Journées de méthodologie statistique de l'Insee 2022

Méthodologie d'échantillonnage de la future Enquête Emploi en Continu à Mayotte

Julien Jamme (*), Cyril Favre-Martinoz (**), Édouard Fabre (***), Joachim Timotéo (***)

(*) Insee, Département des Méthodes Statistiques
(**) Insee, Département de la Démographie
(***) Insee, Direction Interrégionale La Réunion-Mayotte

julien.jamme@insee.fr

Mots-clés. (6 maximum) : Échantillonnage, équilibrage, collecte, variance.

Domaines. Théorie des sondages amont.

Résumé

Après une première expérience menée en 2009, l'enquête Emploi est réalisée tous les ans à Mayotte depuis 2013. Il s'agit d'une enquête annuelle réalisée de début mars à début juillet de chaque année. Aujourd'hui, le nouveau règlement européen (IESS), en vigueur depuis le 1er janvier 2021, impose que tout le territoire d'un pays soit couvert par une enquête Emploi en continu. La France a obtenu pour Mayotte une nouvelle dérogation pour la période 2021-2023. Elle est donc dans l'obligation de mettre en place une EEC opérationnelle au premier trimestre 2024. Des premiers travaux relatifs à l'échantillonnage ont été menés afin de respecter certaines contraintes méthodologiques, réglementaires et de les concilier avec les questions d'organisation de la collecte sur le terrain (nombre d'enquêteurs, charge de travail par enquêteur, secteurs d'enquête, spécificité de la collecte à Mayotte, etc.). Cet article a pour objectif de décrire en détails ces travaux. À Mayotte, la base de sondage retenue est la base cartographique. Des alternatives comme les sources administratives (Fidéli) et les enquêtes annuelles de recensement ont été envisagées mais écartées pour deux raisons : la première EAR date de 2021 et les sources administratives sur ce territoire ne sont pas exhaustives. La base cartographique de Mayotte est une base recensant l'ensemble des adresses présentes sur le territoire de Mayotte. Elle est mise à jour chaque année par l'intermédiaire d'une enquête cartographique qui a lieu sur un cinquième des îlots de Mayotte (correspondant à un groupe de rotation [1] donné et attribué à chaque îlot [2]).

1. Les groupes de rotation des îlots ont été constitués séparément pour les petites communes par le Criem et les grandes communes par la Division des Méthodes et Traitements du Recensement (DMTR) de l'Insee, tout en suivant une démarche et des méthodes similaires. Ils sont constitués successivement par un tirage équilibré sur des variables auxiliaires telles que le nombre de logements, le nombre de logements en dur, le nombre de logements collectifs, la population par âge, la population selon le lieu de naissance et une typologie des villages.

2. On appelle îlot une unité géographique regroupant un ou plusieurs pâtés de maisons et délimité par des voies (rues, routes, chemins) ou des barrières naturelles. Les 811 îlots actuels ont été constitués

Face à l'évolution très rapide du parc des logements, l'échantillonnage de l'enquête Emploi à l'année N se limite aux adresses qui ont été enquêtées par l'enquête cartographique à l'année N-1, donc sur un cinquième du territoire. La mise à jour de celle-ci ayant lieu fin octobre de l'année N-1, on procédera, en fin d'année N-1, au tirage de l'échantillon des trimestres T2, T3 et T4 de l'année N et du trimestre T1 de l'année N+1 dans la partie de la base cartographique qui aura été enquêtée au cours de l'année N-1.

Afin de répartir l'échantillon de logements à collecter dans l'espace et dans le temps, des Secteurs d'Activité Enquêteurs (SAE) ont été constitués par le Criem, Chaque secteur doit recevoir une partie égale de l'échantillon. La méthode consiste à regrouper les îlots en suivant un chemin issu de l'algorithme du voyageur du commerce. Cet algorithme permet de passer d'un îlot à l'un de ses voisins, chaque îlot étant visité une et une seule fois. Le regroupement des îlots est arrêté dès qu'un nombre de logements, fixé à l'avance, est atteint. L'objectif est de construire les SAE les plus homogènes possible en termes de nombre de logements sur chacun des groupes de rotation des îlots. Ensuite, le plan de sondage proposé devra respecter un certain nombre de contraintes :

— Tirer un échantillon dont la taille s'approche le plus possible des 936 logements entrants par an ;

— Obtenir une égale répartition de l'échantillon entre les SAE, soit 36 +/- 4 logements entrants chaque année ;

— Respecter la contrainte européenne de précision ;

— Faire en sorte que le plan de sondage soit autopondéré ;

— Obtenir la meilleure précision possible pour les estimateurs étudiés.

Ne disposant que d'une base de sondage d'adresses issue d'une enquête cartographique annuelle et de peu d'informations dans cette base, si ce n'est le nombre de logements à l'adresse, nous proposons un tirage stratifié des logements en deux étapes. La variable de stratification retenue est la variable de type d'adresse définie comme suit :

— Les Monologements (Mono) : une adresse = un logement ;

— Les Petites Adresses (PA) : une adresse = 2 à X logements ;

— Les Grandes Adresses (GA) : une adresse = plus de X logements.

Le choix du seuil X fera l'objet d'une discussion spécifique. Le tirage se décline en deux étapes :

— un tirage des logements en grandes adresses rassemblées en une seule strate sans distinction du SAE ;

— des tirages indépendants dans chaque croisement SAE*type d'adresse avec des allocations égales par secteur et fixées à l'avance.

Les simulations effectuées montrent que le scénario retenu permet de contrôler la taille de l'échantillon et la charge par enquêteur. La construction des SAE assure empiriquement l'autopondération du plan de sondage. La taille de l'échantillon fixée en amont est suffisante pour respecter la contrainte européenne de précision. Le taux de chômage annuel serait estimé avec une précision de +/-3.4 points de pourcentage. En parallèle, d'autres méthodes d'échantillonnage (tirage équilibré, spatialement équilibré et équilibré) ont été testées et les résultats sur les différents indicateurs d'intérêt (taille fixe, autopondération, précision) seront présentés.

à partir des données du millésime 2017 du recensement de la population. Ce découpage a été réalisé en 2020.

Abstract

In 2024, Insee must organise a Labour Force Survey (LFS) in Mayotte as required by the European Union's statistical office (Eurostat). To prepare the transition between the current annual labour survey to the continuing quarterly LFS survey, we propose a sampling design and we estimate the precision of the key indicators of the LFS. In this work, we perform only cross-sectional estimations, not longitudinal estimates.

As we deal with many constraints and limitations, presented in the second to the fourth sections of the article, we choose a very specific sampling design to especially ensure a good repartition of the sample in time and space. After the presentation of the sampling design in the fifth section, we provide our variance estimations based on Monte-Carlo simulations.

Introduction

En septembre 2021, l'Insee engagera les travaux de refonte de l'enquête Emploi organisée à Mayotte. Présent sur l'île depuis 1996, l'Insee y a réalisé une enquête Emploi ponctuelle en 2009. La départementalisation de l'île en 2011 suivie, en 2014, d'une reconnaissance de ce territoire comme partie intégrante de l'Union européenne[3] a conduit l'institut à mettre en place une enquête Emploi annuelle dès 2013[4]. Le règlement européen n°1991/2002[5] impose aux États d'organiser une Enquête Emploi en Continu (EEC) pour 2003 au plus tard. L'Insee respectera cette échéance pour la métropole. Le passage à l'EEC a eu lieu dans les Dom dits *historiques* (Guyane, Guadeloupe, Martinique et La Réunion) en 2013. Une première dérogation a été obtenue pour Mayotte en raison des difficultés d'organiser une enquête en continu.

Aujourd'hui, le vieillissement du questionnaire et de la chaîne des traitements aval (après-collecte) de l'enquête Emploi mahoraise rend nécessaire sa refonte et un nouveau règlement européen[6], en vigueur depuis le 1^{er} janvier 2021, impose que tout le territoire d'un pays soit couvert par une enquête Emploi en continu. La France a obtenu pour Mayotte une nouvelle dérogation pour la période 2021-2023. Elle est donc dans l'obligation de mettre en place une EEC opérationnelle au premier trimestre 2024. Le projet de refonte débutera en septembre 2021 et organisera des tests sur le terrain au cours de l'année 2022. Pour atteindre l'objectif, une montée en charge du nouvel échantillon débutera dès l'année 2023.

Depuis 2018, un groupe de travail, qu'on nommera par la suite GT EEC Mayotte, réunit, au sein de l'institut, les principaux acteurs de cette refonte pour « instruire et préparer le passage de Mayotte à l'EEC[7] » ce qui recouvre les questions d'organisation de la collecte sur le terrain (nombre d'enquêteurs, charge de travail par enquêteur, secteurs d'enquête, spécificité de la collecte à Mayotte, etc.) et de réflexion sur le plan de sondage à adopter. Dans le cadre de ce groupe de travail, le nôtre s'est concentré particulièrement sur l'échantillonnage de la nouvelle enquête afin d'étudier différents plans de sondage possibles et de proposer une estimation de la précision des estimateurs des paramètres d'intérêt les plus importants de l'enquête Emploi. La contrainte de précision portant sur des estimateurs transversaux, nous ne proposons pas dans ce travail d'estimations de la précision des estimateurs longitudinaux.

Après une présentation de l'enquête Emploi et de son organisation en métropole et dans les Dom, nous décrirons les choix du GT EEC Mayotte qui s'imposent à nous, ainsi que les contraintes de coûts, d'organisation et de méthodologie que les plans de sondage envisagés devront respecter. Il sera temps, alors, de présenter les scénarios envisagés. Nous ne présenterons exhaustivement dans cet article que le scénario de base qui nous a servi de point de départ et de référence et le scénario final retenu par le GT EEC Mayotte. Nous décrirons ensuite le cadre général de nos estimations, ainsi que les mécanismes de réponse simulés et les techniques de redressement des poids de sondage utilisées. Nous terminerons en fournissant les résultats chiffrés de nos estimations par simulation permettant de justifier le choix du plan de sondage et de proposer une estimation de la précision de l'estimation du taux de chômage.

3. Les Dom dont Mayotte sont considérés comme des régions ultra-périphériques (RUP) par le droit européen, c'est-à-dire des territoires extérieurs au continent européen mais sur lesquels les règlements européens doivent s'appliquer.

4. [1]

5. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32002R1991&qid=1621262536746`

6. Règlement n°2019/1700, `https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32019R1700`

7. [2], p. 1

Mayotte en quelques chiffres et deux cartes

Mayotte est une île de l'Océan Indien située dans le canal du Mozambique entre l'île de Madagascar et la côte sud-est de l'Afrique. Le chef-lieu du département est situé à Mamoudzou.

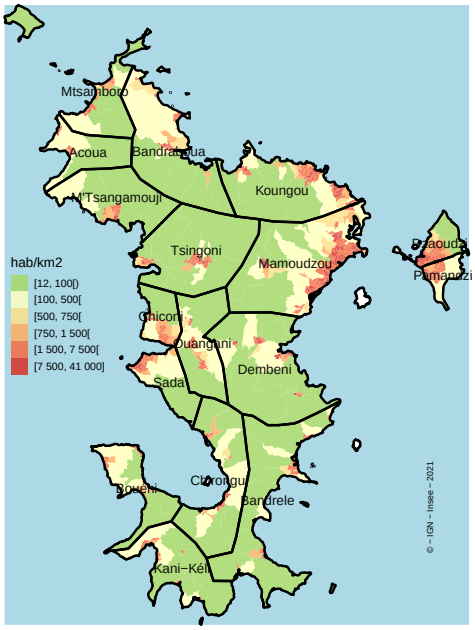
Avec 256 500 habitants au recensement de 2017, Mayotte est le département le plus dense de France hors Île-de-France. La population est très concentrée le long des côtes et particulièrement sur les communes de Mamoudzou et Koungou au nord-ouest (fig.1). Plus de 100 000 habitants résident dans ces deux communes. Le cartogramme de la figure 2, en déformant la taille des îlots en fonction de la population en âge de travailler, montre l'excroissance de la partie nord-ouest et le rétrécissement du sud et de l'est de l'île.

Sur cinq ans, la population a crû de +3,8% par an quand l'ensemble de la population française a augmenté en moyenne de +0,4% par an sur la même période. La croissance démographique provient à la fois d'un fort excédent naturel et d'un excédent migratoire, provenant principalement des îles voisines des Comores.

Parmi les 63 000 résidences principales que compte l'île, « les constructions fragiles (maisons en tôle, bois, végétal ou terre)constituent près de quatre logements sur dix, comme 20 ans auparavant », note le service régional de Mayotte ([3]). Un accès à l'eau courante est absent de près d'un tiers des logements.

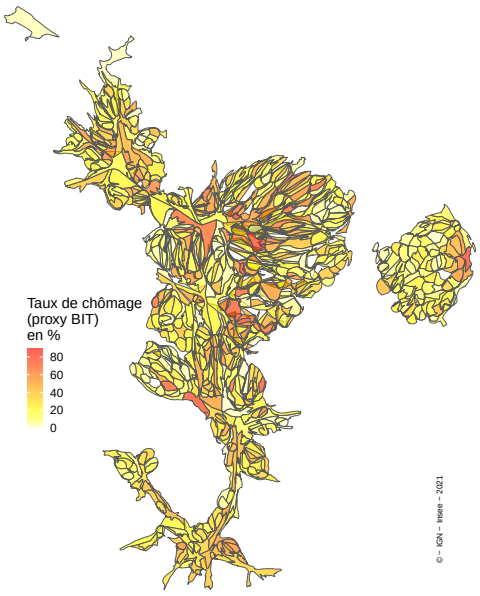
L'île est marquée par un fort taux de pauvreté : en 2018, 77% de la population mahoraise vit en-dessous du seuil de pauvreté. En 2017, 69% de la population de 15 ans ou plus non scolarisée n'a aucun diplôme (28% en métropole). D'après l'enquête Emploi 2019, on dénombre 22 500 chômeurs au sens du BIT et 52 200 actifs en emploi, soit un taux de chômage annuel de 30,1%.

FIGURE 1 – Densité de la population de 15 ans ou plus par îlot en 2017



Source : RP 2017. Note : Les frontières correspondent aux limites communales.

FIGURE 2 – Cartogramme - Taux de chômage par îlot en 2017



Source : RP 2017. Note : Les îlots ont été déformés proportionnellement à la population de 15 ans ou plus qui y réside.

1 L'enquête Emploi

L'enquête Emploi est « l'unique source permettant de mesurer l'emploi et le chômage suivant les concepts du Bureau international du travail (BIT) » [8]. Créée en 1950 sous la forme d'une enquête annuelle voire semestrielle, l'enquête Emploi est une enquête en continu en France métropolitaine depuis 2003 et dans les Dom, hors Mayotte, depuis 2013.

Le champ de l'enquête est constitué de l'ensemble « des personnes vivant en logement ordinaire et résidant habituellement en France » [9]. Les communautés telles que les casernes, les foyers, les établissements hospitaliers, etc. ne font pas partie du champ.

L'échantillon est un ensemble de logements tirés pour former un panel rotatif en 6 vagues d'interrogation trimestrielles. Ainsi, un logement est enquêté six trimestres consécutivement avant de sortir de l'enquête. Chaque trimestre, un sixième de l'échantillon est renouvelé. En tant qu'enquête *en continu*, il est fait en sorte que, chaque semaine de l'année, une partie de l'échantillon soit interrogée.

1.1 Le passage à l'EEC dans les DOM *historiques* **on entend par DOM historiques : La Guadeloupe, La Martinique, La Guyane et La Réunion en 2013, un point de départ**

Le passage d'une enquête annuelle à l'enquête Emploi en continu a été réalisé dans les Dom dits « historiques » en 2013. Nous reprenons de ([1]) [10] les points essentiels de la stratégie d'échantillonnage utilisée dans ces Dom, car elle constitue un point de départ aux réflexions menées pour le passage à l'EEC à Mayotte.

D'une part, les sources fiscales étant de moins bonne qualité dans les Dom, les Enquêtes Annuelles de Recensement (EAR) ont été préférées à la base Fidéli. Chaque année, $1/5^e$ des petites communes est recensé et $1/8^e$ des logements de chaque grande commune est enquêté. Malgré un relatif manque de fraîcheur des EAR les plus anciennes (datant de cinq ans pour la plus ancienne), la base de sondage est constituée de l'empilement des cinq EAR composant un cycle complet de l'enquête du recensement de la population pour les petites communes. Dans les grandes communes, celles-ci étant enquêtées tous les ans, seules deux EAR sont empilées pour constituer la base de sondage. L'enquête Emploi dans les Dom bénéficie ainsi, avec le recensement de la population (RP), d'une base de sondage très riche. La différence avec Fidéli est un certain manque de fraîcheur d'une partie des données.

Le passage d'une enquête annuelle à une enquête trimestrielle en continu augmente la charge d'enquête et génère des coûts supplémentaires importants pour préserver le niveau de précision des estimations annuelles. Pour respecter les contraintes de précision européennes tout en contrôlant la hausse des coûts d'enquête, l'Insee a choisi une stratégie d'échantillonnage différente de celle adoptée en France métropolitaine. L'effet grappes étant une source importante d'imprécision et l'augmentation de la taille de l'échantillon pour le contrôler étant trop coûteuse, il a été décidé de disperser l'échantillon sur l'ensemble du territoire [11].

Néanmoins, autant pour l'organisation de la collecte et la répartition de la charge entre les enquêteurs que pour assurer la continuité de l'enquête, le territoire a été découpé en 26 Secteurs d'Activité des Enquêteurs (SAE) entre lesquels l'échantillon est réparti le plus également possible. Chaque secteur est associé à une semaine de référence de chaque trimestre de l'année et est attribué idéalement à un enquêteur.

8. [4], p.1. Toute cette partie est rédigée à partir de cette note méthodologique mise à jour le 13 juillet 2021.

9. un logement ordinaire étant défini comme « un local séparé et indépendant utilisé pour l'habitation » [4], p.9

10. Notamment la section 4 « *Le nouveau dispositif en septembre 2013 pour les* Dom *« historiques »*, pp.54-65

11. [1], p.54

Ces secteurs, qui forment une partition du territoire, ont été construits « en secteurs réputés équivalents » [12]. Cette construction a ainsi fait face à « une logique théorique de construction de strates de tirage[,] les strates d[evant] abriter des populations aussi homogènes que possible »ainsi qu'à « une logique de construction de secteurs de collecte [,]les secteurs (...) d[evant] être aussi homogènes que possible en charge et en accessibilité » [13]. Nous retrouverons ces deux logiques à l'œuvre dans la construction des SAE à Mayotte.

Dans chaque Dom hors Mayotte, environ 2 000 logements sont enquêtés chaque trimestre.

1.2 L'enquête Emploi annuelle à Mayotte

Après une première expérience menée en 2009, l'enquête Emploi est réalisée tous les ans à Mayotte depuis 2013. Il s'agit d'une enquête annuelle réalisée au second trimestre de chaque année. L'échantillon est un panel rotatif composé de 3 vagues d'interrogation : chaque logement enquêté est interrogé trois années consécutives avant de sortir de l'échantillon. Chaque année, l'échantillon est donc renouvelé au tiers.

La base de sondage utilisée ne provient ni des sources fiscales ni du recensement de la population mais d'une base d'adresses issue d'une enquête cartographique lors de laquelle un cinquième des îlots sont enquêtés. Celle-ci est dépourvue de variables caractérisant les ménages : elle contient uniquement quelques informations sur le bâti (type de construction, nombre de logements) et des caractéristiques géographiques. L'échantillon comprend environ 3 000 logements et le taux de chômage annuel est environ estimé avec une précision de $\pm 2,0$ points de pourcentage. Le taux de collecte, c'est-à-dire la part de logements répondants parmi l'ensemble des logements enquêtés - y compris les logements hors-champ - s'élève à 69% ([5], p.2).

La stratégie adoptée par le groupe de travail est de repartir de la méthodologie employée en 2013 lors du passage à l'enquête Emploi en continu dans les Dom *historiques*. L'échantillonnage nécessite la construction préalable de secteurs d'enquête - les SAE - entre lesquels l'échantillon doit être réparti le plus également possible. Ces secteurs doivent permettre d'assurer une bonne dispersion de l'échantillon sur l'ensemble du territoire tout en facilitant l'organisation de la collecte - en associant un secteur à un enquêteur - et en assurant la bonne répartition de l'échantillon par semaine de référence - en associant un secteur à une semaine du trimestre.

Néanmoins, l'inadéquation de bases de logements telles que le recensement ou Fidéli comme bases de sondage rend inévitable l'utilisation d'une base d'adresses issue de l'enquête cartographique.

2 La base de sondage

La seule base de sondage disponible pour initier l'enquête Emploi en continu, comme les autres enquêtes ménages à Mayotte, est la base issue d'une enquête cartographique réalisée annuellement à Mayotte. Les sources fiscales ne sont en effet pas d'assez bonne qualité : elles présentent une forte sous-couverture en termes de logements et d'individus. Lors du passage à l'enquête Emploi en continu dans les autres Dom en 2013, c'est un empilement d'enquêtes annuelles du recensement qui avait été utilisées. À Mayotte, la première enquête annuelle de recensement n'ayant eu lieu qu'en début d'année 2021, un cycle complet du recensement de la population n'aura pas été réalisé courant 2022 lorsqu'il faudra tirer les premiers échantillons de la nouvelle enquête.

12. [1], p.55
13. *ibid.*

2.1 La fraîcheur de l'information cartographique

L'enquête cartographique permet de répertorier une partie des adresses de l'île chaque année. Le territoire est divisé en cinq groupes de rotation, chacun mobilisé une fois sur un cycle de cinq ans. La base cartographique est donc composée de données plus ou moins fraîches, certaines adresses ayant été enquêtées jusqu'à cinq ans auparavant. Face à l'évolution très rapide du parc des logements, l'échantillonnage de l'enquête Emploi à l'année N se limite aux adresses qui ont été enquêtées par l'enquête cartographique à l'année N-1. On utilisera la même base de sondage pour l'EEC.

La mise à jour de la base cartographique ayant lieu fin octobre de l'année N-1, l'idée est de procéder, en fin d'année N-1, au tirage de l'échantillon des trimestres T2, T3 et T4 de l'année N et du trimestre T1 de l'année N+1 dans la partie de la base cartographique qui aura été enquêtée au cours de l'année N-1.

Ainsi, parmi les échantillons entrants à l'année N, l'information cartographique sur les adresses enquêtées en T1 de l'année N+1 datera de plus d'un an. Chaque échantillon entrant étant interrogé six trimestres consécutifs, cet échantillon entrant au T1 de l'année N+1 sera interrogé jusqu'au T2 de l'année N+2 et donc basé sur une information cartographique vieille de deux ans et demi lors de la dernière interrogation (voir schéma 3).

2.2 Une première phase de tirage : les groupes de rotation des îlots

Pour mener l'enquête cartographique, une première phase de tirage est nécessaire pour déterminer la composition des cinq groupes de rotation d'îlots.

On appelle « îlot »une unité géographique regroupant un ou plusieurs pâtés de maison et délimité par des voies (rues, routes, chemins) ou des barrières naturelles. Les 811 îlots actuels ont été constitués à partir des données du millésime 2017 du recensement de la population [14]. Ce découpage a été réalisé en 2020. La démarche a consisté, à partir de zones élémentaires très fines, à construire des îlots les plus homogènes possibles en termes de taille (autour de 80 résidences principales) [15]. Les îlots ont été divisés en cinq groupes de rotation : ceux-ci sont nécessaires pour la mise en place de l'enquête annuelle de recensement dans les grandes communes [16]où seuls 8% des logements sont enquêtés tous les ans et tirés dans un seul groupe de rotation (soit environ 40% sur un cycle complet, hors habitations de fortune). Le découpage des petites communes en îlots est nécessaire pour mener l'enquête cartographique dont dépend les enquêtes ménages à Mayotte.

