,2 %
Languedoc-Roussillon	2,5	1,1	0,4	13,1 %	1,8	0,7	0,1	20,6 %
Limousin	5,1	2,0	0,7	8,6 %	3,9	1,6	0,5	12,9 %
Lorraine	2,0	1,0	0,4	15,6 %	1,4	0,7	0,2	18,1 %
Midi-Pyrénées	5,0	1,8	0,6	12,8 %	3	1,2	0,2	18,6 %
Nord-Pas-de-Calais	1,7	0,8	0,4	10,8 %	1,1	0,5	0,2	18,5 %
Pays de la Loire	3,7	1,6	0,6	11,7 %	2,1	1	0,3	15,8 %
Picardie	2,3	1,3	0,6	10,1 %	1,4	0,8	0,3	16,3 %
Poitou-Charentes	4,3	2,0	1,0	6,0 %	2,4	1,2	0,6	9,4 %
PACA	2,6	0,7	0,1	23,6 %	1,8	0,4	0	36,9 %
Rhône-Alpes	3,2	1,0	0,1	23,3 %	1,9	0,6	0	32,6 %
France entière	**2,6**	**1,0**	**0,2**	**19,4 %**	**1,7**	**0,6**	**0**	**27,8 %**

Annexe 2 - Choix des variables d'équilibrage pour les scénarios de tirage au second degré

Pour les scénarios de tirage équilibrés et spatialement équilibrés présentés en section 3.2.3 de la partie D, un choix préalable de variables auxiliaires a dû être effectué.

Dans un premier temps, le tirage équilibré sur les résidences principales a été réalisé à l'aide de l'ensemble des variables auxiliaires de niveau logement qui paraissaient pertinentes [18] :

- probabilités d'inclusion
- nombre de personnes dans le logement ;
- revenu fiscal dans le logement ;
- indicatrice de logement collectif ;
- indicatrice de logement social ;
- indicatrice d'appartenance à un quartier prioritaire ;
- nombre de personnes par sexe dans le logement ;
- nombre de personnes par classe d'âge dans le logement.

Dans un second temps, les tirages équilibrés se sont concentrés sur un nombre plus limité de variables auxiliaires. Diverses combinaisons de variables auxiliaires ont été étudiées. Les résultats de ces tirages équilibrés ont été comparés sur diverses variables d'intérêt par ailleurs utilisées pour les comparaisons de scénarios dans la partie D :

- le total d'individus de Poitiers [19] ;
- le taux de pauvreté dans l'unité primaire ;
- le total de la variable de niveau de vie des ménages composant les résidences principales poitevines ;
- le total de la variable de revenus fiscaux des ménages vivant dans les résidences principales de Poitiers avant abattements ;
- le total d'aides aux logements perçues à Poitiers ;
- le total de minima sociaux perçus à Poitiers.

18. Il ne sert à rien que le nombre de modalités ou variables utilisées pour l'équilibrage excède la taille de l'échantillon, c'est-à-dire 50.

19. Notons que, comme il s'agit de tirages de résidences principales, le nombre de résidences principales est parfaitement estimé pour chaque tirage tandis que le nombre d'individus estimé présente une variance non nulle.

TABLE 2 – CV des estimateurs obtenus par un tirage spatialement équilibré sur la totalité des variables auxiliaires

Estimateur	CV (%)
Population	2.21
Taux de pauvreté	31.63
Niveau de vie	7.17
Revenus fiscaux	30.07
Aides logement	20.26
Minima sociaux	36.72

Note de lecture : Le coefficient de variation de l'estimateur du total de niveau de vie à Poitiers est de $7,17$ *lorsqu'on réalise un tirage spatialement équilibré sur l'ensemble des variables d'équilibrage présentées précédemment.*

TABLE 3 – CV pour des estimateurs obtenus par des tirages spatialement équilibré pour différentes combinaisons de variables auxiliaires en utilisant comme base le tirage spatialement équilibré sur l'ensemble des variables auxiliaires

Estimateur	Equilibrage sur nombre de personnes revenus et perception d'allocations chômage	Equilibrage sur nombre de personnes revenus	Equilibrage sur nombre de personnes	Equilibrage sur type de logement nombre de personnes revenus	Equilibrage sur statut d'occupation nombre de personnes	Equilibrage sur distribution de revenus nombre de personnes revenus
Population	1.00	1.04	1.09	1.01	1.05	1.23
Taux de pauvreté	1.01	1.01	1.11	1.00	1.07	0.96
Niveau de vie	1.00	1.01	1.43	1.04	1.29	0.97
Revenus fiscaux	1.01	1.01	1.46	1.00	1.27	1.08
Aides logement	1.04	1.04	1.19	1.01	1.06	0.97
Minima sociaux	1.03	1.02	1.11	1.03	1.08	0.98