Les groupes de rotation des îlots ont été constitués séparément pour les petites communes par le Criem ([6]) et les grandes communes par la Division des Méthodes et Traitements du Recensement (DMTR) de l'Insee ([7]), tout en suivant une démarche et des méthodes similaires. Ils sont constitués successivement par un tirage équilibré sur des variables auxiliaires telles que le nombre de logements, le nombre de logements en dur, le nombre de logements collectifs, la population par âge, la population selon le lieu de naissance et une typologie des villages [17]. La principale différence entre les tirages pour les petites et les grandes communes est que le tirage est effectuée sur l'ensemble des petites communes et séparément dans chacune des grandes communes [18].

14. Le recensement de la population était quinquennal et exhaustif jusqu'au millésime 2017.

15. voir la partie 2 de la note [6], partie 2

16. Au sens du recensement de la population, une grande commune est une commune de 10 000 habitants ou plus.

17. Les variables d'équilibrage ne sont pas exactement identiques pour les petites et les grandes communes. La liste complète est disponible dans [6], p.12, pour les petites communes, et dans [7], p.3, pour les grandes.

18. Deux petites communes, Chiconi et Chirongui, étant à la limite du seuil des 10 000 habitants, ont

2.3 Une base d'adresses pauvre en information

La base cartographique est une base d'adresses et non de logements. En tirant des adresses pour échantillonner des logements, aucune stratégie d'échantillonnage viable sur le terrain ne pourra assurer de tirer un échantillon de logements de taille fixe. Il sera important de maîtriser les fluctuations de la taille de l'échantillon pour que la charge d'enquête soit la plus constante possible dans le temps. Ces fluctuations seront en partie limitées par le fait qu'une grande partie des logements sont des logements individuels, appelés aussi « monologements ».

Les principales informations disponibles dans la base cartographique et qui peuvent s'avérer utiles pour l'échantillonnage sont :

— le nombre de logements à l'adresse ;

— l'aspect de la construction : on distinguera notamment les habitations « en dur », dont les fondations et murs principaux sont fabriqués en béton, briques, parpaings, etc., des autres constructions (en bois, en tôle ou les cases traditionnelles) ;

— la localisation géographique (coordonnées géographiques, îlot, village, commune).

2.3.1 Une stratification par type d'adresse

Dans les enquêtes ménages menées à Mayotte, trois strates d'adresses sont constituées en fonction de leur taille. La variable *type d'adresse* est définie comme suit :

- Les Monologements (Mono) : une adresse = un logement ;

- Les Petites Adresses (PA) : une adresse = 2 à 10 logements ;

- Les Grandes Adresses (GA) : une adresse = plus de 10 logements.

Cette catégorisation n'étant pas vraiment le fruit d'une expertise préalable, nous interrogerons la pertinence du seuil entre petites et grandes adresses dans nos simulations pour proposer son éventuel abaissement.

2.4 Construction d'une information auxiliaire : typologie des îlots

Sur le modèle de la typologie des villages construite par le service régional de l'Insee à Mayotte ([8]), nous avons constitué une typologie des îlots afin de récupérer un peu d'information auxiliaire susceptible d'améliorer la précision de nos estimateurs si cette variable s'avère relativement bien corrélée aux principaux indicateurs de l'enquête Emploi. La typologie des villages ne convenait pas, car le découpage en villages [19] est un peu trop grossier pour nos besoins. En revanche, nous avons repris des indicateurs similaires tels que le taux de chômage au sens du RP, la part des moins de 18 ans et des plus de 65 ans, la part des logements en dur ou encore la part de logements disposant d'un accès à l'eau à l'intérieur du domicile, entre autres [20].

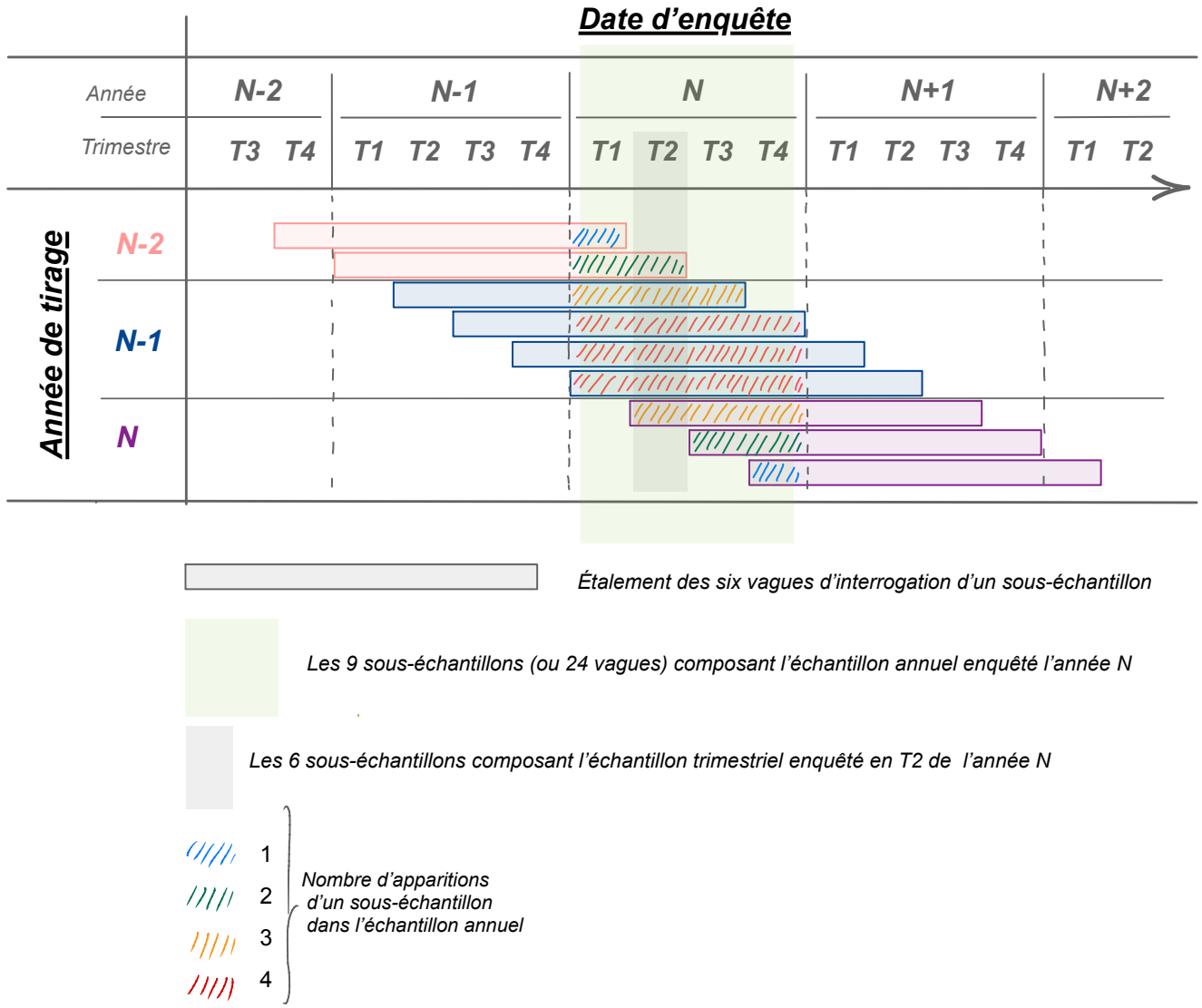
La typologie des îlots utilisée dans la suite de nos travaux est issue d'un partitionnement par la méthode des k-means en 6 classes. Des précisions et des résultats plus précis sur la méthode sont fournis en annexe B. Cette typologie s'avère bien corrélée aux paramètres clés de l'EEC.

également été traitées séparément, afin de limiter les déséquilibres entre les groupes de rotations en cas de franchissement de seuil à la hausse.

19. Chaque commune mahoraise est composée de plusieurs villages. L'île compte 17 communes et 72 villages.

20. La liste exhaustive est fournie en annexe B

FIGURE 3 – Le roulement des échantillons dans l'enquête Emploi en continu



Note de lecture : Le premier rectangle horizontal rose correspond à la partie de l'échantillon tiré l'année N-2 entrée dans l'enquête au T4 de l'année N-2. Ce sous-échantillon est interrogé six trimestres consécutifs, c'est-à-dire jusqu'au T1 de l'année N. Les 3 sous-échantillons tirés en N-1 et entrants dans l'enquête entre le T3 de l'année N-1 et le T1 de l'année N sont enquêtés les 4 trimestres de l'année N. Ils sont donc sollicités chacun 4 fois (hachures rouges) dans l'échantillon annuel (couleur de fond en vert clair).

Nous pouvons retenir pour la suite que :

- un îlot regroupe environ 80 résidences principales ;
- les îlots sont rassemblés en cinq groupes de rotation équilibrés en termes de nombre de logements, de type de construction, de population par âge et selon le lieu de naissance ;
- le tirage des groupes de rotation constitue une première phase d'échantillonnage non modifiable ;
- la base de sondage est constituée de l'ensemble des adresses localisées dans un cinquième des îlots ;
- pour la phase de tirage dans la base d'adresses, peu d'information est disponible ;
- nous reprenons le type d'adresse (MONO, PA et GA) comme principale variable de stratification ;
- la pertinence du seuil séparant petites et grandes adresses sera étudiée ;
- une typologie des îlots a été construite pour bénéficier d'une information auxiliaire liée aux paramètres clés de l'EEC.

3 Des *Secteurs d'Activité des Enquêteurs* (SAE), pour organiser la collecte

Pour assurer une bonne répartition de la charge de l'enquête entre enquêteurs et par semaine de l'année, le territoire de chaque Dom historique a été partitionné en 26 secteurs les plus homogènes possibles. Le même type de partitionnement a été réalisé pour Mayotte. Un secteur d'activité est ainsi une zone au sein de laquelle l'enquêteur aura un ensemble de logements à enquêter une semaine de référence donnée et sur laquelle il aura trois semaines pour réaliser la collecte.

3.1 Les contraintes pour leur constitution

Les SAE doivent remplir un certain nombre de contraintes. D'un point de vue organisationnel, les SAE doivent pouvoir faciliter le suivi de la collecte par la division Enquête Ménages du Service Régional, par exemple pour assurer les remplacements des enquêteurs.

Il est également important de prendre en compte les contraintes du terrain comme tous les obstacles qui réduiraient la vitesse de déplacement des enquêteurs. Un SAE étant une zone de collecte pour un enquêteur, elle doit lui permettre de réaliser son travail dans le temps qui lui est imparti, soit trois semaines pour une semaine de référence donnée. Elle devra ainsi être suffisamment compacte pour lui faciliter le travail. En l'absence de distancier de qualité disponible pour Mayotte, nous nous sommes appuyés sur une matrice de distance routière entre villages calculées à partir de l'API de *Google Maps*[21] et corrigée par le service régional de Mayotte.

Sur le plan méthodologique, les SAE doivent avoir des tailles semblables en termes de nombre de logements pour chacun des groupes de rotation d'îlots. Enfin, pour anticiper sur de futurs travaux ou enquêtes, on cherchera à produire des secteurs équilibrés aussi en termes de nombre de résidences principales et de résidences louées par groupe de rotation.

21. *Distance Matrix API*, `https://developers.google.com/maps/documentation/distance-matrix`

3.2 La construction des SAE

La construction des Secteurs d'Activité des Enquêteurs a été menée par le Criem en collaboration avec le service régional de Mayotte qui a été sollicité pour les arbitrages finaux [22].

La méthode consiste à regrouper les îlots en suivant un chemin issu de l'algorithme du voyageur du commerce. Cet algorithme permet de passer d'un îlot à l'un de ses voisins, chaque îlot étant visité une et une seule fois. Le regroupement des îlots est arrêté dès qu'un nombre de logements par groupe de rotation, fixé à l'avance, est atteint. L'objectif est de construire les SAE les plus homogènes possible sur chacun des groupes de rotation des îlots. La méthode étant sensible au point de départ et au seuil utilisé, de nombreuses itérations sont réalisées. Sont retenus les jeux de 26 secteurs les plus homogènes autour du seuil sur chaque groupe de rotation. L'étendue spatiale des SAE et les temps de collecte estimés sont deux autres critères que le jeu de SAE retenu devra minimiser. La contrainte de contiguïté a été relâchée, essentiellement pour atteindre les cibles par groupe de rotation : les SAE sont donc composés d'îlots non nécessairement contigus.

Le service régional de Mayotte a alors fait parler sa connaissance du terrain pour déterminer le meilleur découpage parmi les solutions retenues, proposer des retouches manuelles et améliorer le découpage final. Celui-ci a été validé courant mai par le GT EEC Mayotte et il est représenté sur la carte 4. La carte 5 montre la répartition des îlots du groupe de rotation 1 et leur SAE d'appartenance (en couleur).

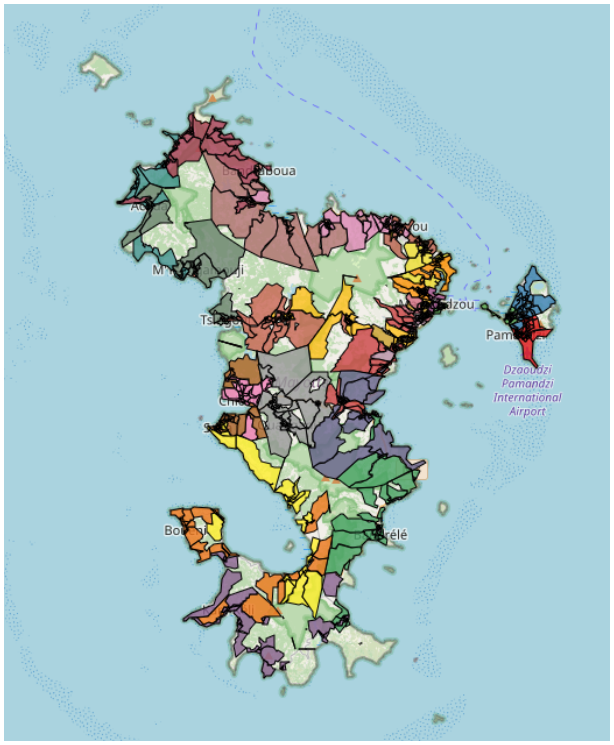


Figure 4 – Les Secteurs d'activité des enquêteurs (jeu retenu)
Source : Insee, Criem, *2021* - Les couleurs correspondent aux SAE. Les frontières des îlots correspondent aux bordures noires. Les zones hors SAE sont des zones inhabitées de l'île.

Figure 5 – Les Secteurs d'activité des enquêteurs (jeu retenu), restreints aux îlots du premier groupe de rotation
Source : Insee, Criem, *2021*

22. La méthode a été explicitée lors de la restitution des travaux du groupe de travail GT EEC Mayotte le 25 juin 2021.

4 Les contraintes sur le plan de sondage

Le plan de sondage retenu doit remplir un certain nombre de contraintes.

4.1 Contrainte de précision

Pour l'enquête menée à Mayotte, la contrainte européenne porte sur l'estimateur trimestriel de la part de chômeurs dans la population âgée de 15 à 74 ans. La contrainte est présentée sous la forme d'un écart-type maximal qui dépend à la fois de la valeur de l'estimateur et d'une fonction de la population d'intérêt (voir Annexe A, pour plus de détail).

Avec l'enquête Emploi annuelle, la part de chômeurs parmi les 15-74 ans est estimée autour de 12,5% et on dénombre 140 860 personnes âgée de 15 à 74 ans à Mayotte selon le recensement de la population 2017. La contrainte européenne correspond alors à un écart-type de 1,35 points de pourcentage soit une précision de $\pm 2,7$ points de pourcentage, pour un niveau de confiance de 95%. Cette contrainte étant définie en fonction de la part estimée de chômeurs, nous serons amené à la recalculer pour convenir au contexte de nos simulations.

4.2 Taille de l'échantillon

Lors de travaux simulatoires menés pour le groupe de travail en charge d'explorer le passage à l'EEC à Mayotte, le Criem a estimé la taille d'échantillon nécessaire pour atteindre strictement cet objectif de précision. À l'aide d'un plan de sondage rudimentaire, les simulations menées par le Criem ont conduit à estimer que 1 400 logements échantillonnés sur un trimestre étaient nécessaires pour atteindre l'objectif européen, en appliquant le taux de collecte de l'enquête annuelle (69%).

Pour des raisons pratiques, nous avons choisi de travailler avec un échantillon de 1 404 logements. Ce chiffre a l'avantage d'être divisible par 26 - nombre de secteurs enquêteurs (SAE) - et par 6 - nombre de vagues d'interrogation.

Chaque trimestre, un sixième de l'échantillon est renouvelé, soit $1404/6 = 234$ logements entrants. Le tirage aura lieu une fois par an. Le producteur devra tirer les quatre sous-échantillons d'entrants, soit $234 * 4 = 936$ logements au total (voir schéma 3).

Table 1 – Taille de l'échantillon de l'enquête Emploi en continu à Mayotte

Échantillon trimestriel global	1404
Échantillon annuel global	5616
Échantillon trimestriel entrant	234
Échantillon annuel entrant	936
Échantillon trimestriel entrant par SAE	9
Échantillon annuel entrant par SAE	36

Des pistes écartées

L'idée, par exemple, d'atteindre le même niveau de précision que l'enquête annuelle sur le taux de chômage n'a pas été retenue ($\pm 2,0$ points). Elle conduisait, en effet, à plus que doubler la taille de l'échantillon annuel actuel en passant de 3 000 à plus de 7 000 [23], ce qui n'était pas raisonnable pour des raisons principalement budgétaires et d'organisation de la collecte.

L'idée de travailler à égalité de moyens avec les autres Dom a aussi été discutée. Ceux-ci disposent de 5 équivalents temps plein (ETP) enquêteurs. Une telle ressource aurait conduit à

23. [5], p.2

tirer un échantillon annuel de 8 000 logements [24], ce qui était encore moins raisonnable d'un point de vue budgétaire et d'organisation de la collecte.

4.3 Contraintes méthodologiques

Une forte contrainte d'organisation de la collecte est imposée par le règlement européen régissant l'organisation de la *Labor Force Survey* (LFS) dont l'enquête Emploi en continu est la déclinaison française. En effet, ce règlement impose que chaque semaine de l'année un nombre de logements équivalents soient enquêtés. Un trimestre étant composé de 13 semaines en général, 2 SAE pourront être enquêtés chaque semaine de référence. Cette contrainte est donc absorbée par la nécessité de répartir de manière égale l'échantillon d'entrants entre SAE, construits en partie pour répondre à cette contrainte.

Enfin, on cherchera à construire *un plan de sondage autopondéré*, c'est-à-dire un plan dans lequel les logements ont la même probabilité d'être tiré. L'autopondération permet d'assurer un bon équilibre de la base de sondage dans le temps. En effet, un logement enquêté six trimestres consécutifs ne peut plus être rééchantillonné. Si tous les logements ont le même le poids, la partie restante de la base de sondage représente correctement l'ensemble de la population de départ, c'est-à-dire tous les logements y compris ceux déjà enquêtés. En outre, avec une moindre dispersion des poids de sondage (voire aucune dans le cas d'une parfaite autopondération), les estimateurs gagnent en général en précision.

4.4 Priorité des contraintes

Toutes ces contraintes ne sont pas facilement compatibles entre elles : répartir également l'échantillon sur l'ensemble des SAE ne facilite pas la recherche d'un plan autopondéré, par exemple. C'est pourquoi, nous sommes tenus d'ordonner ces contraintes pour indiquer l'ordre de priorité qui, *in fine*, permettra de décider entre les différents scénarios envisagés.

Dans un contexte purement théorique, la contrainte de précision minimale serait bien entendu la plus importante. Elle nous paraît, ici, passer après le respect de certaines contraintes liées à l'organisation de la collecte. La répartition équilibrée de l'échantillon sur l'ensemble des semaines de référence est requise par le règlement européen fixant la méthode de l'EEC. Ayant fait le choix d'associer une semaine de référence à un secteur d'enquête, la maîtrise de la taille de l'échantillon par SAE est dès lors prioritaire sur toutes les autres. L'autre contrainte européenne requise est une contrainte de précision. Si sa priorité est très élevée, cette contrainte pèsera peu car nous verrons qu'elle est largement respectée par l'ensemble des scénarios envisagés. La recherche d'un plan autopondéré vient ensuite et est prioritaire sur l'objectif d'atteindre le meilleur plan de sondage en termes de précision des estimateurs afin de maintenir de façon pérenne l'équilibrage de la base de sondage.

L'ordre décroisssant de priorité des objectifs est donc le suivant :

- 1. Tirer un échantillon dont la taille s'approche le plus possible des 936 logements entrants par an ;
- 2. Obtenir une égale répartition de l'échantillon entre les SAE, soit 36 ± 4 logements entrants chaque année ;
- 2. Respecter la contrainte européenne de précision ;
- 3. Faire en sorte que le plan de sondage soit autopondéré ;
- 4. Obtenir la meilleure précision possible pour les estimateurs étudiés.

24. *ibid.*, p.3

Principales notations

La présentation des différents plans de sondage envisagés nécessite quelques notations dont les plus utilisées sont rassemblées ici.

Les notations principales utilisées dans ce document sont les suivantes :

- L, seuil de logements pour différencier petites et grandes adresses ;
- U^a, la population d'adresses, de taille N^a ;
- U^a_{mo}, la population des monologements, de taille N^a_{mo} ;
- U^a_{pa}, la population des petites adresses, de taille N^a_{pa} ;
- U^a_{ga}, la population des grandes adresses, de taille N^a_{ga} ;
- S^a, un échantillon d'adresses, de taille n^a ;
- S^a_{mo}, un échantillon de monologements, de taille n^a_{mo} ;
- S^a_{pa}, un échantillon de petites adresses, de taille n^a_{pa} ;
- S^a_{ga}, un échantillon de grandes adresses, de taille n^a_{ga} ;
- $\pi^a_k = \mathbb{P}(k \in S^a)$, la probabilité d'inclusion de l'adresse k dans l'échantillon S^a ;
- $\mathbf{I}^a_k = \mathbb{1}(k \in S^a)$, l'indicatrice d'appartenance de l'adresse k à l'échantillon S^a ;
- N^l_k, le nombre de logements à l'adresse k ;

Nous utilisons des notations similaires pour les logements :

- U^l, la population d'adresses, de taille N^l ;
- U^l_{mo}, la population des monologements, de taille N^l_{mo} ;
- U^l_{pa}, la population des logements en petites adresses, de taille N^l_{pa} ;
- U^l_{ga}, la population des logements en grandes adresses, de taille N^l_{ga} ;
- S^l, un échantillon de logements, de taille n^l ;
- S^l_{mo}, un échantillon de monologements, de taille n^l_{mo} ;
- S^l_{pa}, un échantillon de logements en petites adresses, de taille n^l_{pa} ;
- S^l_{ga}, un échantillon de logements en grandes adresses, de taille n^l_{ga} ;
- $\overline{N}^l_{pa}$, le nombre moyen de logements par petite adresse ;
- π^l_l, la probabilité d'inclusion du logement l dans l'échantillon S^l ;
- $\mathbf{I}^l_l$, l'indicatrice d'appartenance du logement l à l'échantillon S^l ;

Nous déclinons toutes ces notations par SAE en indexant par un s. Par exemple :

- U^a_s, la population d'adresses dans le SAE s, de taille N^a_s ;
- U^l_s, la population de logements dans le SAE s, de taille N^l_s ;
- S^a_s, un échantillon d'adresses du SAE s, de taille n^a_s ;
- $\pi^a_{s,k}$, la probabilité d'inclusion de l'adresse k du SAE s dans l'échantillon S^a ;
- etc. ;

Des notations complémentaires seront signalées quand leur usage sera nécessaire.

5 Les plans de sondage envisagés

Nous présentons ici les plans de sondage envisagés pour échantillonner l'enquête Emploi en continu à Mayotte en respectant au mieux les contraintes exposées plus haut.

Pour nous servir de point de référence, nous reprenons le plan de sondage utilisé lors des premières simulations lancées par le CRIEM pour l'estimation de la taille de l'échantillon permettant d'atteindre la précision européenne requise.

Dans la présentation des scénarios, *nous ne revenons pas sur la première phase de tirage* des groupes de rotation des îlots, décrite plus haut et sur laquelle nous n'avons pas la main. Néanmoins, nous la réintègrerons dans nos simulations pour mesurer toutes les sources de variabilité. Pour l'ensemble de ce chapitre, nous raisonnons donc conditionnellement à la réalisation de la première phase de tirage.

L étant le seuil de logements séparant petites et grandes adresses , on définit trois types d'adresses :

- les monologements (MONO) ou adresses d'un seul logement ;
- les petites adresses (PA), ayant de 2 à L logements ;
- les grandes adresses (GA), de plus de L logements.

5.1 Le scénario de base

Le scénario de base permet de générer un plan de sondage autopondéré, mais aucun effort n'est fait pour contrôler précisément la taille de l'échantillon ni la répartition de l'échantillon par SAE. Nous présentons ce plan de sondage comme *un tirage stratifié à deux degrés* : les unités primaires (UP) tirées au premier degré correspondent aux adresses ; les unités secondaires (US) tirées au second degré correspondent aux logements de ces adresses. En utilisant les notations présentées plus haut, la population des UP correspond à U^a et celle des US à U^l.

Un tirage à deux degrés consiste à tirer, au premier degré, un échantillon d'UP (S^a) dans U^a, puis, au second degré, à tirer des US au sein de chacune des UP échantillonnée. L'échantillon complet des US correspond à S^l, avec nos notations.

1^{er} degré : Constitution de l'échantillon d'adresses S^a

L'échantillon d'adresses est constitué par le moyen d'un *tirage stratifié par type d'adresses* (schéma 6, étape (A)). Au sein des strates de monologements et de petites adresses, *un tirage aléatoire simple à probabilités d'inclusion égales* est effectué. Le tirage des grandes adresses est effectué selon *un tirage à probabilités inégales proportionnelles à la taille de l'adresse* (schéma 6, étape (B)).

2^{nd} degré : Consitution de l'échantillon de logements S^l

Dès lors que l'échantillon d'adresses est tiré dans chacune des strates, l'échantillon de logements est constitué :

- de l'ensemble des logements des adresses de S^a_{mo} (1 adresse = 1 logement) ;
- de l'ensemble des logements de chacune des adresses de S^a_{pa} (1 adresse = grappe de logements) ;
- de L logements tirés parmi les logements de chacune des adresses S^a_{ga}.

Ainsi, le tirage des logements individuels est obtenu directement du tirage des adresses. On peut le voir comme un tirage par grappes d'un seul logement. Le tirage des logements en petites adresses est *un tirage par grappes de logements* : tous les logements d'une adresse échantillonnée sont enquêtés. Au sein de chacune des grandes adresses de S^a, L logements sont tirés selon un plan de sondage aléatoire simple à probabilités égales (schéma 6, étape (C)).

Les allocations

On définit d'abord les allocations en termes de nombre de logements à tirer au sein de chacune des strates, pour en déduire les allocations en termes de nombre d'adresses. Ces allocations de logements sont théoriques car seulement espérées, puisque le tirage des logements est indirect.

Les allocations théoriques souhaitées sont proportionnelles à la taille de chaque strate. On définit ainsi :

$$\begin{cases} n^l_{mo} = n^l \frac{N^l_{mo}}{N^l} \\ n^l_{pa} = n^l \frac{N^l_{pa}}{N^l} \\ n^l_{ga} = n^l \frac{N^l_{ga}}{N^l} \end{cases} \tag{1}$$

La taille de l'échantillon S^l_{mo} est parfaitement maîtrisée par le processus de tirage. Pour tirer n^l_{mo} monologements, il faudra tirer $n^a_{mo} = n^l_{mo}$ adresses dans la strate de monologements. Si on néglige le problème des arrondis, pour tirer n^l_{ga} logements en grandes adresses, il suffira de tirer $n^a_{ga} = \frac{n^l_{ga}}{L}$ adresses. En revanche, l'allocation de logements en petites adresses n^l_{pa} est clairement une valeur théorique qui n'a pas de raison d'être atteinte. Si on note $\hat{n}^l_{pa}$ la taille observée de l'échantillon S^l_{pa} et N^l_k le nombre de logements d'une adresse k, on a :

$$\hat{n}^l_{pa} = \sum_{k \in S^a_{pa}} N^l_k$$

L'allocation d'adresses n^a_{pa} que nous recherchons doit satisfaire l'équation suivante [25] :

$$\mathbb{E}(\hat{n}^l_{pa}) = n^l_{pa} \tag{2}$$

Nous montrons en annexe E que cette condition est vérifiée si et seulement si $n^a_{pa} = \frac{n^l_{pa}}{\overline{N}^l_{pa}}$, où $\overline{N}^l_{pa}$ est le nombre moyen de logements par petites adresses.

Pour récapituler, nous obtenons les allocations d'adresses suivantes :

$$\begin{cases} n^a_{mo} = n^l_{mo} \\ n^a_{pa} = \frac{n^l_{pa}}{\overline{N}^l_{pa}} \\ n^a_{ga} = \frac{n^l_{ga}}{L} \end{cases} \tag{3}$$

Fluctuation de la taille de l'échantillon

Deux sources de fluctuations de la taille de l'échantillon sont discernables : celle dûe à l'arrondi des allocations de grandes adresses et celle issue du tirage par grappes des petites adresses.

Supposons que le seuil soit fixé à $L = 10$ et que nos calculs nous mènent à une allocation de logements en grandes adresses $n^l_{ga} = 17,3$ logements. Cela implique que nous devons tirer $n^a_{ga} = \frac{17,3}{10} = 1,73$ adresses. Ainsi, en choisissant de ne pas arrondir l'allocation, un tirage nous conduira à obtenir 1 GA, quand un autre conduira à en tirer 2. On peut même s'assurer à ce qu'en moyenne, les échantillons aient $1,73$ adresses. Ne pas arrondir les allocations d'adresses permet de respecter, en moyenne, les poids de départ. Néanmoins, la taille de l'échantillon S^l_{ga} variera entre 10 et 20 logements en grandes adresses. Ce problème ne pourra être résolu complètement mais l'abaissement du seuil L permettra de contribuer à limiter ces variations.

La taille de l'échantillon des logements en petites adresses $\hat{n}^l_{pa}$ est également sujette à des fluctuations par la présence du nombre moyen de logements par adresse dans le calcul de l'allocation, même si nous avons choisi des allocations d'adresses qui assurent d'atteindre, en moyenne, la taille souhaitée n^l_{pa}.

25. $\mathbb{E}$ désigne l'espérance sous le plan de sondage conditionnellement à la première phase de tirage.

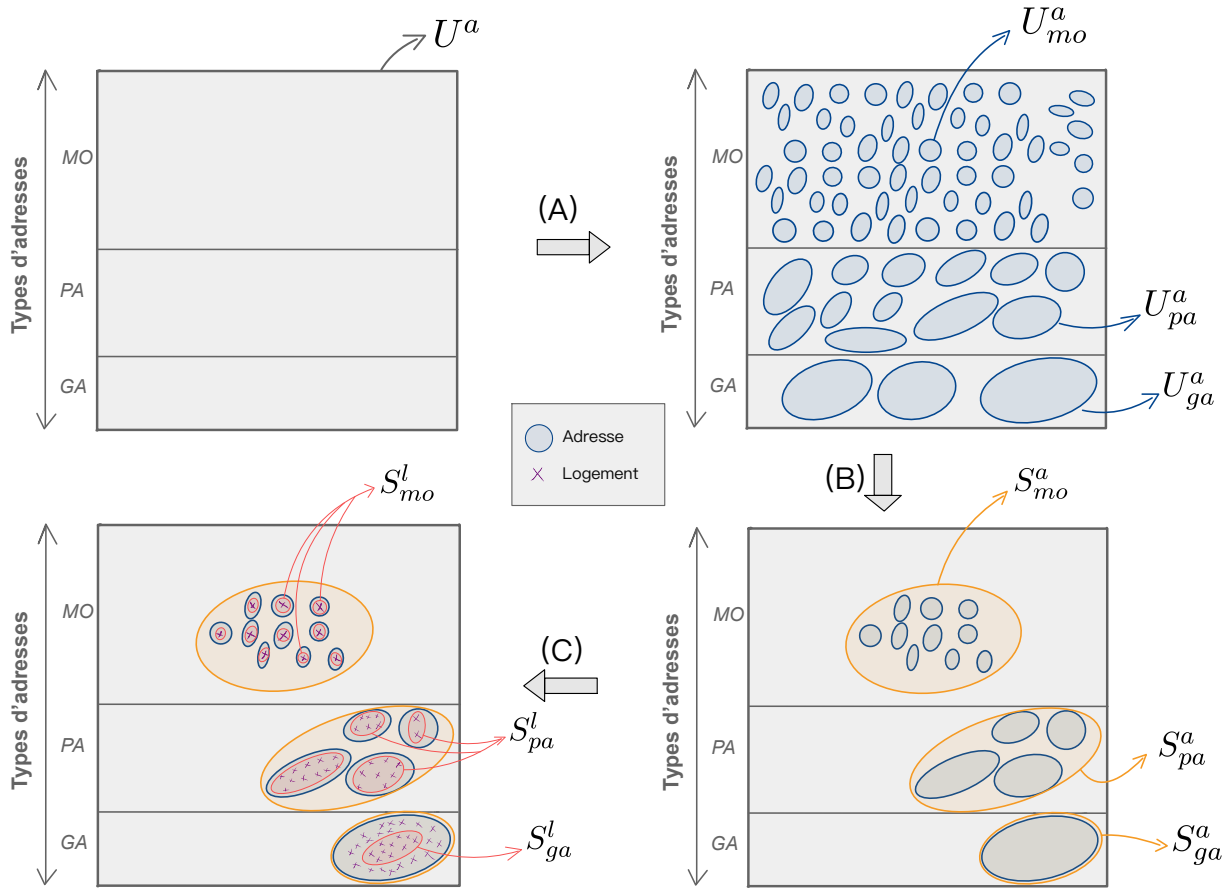


FIGURE 6 – Le tirage de l'échantillon selon le scénario de base
(A) : Stratification des adresses par taille ; (B) : Tirage stratifié des adresses (1er degré) ; (C) : Tirage des logements (2nd degré).

Le plan de sondage de base, de même que ses alternatives que nous présenterons dans la suite, n'est donc pas un plan de taille fixe.

Nous présentons un résumé des caractéristiques de ce plan de sondage de référence dans le tableau 2. Nous en profitons pour y démontrer que le plan est autopondéré, en montrant que, quelle que soit la strate, la probabilité d'inclusion d'un logement est égale à $\frac{n^l}{N^l}$. Le tableau utilise les notations complémentaires suivantes :

— $\pi^l_{l|k,mo}$, la probabilité d'inclusion du logement l conditionnellement au tirage de l'adresse k à laquelle il appartient, dans la strate de monologements ;

— $\pi^l_{l,mo} = \pi^a_{k,mo}\pi^l_{l|k,mo}$, la probabilité d'inclusion finale du logement l dans la strate des monologements ;

— $N^l_{k,ga}$, le nombre de logements à la grande adresse k.

Strates	Monologements (MONO)	Petites adresses (PA)	Grandes adresses (GA)
Allocations logements	$n^l_{mo} = \frac{N^l_{mo}}{N^l}$	$n^l_{pa} = \frac{N^l_{pa}}{N^l}$	$n^l_{ga} = \frac{N^l_{ga}}{N^l}$
Allocations adresses	$n^a_{mo} = n^l_{mo}$	$n^a_{pa} = \frac{n^l_{pa}}{\bar{N}^l_{pa}}$	$n^a_{ga} = \frac{n^l_{ga}}{10}$
Tirage des adresses	SRS	SRS	Systématique
Tri des adresses	-	-	Nb de logts à l'adresse
Tirage des logements	-	grappe	SRS(10)/GA tirée
Prob. d'incl. - 1er degré	$\pi^a_{k,mo} = \frac{n^a_{mo}}{N^a_{mo}}$ $\pi^a_{k,mo} = \frac{n^l_{mo}}{N^l_{mo}}$	$\pi^a_{k,pa} = \frac{n^a_{pa}}{N^a_{pa}}$ $\pi^a_{k,pa} = \frac{n^l_{pa}}{\bar{N}^l_{pa}N^a_{pa}}$ $\pi^a_{k,pa} = \frac{n^l_{pa}}{N^l_{pa}}$	$\pi^a_{k,ga} = n^a_{ga}\frac{N^l_{k,ga}}{N^l_{ga}}$
Prob. d'incl. - 2nd degré	$\pi_{l\|k,mo} = 1$	$\pi_{l\|k,pa} = 1$	$\pi_{l\|k,ga} = \frac{N^l_{k,ga}}{10}$
Prob. d'incl. - logements	$\pi^l_{l,mo} = \frac{n^l_{mo}}{N^l_{mo}}$ $\pi^l_{l,mo} = \frac{n^l}{N^l}$	$\pi^l_{l,pa} = \pi^a_{k,pa}$ $\pi^l_{l,pa} = \frac{n^l_{pa}}{N^l_{pa}}$ $\pi^l_{l,pa} = \frac{n^l}{N^l}$	$\pi^l_{l,ga} = \pi^a_{k,ga}\pi_{l\|k,ga}$ $\pi^l_{l,ga} = n^a_{ga}\frac{N^l_{k,ga}}{N^l_{ga}}\frac{N^l_{k,ga}}{10}$ $\pi^l_{l,ga} = \frac{n^a_{ga}}{10N^l_{ga}}$ $\pi^l_{l,ga} = \frac{n^l_{ga}}{N^l_{ga}}$ $\pi^l_{l,ga} = \frac{n^l}{N^l}$

TABLE 2 – Le plan de sondage de base

- SRS : Sondage aléatoire simple
- tirage d'une grappe : les logements ont la même probabilité d'inclusion que leur adresse
- Le nombre de logements dans la strate des petites adresses est égal au nombre moyen de logements par petite adresse multiplié par le nombre de petites adresses : $N^l_{pa} = \bar{N}^l_{pa}N^a_{pa}$;
- La probabilité d'inclusion d'une grande adresse est proportionnelle à la taille de l'adresse

5.2 Des alternatives

Le scénario de base a été conçu pour générer un plan autopondéré, sans attention particulière à la maîtrise de la taille de l'échantillon de logements ni à sa répartition par secteurs d'activité des enquêteurs (SAE). Et il échoue sur ces deux points essentiels, comme nous le verrons lors de la présentation des résultats de nos simulations. Des alternatives ont été testées pour, à partir de ce scénario, corriger ces deux manques, parfois au détriment de l'autopondération, parfois sans

parvenir à maîtriser suffisamment la taille de l'échantillon par SAE.

Une première tentative : stratification par SAE et ajustement des allocations par la méthode du raking ratio

La première chose a été de stratifier le tirage par SAE et par type d'adresses. Les SAE ayant été conçus pour être homogènes, notamment en termes de nombre de logements, cette stratification permet de mieux répartir l'échantillon par SAE. De plus, en reprenant les modalités de tirage dans les strates du plan de base présenté plus haut, il est aisé de montrer que ce plan de sondage stratifié par SAE*type d'adresses est également un plan de sondage autopondéré.

Mais l'homogénéité des SAE n'est pas parfaite et la répartition de l'échantillon par SAE demeure insatisfaisante pour le niveau d'exigence requis par l'organisation de l'EEC. Nous avons alors ajusté les allocations de logements par strate en utilisant *l'algorithme du raking ratio*, présenté notamment dans [9] [26]. La méthode modifie itérativement les allocations initiales par SAE*type d'adresses afin que les allocations marginales par SAE, d'une part, et par type d'adresses, d'autre part, soient respectées. Les marges par type d'adresses correspondent aux proportions observées dans la population, pendant que nous choisissons les marges par SAE égales à la taille de l'échantillon de logements souhaitée, soit 36 logements par SAE, pour une taille d'échantillon de 936 logements entrants et 26 secteurs d'activité des enquêteurs.

Cette méthode s'avère relativement efficace pour contrôler la taille de l'échantillon et sa répartition par SAE, même si une phase réjective s'avère nécessaire pour s'en assurer pleinement. Le seuil choisi entre petites et grandes adresses joue également un rôle dans la facilité à respecter la contrainte de taille de l'échantillon et de sa répartition par SAE. En effet, plus le seuil est bas, plus facilement la contrainte est respectée.

Malheureusement, en modifiant les allocations par strate, cette technique déforme sensiblement les probabilités d'inclusion initiales des adresses et ne permet plus d'assurer l'autopondération du plan de sondage, même approximativement.

Les tirages équilibrés

Nous avons également testé des tirages équilibrés avec des pondérations proportionnelles à la taille de la strate SAE*Type d'adresses. Si ce type d'échantillonnage s'est avéré le plus précis des plans de sondage envisagés et s'il permet également de respecter la contrainte d'autopondération, il ne permet pas de construire des échantillons dont la taille par SAE soit suffisamment précise autour de la cible pour tous les secteurs d'enquête en même temps (voir Annexe C). En effet, alors que nous recherchons un plan de sondage qui construit des échantillons de 36 logements par SAE chaque année, avec une tolérance de ± 4, les différents tirages équilibrés envisagés ne permettent de construire aucun échantillon respectant cette condition pour tous les secteurs en même temps, sauf au prix d'une perte d'autopondération en utilisant des poids ajustés par la méthode du raking ratio décrite plus haut.

Le problème des grandes adresses

Le tirage des grandes adresses pose quelques problèmes lorsque le tirage est stratifié par SAE*type d'adresses. En effet, les grandes adresses sont très peu nombreuses et mal réparties sur le territoire. Avec un seuil séparant petites et grandes adresses fixé à $L = 10$, parmi les 56 800 adresses de Mayotte, seules 62 sont des GA et rassemblent 1 201 logements, soit moins de 2% des logements de l'île. En abaissant le seuil, la problématique reste la même (700 adresses et environ 7% des logements, avec un seuil $L = 4$). En outre, les GA sont essentiellement concentrées sur quelques îlots des communes de Mamoudzou et Koungou (voir carte en figure 7).

26. L'algorithme est présenté p.283-288.

Lorsqu'on envisage un tirage stratifié par SAE*type d'adresses avec des allocations proportionnelles à la taille de la strate, les strates de grandes adresses, à cause du faible nombre de logements qui s'y trouvent, se voient souvent alloués un nombre de logements inférieur au seuil L, soit, en termes d'adresses, une allocation comprise entre 0 et 1. Or, pour éviter un défaut de couverture, nous sommes contraints d'arrondir ces allocations à 1 adresse, soit L logements. En effet, ne pas tirer une GA dans la strate reviendrait à ne pas enquêter une partie de la population cible ou champ de l'enquête. Le tableau 3 montre que la population présente dans ces grandes adresses n'est pas tout à fait la même que dans les autres adresses : moins souvent au chômage et très majoritairement actifs, ils sont également plus souvent nés en France hors Mayotte ainsi que plus diplômés. Dans notre cas, un défaut de couverture conduirait à surestimer le taux de chômage à Mayotte.

Arrondir les allocations entraîne une surreprésentation des logements en grandes adresses dans l'échantillon : pour un seuil fixé à $L = 10$, ces logements représenteraient entre 2 et 9% de l'échantillon (selon le groupe de rotation considéré) contre $1,6\%$ dans la population. Les allocations arrondies étant plus grandes que les allocations théoriques, les pondérations de ces logements sont inférieures au poids théorique du plan de sondage autopondéré imaginé dans un premier temps. L'autopondération initiale du plan de sondage n'est donc plus assurée.

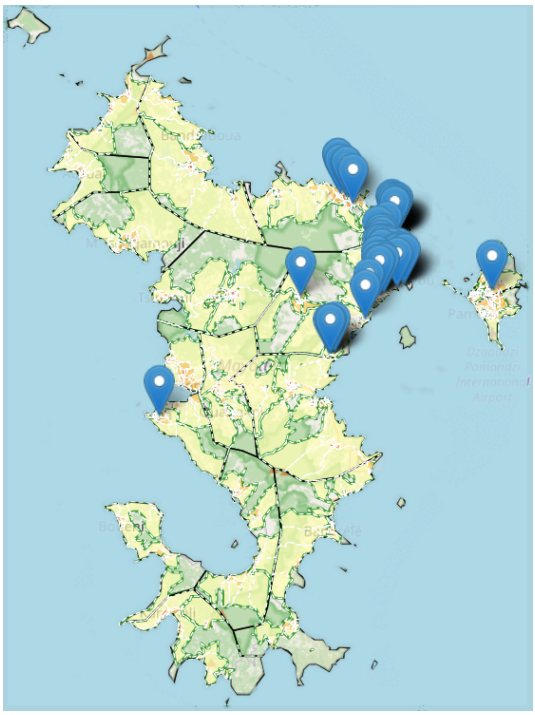


FIGURE 7 – Localisation des grandes adresses pour un seuil fixé à 10 logements

Source : Appariement RP 2017 et Base cartographique 2017

Pour toutes ces raisons, nous avons choisi de rassembler les grandes adresses en une seule strate, tous SAE confondus, et de les tirer lors d'une première étape de tirage.

TABLE 3 – Caractéristiques des 15 ans ou + selon le seuil L et le type d'adresses en 2017

Seuil	Type d'adr.	Nb de logts	Taux (sens RP, en %) Chômage	Emploi	Activité	Lieu de naiss. (en %) Étr.	May.	Fr. hors May.	Bac ou + (en %)
	1-mo	46 661	44	25	45	49	47	4	19
10	2-pa	25 928	40	32	54	55	38	7	26
	3-ga	1 201	23	59	77	52	15	33	57
	1-mo	46 661	44	25	45	49	47	4	19
4	2-pa	22 022	40	32	53	53	41	6	25
	3-ga	5 107	39	38	62	65	19	16	34
	Ens.	73 790	42	28	49	51	44	5	22

Source : Recensement de la population de Mayotte, 2017
Champ : Individus de 15 ans ou + résidant dans un logement ordinaire.

5.3 Le plan de sondage retenu

Nous présentons le scénario retenu qui s'est révélé le plus apte à répondre, parmi l'ensemble des alternatives que nous avons testées, aux différents problèmes rencontrés et à l'ensemble des contraintes imposées.

Une allocation fixe et égale par SAE

Pour rappel, l'échantillon d'entrants comprend en théorie 936 logements qui doivent être répartis équitablement entre les 26 SAE. La solution que nous retenons part du principe d'effectuer un tirage indépendant de 36 logements dans chacun des SAE. Il s'agit d'une sorte de stratification où les allocations sont fixées - ici égales - et non calculées proportionnellement à la taille de la strate.