Note de lecture : Le coefficient de variation de l'estimateur du total de niveau de vie à Poitiers dans le scénario d'équilibrage spatial sur le nombre de personnes est $1,43$ *fois celui de l'estimateur du total de niveau de vie lorsqu'on réalise un tirage spatialement équilibré sur l'ensemble des variables d'équilibrage présentées précédemment.*

Le tableau 2 indique que, en raison du nombre d'unités tirées (50), l'équilibrage ne permet pas une grande précision y compris sur les premières variables d'équilibrage. Le tableau 3 montre que l'équilibrage spatial sur des variables de revenus et de nombre de personnes suffit à atteindre la performance d'un algorithme d'équilibrage spatial utilisant toutes les variables auxiliaires sus-mentionnées dans l'équilibrage. Ceci s'explique par le faible nombre d'unités tirées (50), ce qui ne permet pas un équilibrage de qualité à partir de nombreuses variables auxiliaires. Ainsi, les variables d'équilibrage retenues pour les travaux de simulations du chapitre 3 de la partie D sont le nombre de personnes dans le logement et les revenus du logement (sous forme de total ou de distribution (quartiles) selon les scénarios).

Annexe 3 - Analyse de la structure des échantillons de second degré obtenus par des tirages systématiques et par des tirages équilibrés

Les résultats présentés en section 3.2.6 de la partie D montrent que le tirage systématique engendre des estimateurs avec une moindre variance que le tirage équilibré. Il s'agit donc ici de comprendre les raisons pour lesquelles le tirage systématique présente de meilleures performances que le tirage équilibré.

Si les logements ont des probabilités de sélection égales d'une méthode à l'autre, rien n'assure que les probabilités d'inclusion doubles soient uniformément distribuées au sein des couples de résidences principales. Ce n'est de fait pas le cas pour le tirage systématique où l'on sélectionne des unités distantes les unes des autres dans la base triée au moyen d'un pas de tirage (cf. encadré de la section 3.1 de la partie D). Or, la composition de l'échantillon n'est pas anodine pour la déformation des poids dans l'algorithme de calage sur marges. Pour comprendre pourquoi, après calage sur marges, le tirage systématique présente de meilleurs résultats que le tirage équilibré, il convient donc d'étudier la structure des probabilités d'inclusion double.

Plutôt que de calculer les probabilités d'inclusion doubles des unités, ce sont les probabilités d'inclusion doubles d'unités selon leur tranche de revenus qui ont été calculées par simulation, c'est-à-dire la probabilité qu'une unité d'une tranche de revenus j se trouve dans le même échantillon qu'une autre unité appartenant à une tranche j' [20]. Pour chaque échantillon simulé, on comptabilise le nombre de couples d'une unité d'une tranche j et d'une unité d'une tranche j' rapporté à la taille des tranches. La moyenne obtenue sur les M simulations permet de fournir une estimation de la probabilité d'inclusion double moyenne entre les unités des tranches j et j'.

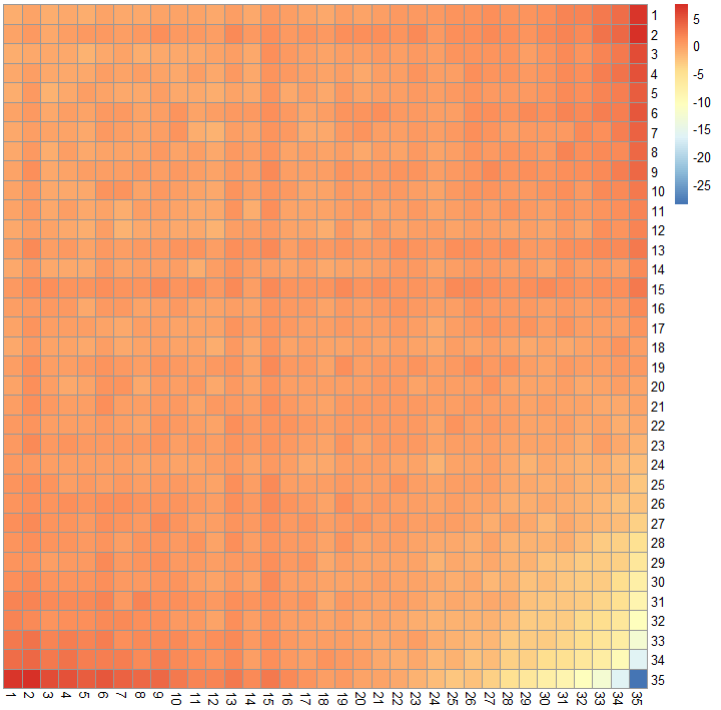
Dans le cas du sondage aléatoire simple, les probabilités d'inclusion double sont toutes égales à $\frac{n(n-1)}{N(N-1)}$ où n est la taille de l'échantillon et N la taille de la base de sondage. Nous pouvons alors mesurer les écarts relatifs des probabilités d'inclusion double estimées par rapport au sondage aléatoire simple afin d'identifier s'il y a une certaine attraction ou répulsion entre des tranches de revenus.