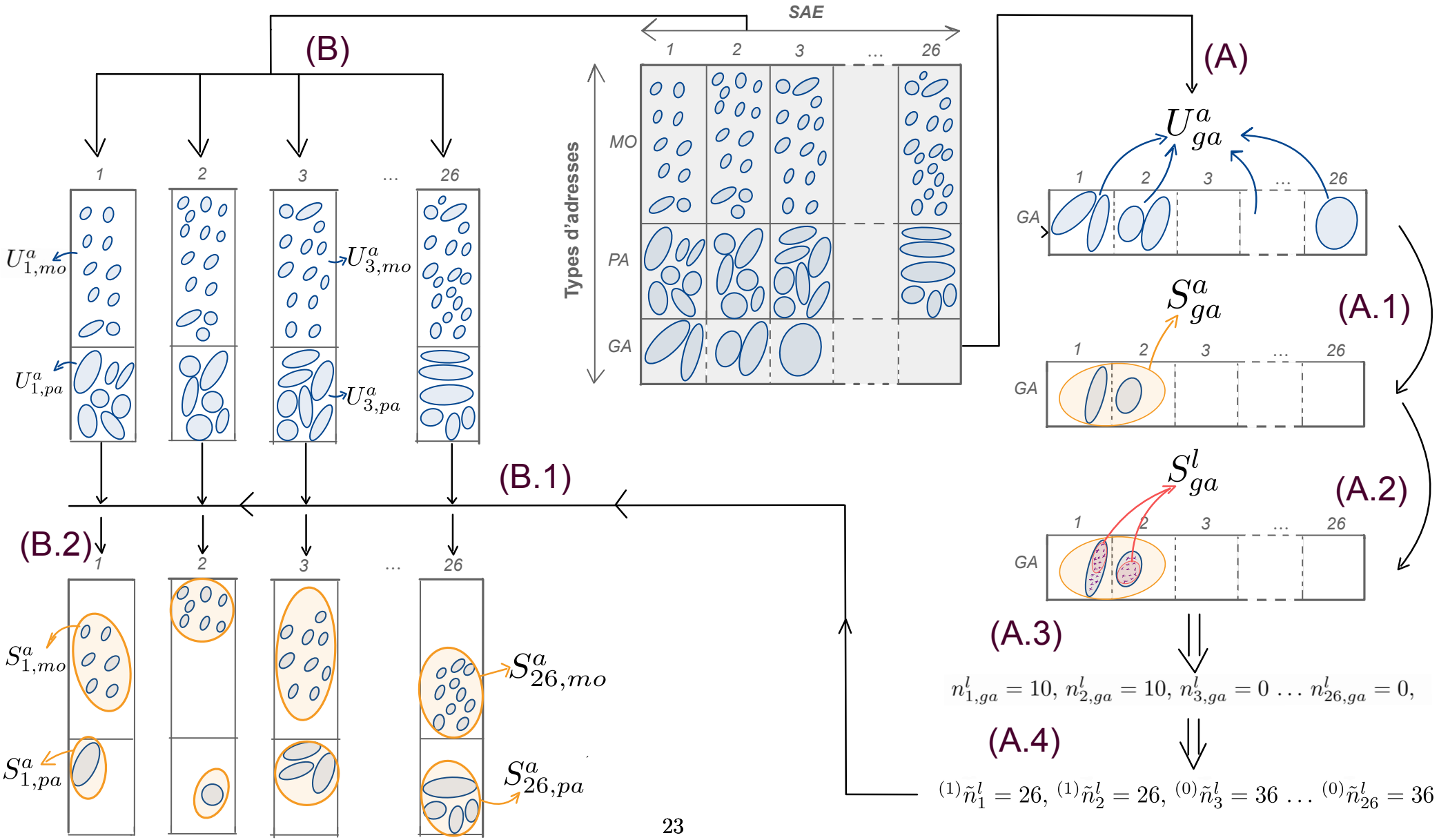
Un tirage des adresses en deux étapes

Pour gérer le problème des grandes adresses évoqué précédemment, un tirage en deux étapes s'est avéré plus efficace, notamment pour maîtriser la taille de l'échantillon. Cela consiste à rassembler toutes les grandes adresses dans une seule et même strate et à opérer, dans une première étape, au tirage des grandes adresses. La seconde étape consiste à combler l'échantillon par un tirage de monologements et de petites adresses par des tirages indépendants au sein de chacun des SAE.

Pour décrire l'ensemble du processus de tirage, nous utilisons les notations complémentaires suivantes :

- s, un SAE ;
- n_s^l, la taille cible de l'échantillon de logements à obtenir dans le SAE s ;
- $\tilde{N}_s^l$, le nombre de monologements et de logements en petites adresses dans le SAE s ;
- $N_{s,mo}^l$, le nombre de monologements dans le SAE s ;
- $N_{s,pa}^l$, le nombre de logements en petites adresses dans le SAE s ;
- $\overline{N}_{s,pa}^l$, le nombre moyen de logements par petites adresses dans le SAE s ;
- α, le nombre de grandes adresses tirées et qui sont localisées dans le SAE s ;
- ${}^{(\alpha)}\tilde{n}_s^l$, la taille de l'échantillon de monologements et de logements en petites adresses dans le SAE s, dans lequel α grandes adresses ont été préalablement tirées ;
- ${}^{(\alpha)}\tilde{n}_{s,mo}^l$, la taille de l'échantillon de monologements dans le SAE s, dans lequel α grandes adresses ont été préalablement tirées ;
- ${}^{(\alpha)}\tilde{n}_{s,pa}^l$, la taille de l'échantillon de logements en petites adresses dans le SAE s, dans lequel α grandes adresses ont été préalablement tirées ;
- ${}^{(\alpha)}\pi_{l,s}$, la probabilité d'inclusion du logement l du SAE s, dans lequel α grandes adresses ont été préalablement tirées.

FIGURE 8 – Principales étapes du tirage sous le plan de sondage retenu



L'idée principale étant de tirer $n_s = \frac{n^l}{26}$ logements dans chacun des SAE (soit 36 avec la taille d'échantillon retenue de 936 entrants par an), on peut décrire les deux étapes de tirage et leurs sous-étapes de la façon suivante :

1. Le tirage à deux degrés des logements en grandes adresses (étapes (A) sur le schéma 8) :
   (a) On rassemble les grandes adresses dans une seule et même strate ;
   (b) La taille de l'échantillon de logements à obtenir dans cette strate est : $n^l_{ga} = n\frac{N^l_{ga}}{N^l}$;
   (c) On en déduit la taille de l'échantillon d'adresses à tirer dans cette strate est : $n^a_{ga} = n^l \frac{n^l_{ga}}{L}$;
   (d) Premier degré de tirage : Tirage systématique des GA avec des probabilités d'inclusion proportionnelles à leur taille (A.1) ;
   (e) Second degré de tirage : Tirage des logements par SRS(L) au sein de chaque GA tirée (A.2).
2. Le tirage des monologements et des petites adresses :
   (a) On retire de la taille de l'échantillon n^l_s des SAE concernés le nombre de logements des grandes adresses préalablement tirées, soit αL logements (A.3) ;
   (b) Ainsi, dans le SAE s, il reste à tirer ${}^{(\alpha)}\tilde{n}^l_s = n^l_s - \alpha L$ monologements et logements en petites adresses (A.4) ;
   (c) Au sein du SAE s, les allocations par strate d'adresses sont : $n^l_{s,mo} =^{(\alpha)} \tilde{n}^l_s \frac{N^l_{s,mo}}{\tilde{N}^l_s}$ et $n^l_{s,pa} =^{(\alpha)} \tilde{n}^l_s \frac{N^l_{s,pa}}{\tilde{N}^l_s}$ (B.1) ;
   (d) On en déduit la taille de l'échantillon d'adresses à tirer dans cette strate : $n^a_{s,mo} = n^l_{s,mo}$ et $n^a_{s,pa} = \frac{n^l_{s,pa}}{\overline{N}^l_{s,pa}}$ (B.1) ;
   (e) Un tirage stratifié par type d'adresses est effectué, indépendamment, dans chacun des SAE (B.2).

Nous présentons un résumé des caractéristiques de ce plan de sondage dans le tableau 4.

Strates	Dans le SAE s		Tous SAE confondus
	Monologements (MONO)	Petites adresses (PA)	Grandes adresses (GA)
Allocations logements	$n^l_{s,mo} = \frac{N^l_{s,mo}}{\tilde{N}^l}$	$n^l_{s,pa} = \frac{N^l_{s,pa}}{\tilde{N}^l}$	$n^l_{ga} = \frac{N^l_{ga}}{N^l}$
Allocations adresses	$n^a_{s,mo} = n^l_{s,mo}$	$n^a_{s,pa} = \frac{n^l_{s,pa}}{\bar{N}^l_{s,pa}}$	$n^a_{ga} = \frac{n^l_{ga}}{10}$
Tirage des adresses	Systématique	Systématique	Systématique
Tri des adresses	typologie	taille et typologie	taille et typologie
Tirage des logements	-	grappe	SRS(10)/GA tirée
Prob. d'incl. - 1er degré	$\pi^a_{k,s,mo} = \frac{n^a_{s,mo}}{N^a_{s,mo}}$ $\pi^a_{k,s,mo} = \frac{n^l_{s,mo}}{N^l_{s,mo}}$	$\pi^a_{k,s,pa} = \frac{n^a_{s,pa}}{N^a_{s,pa}}$ $\pi^a_{k,s,pa} = \frac{n^l_{s,pa}}{\bar{N}^l_{s,pa} N^a_{s,pa}}$ $\pi^a_{k,s,pa} = \frac{n^l_{s,pa}}{N^l_{s,pa}}$	$\pi^a_{k,ga} = n^a_{ga} \frac{N^l_{k,ga}}{N^l_{ga}}$
Prob. d'incl. - 2nd degré	$\pi_{l\|k,s,mo} = 1$	$\pi_{l\|k,s,pa} = 1$	$\pi_{l\|k,ga} = \frac{N^l_{k,ga}}{10}$
Prob. d'incl. - logements	$\pi^l_{l,s,mo} = \frac{n^l_{s,mo}}{N^l_{s,mo}}$ $\pi^l_{l,s,mo} = \frac{n^l_s}{N^l_s}$	$\pi^l_{l,s,pa} = \pi^a_{k,s,pa}$ $\pi^l_{l,s,pa} = \frac{n^l_{s,pa}}{N^l_{s,pa}}$ $\pi^l_{l,s,pa} = \frac{n^l_s}{N^l_s}$	$\pi^l_{l,k,ga} = \pi^a_{k,ga} \pi_{l\|k,ga}$ $\pi^l_{l,k,ga} = n^a_{ga} \frac{N^l_{k,ga}}{N^l_{ga}} \frac{N^l_{k,ga}}{10}$ $\pi^l_{l,k,ga} = \frac{n^a_{ga}}{10 N^l_{ga}}$ $\pi^l_{l,k,ga} = \frac{n^l_{ga}}{N^l_{ga}}$ $\pi^l_{l,k,ga} = \frac{n^l}{N^l}$

TABLE 4 – Le plan de sondage retenu

Le nombre de logements dans la strate des petites adresses au sein du SAE s est égal au nombre moyen de logements par petite adresse multiplié par le nombre de petites adresses : $N^l_{s,pa} = \bar{N}^l_{s,pa} N^a_{s,pa}$. La probabilité d'inclusion d'une grande adresse est proportionnelle à la taille de l'adresse.

Un plan de sondage approximativement autopondéré

Ce plan de sondage est celui, parmi l'ensemble des scénarios envisagés lors de nos travaux, qui assure la meilleure maîtrise de la taille de l'échantillon de logements entrants par SAE, et donc sa taille globale. Néanmoins, rien n'assure théoriquement qu'il fournisse un plan de sondage autopondéré.

Nous avons étudié les conditions suffisantes pour qu'un tel plan soit autopondéré (voir Annexe F) et si ces conditions s'appliquaient aux données dont nous disposons. Selon ces conditions,

le nombre de monologements et de petites adresses dans chaque SAE doit être dans une certaine relation proportionnelle avec le nombre de logements total. Cette relation dépend du nombre de grandes adresses tiré en première étape (voir Annexe F, équation 40b).

Sur le groupe de rotation d'îlots enquêtés en 2020 pour l'enquête cartographique, nous réalisons 10 000 tirages et observons que ces conditions sont relativement bien remplies pour 24 SAE sur 26. Pour ces SAE, les poids des logements s'écartent au plus de 25% du poids théorique recherché. Les SAE 5 et 16, sur le groupe de rotation étudié, disposent d'un nombre de logements et d'un nombre de logements en grandes adresses trop grands pour bien remplir la condition. Néanmoins, les poids des logements ne s'écartent jamais plus de deux fois le poids théorique souhaité (Annexe F, figure 19).

Un tirage systématique dans chaque strate

Pour réaliser concrètement le tirage des adresses au sein de chaque strate, nous avons utilisé l'algorithme de tirage systématique. Ce type de tirage est adapté aux tirages à probabilités inégales, comme dans la strate de grandes adresses, ou aux tirages à probabilités égales réalisés dans les autres strates.

Si de nombreux algorithmes de tirage existent, notre choix s'est porté sur le tirage systématique car il permet d'améliorer la précision des estimateurs quand les unités sont triées selon une variable auxiliaire bien corrélée aux variables d'intérêts. Sur ce point, P. Ardilly note qu'il est plus intéressant qu'un tirage équilibré si une seule variable auxiliaire sert au tri (systématique) ou à l'équilibrage (tirage équilibré).

Celui-ci fera en sorte de respecter la moyenne de la variable d'équilibrage, quand celui-là permettra aussi de reproduire la forme de la distribution de la variable auxiliaire [27]. Or, la pauvreté de la base de sondage ne nous permet pas d'envisager d'utiliser une information auxiliaire abondante.

Nous avons choisi de trier :

— les monologements par la typologie des îlots que nous avons construites ;

— les petites adresses par leur taille et la typologie des îlots ;

— les grandes adresses par leur taille et la typologie des îlots, également.

Principe de fonctionnement :

Le tirage systématique à probabilités égales consiste à tirer les unités séparées entre elles d'un certain pas, fixé à l'avance et dépendant de la taille de l'échantillon souhaité (en fait $\frac{N}{n}$). La première unité tirée est choisie au hasard parmi les N/n premières unités. Ainsi, le tirage systématique est proche d'un tirage par grappes : une fois la première unité sélectionnée, le reste de l'échantillon est déterminé. Le défaut majeur du tirage par grappes est qu'il consiste en général à tirer des unités qui se ressemblent - comme des logements situés à la même adresse par exemple. Mais, en triant préalablement les unités selon une variable auxiliaire liée aux variables d'intérêt, le tirage systématique permet d'éviter cet écueil et de gagner en précision. Par principe, de nombreuses unités ne pourront pas être tirées en même temps. Dans notre schéma (figure 9), les unités v_1 et v_2 ne seront jamais présentes dans le même échantillon : leur probabilité d'inclusion d'ordre 2 sera nulle.

27. [9], p.125 : P. Ardilly ne démontre pas cette propriété mais son raisonnement nous convainc. Une simulation menée plus loin dans le livre (p.198) tend à en apporter une première confirmation empirique.

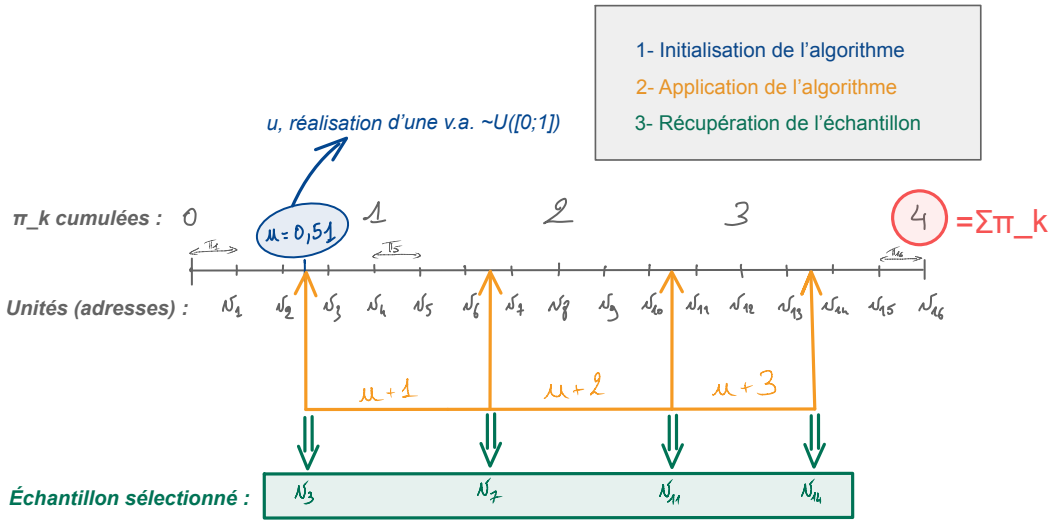


FIGURE 9 – Le tirage systématique à probabilités égales

Source : Ce schéma est inspiré du cours de G. Chauvet en Master de Statistique publique, *Techniques avancées d'échantillonnage*, dispensé à l'Ensai en Novembre 2020

5.4 Bilan

De nos travaux, nous retenons donc deux plans de sondage :

— un plan de référence, autopondéré mais n'intégrant aucune adaptation pour répondre aux contraintes de taille et de répartition de l'échantillon ;

— le plan qui a le mieux répondu à l'ensemble des contraintes, plan approximativement autopondéré.

À quels gains et quelles pertes de précision pouvons-nous nous attendre ?

Malgré une base de sondage frustre, la stratification par type d'adresses devrait permettre de gagner en précision par rapport à un sondage aléatoire simple. En effet, les populations ont des profils un peu différents selon la taille de leur adresse (tableau 3), notamment sur l'emploi, le chômage et l'activité. Les écarts sont moins nets quand le seuil entre petites et grandes adresses est abaissé.

Dans un tirage à deux degrés, la variance dûe au premier degré est en général très supérieure à la variance dûe au second degré [28]. Ainsi, il est plus facile de faire des gains en précision en améliorant le premier degré de tirage. Dans notre cas, même si le plan retenu est un peu plus complexe qu'un tirage à deux degrés classique, on peut s'attendre à ce que le tirage systématique des adresses triées selon une variable (typologie des îlots) liée aux paramètres d'intérêts (chômage, emploi, activité) permettra des gains notables en précision.

Les SAE ont été conçus pour fournir des ensembles les plus semblables possibles. Stratifier par SAE n'est pas une stratégie efficace pour gagner en précision. En effet, la stratification est

28. Quand un plan de sondage à deux degrés respecte les conditions d'invariance du tirage au second degré et d'indépendance des tirages dans les UP conditionnellement au premier degré, les ordres de grandeur sont $\mathbf{V}_{UP}(\hat{t}_{y\pi}) = O(\frac{N^2}{n_I})$ et $\mathbf{V}_{US}(\hat{t}_{y\pi}) = O(\frac{N^2}{n_I n_0})$. Nous reprenons ceci du cours d'*Échantillonnage à plusieurs degrés* de G. Chauvet dispensé à l'Ensai en mars 2020

d'autant plus efficace que les strates sont les plus dissemblables entre elles [29]. Si les alternatives rejetées étaient basées sur une stratification par SAE, nous trouvons que le plan retenu assume plus le fait que les SAE sont davantage une variable de répartition qu'une variable de stratification : les allocations par SAE étant fixées par avance.

Le tirage par grappes conduit à tirer des logements et donc des ménages qui ont beaucoup plus de chances de se ressembler que s'ils étaient tirés de manière dispersée. Les phénomènes économiques mesurés par l'enquête Emploi (le chômage, l'activité et l'emploi) sont, en outre, spatialement autocorrélés [30] : les ménages ont plus de chances d'être entourés de ménages avec des caractéristiques similaires (autocorrélation positive). L'avantage des secteurs d'activité des enquêteurs est qu'ils dispersent spatialement l'échantillon sur toute l'île. Cela peut contenir la perte de précision dûe à l'effet de grappes des tirages des logements en petites et grandes adresses.

6 Estimations

Le propos de cette partie est de fixer le cadre général des estimations dont nous présenterons les résultats par simulation dans la partie suivante. Rappelons que l'objectif de notre travail est de fournir une estimation de la précision des estimateurs transversaux des paramètres clés de l'enquête Emploi. Nous avons particulièrement ciblés les paramètres suivants :

— Le taux de chômage au sens du BIT ;

— Le taux d'emploi au sens du BIT ;

— Le taux d'activité au sens du BIT ;

— La part de chômeurs parmi la population âgée de 15 à 74 ans.

Le dernier paramètre est surtout utile pour vérifier si la contrainte européenne de précision minimale est vérifiée.

Nous limitons, ici, à présenter le formalisme et les résultats obtenus pour le taux de chômage et la part de chômeurs parmi les 15-74 ans. Nous nous attachons aux estimations transversales trimestrielles (rythme de publication des résultats et contrainte européenne) et annuelle (point de comparaison avec l'enquête actuelle).

6.1 Notations

Nous nous sommes attachés à construire un échantillon de logements, dont l'ensemble des personnes y résidant seront interrogées. Ainsi, en supposant que, si un ménage répond, tous les individus qui le composent répondent, les poids des individus sont identiques au poids du logement dans lequel ils résident. Nous travaillons ainsi dans cette partie au niveau individuel tant que cela est possible afin de simplifier l'exercice formel.

Nous utiliserons les notations suivantes :

— U, la population des individus dans le champ de l'enquête (résidant dans un logement ordinaire) ;

— N, la taille de la population U ;

— S, l'échantillon trimestriel des individus ;

— n, la taille de l'échantillon S ;

29. On peut montrer dans le cas extrême de strates homogènes en taille et en moyenne, qu'un tirage stratifié avec SRS dans chaque strate sera moins efficace qu'un tirage SRS direct. L'effet de plan est $D_{eff} = \frac{N-1}{N-H} > 1$ dès que le nombre de strates H est plus grand que 1 (calcul réalisé avec une variable d'intérêt de type indicatrice).

30. L'autocorrélation spatiale est définie « comme la corrélation, positive ou négative, d'une variable avec elle-même du fait de la localisation spatiale des observations. », [10], p.54

— U_D, la population de U âgée entre 15 et 74 ans compris (domaine) ;

— N_D, la taille du domaine U_D ;

— $\delta_k = \mathbb{1}(k \in U_D)$;

— $y_k = \mathbb{1}(k$ est au chômage), la variable d'intérêt indiquant si l'individu $k \in U$ est au chômage ;

— $v_k = \delta_k y_k$, la variable indiquant si un individu est au chômage et âgé entre 15 et 74 ans ;

— $z_k = \mathbb{1}(k$ est en activité), la variable d'intérêt indiquant si l'individu $k \in U$ est en activité (en emploi ou au chômage) ;

— $\pi_k = \mathbb{P}(k \in S)$, la probabilité d'inclusion de l'individu k dans S, en prenant en compte les deux phases du tirage ;

— $d_k = \frac{1}{\pi_k}$, le poids de l'individu k dans S ;

On définit également les paramètres suivants :

— le nombre total de chômeurs dans U :

$$t_y = \sum_{k \in U} y_k$$

— le nombre total de chômeurs dans U_D :

$$t_{y_D} = \sum_{k \in U_D} y_k = \sum_{k \in U} \delta_k y_k = \sum_{k \in U} v_k = t_v$$

— la taille de U_D :

$$N_D = \sum_{k \in U} \delta_k = t_\delta$$

6.2 Les variables d'intérêt et leurs estimateurs

Une première version avec les estimateurs Horvitz-Thompson

Dans un premier temps, nous supposons les π_k parfaitement connues, pour toutes les unités k.

Le premier paramètre d'intérêt, la **part de chômeurs** parmi les 15-74 ans, est défini par le ratio :

$$P = \frac{t_{y_D}}{N_D} = \frac{t_v}{t_\delta}$$

Les deux paramètres t_v et t_δ sont définis sur U et peuvent être estimés sans biais par leurs estimateurs Horvitz-Thompson respectifs $\hat{t}_{v,\pi}$ et $\hat{t}_{\delta,\pi}$:

$$\begin{cases} \hat{t}_{v\pi} & = \sum_{k \in S} \frac{v_k}{\pi_k} \\ \hat{t}_{\delta\pi} & = \sum_{k \in S} \frac{\delta_k}{\pi_k} \end{cases}$$

Nous estimons le ratio P par l'estimateur par substitution $\hat{P}$ défini par :

$$\hat{P}_\pi = \frac{\hat{t}_{v\pi}}{\hat{t}_{\delta\pi}}$$

Dans le cas d'un plan simple sans remise, le biais d'un estimateur par le ratio est d'un ordre de grandeur en $1/n$ et peut donc être considéré comme « négligeable quand la taille de l'échantillon est grande » [31]. Avec un échantillon trimestriel de 1 400 logements, l'estimateur trimestriel devrait

31. [11],p.69

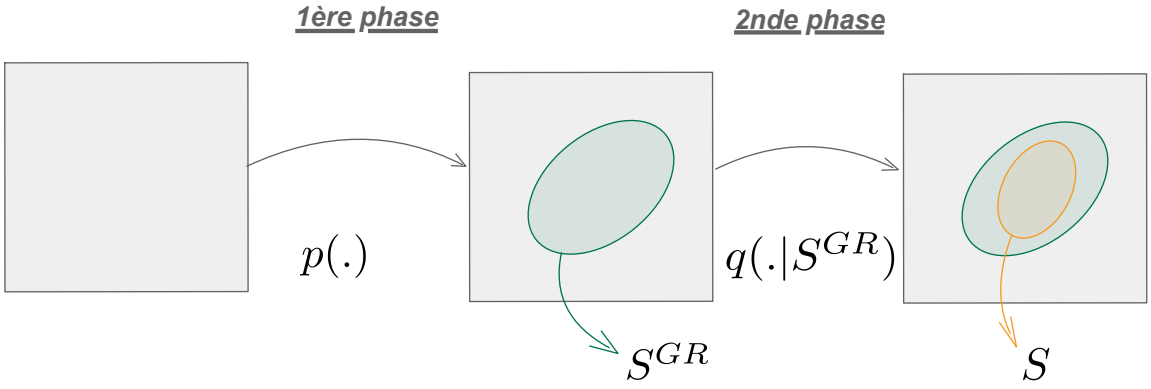


FIGURE 10 – Les deux phases de tirage

être approximativement sans biais. Les simulations permettront de vérifier si c'est bien le cas dans le cadre du plan complexe que nous utilisons.

Le **taux de chômage au sens du BIT** est défini comme le rapport entre le nombre de chômeurs au sens du BIT et le nombre d'actifs au sens du BIT :

$$C = \frac{t_y}{t_z} \tag{4}$$

Pour l'estimer, on utilise directement l'estimateur par le ratio, soit :

$$\hat{C}_\pi = \frac{\hat{t}_{y,\pi}}{\hat{t}_{z,\pi}} \tag{5}$$

Les estimateurs par expansion

Pour cette partie, nous utilisons les notations complémentaires suivantes :

— S^{GR}, l'échantillon tiré en première phase (groupe de rotation d'îlots) ;
— $p(.)$, le plan de sondage de première phase ;
— $\pi_{1k} = \mathbb{P}(k \in S^{GR})$, la probabilité d'inclusion de première phase de l'unité k ;
— $q(.|S^{GR})$, plan de sondage de deuxième phase ;
— $\pi_{2k} = \mathbb{P}(k \in S|S^{GR})$, la probabilité d'inclusion de deuxième phase de l'unité k ;
— $\mathbb{E}_p$ désigne l'espérance sous le plan de sondage $p(.)$.

Chaque année, le CRIEM tirera un échantillon de logements entrants dans l'enquête en utilisant des programmes qui ne réalisent que la seconde phase de tirage [32], la première phase ayant été réalisée préalablement à notre stage. Les probabilités d'inclusion issues de ce tirage annuel ne correspondent donc qu'aux probabilités d'inclusion conditionnelles de deuxième phase π_{2k}. Le tirage de première phase consiste en la succession de quatre tirages équilibrés puis de la sélection d'un des cinq groupes formés. Les probabilités d'inclusion théoriques des îlots sont égales et on considèrera que les tirages équilibrés n'affectent pas trop ces probabilités d'inclusion de départ. On posera ainsi que $\pi_{1k} = 1/5$, où k désigne ici indistinctement un îlot (unité échantillonnée à ce stade) ou un logement, car tous les logements d'un îlot de S^{GR} sont retenus dans l'échantillon obtenu à l'issue de la première phase.

En toute rigueur, la probabilité d'inclusion conditionnelle à la seconde phase est une variable aléatoire, puisqu'elle est dépendante d'un événement aléatoire qu'est le tirage de S^{GR}. On a :

32. Lors de nos simulations, nous simulerons l'ensemble du processus en reproduisant le tirage de première phase.

$$\pi_k = \pi_{1k}\mathbb{E}_p(\pi_{2k}) \tag{6}$$

$\mathbb{E}_p(\pi_{2k})$ étant inconnue [33], nous l'estimons par π_{2k} et nous obtenons un estimateur des π_k :

$$\hat{\pi}_k = \pi_{1k}\pi_{2k} = \frac{\pi_{2k}}{5} \tag{7}$$

Ainsi, l'estimateur de t_y que nous utilisons ne serait un estimateur de Horvitz-Thompson que si les π_k étaient connues. Devant les estimer par $\hat{\pi}_k$, nous utilisons en réalité un estimateur « par expansion » :

$$\hat{t}_{ye} = \sum_{k\in S} \frac{y_k}{\hat{\pi}_k} \tag{8}$$

$$= \sum_{k\in S} \frac{y_k}{\pi_{1k}\pi_{2k}} \tag{9}$$

$$= \sum_{k\in S} 5\frac{y_k}{\pi_{2k}} \tag{10}$$

On a, de même, les estimateurs par expansion de t_z, t_v et t_δ :

$$\begin{cases} \hat{t}_{ze} &= \sum_{k\in S} 5\frac{z_k}{\pi_{2k}} \\ \hat{t}_{ve} &= \sum_{k\in S} 5\frac{v_k}{\pi_{2k}} \\ \hat{t}_{\delta e} &= \sum_{k\in S} 5\frac{\delta_k}{\pi_{2k}} \end{cases} \tag{11}$$

Ainsi, les estimateurs de P et C que nous utilisons sont, en réalité, les estimateurs par le ratio exprimés en fonction des estimateurs par expansion des totaux de y, z, v et δ, soit :

$$\begin{cases} \hat{P}_e = \frac{\hat{t}_{ve}}{\hat{t}_{\delta e}} \\ \hat{C}_e = \frac{\hat{t}_{ye}}{\hat{t}_{ze}} \end{cases} \tag{12}$$

6.3 Mécanisme de réponse

Dans cette partie, nous utilisons les notations complémentaires suivantes :

— S_r, l'échantillon des répondants ;

— $r_k = \mathbb{1}(k \in S_r)$, la variable indiquant que le logement k a répondu ou non ;

— $p_k = \mathbb{P}(k \in S_r|S)$, la probabilité de réponse du logement [34] k ;

— $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{mk})$, un vecteur de variables auxiliaires, disponibles dans la base de sondage.

Pour envisager des situations plus réalistes, nous ajoutons un mécanisme générant de la non-réponse. Nous ne nous attachons qu'à la non-réponse totale, supposant qu'aucune non-réponse partielle ne vienne entâcher les données d'enquête. La non-réponse totale consiste en l'incapacité de joindre le ménage ou dans le refus de l'ensemble de ses membres de répondre à l'enquête. Dans ce cas, nous ne disposons à propos de ce ménage d'aucune autre information que celles présentes dans la base de sondage.

33. « [C]ette probabilité ne peut jamais être calculée. »[11], p.182

34. Dans notre cas, n'étudiant que la non-réponse totale, la probabilité de réponse d'un individu est identique à celle de son logement.

Un mécanisme de réponse aléatoire (MAR)

Hypothèse

Nous modélisons un mécanisme de réponse ignorable (*Missing At Random (MAR)*) : on suppose que la probabilité de réponse d'un logement peut être expliquée par de l'information auxiliaire dont on dispose et qu'une fois prise en compte cette information, la probabilité de réponse n'est plus liée à la variable d'intérêt. Ce qui peut s'écrire, pour la variable d'intérêt y et l'information auxiliaire $\mathbf{x}_k$:

$$\mathbb{P}(r_k = 1|y_k, \mathbf{x}_k) = \mathbb{P}(r_k = 1|\mathbf{x}_k) \tag{13}$$

Nous supposons que cela est vrai pour toutes les variables d'intérêt utilisées ici, le chômage et l'activité en particulier. Cette hypothèse est très forte et nécessite, pour être la plus réaliste possible, de disposer d'une information auxiliaire relativement bien liée aux variables d'intérêt.

Or, l'information auxiliaire disponible dans la base de sondage est assez pauvre : il est peu vraisemblable qu'on puisse capter avec elle l'ensemble du lien existant entre non-réponse et les variables d'intérêt. Mais, nous nous en contenterons devant la difficulté de modéliser un mécanisme de réponse non-ignorable (NMAR) dans nos travaux.

Nous modélisons la non-réponse par un modèle logit :

$$\text{logit}\Big(\mathbb{E}(r_k|\mathbf{X} = \mathbf{x})\Big) = \text{logit}\Big(\mathbb{P}(r_k = 1|\mathbf{X} = \mathbf{x})\Big) = \beta_0 + \sum_{i=1}^{m} \beta_i x_i \tag{14}$$

Après une étape de sélection de variables, quatre variables de la base de sondage sont retenues (le type d'adresse, l'aspect du bâti, le regroupement géographique de la commune et la localisation de l'adresse dans un quartier de la politique de la ville (QPV)). À partir des probabilités estimées par le modèle, nous construisons des groupes de réponse homogène (GRH) par croisement des modalités des quatre variables explicatives retenues. Le modèle est entraîné sur les données de l'enquête annuelle 2019. Il permet d'estimer la probabilité de répondre de chaque unité échantillonnée. On attribue alors à chaque GRH un score de propension à répondre à l'enquête calculé comme la moyenne des probabilités de réponse estimées des unités échantillonnées.

À chaque tirage, on applique ces probabilités de réponse sur l'échantillon tiré pour obtenir l'échantillon de répondants S_r. Les probabilités de réponse sont, en réalité, corrigées d'un certain facteur afin d'obtenir un taux de collecte souhaité.

6.4 Redressement des poids de sondage en deux étapes

Pour corriger les erreurs dûes à la non-réponse et à un éventuel défaut de couverture de certaines sous-populations, Haziza et Lesage distinguent deux approches ([12]). *Une approche en deux étapes* consiste à corriger la non-réponse en redressant le poids des répondants de l'inverse d'une probabilité estimée de réponse suivie d'une étape de calage sur marges. *Une approche en une étape* mène de front correction de la non-réponse et calage. Afin de bien distinguer les effets sur nos estimations des deux sources de redressement, nous préférons une approche en deux étapes.

La correction de la non-réponse

La correction de la non-réponse totale consiste à réduire le biais des estimateurs provenant d'une différence entre les profils des répondants et des non-répondants.

Pour corriger la non-réponse totale, nous corrigeons l'erreur dûe à la non-réponse par une méthode développée par Haziza et Beaumont [35] ([14], pp.32-33). À partir d'une modélisation de la probabilité de réponse, les deux auteurs proposent un algorithme pour construire des GRH et

35. Nous nous sommes aussi appuyés sur [13].

en choisir le nombre optimal. En partant de deux classes, l'idée consiste à incrémenter le nombre de classes tant qu'elles ne sont pas jugées suffisamment homogènes en leur sein. Le critère choisi est le coefficient de détermination (R^2) qui rapporte la dispersion interclasse (dispersion entre les moyennes des probabilités de réponse par classe) à la dispersion totale des probabilités de réponse. Plus le R^2 est grand, plus l'homgénéité au sein des classes est grande. Dans nos simulations, un R^2 de $0,85$ s'est avéré une valeur pertinente.

Quand ces GRH sont construites, les poids de sondage des unités répondantes (S_r) sont corrigés de l'inverse de la probabilité de réponse moyenne observée dans le GRH à laquelle l'unité appartient. En notant $\hat{p}_k$ cette probabilité, les estimateurs par expansion de t_y, t_z, t_v et t_δ subissent une nouvelle expansion :

$$\begin{cases} \hat{t}_{ye2} &= \sum_{k\in S} 5\frac{z_k}{\pi_{2k}\hat{p}_k} \\ \hat{t}_{ze2} &= \sum_{k\in S} 5\frac{z_k}{\pi_{2k}\hat{p}_k} \\ \hat{t}_{ve2} &= \sum_{k\in S} 5\frac{v_k}{\pi_{2k}\hat{p}_k} \\ \hat{t}_{\delta e2} &= \sum_{k\in S} 5\frac{\delta_k}{\pi_{2k}\hat{p}_k} \end{cases} \tag{15}$$

Ainsi, à la suite de cette correction, nos deux estimateurs par le ratio sont :

$$\begin{cases} \hat{P}_{e2} = \frac{\hat{t}_{ve2}}{\hat{t}_{\delta e2}} \\ \hat{C}_{e2} = \frac{\hat{t}_{ye2}}{\hat{t}_{ze2}} \end{cases} \tag{16}$$

Le modèle utilisé est identique à celui qui a servi à la construction d'un mécanisme de réponse ignorable (équation (14)). Si le mécanisme de réponse a été entraîné sur les données de l'enquête Emploi annuelle 2019, il est, ici, appliqué sur les échantillons générés dans nos simulations.

Le fait de connaître parfaitement le mécanisme de génération de la non-réponse est une situation idéale. Cela aura tendance à présenter une correction de la non-réponse plus efficace dans nos simulations que celle à laquelle on peut s'attendre dans la réalité, ce mécanisme étant inconnu.

Une étape de calage

Le calage sur marges modifient les poids de sondage pour que les totaux connus au niveau de la population soient parfaitement estimés par le système de pondération final, tout en modifiant le moins possible les poids initiaux (ici les poids issus de la correction de la non-réponse). Si $\mathbf{X}$ est un vecteur de variables auxiliaires dont on connaît les totaux sur la population U, en notant le vecteur des totaux $\mathbf{t_x}$, on cherchera un nouveau jeu de poids $(w_k)_{k\in S_r}$ tels que :

$$\begin{cases} \sum_{k\in S_r} w_k \mathbf{x}_k = \mathbf{t_x} \\ \arg\min_{w_k} \left(\sum_{k\in S_r} G_k(w_k, d_k) \right) \end{cases} \tag{17}$$

où $G_k(w_k, d_k)$ est une fonction mesurant la distance entre anciens et nouveaux poids. Le système aboutit aux équations de calage suivantes [36] :

$$\sum_{k\in S_r} d_k \mathbf{x}_k F_k(\mathbf{x}_k'\lambda) = \mathbf{t_x} \tag{18}$$

où F_k est appelée *fonction de calage* et définie à partir de la dérivée de G_k et λ le vecteur des multiplicateurs de Lagrange.

Pour mener cette étape nous utilisons la fonction `icarus::calibration` [37] dans le langage de programmation `R`, fonction qui reprend la démarche utilisée par la macro `SAS CALMAR`. Elle nécessite de choisir les variables auxiliaires et une fonction de calage.

36. On reprend les notations de [11], p.217-219
37. Le package `icarus` a été développé par Antoine Rebecq ([15]).

Parmi les fonctions de calage, seule la fonction linéaire ne nécessite pas de méthodes numériques pour résoudre les équations de calage. De plus, [12] remarquent que « toutes les méthodes de calage sont asymptotiquement équivalentes dans le sens où elles mènent toutes à l'estimateur calé de la méthode linéaire » [38]. Pour éviter d'éventuels problèmes de convergence d'algorithmes d'optimisation lors de nos simulations, nous optons pour la méthode linéaire, qui repose sur la fonction de calage : $F(u) = 1 + u$. Cette méthode a néanmoins le désavantage de pouvoir produire des poids négatifs. L'utilisation d'une méthode permettant de borner les poids serait plus judicieuse dans un contexte de production. Nous nous accommodons pour nos travaux de simulations de cet inconvénient.

Les variables auxiliaires utilisées pour le calage sont :

— la population par sexe ;

— la population par tranche d'âge ;

— la population selon le lieu de naissance ;

— l'aspect du bâti (logement construit en dur ou non) ;

— la population selon qu'elle réside ou non en QPV ;

— la situation géographique de la commune (Nord, Sud, Petite Terre et Mamoudzou-Koungou).

Toutes ces variables sont *a minima* disponibles pour l'échantillon des répondants au travers du questionnaire de l'enquête Emploi. Les variables géographiques ou sur la structure du bâti sont directement disponibles dans la base de sondage. Les trois dernières variables de la liste sont celles qui nous servent à expliquer la non-réponse (voir *supra*) . Leur intégration dans l'étape de calage doit pouvoir réduire le biais de non-réponse si l'étape précédente ne l'a pas résorbé complètement. Le lieu de naissance est intégré car, à Mayotte, il est assez bien lié aux phénomènes économiques observés par l'EEC.Nous remarquons, au moment de conclure nos travaux, que nous n'avons pas intégré le type d'adresses dans nos variables auxilaires. Il est pourtant recommandé d'introduire les variables participant à la construction du plan de sondage lors de la phase de calage [39]. Mais, en réalité, nous ne disposerons pas de ces valeurs au moment d'appliquer le calage.

6.5 Remarque sur les probabilités d'inclusion et les poids dans les estimations trimestrielles et annuelles

L'échantillon trimestriel S, utilisé pour les estimations trimestrielles du taux de chômage ou de la part des chômeurs parmi les 15-74 ans, est composé de six sous-échantillons correspondant chacun à une des six vagues d'interrogation comme le veut l'enquête Emploi en continu. Ces six sous-échantillons sont entrés dans l'enquête à des moments différents et peuvent avoir été tirés à des années différentes. Par exemple, au deuxième trimestre de l'année N, l'échantillon trimestriel est composé des sous-échantillons entrés entre le T1 de l'année N-1 et le T2 de l'année N (figure 3, zone grisée verticale). Nous ajustons les probabilités d'inclusion brute (notons les π_{2k}^{tir}), c'est-à-dire celles issues directement du tirage de l'échantillon annuel, comme dans le chapitre précédent, d'un facteur 6/4 pour obtenir les probabilités d'inclusion conditionnelles dans l'échantillon trimestriel enquêté (notons les provisoirement π_{2k}^{enq} [40] : $\pi_{2k}^{tir} = \frac{6}{4}\pi_{2k}^{enq}$, soit du côté des poids : $d_{2k}^{tir} = \frac{4}{6}d_{2k}^{enq}$ où 6 est le nombre de sous-échantillons qui composent un échantillon trimestriel et 4 le nombre de trimestres dans une année (figure 3).

L'échantillon annuel est quant à lui composé de 9 sous-échantillons utilisés entre 1 et 4 fois, ce qui fait en tout 24 vagues d'interrogation (figure 3, zone sur fond vert clair). Le poids de

38. [12], p.132

39. Conseil prodigué dans le cours de *Techniques avancées d'échantillonnage* dispensé en octobre 2020 à l'Ensai par G. Chauvet

40. Elles correspondent en effet à nos π_{2k}

chaque vague de chacun des sous-échantillons se verra ainsi affecter d'un facteur $\frac{1}{24}$. Certains sous-échantillons, interrogés les quatre trimestres d'une même année, se verront ainsi attribuer d'un facteur de $\frac{4}{24}$.

Pour résumer

Le plan de sondage en deux phases nécessite d'utiliser des estimateurs par expansion pour estimer les totaux de nos variables d'intérêt. Les paramètres P (part de chômeurs parmi les 15-74 ans) et C (taux de chômage au sens du BIT) sont alors estimés par un estimateur par substitution (ratio) à partir de ces estimateurs par expansion. L'ajout d'un mécanisme de réponse ignorable ajoute un second degré d'expansion, comme si on ajoutait une troisième phase de tirage. La première étape du redressement consiste à corriger la non-réponse en estimant un score de propension à répondre à l'aide d'une méthode proposée par [14]. Cette étape cherche à réduire le biais dûe à la non-réponse totale. Enfin, un calage sur marges doit permettre de réduire encore ce biais de non-réponse en réutilisant les variables expliquant la non-réponse parmi les variables de calage, et réduire la variance des estimateurs.

Nous n'avons rien dit dans cette partie de la variance de nos estimateurs, ni exhiber de potentiels estimateurs de cette variance. Devant la complexité du plan de sondage retenu et l'empilement des phases de tirage et de redressements, nous nous sommes contentés d'estimer la variance de nos estimateurs par des travaux de simulations.

7 Analyse des résultats des simulations

On présente ici les principaux résultats des travaux de simulations menés pour estimer la variance des estimateurs de la part du chômage parmi les 15-74 ans et du taux de chômage.

7.1 Reproduire l'ensemble du processus d'échantillonnage

Pouvoir formuler une proposition d'estimation de la précision d'un plan de sondage demande de répliquer l'ensemble du processus d'échantillonnage. Pour cela, il nous faut reproduire la première phase de tirage, c'est-à-dire le tirage des groupes de rotation d'îlots. La reproduction n'a pas posé de problème puisque le Criem a pu nous fournir les programmes qui ont servi lors de la refonte des groupes de rotation en 2020. Néanmoins, le plan de sondage retenu, comme toutes les alternatives que nous avons envisagées, repose sur la construction de SAE optimisés sur les groupes de rotation d'îlots effectivement tirés lors de la première phase. Mais, nous ne disposons que d'un seul jeu de SAE : celui qui a été jugé optimal pour les groupes de rotation tirés en 2020 par le Criem. Cette étape est relativement lourde en temps de calcul : plusieurs heures, voire jours, sont nécessaires pour que tournent les algorithmes suffisamment de fois pour obtenir un jeu de SAE convenable. En termes de temps de calcul, il était impossible d'envisager intégrer cette étape dans nos simulations : si la construction d'un seul jeu de SAE ne prenait ne serait-ce qu'une heure (ce qui est loin d'être la réalité), 10 000 tirages auraient nécessité plus de 400 jours de calculs.

Rappel : Ensemble des étapes pour reproduire tout le processus d'échantillonnage :

1. Tirage des groupes de rotation des îlots ;
2. Construction d'un jeu de SAE optimisé sur les groupes de rotation générés en 1ere étape ;
3. Tirage de l'échantillon de logements :
   (a) Tirage des adresses (1er degré) ;
   (b) Tirage des logements (2nd degré).
4. Tirage des répondants

En revanche, le plan de sondage qui nous sert de référence ne fait pas intervenir les SAE. La deuxième étape n'affecte donc en rien l'échantillonnage. C'est pourquoi nous pouvons envisager d'estimer une précision globale des estimateurs sans craindre de devoir attendre une année pour disposer des résultats.

Pour le plan de sondage retenu, nous proposons deux approches. La première consiste à estimer un proxy de la précision globale en utilisant le jeu de SAE construit par le Criem. Cette approche a l'avantage de tenir compte des deux phases de tirage mais rend inadaptés les SAE en les déséquilibrant, parfois fortement, et jouera négativement sur la précision. La seconde approche consiste à raisonner conditionnellement à la première phase de tirage en utilisant les groupes de rotation tirés en 2020. Cette approche a l'avantage de rendre leur plein potentiel d'efficacité aux plans alternatifs et l'inconvénient de ne permettre d'estimer que la précision de la seconde phase de l'échantillonnage.

Avec la première approche, nous pourrons fournir une bonne estimation de la précision globale obtenue avec le plan de référence et une estimation plutôt conservatrice pour le plan de sondage retenu. La seconde approche permettra de comparer les deux plans de sondage *à égalité de traitements*, et vérifier si le plan retenu s'avère meilleur que le plan de référence.

7.2 Contexte simulatoire : aspects pratiques

L'enquête Emploi en continu n'étant pas encore lancée, il est impossible de travailler avec des données d'enquête réelles. Nous utiliserons la base du recensement de la population 2017 - dernier recensement exhaustif de Mayotte - comme un proxy de la « réalité »de la situation des individus vis-à-vis de l'emploi, du chômage et de l'activité. Nous supposons donc disposer d'une connaissance totale des caractéristiques de la population U.