Les figures 4 et 5 correspondent, respectivement pour le cube 1 et le cube 2 présentés en section 3.2.3 de la partie D, aux matrices de chaleur des écarts relatifs à l'indépendance

20. Un quart des logements de la base de sondage ont un revenu fiscal nul, ils sont inclus dans la même tranche de revenu, la tranche 1. Les autres logements sont ensuite répartis en 34 tranches de revenus (tranche 2 à tranche 35) comprenant chacune environ 1 000 logements.

des probabilités d'inclusion double des unités appartenant aux différentes tranches de revenus [21].

FIGURE 4 – Probabilités d'inclusion double selon les tranches de revenus pour un tirage équilibré sur les niveaux de revenus, en écarts relatifs par rapport au sondage aléatoire simple

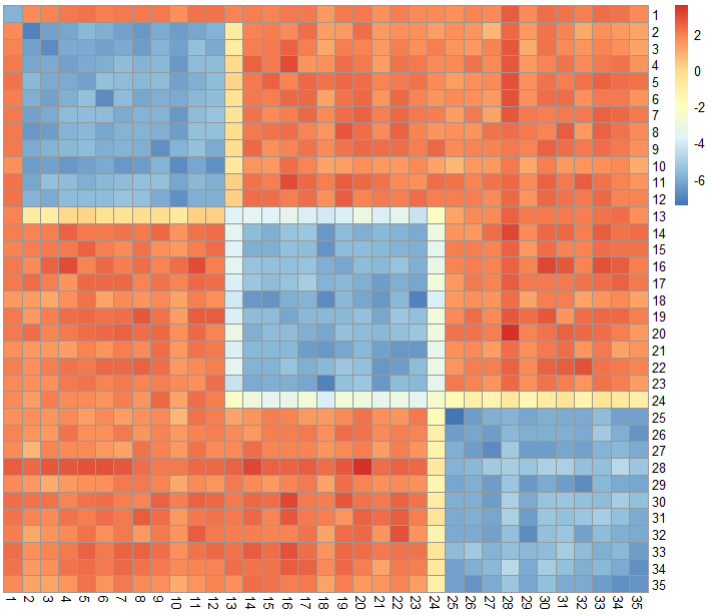


Note de lecture : Lorsqu'on réalise un tirage équilibré sur les niveaux de revenus, la probabilité qu'une unité au revenu nul (tranche 1) et qu'une unité dans la tranche de revenus la plus élevée (tranche 35) soient tirées simultanément dans l'échantillon est supérieure à 5% par rapport à la probabilité d'inclusion double qu'on observerait dans un sondage aléatoire simple.

Dans le cas du cube avec le revenu en niveau comme variable d'équilibrage (figure 4), on note un fort effet répulsif entre unités de la dernière tranche et répulsif en général entre les unités des tranches 30 et au delà. Dès lors qu'une unité de ces tranches est tirée, l'algorithme du cube a tendance à plus sélectionner les autres unités en dehors de ces tranches. Il y a également une nette attraction entre les unités des tranches 1 et 2 et celle de la tranche 35. En dehors de ces cas, les interactions entre tranches sont relativement proches de tirages indépendants (zone orangée). Le tirage équilibré sur les revenus en niveau conduit donc à des tirages qui ont soit tendance à surreprésenter les extrêmes de la distribution de revenus simultanément, soit tendance à les sous-représenter.