Le chômage et l'activité mesurés par le recensement de la population ne respectent pas la définition que le BIT leur donne. Au recensement, il s'agit d'une situation déclarée par l'individu recensé alors que de nombreuses questions de l'enquête Emploi sont nécessaires pour mesurer le taux de chômage BIT. On observe en général que le taux de chômage au RP est très supérieur au taux de chômage BIT. Mayotte ne fait pas exception : au RP 2017, le taux de chômage s'élève à 42,3% de la population active, quand l'enquête Emploi annuelle estime le taux de chômage BIT autour de 30% en 2019 ([3]).

La variance d'une proportion p étant une fonction de $p(1-p)$ qui atteint son maximum quand la proportion est proche de 1/2, on risque de surestimer la variance de nos estimateurs en travaillant à partir des données brutes du RP. Pour une estimation plus réaliste, nous construisons, à partir des données du RP, *un proxy du chômage BIT* : un individu du RP est considéré comme chômeur s'il réside dans un logement ordinaire, déclare être chômeur à la question sur le type d'activité, être à la recherche d'un emploi et dont la situation principale déclarée est d'être au chômage. Cette dernière précision permet notamment de retirer certains étudiants. Nous construisons également *un proxy de l'emploi BIT* : Une personne en emploi est une personne de

15 ans ou plus résidant dans un logement ordinaire qui déclare être 'actif en emploi' et dont la situation principale déclarée est d'être en emploi ou au chômage. Ceci permet de construire un proxy BIT de la population active par réunion de population au chômage et celle en emploi. Avec ces définitions, nous obtenons *un proxy du taux de chômage BIT* évalué à $27,7\%$, plus proche de la mesure BIT de l'enquête Emploi.

7.3 Contexte simulatoire : aspect formel de la méthode de Monte-Carlo

Estimateurs de Monte-Carlo

On note θ un paramètre de $[0,1]$, correspond à C ou P, qu'on estimera par $\hat{\theta}$ ($\hat{P}$ ou $\hat{C}$). On réalise m simulations. Á la j^e simulation, on tirera l'échantillon $S^{(j)}$ à partir duquel on estimera le paramètre θ par $\hat{\theta}^{(j)}$ ($\hat{P}^{(j)} = \frac{\hat{t}_u^{(j)}}{\hat{t}_\delta^{(j)}}$ ou $\hat{C}^{(j)} = \frac{\hat{t}_y^{(j)}}{\hat{t}_z^{(j)}}$). L'estimateur Monte-Carlo de $\hat{\theta}$ est défini par la moyenne empirique des m estimations ponctuelles $\hat{\theta}^{(1)} \dots \hat{\theta}^{(m)}$:

$$\bar{\hat{\theta}} = \frac{1}{m} \sum_{j=1}^{m} \hat{\theta}^{(j)} \tag{19}$$

Les tirages étant issus d'un même plan de sondage et indépendants entre eux, la loi forte des grands nombres assure que $\bar{\hat{\theta}} \xrightarrow[m\to\infty]{p.s.} \mathbb{E}(\hat{\theta})$. Pour vérifier si $\hat{\theta}$ est un estimateur sans biais, on estime le biais relatif $\frac{\mathbb{E}(\hat{\theta})-\theta}{\theta}$ de l'estimateur $\hat{\theta}$ par :

$$\hat{\mathrm{BR}}(\hat{\theta}) = \frac{\bar{\hat{\theta}} - \theta}{\theta} \tag{20}$$

L'objectif principal de notre travail consiste à mesurer la variance de $\hat{\theta}$, $\mathbb{V}(\hat{\theta}) = \mathbb{E}\Big(\{\hat{\theta} - \mathbb{E}(\hat{\theta})\}^2\Big)$. On l'estimera à l'aide de la variance empirique :

$$\hat{\mathbb{V}}(\hat{\theta}) = \frac{1}{m} \sum_{j=1}^{m} \left(\hat{\theta}^{(j)} - \bar{\hat{\theta}}\right)^2 \tag{21}$$

Nous pourrons également exhiber un intervalle de confiance à 95% de l'estimation de θ. Nous choisissons de construire cet intervalle à partir des quantiles de la série d'estimations $\hat{\theta}^{(1)} \dots \hat{\theta}^{(m)}$. Ainsi, l'estimation de l'intervalle de confiance à 95% est défini par :

$$\hat{I}(\hat{\theta})_{0,95} = \Big[\hat{q}_{0,025}\,;\,\hat{q}_{0,975}\Big] \tag{22}$$

où $\hat{q}_{0,025}$ (resp. $\hat{q}_{0,975}$) est le quantile d'ordre $2,5\%$ (resp. $97,5\%$) de la série $\hat{\theta}^{(1)} \dots \hat{\theta}^{(m)}$.

L'erreur standard de Monte-Carlo

Les estimations par méthode de Monte-Carlo sont utiles si on s'asssure qu'on effectue un nombre suffisant de simulations pour assurer la bonne convergence des estimateurs. On estimera à cet effet l'erreur standard de Monte-Carlo :

$$\mathrm{se}(\hat{\theta}) = \sqrt{\frac{\mathcal{S}^2}{n}} \tag{23}$$

où $\mathcal{S}^2$ est la variance empirique corrigée de $\hat{\theta}$: $\mathcal{S}^2 = \frac{1}{m-1} \sum_{j=1}^{m} \left(\hat{\theta}^{(j)} - \bar{\hat{\theta}}\right)^2$.

Si le nombre de simulations est suffisant, l'erreur standard de Monte-Carlo doit être très faible.

7.4 Résultats sur la précision des estimations

Nous présentons uniquement les résultats des estimations issues des paramètrages suivants - sauf précision contraire dans le corps du texte :

— Taille de l'échantillon annuel d'entrants : 936 logements ;
— Taux de collecte : 69% ;
— Seuil entre petites et grandes adresses : $L = 4$;
— Nombre de simulations : entre 12 et 15 000.

Les tableaux de résultats 5, 6 et 7 font apparaître, dans l'ordre des colonnes, la valeur de l'estimation trimestrielle, son biais relatif en %, son écart-type et son coefficient de variation. Enfin, sont présentées les précisions trimestrielle [41] et annuelle, exprimées en pts de % avec un niveau de confiance à 95%.

Une estimation directe de la précision globale

Pour estimer la précision globale de nos estimateurs, c'est-à-dire une précision qui prend en compte les deux phases du tirage des échantillons, nous réalisons, dans un premier temps, une estimation directement issue des simulations selon la première approche décrite plus haut (voir section 7.1) et qu'on appelera estimation *directe*. Les résultats de ces estimations directes sont présentés dans le tableau 5 pour l'estimation de la part de chômage parmi les 15-74 ans et dans le tableau 6 pour l'estimation du taux de chômage BIT (proxy). En comparant les estimations directes issues des deux scénarios, nous observons que le scénario retenu semble faire moins bien que le scénario de référence : pour les deux indicateurs, la précision est meilleure dans celui-ci que dans celui-là. Cette approche ne permet en effet pas de rendre justice à l'étape d'optimisation de la construction des SAE sur les groupes de rotation, étape que nous n'avons pas pu reproduire.

Des estimateurs sans biais

Quel que soit l'indicateur ou le scénario considéré, les estimations sont bien corrigés du biais issu de la non-réponse. Les estimateurs présentent un biais relatif très inférieur à 1%.

Respect de la contrainte européenne

Le premier tableau (tab.5) permet de vérifier que la contrainte européenne est bien respectée : en effet, la part de chômage (proxy BIT) parmi les 15-74 ans est estimée à 10,28% avec un écart-type estimé à $0,87$ points de % pour le scénario retenu, alors même que, comme nous l'avons signalé, cette estimation directe est défavorable à notre scénario. En annexe A, nous avons mesuré l'écart-type maximal attendu par Eurostat pour notre estimateur à $1,22$ points de %. La contrainte européenne est donc respectée dans le cadre de nos simulations.

Table 5 – Précision globale de l'estimation de la part de chômage parmi les 15-74 ans

	Estimation (en %)	Biais rel. (en %)	Écart-type (en pts de %)	CV (en %)	Précision à 95% (en pts de %) trim.	annuelle
Scénario de base	**10,28**	**0,18**	**0,86**	**8,35**	**1,71**	**1,52**
Scénario retenu	10,28	0,19	0,87	8,50	1,74	1,54

Note de lecture : Avec le scénario de base, la part trimestrielle de chômeurs BIT (Proxy) parmi les 15-74 ans est estimée à $10,28\% \pm 1,71$ *points. Notes : Résultats pour* 12 000 *simulations ; Taux de collecte = 69% ; Seuil = 4 logements. L'erreur standard de Monte-Carlo est estimée à* 8.10^{-5} *dans les deux cas.*

41. La précision trimestrielle à 95% correspond à environ 2 fois l'écart-type de l'estimation trimestrielle, présenté en troisième colonne du tableau.

L'estimation du taux de chômage

La diffusion des résultats de l'EEC porte davantage son attention sur les taux de chômage, d'emploi et d'activité. Si Eurostat ne fixe pas de contrainte de précision au niveau départemental pour ces indicateurs, il n'en demeure pas moins qu'une précision raisonnable est nécessaire pour assurer la qualité des indicateurs publiés.

L'approche globale choisie ici étant plus adaptée au scénario de base, nous nous focalisons pour le moment sur ce scénario. Les estimations directes nous fournissent ainsi une estimation conservatrice de la précision des estimateurs. Le taux de chômage (proxy BIT) est estimé à 27,7% avec un faible biais relatif (0,13%). L'écart-type estimé vaut 1,91, soit une estimation trimestrielle du taux de chômage à $27,7\% \pm 3,8$ points, pour un niveau de confiance à 95%. Le dispositif envisagé pour la mise en place de l'enquête Emploi en continu à Mayotte induit, dans l'état actuel des choses, une perte de précision importante. Avec l'enquête Emploi annuelle menée à Mayotte aujourd'hui, le niveau de précision de l'estimation du taux de chômage annuel est estimé à $\pm 2,0$ points de pourcentage alors que nos simulations de tirage de l'EEC nous permettent d'estimer cette précision annuelle à $\pm 3,4$ points de % (tab. 6). Même si cette estimation est conservatrice, l'écart ne saurait être résorbé.

Pour chacun des deux scénarios étudiés et des deux paramètres présentés ici, la variance de l'estimation annuelle est dans un rapport quasi identique avec la variance trimestrielle. Le ratio des écarts-types est approximativement de 0,89. Une démonstration de cette relation, sous certaines hypothèses peut être fournie par les auteurs.

Table 6 – Précision globale de l'estimation du taux de chômage

	Estimation (en %)	Biais rel. (en %)	Écart-type (en pts de %)	CV (en %)	Précision à 95% (en pts de %) trim.	annuelle
Scénario de base	**27,70**	**0,13**	**1,91**	**6,91**	**3,82**	**3,41**
Scénario retenu	27,73	0,24	1,97	7,11	3,93	3,47

Note de lecture : Avec le scénario de base, le taux de chômage trimestriel BIT (Proxy) est estimé à $27,70\% \pm 3,82$ *points. Notes : Résultats pour* 12 000 *simulations ; Taux de collecte = 69% ; Seuil = 4 logements. L'erreur standard de Monte-Carlo est estimée à* 2.10^{-4} *dans les deux cas.*

Comparaison des scénarios sur la seconde phase de tirage

Pour comparer honnêtement nos deux plans de sondage, nous estimons la précision de l'estimateur du taux de chômage à partir de notre seconde approche qui consiste à tirer l'échantillon conditionnellement aux groupes de rotation tirés en 2020. Par ce procédé, nous annulons toute variabilité issue de la première phase pour estimer la seule variabilité issue de la seconde phase. Les résultats de l'estimation du taux de chômage sont présentés dans le tableau 7. Pour comparer les deux scénarios, nous mesurons un effet de plan, défini comme le rapport de l'écart-type obtenu pour le scénario retenu sur celui obtenu avec notre scénario de base. Avec un effet de plan estimé à 0,96, le scénario retenu améliore la précision de l'estimateur de 4%.

TABLE 7 – Précision de phase 2 de l'estimation du taux de chômage

	Estimation (en %)	Biais rel. (en %)	Écart-type (en pts de %)	CV (en %)	Précision à 95% (en pts de %) trim.	Effet de plan
Scénario de base	27,75	0,31	1,85	6,70	3,71	1,00
Scénario retenu	**27,74**	**0,29**	**1,78**	**6,42**	**3,55**	**0,96**

Note de lecture : L'effet de plan rapporte l'écart-type du scénario à l'écart-type du scénario de base. Le scénario retenu améliore la précision dûe à la seconde phase de tirage de 4% environ par rapport au scénario de base. Notes : Résultats obtenus avec 15 000 *simulations, un taux de collecte de 69% et un seuil de 4 logements. L'erreur standard de Monte-Carlo est estimée à* 2.10^{-4} *dans les deux cas.*

7.5 Une estimation indirecte de la précision globale

L'estimation directe est très conservatrice. Ne pouvant simuler correctement les deux phases pour le scénario retenu, nous utilisons le scénario de référence pour estimer la précision globale de nos estimateurs, tout en sachant que sur la seconde phase, notre scénario fait légèrement mieux que lui.

Pour fournir une autre estimation de la précision du taux de chômage, nous proposons une méthode indirecte. Il s'agit de calculer la variance totale du plan de sondage retenu ($\mathbb{V}_{tot}^{ret}(\hat{C})$) à partir de la variance de sa phase 2 ($\mathbb{V}_{P2}^{ret}(\hat{C})$) et en supposant que le rapport des variances entre l'estimation globale et l'estimation de phase 2 est identique pour les deux scénarios. Formellement, nous supposons que :

$$\frac{\mathbb{V}_{tot}^{ret}(\hat{C})}{\mathbb{V}_{P2}^{ret}(\hat{C})} = \frac{\mathbb{V}_{tot}^{base}(\hat{C})}{\mathbb{V}_{P2}^{base}(\hat{C})} \tag{24}$$

ce qui permet d'en déduire une estimation indirecte de la variance globale de l'estimation du taux de chômage pour le scénario retenu ($v_{tot}^{ret}(\hat{C})$) :

$$v_{tot}^{ret}(\hat{C}) = v_{P2}^{ret}(\hat{C}) \frac{v_{tot}^{base}(\hat{C})}{v_{P2}^{base}(\hat{C})} \tag{25}$$

D'après le tableau 8, le scénario retenu permettrait d'estimer le taux de chômage annuel à $\pm 3,3$ points de pourcentage près, l'estimation indirecte de la variance globale étant près de 10% plus basse que l'estimation directe.

TABLE 8 – Une estimation indirecte de la précision globale du plan de sondage retenu

	Variance			Précision (en pts de %)	
	Est. directe	Est. indirecte	Écart rel. (en %)	Est. directe	**Est. indirecte**
est. trimestrielle	3,87e-04	3,35e-04	-13,33	3,93	**3,66**
est. annuelle	3,01e-04	2,74e-04	-9,22	3,47	**3,31**

Note de lecture : L'estimation directe de la variance correspond à l'estimation obtenue à partir des simulations du tirage des deux phases. L'estimation indirecte de la variance correspond à l'estimation de la varinace totale recalculée à partir de l'estimation de la variance de la seconde phase du scénario alternatif et de l'estimation sur le scénario de base de la part de la variance de la phase 1 dans la variance totale. Note : Résultats obtenus à partir de 12 000 simulations pour les estimations directes et 15 000 simulations pour les estimations indirectes.

7.6 Vérification des autres contraintes

La question de l'autopondération a été réglée précédemment. En effet, le plan de sondage de base est autopondéré par principe et nous avons montré que le plan de sondage que nous retenons l'est sous des conditions qui s'avèrent, empiriquement, globalement satisfaites (voir section 5.3 et les résultats détaillés en annexe F).

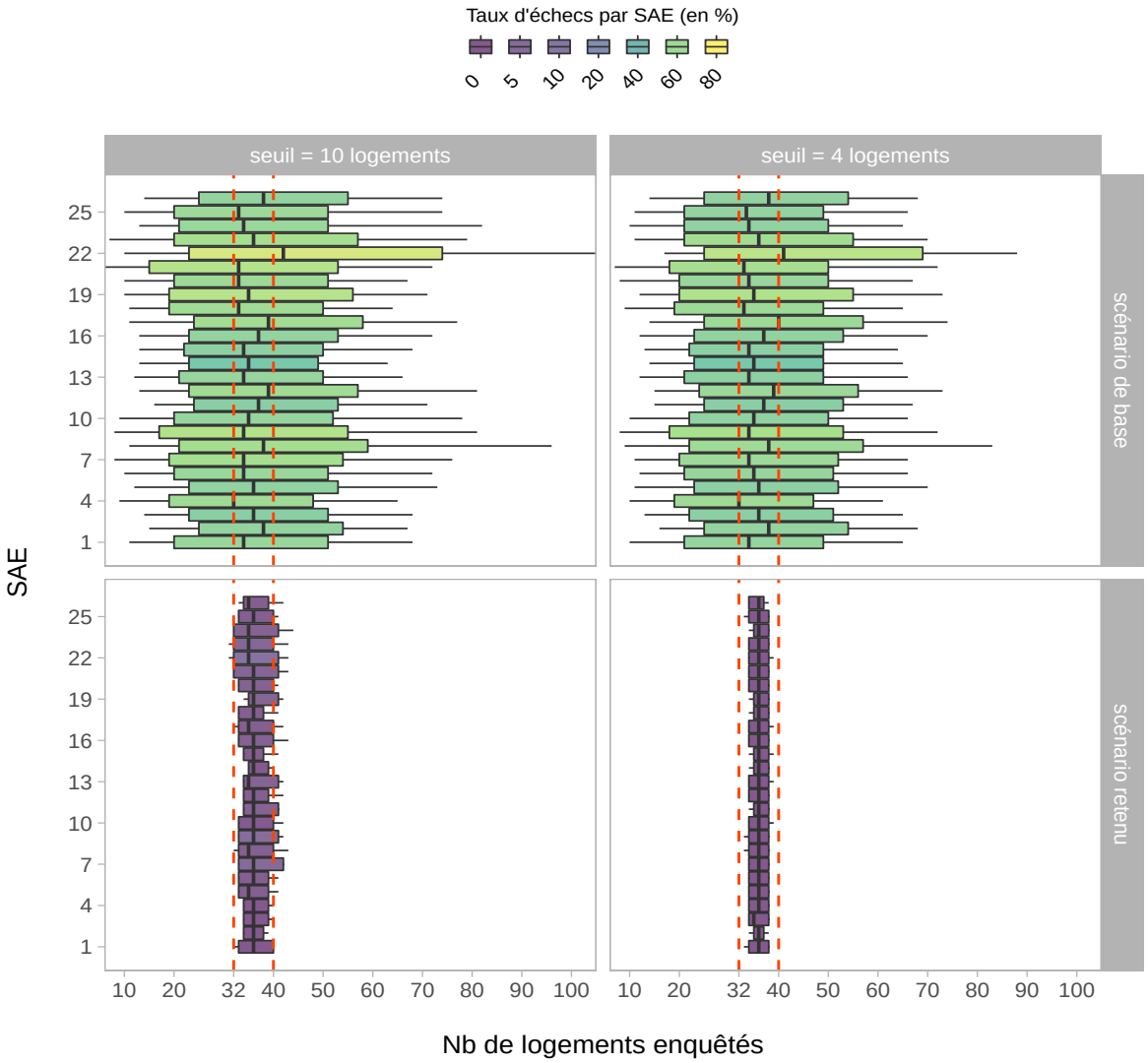
La taille de l'échantillon

Dans l'ordre des priorités, le respect d'une égale répartition de l'échantillon par SAE était tout en haut de notre liste. Selon cette contrainte, un échantillon annuel entrant, composé de 936 logements environ, doit pouvoir être réparti en 36 logements par SAE. Nous avons fixé un intervalle cible : un tirage est convenable si chaque SAE contient entre 32 et 40 logements, bornes comprises.

Le scénario de base n'a pas été conçu pour respecter cette contrainte et nos simulations montrent qu'il échoue systématiquement : sur 10 000 simulations, aucun tirage ne permet de remplir la condition : la cible n'est pas atteinte pour au moins un SAE à chaque tirage. Ceci est vrai quel que soit le seuil. Le scénario retenu, au contraire, réussit systématiquement dès lors que le seuil entre petites et grandes adresses est suffisamment abaissé. Pour un seuil fixé à 10, 40% des tirages seraient rejetés, alors qu'un seuil à 4 ou moins rend infime la part de rejet [42]. La figure 11 illustre le fait que le changement de scénario est la principale cause du respect de la contrainte de taille en limitant les fluctuations de la taille des échantillons par SAE.

42. Dans nos simulations, nous n'avons aucun rejet d'échantillon, mais cela n'implique pas nécessairement qu'il soit impossible de tirer un mauvais échantillon.

FIGURE 11 – Taille de l'échantillon par SAE, selon le scénario et le seuil fixé

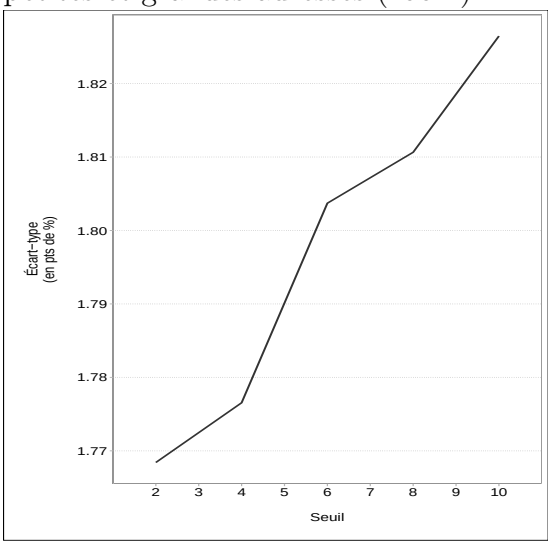


Note de lecture : Les boîtes à moustaches représentées sont construites à partir des tailles d'échantillon minimale et maximale aux extrémités des moustaches et des quantiles à 2,5%, 50% et 97,5%. Ainsi, chaque boîte contient 95% des tirages effectués. Pour un seuil fixé à 10, 95% des échantillons entrants annuels tirés dans le SAE 1 ont une taille comprise entre 20 et 51 logements lorsqu'on utilise le scénario de base, et entre 33 et 40 logements lorsqu'on utilise le scénario retenu. Note : Résultats obtenus à partir de 10 000 simulations, sans phase réjective additionnelle pour le plan de sondage retenu. Les lignes verticales en traits pointillés orange correspondent à la taille cible de l'échantillon par SAE (32-40 logements).

7.7 Le choix du seuil : pourquoi 4 ?

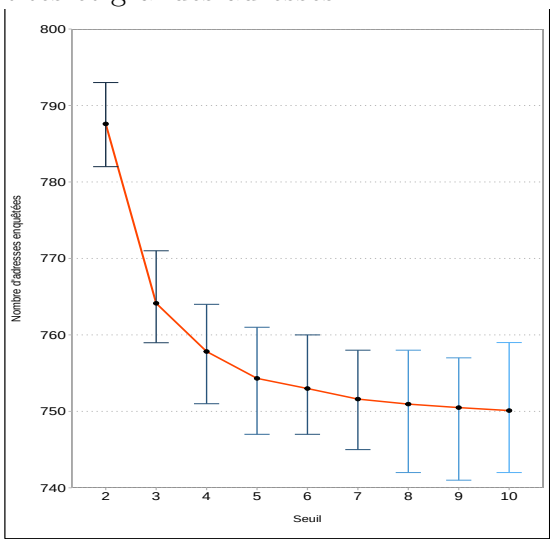
Aujourd'hui, la distinction entre les trois types d'adresses repose sur un seuil séparant petites et grandes adresses fixé à 10 logements. Or, ce seuil fût fixé sans une réelle expertise préalable. Nous proposons de conclure cet article en donnant quelques éléments justifiant la proposition d'abaisser le seuil à 4 logements.
Le premier argument en faveur de cet abaissement est la capacité à mieux contrôler la taille de l'échantillon quand le seuil de logements est plus faible, comme nous l'a montré la figure 11.
Le second argument est visible sur le graphique 12 : la précision s'améliore avec l'abaissement du seuil, même si, avec l'effet de zoom, le graphique exagère la réalité de cette amélioration : en passant le seuil de 10 à 4, l'écart-type de phase 2 de l'estimation du taux de chômage est diminué d'un peu moins de 3%. Ainsi, l'abaissement du seuil a un effet sur la précision du même ordre de grandeur que le passage du scénario de base au scénario retenu (-4%).

FIGURE 12 – Estimation de la précision de phase 2 de l'estimation du taux de chômage en fonction du seuil choisi entre petites et grandes adresses (zoom)



Résultats obtenus à partir de 15 000 simulations sous le plan de sondage retenu. Estimation de l'imprécision de la 2nde phase du tirage uniquement.

FIGURE 13 – Nombre d'adresses enquêtées en fonction du seuil choisi entre petites et grandes adresses.



Résultats obtenus à partir de 1 000 tirages sous le plan de sondage retenu.

Une solution extrême aurait pu consister à fixer un seuil à 2 logements, voire à supprimer la catégorie des petites adresses en le fixant à 1. Pourquoi s'arrêter à 4 ?
D'abord, l'abaissement du seuil conduit à augmenter le nombre d'adresses à enquêter et donc les temps de déplacement et de collecte. Avec un seuil fixé à 10, le nombre moyen d'adresses à enquêter est de 750. En abaissant le seuil à 4, le nombre d'adresses augmente peu (758 en moyenne), mais il faut compter près de 40 adresses supplémentaires si le seuil était fixé à 2 (figure 13).
Le seuil de 4 a finalement été préconisé par le GT EEC MAYOTTE, car il augmente relativement peu le nombre d'adresses à enquêter tout en permettant de gagner en précision et en contrôle de la taille de l'échantillon.

7.8 Bilan

A partir des estimations par la méthode de Monte-Carlo que nous avons menées, nous pouvons retenir que :

- La contrainte européenne de précision est satisfaite dans toutes les configurations envisagées ;
- Le scénario retenu respecte la taille d'échantillon cible par SAE pour un seuil fixé à 4, contrairement au scénario de base ;
- Nos estimateurs ne sont pas biaisés, après correction de la non-réponse et calage sur marges ;
- Conditionnellement à la réalisation d'un jeu de groupes de rotation d'îlots, le scénario retenu permet un gain de précision de 4% environ par rapport au scénario de base ;
- À partir d'une estimation directe et conservatrice, le taux de chômage annuel est estimé à $\pm 3,4$ points de pourcentage (contre $\pm 2,0$ points avec l'enquête annuelle actuelle) ;
- À partir d'une estimation indirecte rendant un peu plus justice à notre scénario, le taux de chômage annuel est estimé à $\pm 3,3$ points de pourcentage.

Annexes

A La contrainte de précision européenne

Le document ci-dessous (fig. 14), extrait d'un document d'Eurostat, décrit la façon de calculer la précision souhaitée pour la *Labor Force Survey*, dont l'enquête Emploi en continu est la version française. La première formule est une fonction de l'écart-type d'une variable binaire ($\sqrt{\hat{p}(1-\hat{p})}$ et d'une fonction de la taille de la population. Cette dernière reçoit différentes valeurs selon la taille de la population d'intérêt (ici la population des 15-74 ans). Cette population étant inférieure à 300 000 à Mayotte, on utilise la formule : $f(N) = \frac{1300}{0.3}N$.

En utilisant le proxy du chômage BIT que nous définissons en **??**, nous obtenons une part de chômage parmi les 15-74 ans de $10,26\%$ à partir des données du recensement 2017. La population des 15-74 ans à Mayotte s'élevant à $N = 142\ 771$ individus au RP 2017, nous pouvons appliquer la formule présentée dans le document européen et calculer σ_{max}, le seuil de précision maximal attendu :

$$\begin{aligned}
\sigma_{max} &= \sqrt{\frac{\hat{p}(1-\hat{p})}{f(N)}} \\
&= \sqrt{\frac{\hat{p}(1-\hat{p})}{1300\frac{142771}{300000}}} \\
&= \sqrt{\frac{10,26(1-10,26)}{618,7}} \\
&= 0,0122
\end{aligned} \tag{26}$$

Ainsi, nous devrons retenir un plan de sondage dont l'estimation de l'écart-type associé à l'estimation de la part de chômage (au sens du proxy BIT) parmi les 15-74 ans est inférieure à $1,22$ points de %, soit une précision de $\pm 2,44$ points de % avec un niveau de confiance à 95%.

Precision requirements

1. Precision requirements for all data collections are expressed in standard errors and are defined as continuous functions of the actual estimates and of the size of the statistical population in a country or in a NUTS 2 region.
2. The estimated standard error of a particular estimate $\widehat{SE}(\hat{p})$ shall not be bigger than the following amount:

$$1.\ \sqrt{\frac{\hat{p}(1-\hat{p})}{f(N)}}$$

3. The function *f(N)* shall have the form of *f(N)=a√N+b*
4. The following values for parameters *N*, a and b shall be used.

$\hat{p}$	*N*	*a*	*b*
Labour market domain: 3 precision requirements			
Estimated (national) quarterly unemployment-to-population 15-74 ratio	Country population aged 15-74 residing in private households, in million persons and rounded to 3 decimal digits	7800	-4500
Estimated (national) quarterly employment-to-population 15-74 ratio	Country population aged 15-74 residing in private households, in million persons and rounded to 3 decimal digits	7800	-4500
Estimated quarterly unemployment-to-population 15-74 ratio in each NUTS II region	Population aged 15-74 in the NUTS II region residing in private households, in million persons and rounded to 3 decimal digits	See footnote	

5. Should countries have negative *f(N)* value with the parameters expressed above, they will be exempted from the corresponding requirement.

For the estimated ratio unemployment to population 15-74 in each NUTS II region, the function *f(N)* is defined as follows:

$$f\left(N_{r,15-74}\right) = \begin{cases} 1300, & \text{if } N_{r,15-74} \geq 0.300 \text{ million inhabitants} \\ \frac{1300}{0.3} N_{r,15-74}, & \text{if } N_{r,15-74} < 0.300 \text{ million inhabitants} \end{cases}$$

FIGURE 14 – Calcul de la contrainte européenne de précision dans l'enquête Emploi

Source : Document n°Eurostat/F3/LAMAS/38/14, WORKING GROUP LABOUR MARKET STATISTICS, Document for item 2.1 of the agenda - IESS (Integrated European Social Statistics) Framework regulation : state of play and impact on the LFS, Décembre 2014, p.8

B Typologie des îlots

B.1 Description des classes

On peut décrire les six groupes de la manière suivante :

- **Groupe 1 :** La population est pour moitié née à Mayotte et pour un tiers née à l'étranger. Le taux d'activité de la population est le plus élevé de tous les groupes (1, 3, 4, 5 et 6) où la population métropolitaine est très minoritaire ;
- **Groupe 2 :** Ce groupe rassemble une population née plus souvent en métropole ou dans les DOM. Les agents publics, les diplômés d'un bac ou plus sont plus nombreux en proportion. Les habitations sont quasi exclusivement construites en dur et disposent des commodités essentielles (eau, électricité, toilettes) ;
- **Groupe 3 :** Ce groupe est le seul qui se caractérise par une population très majoritairement née à Mayotte. Les conditions de logement y sont plutôt bonnes avec un accès aux commodités très majoritaire ;
- **Groupe 4 :** Ce groupe se distingue du groupe 5 par un taux de chômage moins élevé et des conditions de logements un peu meilleur ;

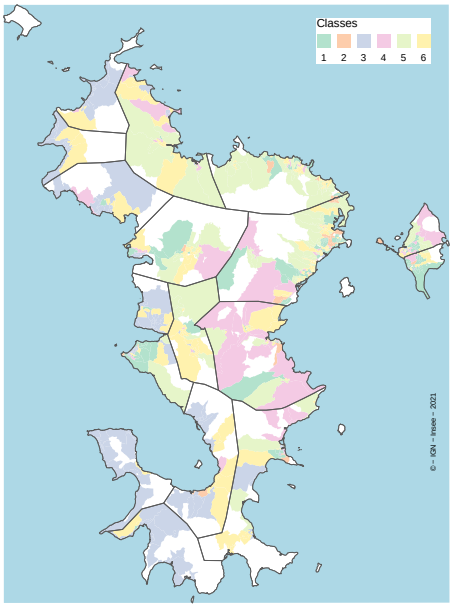


FIGURE 15 – Typologie des îlots à Mayotte

Source : Recensement de la population 2017

Les zones en blanc correspondent aux zones inhabitées de l'île.

- **Groupe 5 :** La population est pour moitié née à l'étranger, l'autre moitié étant née à Mayotte. Six personnes sur 10 ont moins de 18 ans. Le taux d'activité est plutôt faible et les actifs sont en majorité au chômage. Avec 4.69 individus par ménage, les ménages sont en moyenne plus grands que dans les autres groupes. La précarité en conditions de vie est assez forte : 20% des logements ne sont pas construits en dur, seul un logement sur 10 dispose de toilettes, 4 sur 10 n'ont pas un accès intérieur à l'eau ;
- **Groupe 6 :** Ce groupe est proche du groupe 5, mais les conditions de logements y sont meilleures : majoritairement construits en dur, ayant un peu plus accès aux commodités. La population est aussi plus active et plus souvent née à Mayotte qu'à l'étranger. Ce groupe se distingue du groupe 4 par un taux d'activité et de chômage plus fort : Un actif de 15-64 ans sur deux y est au chômage.

On peut remarquer sur la carte 15 que la composition des groupes n'est pas indépendante de la situation géographique des îlots : le fait d'appartenir à un groupe semble être un phénomène spatialement autocorrélé. Ceci peut réduire les gains de précision obtenus grâce au tri des adresses selon leur groupe. En effet, dans certains SAE, notamment dans le sud de l'île, nombreuses sont les adresses à appartenir au même groupe de la typologie.

Peut-être qu'une typologie des îlots construite indépendamment dans chacun des SAE serait plus pertinente et efficace pour réduire la variance de nos estimateurs.

TABLE 9 – Centre des classes de la typologie des îlots

Classe	1	2	3	4	5	6
Part de la population de moins de 18 ans	0.46	0.34	0.46	0.51	0.57	0.51
Part de la population de 65 ans ou plus	0.03	0.02	0.06	0.03	0.01	0.03
Taux de chômage des 15-64 ans	0.30	0.11	0.35	0.30	0.57	0.53
Taux de chômage des 15-24 ans	0.58	0.33	0.60	0.44	0.71	0.74
Taux d'emploi des 15-64 ans	0.44	0.73	0.38	0.24	0.13	0.26
Taux d'emploi des 15-24 ans	0.11	0.24	0.10	0.07	0.04	0.08
Taux d'activité des 15-64 ans	0.64	0.82	0.59	0.34	0.33	0.56
Taux d'activité des 15-24 ans	0.28	0.36	0.26	0.13	0.18	0.30
Part de salariés en contrat autre que le CDI	0.24	0.21	0.28	0.24	0.36	0.33
Part d'agents publics parmi les salariés	0.32	0.46	0.30	0.30	0.15	0.22
Part d'ouvriers et de techniciens parmi les salariés	0.11	0.07	0.12	0.13	0.21	0.18
Part de personnes à temps partiel parmi les salariés	0.11	0.06	0.12	0.13	0.24	0.20
Part de personnes nées à l'étranger	0.36	0.26	0.17	0.39	0.47	0.38
Part de personnes nées à Mayotte	0.52	0.29	0.77	0.58	0.52	0.59
Part de personnes nées en France hors Mayotte	0.12	0.45	0.06	0.04	0.01	0.03
Part de personnes ayant au moins le bac	0.39	0.72	0.29	0.21	0.12	0.19
Part de logements construits en dur	0.78	0.92	0.87	0.57	0.20	0.66
Part de logements avec un accés à l'eau à l'intérieur	0.86	0.93	0.87	0.70	0.39	0.74
Part de logements dipsosant de l'électricité	0.95	0.99	0.96	0.89	0.75	0.93
Taille moyenne des ménages	3.57	2.89	3.81	4.18	4.69	4.08
Part de logements dont le sol est en terre battue	0.21	0.07	0.20	0.37	0.67	0.37
Part de logements disposant de toilettes	0.66	0.89	0.66	0.40	0.13	0.44
Part de la population des 15 ans ou plus (en %)	15.7	2.7	20.2	10.8	23.0	27.6

B.2 Lien de la typologie avec l'activité, l'emploi et le chômage

Les six classes se distinguent assez bien sur les trois paramètres d'intérêt de l'enquête Emploi que nous avons retenus : le taux d'activité, le taux de chômage et le taux d'emploi. Même quand deux classes sont très proches sur un indicateur tel que le chômage pour les groupes 3 et 4, ils sont relativement bien distincts sur les deux autres (fig.16).

Pour vérifier si la différence observée des taux par classe est significative, on mène une analyse de la variance à partir d'un modèle à effet fixe. Soit pour la variable à expliquer y, i un îlot et j une classe de la typologie, on suppose que les 6 échantillons $y_{1j} \dots y_{n_j j}$ sont des échantillons indépendants issus d'une distribution gaussienne de même variance et de moyenne possiblement différente.

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij} \tag{27}$$

μ est la moyenne du taux observé toutes classes confondues et les termes d'erreur suivent une loi normale centrée. On procède au test de Fisher d'égalité des moyennes de chaque classe :

$$\text{H0} : \alpha_1 = \alpha_2 = \dots = \alpha_6 = 0 \text{ vs H1} : \exists j, \alpha_j \neq 0$$

Le résultat du test mené pour chaque variable d'intérêt est présenté dans le tableau et conduit à rejeter l'hypothèse nulle (tab. 10).

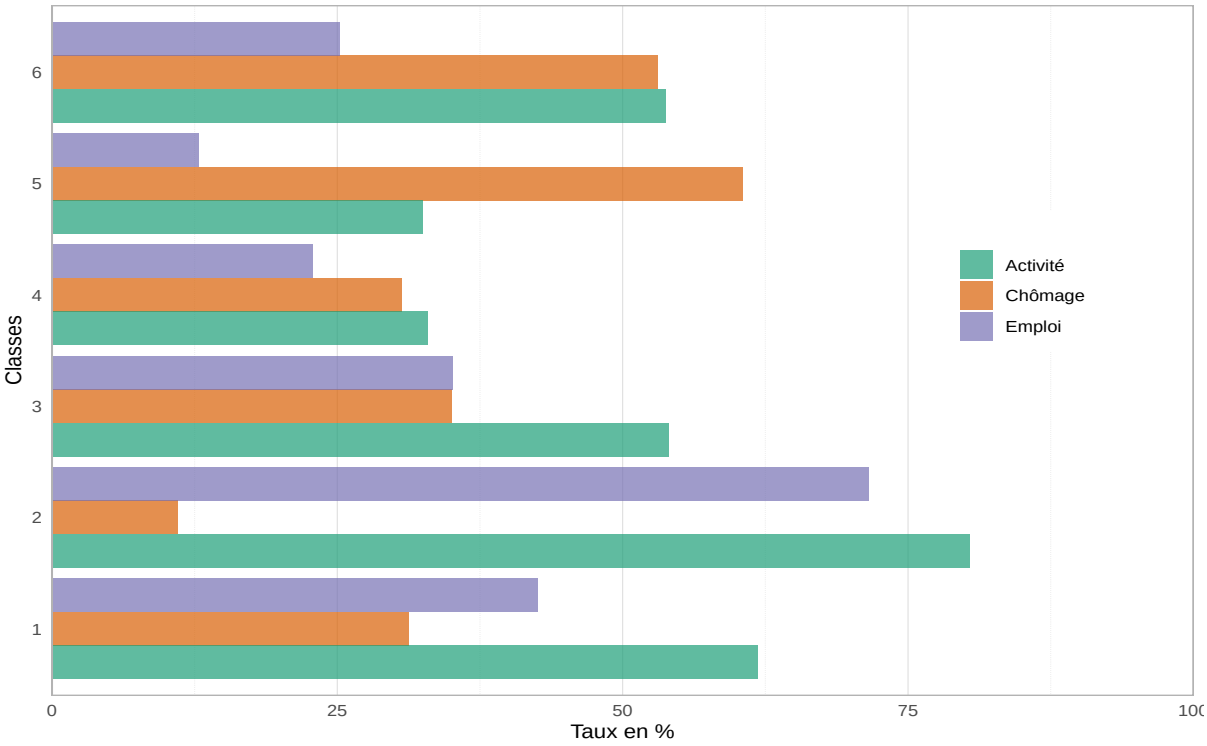


FIGURE 16 – Activité, emploi et chômage selon la classe de la typologie

Source : Recensement de la population 2017

	taux de chômage	taux d'emploi	taux d'activité
F-stat	179.5	561.6	247.7
p-valeur	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$
R^2	0.538	0.744	0.622

TABLE 10 – Résultats du test de Fisher

Note : Le R^2 correspond au coefficient de détermination soit le rapport entre la dispersion interclasse sur la dispersion totale.

C Comparaison de la taille des échantillons générés par le scénario retenu et un tirage équilibré

Le graphique 17 représente la fluctuation de la taille des échantillons simulés (abscisses) par secteur d'activité (ordonnées), selon le scénario envisagé (colonnes) et le seuil utilisé pour distinguer petites et grandes adresses (lignes). Pour chaque SAE, la fluctuation d'échantillonnage est représentée par une boîte à moustaches classique où la boîte proprement dite regroupe 50% des valeurs. Les deux lignes pointillées vertes représentent l'intervalle de fluctuation autorisé, soit entre 32 et 40 logements, pour une cible à 36 logements. Plus une boîte est de couleur foncée plus elle est problématique car moins d'échantillons atteignent la cible. Enfin, les deux scénarios comparés sont le scénario retenu (*V2-alloc-alter*) et un scénario dans lequel on effectue un tirage équilibré avec des probabilités proportionnelles à la taille de la strate SAE*type d'adresse (*V3-equi-pro*).

Un tirage est équilibré permet ainsi de générer des échantillons dont la taille par SAE est relativement bien centrée sur l'intervalle cible. Néanmoins, le résultat n'est pas homogène : dans les SAE 4 ou 24, la médiane est située au bord supérieur de l'intervalle, voire au-delà pour les seuils 5 et 10. De plus, de nombreux tirages ont des valeurs très éloignées de la cible. Au contraire, le scénario retenu est construit pour respecter la cible à chaque coup. Enfin, ce que ne montre pas ce graphique et qui finit de nous convaincre d'abandonner les tirages équilibrés, c'est l'incapacité des ces tirages de nous présenter des tirages où la condition est vérifiée pour tous les SAE en même temps.

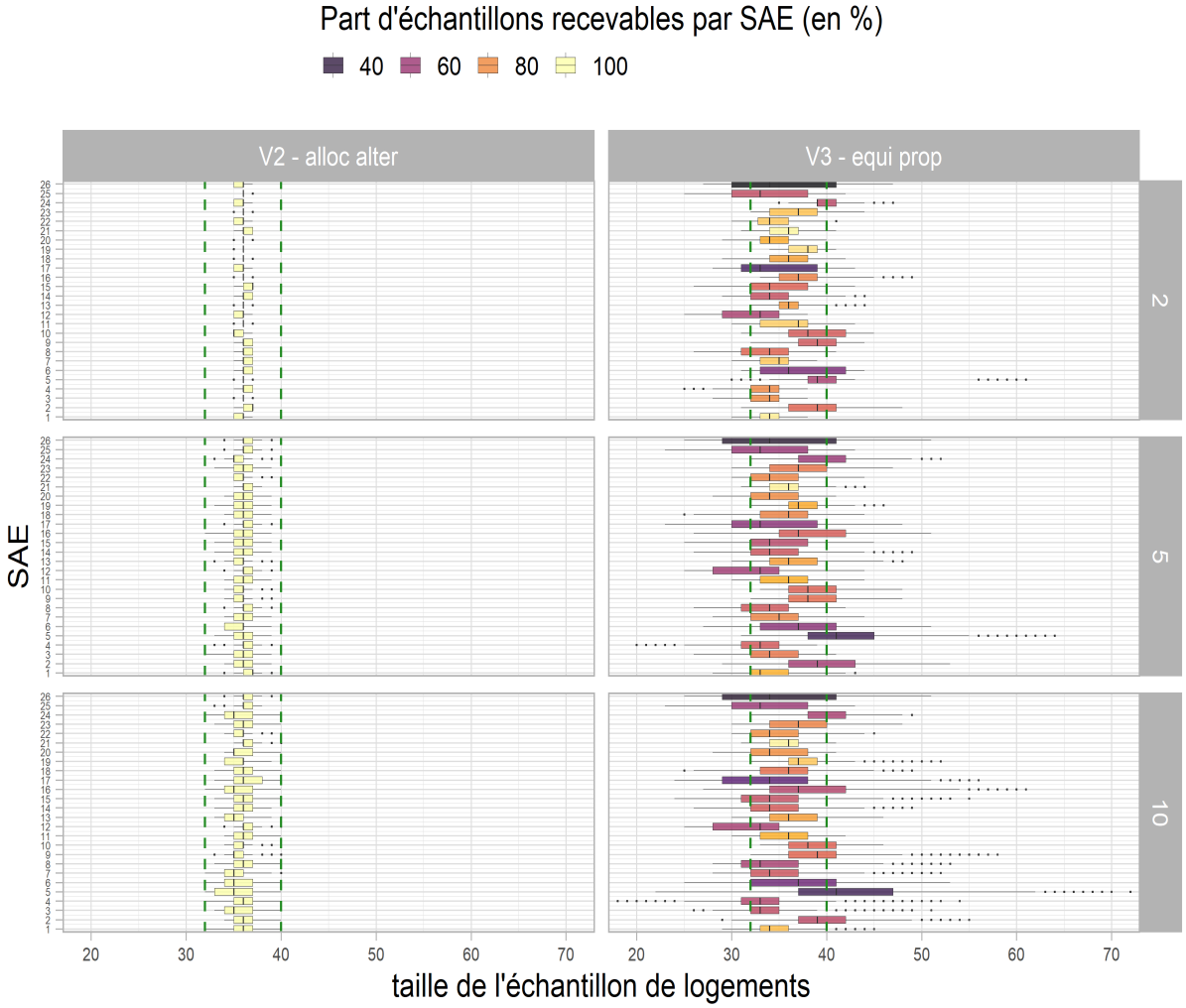


FIGURE 17 – Comparaison de la taille de l'échantillon dans chacun des SAE selon le seuil et le scénario envisagé, entre le scénario retenu et un tirage équilibré à probabilités proportionnelles.

D Scénario de base : détermination des allocations d'adresses

L'échantillon de l'enquête emploi en continu sera tiré dans une base d'adresses, la base cartographique. Néanmoins, ce sont les logements qui sont enquêtés et c'est à ce niveau que sont fixés les objectifs de précision. Ainsi, pour une taille de logements cible, il est nécessaire de déterminer le nombre d'adresses à échantillonner.

Nous nous proposons ici de déterminer la taille de l'échantillon d'adresses étant donnée la taille cible de l'échantillon de logements tout en obtenant, *in fine*, un plan de sondage autopondéré. Un tirage autopondéré de logements consiste en un tirage par lequel les logements reçoivent tous la même probabibilité d'être tiré (probabilité d'inclusion dans l'échantillon).