21. Les tirages équilibrés par la méthode du cube local présentent des structures similaires.

FIGURE 5 – Probabilités d'inclusion double selon les tranches de revenus pour un tirage équilibré sur la distribution de revenus, en écarts relatifs par rapport au sondage aléatoire simple



Note de lecture : Lorsqu'on réalise un tirage équilibré sur la distribution de revenus, la probabilité que deux unités au revenu non nul dans la tranche de revenus la plus faible (tranche 2) soient tirées simultanément dans l'échantillon est inférieure d'environ 6% par rapport à la probabilité d'inclusion double qu'on observerait dans un sondage aléatoire simple.

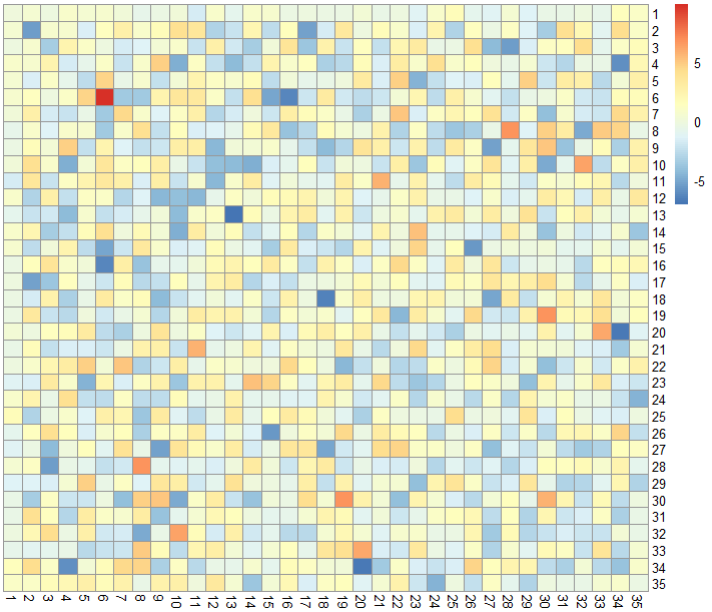
Lorsque qu'on s'intéresse au tirage équilibré sur la distribution des revenus (figure 5), la matrice de chaleur prend une forme très caractéristique où l'on identifie nettement les 4 quartiles de revenus. Dans ce cas l'algorithme du cube cherche à sélectionner des échantillons respectant la représentativité des quartiles de revenus de la population. De fait l'inclusion dans un échantillon d'une unité appartenant à un quartile donné réduit les chances de sélectionner une autre unité appartenant au même quartile de revenus dans la population. Il n'y a pas d'interaction particulière entre tranches de revenus appartenant à des quartiles différents.

En comparaison aux tirages équilibrés, la matrice de chaleur relative au tirage systématique (figure 6) ne laisse pas apparaître de phénomène d'interaction notable [22].

Ainsi, sans connaître l'impact de l'écart de structures de probabilités d'inclusion double dans ces différents tirages, on peut supposer que cela a une influence sur la pré-

22. Cette figure n'est toutefois pas sans évoquer la série de peintures "Colour Charts" réalisée par Gerhard Richter, bien que le lien avec la théorie des sondages reste encore à démontrer.

FIGURE 6 – Probabilités d'inclusion double selon les tranches de revenus pour un tirage systématique trié par revenus, en écarts relatifs par rapport au sondage aléatoire simple



Note de lecture : Lorsqu'on réalise un tirage systématique trié par revenus, et qu'on sépare les unités en 34 quantiles de revenus, la probabilité qu'une unité du 3^e quantile (tranche 4) et qu'une unité du 33^e quantile (tranche 34) soient tirées simultanément dans l'échantillon est inférieure d'environ 5% par rapport à la probabilité d'inclusion double qu'on observerait dans un sondage aléatoire simple.

cision de ces tirages. Les tirages systématiques et équilibrés présentent ici des variances similaires avant calage (cf. tableau 3.1). Empiriquement, d'après des résultats qui ne sont pas présentés dans ce document, il apparaît que, lors du calage sur marges, les rapports de poids sont plus déformés si l'échantillon a été obtenu par tirage systématique plutôt que par tirage équilibré. Ainsi, le gain de variance pour le tirage systématique semble se doubler d'une plus forte déformation de poids lors du calage sur marges que pour les échantillons issus d'un tirage équilibré. *In fine*, l'homogénéité des échantillons systématiques (figure 6), la performance similaire des algorithmes de tirage équilibré et de tirage systématique avant calage et la plus grande déformation de poids des échantillons obtenus par tirage systématique apparaissent interagir pour générer une variance moindre dans le tirage systématique combiné à du calage sur marges. Cette piste mériterait cependant d'être approfondie.