Notations utilisées :

- U^{adr}, respectivement U^{lgt}, l'ensemble des adresses de la base de tirage, respectivement l'ensemble des logements à ces mêmes adresses ;
- S^{adr}, respectivement S^{lgt}, l'échantillon d'adresses, respectivement l'échantillon de logements ;
- N^{adr}, respectivement N^{lgt}, le nombre d'adresses dans la base de tirage, respectivement le nombre total de logements à ces mêmes adresses ;
- n^{adr}, respectivement n^{lgt}, le nombre d'adresses échantillonnées, respectivement le nombre total de logements échantillonnés ;
- π^{adr}, respectivement π^{lgt}, la probabilité d'inclusion d'une adresse, respectivement d'un logement ;
- $h \in \{1 \dots H\}$, le numéro de la strate considérée ;
- S_h^{adr}, respectivement S_h^{lgt}, l'échantillon d'adresses dans la strate h, respectivement l'échantillon de logements dans la strate h ;
- N_h^{adr}, respectivement N_h^{lgt}, le nombre d'adresses dans la strate h dans la base de tirage, respectivement le nombre total de logements à ces mêmes adresses ;
- n_h^{adr}, respectivement n_h^{lgt}, le nombre d'adresses échantillonnées dans la strate h, respectivement le nombre total de logements échantillonés dans la strate h ;
- parmi les adresses, on distingue trois types :
  - MONO : adresses monologements où une adresse est composée d'un seul logement ;
  - PA : petites adresses où une adresse est composée de 2 à 10 logements ;
  - GA : grandes adresses composées d'au moins 11 logements ;
- ainsi, on notera, par exemple, $N_{h,PA}^{adr}$ le nombre total de petites adresses dans la strate h, $S_{h,PA}^{adr}$ l'échantillon de petites adresses tiré et $n_{h,PA}^{adr}$, la taille de cet échantillon.

Déroulement du tirage :

On effectue un tirage stratifié par strates composées du croisement $h\times$ type d'adresse. Le tirage effectué dans chaque strate est différent selon le type des adresses rencontré. Du point de vue 'logements', on peut décrire ces tirages ainsi :

- les tirages dans les strates de monologements sont des tirages directs : tirer une adresse, c'est tirer un logement ;
- les tirages dans les strates de petites adresses sont envisagés comme des tirages par grappes : dans une adresse tirée, tous les logements de l'adresse sont enquêtés ;

— les tirages dans les strates de grandes adresses sont conçus comme des tirages à deux degrés : dans un premier temps, un tirage des unités primaires (UP, les adresses) est effectué, suivi, dans un second temps, d'un tirage aléatoire simple de 10 logements dans chacune des adresses tirées au premier degré.

Étant donnée la taille cible de l'échantillon de logements (n^{lgts}), il reste à déterminer la taille des échantillons d'adresses par strate à tirer pour atteindre l'autopondération des logements souhaitée. En réalité, il reste également à déterminer le type de tirage des grandes adresses.

Autopondération des logements

L'objectif d'autopondération des logements consiste à obtenir des probabilités d'inclusion égales pour tous les logements :

$$\forall h \in \{1 \dots H\}, \pi^{lgt}_{h,MONO} = \pi^{lgt}_{h,PA} = \pi^{lgt}_{h,GA} = \frac{n^{lgt}}{N^{lgt}} \tag{28}$$

Ainsi, cela revient à tirer un nombre de logements dans chaque strate proportionnel à la taille de la strate dans U^{lgt} :

$$\forall h \in \{1 \dots H\}, n^{lgt}_{h,MONO} = \pi^{lgt}_{h,PA} = \pi^{lgt}_{h,GA} = \frac{n^{lgt}}{N^{lgt}} \tag{29}$$

$$\forall h \in \{1 \dots H\}, \begin{cases} n^{lgt}_{h,MONO} = n^{lgt} \frac{N^{lgt}_{h,MONO}}{N^{lgt}} \\ n^{lgt}_{h,PA} = n^{lgt} \frac{N^{lgt}_{h,PA}}{N^{lgt}} \\ n^{lgt}_{h,GA} = n^{lgt} \frac{N^{lgt}_{h,GA}}{N^{lgt}} \end{cases} \tag{30}$$

Échantillonnage des monologements

L'échantillonnage des adresses ne contenant qu'un seul logement ne comporte aucune difficulté. La probabilité d'inclusion d'un logement est égale à la probabilité d'inclusion de son adresse, dans la strate considérée :

$$\begin{aligned} \forall h \in \{1 \dots H\}, \pi^{adr}_{h,MONO} &= \pi^{lgt}_{h,MONO} \\ &= \frac{n^{lgt}_{h,MONO}}{N^{lgt}_{h,MONO}} \\ &= \frac{n^{adr}_{h,MONO}}{Nh, MONO^{adr}}, \text{ car une adresse = 1 logement} \end{aligned} \tag{31}$$

Ainsi,

$$\begin{aligned} \forall h \in \{1 \dots H\}, n^{adr}_{h,MONO} &= n^{lgt}_{h,MONO} \frac{Nh, MONO^{adr}}{N^{lgt}_{h,MONO}} \\ &= n^{lgt}_{h,MONO}, \text{ car } \frac{N^{adr}_{h,MONO}}{N^{lgt}_{h,MONO}} = 1 \end{aligned} \tag{32}$$

D'où $\boxed{n^{adr}_{h,MONO} = n^{lgt}_{h,MONO}}$

Échantillonnage des petites adresses

Pour atteindre la taille cible de logements en petites adresses de $n^{lgt}_{h,PA} = n^{lgt}\frac{N^{lgt}_{h,PA}}{N^{lgt}}$, quel nombre de petites adresses doit-on échantillonner dans une perspective d'autopondération ?

Le tirage des logements en petites adresses est un tirage en grappes : on enquête tous les logements d'une petite adresse échantillonnée. Ainsi,

$$\begin{aligned}\forall h \in \{1 \dots H\}, \pi^{adr}_{h,PA} &= \pi^{lgt}_{h,PA} \\ &= \frac{n^{lgt}_{h,PA}}{N^{lgt}_{h,PA}} \end{aligned} \tag{33}$$

Le nombre de logements par adresse n'est pas constant ici. Mais, nous pouvons constater que $N^{lgt}_{h,PA} = \bar{N}^{lgt}_{h,PA} N^{adr}_{h,PA}$, où $\bar{N}^{lgt}_{h,PA}$ est le nombre moyen de logements par petite adresse dans la strate h (ou, ce qui est équivalent, le nombre moyen de logements par adresse dans la strate (h,PA)). Ainsi,

$$\forall h \in \{1 \dots H\},$$

$$\pi^{adr}_{h,PA} = \frac{n^{lgt}_{h,PA}}{N^{lgt}_{h,PA}} \tag{34}$$

$$= \frac{n^{lgt}_{h,PA}}{\bar{N}^{lgt}_{h,PA} N^{adr}_{h,PA}} \tag{35}$$

En posant $\boxed{n^{adr}_{h,PA} = \frac{n^{lgt}_{h,PA}}{\bar{N}^{lgt}_{h,PA}}}$, on obtient un échantillon de petites adresses qui assure l'autopondération des logements en petites adresses.

Échantillonnage des grandes adresses

Le tirage des logements en grande adresse est un tirage à deux degrés dont les unités primaires sont les adresses de la strate et les unités secondaires les logements.

Au premier degré, on souhaite tirer $n^{adr}_{h,GA}$ grandes adresses dans la strate h, selon un plan de sondage à définir et qui asurera l'autopondération des logements notamment.

Au second degré, il s'agit de tirer 10 logements dans chaque grande adresse échantillonnée par un sondage aléatoire simple sans remise. Comme nous ne disposons pas de base de sondage des logements, ceux-ci ne pourront être tirés que par l'enquêteur une fois sur place. Ainsi, il est difficile d'envisager un autre plan de sondage qu'un sondage aléatoire simple à probabilités égales au sein de l'adresse.

De ce fait, comme nous tirons 10 logements dans chaque grande adresse échantillonnée, nous avons, pour toute strate de grande adresse h :

$$n^{adr}_{h,GA} = \frac{n^{lgt}_{h,GA}}{10}$$

$$\boxed{n^{adr}_{h,GA} = \frac{1}{10} n^{lgt} \frac{N^{lgt}_{h,GA}}{N^{lgt}}} \tag{36a}$$

Pour assurer que le plan conduise à une autopondération des logements, il nous faut préciser le plan de sondage à utiliser lors du tirage des unités primaires. Comme le nombre d'adresses à tirer est déjà connu car dépendant du nombre de logements à obtenir, il reste à préciser les probabilités d'inclusion des différentes adresses.

Les probabilités d'inclusion à chaque étape du tirage sont les suivantes :

$$\begin{cases} \forall h \in \{1 \dots H\}, & \\ \pi^{adr}_{a \in (h,GA)} & \text{, probabilité d'inclusion d'une adresse au premier degré à déterminer} \\ \pi^{lgt}_{l|a} & = \frac{10}{N^{lgt}_a}, \text{ probabilité d'inclusion conditionnelle au second degré.} \end{cases}$$

où $a \in (h, GA)$ désigne une adresse dans la strate et $N^{lgt}_{a \in (h,GA)}$, le nombre de logements à cette adresse.

La probabilité d'inclusion d'un logement dans l'échantillon issu d'un plan à deux degrés est :

$$\forall h \in \{1 \dots H\}, \text{pour toute adresse } a \in (h, GA), \text{pour tout logement } l \in a, \pi^{lgt}_l = \pi^{adr}_a \pi^{lgt}_{l|a} \quad (37)$$

Pour obtenir un plan de sondage autopondéré en termes de logements, nous devons déterminer π^{adr}_a telle que $\pi^{lgt}_l = \frac{n^{lgt}_{h,GA}}{N^{lgt}_{h,GA}} = \frac{n^{lgt}}{N^{lgt}}$, par l'équation (30), qui assure l'allocation proportionnelle des logements dans chaque strate.

$$\begin{aligned} \forall h \in \{1 \dots H\}, \pi^{lgt}_l &= \pi^{adr}_a \pi^{lgt}_{l|a} \\ \frac{n^{lgt}_{h,GA}}{N^{lgt}_{h,GA}} &= \pi^{adr}_a \frac{10}{N^{lgt}_a} \\ \pi^{adr}_a &= \frac{n^{lgt}_{h,GA} N^{lgt}_a}{10 N^{lgt}_{h,GA}} \end{aligned}$$

$$\boxed{\pi^{adr}_a = n^{adr}_{h,GA} \frac{N^{lgt}_a}{N^{lgt}_{h,GA}}} \quad (38a)$$

Ainsi, on assure l'autopondération des logements en tirant les adresses avec des probabilités proportionnelles à leur taille (en nombre de logements). $\boxed{n^{adr}_{h,GA} = \frac{n^{lgt}_{h,GA}}{10}}$. Le tirage des grandes adresses au premier degré est ainsi un tirage à probabilités d'inclusion proportionnelles à leur nombre de logements.

Conclusion

Un tirage autopondéré de logements étant donnée la taille cible de l'échantillon de logements à tirer est envisageable si l'échantillonnage des adresses est tel que :

— on tire $n^{adr}_{h,MONO} = n^{lgt}_{h,MONO} = n^{lgt} \frac{N^{lgt}_{h,MONO}}{N^{lgt}}$ adresses de type monologements ;

— on tire $n^{adr}_{h,PA} = \frac{n^{lgt}_{h,PA}}{\bar{N}^{lgt}_{h,PA}} = n^{lgt} \frac{N^{lgt}_{h,PA}}{\bar{N}^{lgt}_{h,PA} N^{lgt}}$ petites adresses ;

— on tire $n^{adr}_{h,GA} = n^{lgt} \frac{N^{lgt}_{h,GA}}{10 N^{lgt}}$ grandes adresses, par sondage aléatoire simple à probabilités proportionnelles à la taille de l'adresse.

Synthèse

Avec le plan de sondage décrit ci-dessus, on se propose ici de faire le chemin inverse : calculer les pondérations des logements et montrer qu'on obtient effectivement un plan de sondage autopondéré.

Au sein des strates de monologements, on tire par sondage aléatoire simple un nombre de logements proportionnels à la taille de la strate. Ainsi, la pondération d'un monologement est :

$$d^{lgt}_{h,MONO} = d^{adr}_{h,MONO} = \frac{1}{\pi^{adr}_{h,MONO}} = \frac{N^{lgt}_{h,MONO}}{n^{lgt}_{h,MONO}} = \frac{N^{lgt}}{n^{lgt}}$$

Au sein des strates de petites adresses, on réalise un sondage par grappes : tous les logements d'une adresse échantillonnée sont enquêtés, ainsi :

$$d^{lgt}_{h,PA} = d^{adr}_{h,PA} = \frac{1}{\pi^{adr}_{h,PA}} = \frac{n^{lgt}_{h,PA}}{N^{lgt}_{h,PA}} = \frac{N^{lgt}}{n^{lgt}}$$

Au sein des grandes adresses, on réalise un tirage à deux degrés : - au premier degré, les adresses sont tirées proportionnellement à leur taille ; - au second degré, 10 logements sont tirés par un sondage aléatoire simple au sein de chaque adresse tiré au premier degré.

$$\begin{aligned}
d^{lgt}_{h,GA} &= d^{adr}_{a\in(h,GA)} \dot{d}^{lgt}_{l|a} \\
&= \frac{1}{\pi^{adr}_{a\in(h,GA)}} \frac{1}{\pi^{lgt}_{l|a}} \\
&= \frac{N^{lgt}_{h,GA}}{n^{adr}_{h,GA} N^{lgt}_{a}} \frac{N^{lgt}_{a}}{10} \\
&= \frac{N^{lgt}_{h,GA}}{10 \dot{n}^{adr}_{h,GA}} \\
&= \frac{N^{lgt}_{h,GA}}{n^{lgt}_{h,GA}} \\
&= \frac{N^{lgt}}{n^{lgt}}
\end{aligned}$$

Remarques

1- A propos du plan à deux degrés utilisé pour les grandes adresses, on peut préciser que ce plan de sondage respecte les deux propriétés d'indépendance et d'invariance. En effet, conditionnellement à l'échantillon d'unités primaires sélectionné, les tirages au sein des UP sont indépendants. En outre, le plan de sondage au second degré ne dépend pas du plan de sondage du premier degré.

2- Du fait du mode d'échantillonnage des petites adresses, l'échantillon de logements obtenu n'est pas de taille fixe. Il atteindra la cible de logements seulement en moyenne.

E Allocation en petites adresses dans le scénario de base

En notant,

— U_{pa}^a, la population formée des adresses de la strate des petites adresses ;
— N_{pa}^a, la taille de la population U_{pa}^a ;
— S_{pa}^a, l'échantillon d'adresses tiré dans la strate des petites adresses ;
— $\hat{n}_{pa}^l$, la taille effective de l'échantillon de logements en petites adresses ;
— N_k^l, le nombre de logements à l'adresse k ;
— $\mathbf{I}_k^a$, l'indicatrice d'appartenance de l'adresse k à l'échantillon S_{pa}^a ;
— π_k^a, la probabilité d'inclusion de l'adresse k dans l'échantillon S_{pa}^a ;
— $\bar{N}_{pa}^l$, le nombre moyen de logements par petite adresse (rappel) ;

on a :

$$
\begin{aligned}
\mathbb{E}(\hat{n}_{pa}^l) &= \mathbb{E}(\sum_{k \in S_{pa}^a} N_k^l) \\
&= \mathbb{E}(\sum_{k \in U_{pa}^a} N_k^l \mathbf{I}_k^a) \\
&= \sum_{k \in U_{pa}^a} N_k^l \mathbb{E}(\mathbf{I}_k^a) \\
&= \sum_{k \in U_{pa}^a} N_k^l \pi_k^a \\
&= \sum_{k \in U_{pa}^a} N_k^l \frac{n_{pa}^a}{N_{pa}^a}, \text{ car le tirage des petites adresses est un tirage à probabilités égales} \\
&= n_{pa}^a \frac{\sum_{k \in U_{pa}^a} N_k^l}{N_{pa}^a} \\
&= n_{pa}^a \bar{N}_{pa}^l
\end{aligned}
$$

Ainsi,

$$
\begin{aligned}
\mathbb{E}(\hat{n}_{pa}^l) = n_{pa}^l & \\
\iff n_{pa}^a \bar{N}_{pa}^l = n_{pa}^l & \\
\iff n_{pa}^a = \frac{n_{pa}^l}{\bar{N}_{pa}^l} & \qquad (39)
\end{aligned}
$$

F L'autopondération du scénario alternatif, sous quelles hypothèses ?

Objectif : Rechercher quelles hypothèses sont nécessaires pour assurer que la stratégie d'échantillonnage retenue permet de construire un plan de sondage autopondéré.

Notations :

- L, seuil de logements pour différencier petites et grandes adresses ;
- N, le nombre total de logements (tout type d'adresse, y compris les grandes adresses) ;
- n, la taille de l'échantillon entrant global ;
- s, un SAE ;
- n^s, la taille cible de l'échantillon compris dans le SAE s ;
- $\tilde{N}^s$, le nombre de monologements et de logements en petites adresses dans le SAE s ;
- N^s_{mo}, le nombre de monologements dans le SAE s ;
- N^s_{pa}, le nombre de logements en petites adresses dans le SAE s ;
- α, le nombre de grandes adresses tirées et qui sont localisées dans le SAE s ;
- ${}^{(\alpha)}\tilde{n}^s$, la taille de l'échantillon de monologements et de logements en petites adresses dans le SAE s, dans lequel α grandes adresses ont été préalablement tirées ;
- ${}^{(\alpha)}\tilde{n}^s_{mo}$, la taille de l'échantillon de monologements dans le SAE s, dans lequel α grandes adresses ont été préalablement tirées ;
- ${}^{(\alpha)}\tilde{n}^s_{pa}$, la taille de l'échantillon de logements en petites adresses dans le SAE s, dans lequel α grandes adresses ont été préalablement tirées ;
- ${}^{(\alpha)}\pi^s_l$, la probabilité d'inclusion du logement l du SAE s, dans lequel α grandes adresses ont été préalablement tirées.

Le tirage des adresses puis des logements, dans le scénario retenu, procède en plusieurs étapes :

1. L'idée principale est de tirer $n^s = \frac{n}{26}$ logements dans chacun des SAE (soit 36 avec la taille d'échantillon retenue de 936 entrants par an) ;
2. On tire les grandes adresses rassemblées dans une seule strate quelque soit le SAE. Ce tirage est effectué avec des probabilités d'inclusion des adresses proportionnelles à leur taille (en nombre de logements). Ceci permet *in fine* de tirer les logements des grandes adresses à probabilité égale, soit : $\frac{n}{N}$;
3. On retire de la taille de l'échantillon n^s des SAE concernés le nombre de logements des grandes adresses préalablement tirées, soit αL logements ;
4. Ainsi, dans le SAE s, il reste à tirer ${}^{(\alpha)}\tilde{n}^s = n^s - \alpha L$ monologements et logements en petites adresses ;
5. Au sein du SAE s, on effectue un tirage d'adresses stratifié dont les strates sont les deux types d'adresse restants (monologements et petites adresses), et dont les probabilités d'inclusion sont proportionnelles à la taille de la strate (en nombre de logements) dans le SAE restreint aux monologements et petites adresses ;
6. Ainsi, la probabilité d'inclusion d'un monologement l dans le SAE s est :

$$ {}^{(\alpha)}\pi^s_l = \frac{{}^{(\alpha)}\tilde{n}^s_{mo}}{N^s_{mo}} = \frac{{}^{(\alpha)}\tilde{n}^s}{\tilde{N}^s} $$

7. Le tirage des logements en petites adresses étant un tirage par grappes, nous obtenons une probabilité d'inclusion identique pour les logements en petites adresses.

Notre plan de sondage est autopondéré si les poids de tirage des logements sont égaux. On raisonne ici en termes de probabilités d'inclusion ce qui revient au même. Ainsi, notre plan de sondage est autopondéré si, quel que soit le logement l, sa probabilité d'inclusion, π_l est égale à $\frac{n}{N}$.

Par construction, les logements des grandes adresses répondent strictement à l'objectif, mais ce n'est pas le cas des autres logements. D'où les deux questions :

1. À quelle(s) condition(s) l'autopondération est-elle atteinte ?
2. Les hypothèses nécessaires sont-elles réalistes ?

Les hypothèses nécessaires

D'après ce qui précède, une condition suffisante pour que le plan de sondage respecte la condition d'autopondération des monologements et des logements des petites adresses est que, pour chaque SAE s, pour tout logement l dans ce SAE,

$$
\begin{aligned}
& {}^{(\alpha)}\pi_l^s = \frac{n}{N} && \text{(40a)}\\
\iff \quad & \frac{{}^{(\alpha)}\tilde{n}^s}{\tilde{N}^s} = \frac{n}{N} &&\\
\iff \quad & \tilde{N}^s = {}^{(\alpha)}\tilde{n}^s \frac{N}{n} &&\\
\iff \quad & \tilde{N}^s = (n^s - \alpha L)\frac{N}{n} && \text{(40b)}
\end{aligned}
$$

Pour une taille d'échantillon d'entrants $n = 936$ et un seuil de logements entre petites et grandes adresses $L = 4$, la condition (40a) revient à la suivante : $\boxed{\tilde{N}^s = (36 - 4\alpha)\frac{N}{936}}$.

Réalité des hypothèses

Rien ne permet, dans le plan de sondage retenu, d'assurer que la condition suffisante est respectée dans les faits. Néanmoins, les groupes de rotation des îlots ont été construits de telle sorte qu'ils soient équilibrés en termes de nombre de logements et de logements collectifs notamment. De même, les SAE ont également été construits afin d'assurer une certaine homogénéïté en termes de nombre de logements et de types d'adresses. Ces deux aspects nous permettent d'être optimiste. En revanche, les grandes adresses, quel que soit le seuil utilisé, sont très mal réparties sur le territoire : le déséquilibre entre SAE, malgré les efforts d'équilibrage, pourrait nuire au respect de la condition (40b).

Nous nous proposons ici de vérifier ce qu'il en est en menant quelques simulations. Nous mesurons l'écart existant entre $\tilde{N}^s$, c'est-à-dire le nombre de monologements et de logements en petites adresses dans le SAE s, et $(36 - 4\alpha)\frac{N}{936}$, c'est-à-dire le nombre théorique attendu pour que soit respecté la condition d'autopondération des logements du SAE. Nous présentons, dans le tableau 11, les valeurs théoriques pour différentes valeurs de α. Par l'équation (40a), l'écart peut aussi être mesuré entre le poids de tirage réel utilisé $\frac{1}{{}^{(\alpha)}\pi_l^s}$ et le poids théorique $\frac{N}{n}$.

Nous tirons 10 000 échantillons de logements selon le plan de sondage retenu à partir de la base cartographique 2020 qui a servi au tirage des enquêtes ménages en 2021. Cette base

Table 11 – Valeur théorique attendue de $\tilde{N}^s$ pour atteindre l'autopondération

α	0	1	2	3	4	5	6	7	8	9
$\tilde{N}^s$ attendue	553	492	430	369	307	246	184	123	61	0

Note de lecture : Si, dans un SAE donné, une grande adresse est tirée, le nombre total de monologements et de petites adresses qui assure l'autopondération des logements de ce SAE est de 492.
La valeur théorique attendue est donnée par la formule $(36 - 4\alpha)\frac{N}{936}$, où $N = 14\,380$ logements.

est donc restreinte à un seul groupe de rotation d'îlots. Nous présentons les résultats de l'écart absolu relatif (en %) :

$$\frac{|\tilde{N}^s - (36 - 4\alpha)\frac{N}{936}|}{(36 - 4\alpha)\frac{N}{936}} * 100$$

Les résultats présentés dans le graphique 18 et les tableaux 12 montrent que pour 24 SAE sur 26 les écarts entre le nombre de logements attendus et le nombre réel sont inférieurs à 25% pour la très grande majorité des tirages. C'est-à-dire que les poids des monologements et des petites adresses de ces SAE s'écartent de 25% de la valeur théorique attendue dans plus des trois quarts des tirages et de moins de 50% pour tous les tirages. En revanche, les SAE 5 et 16 sont problématiques. Dans le SAE 16, les écarts varient de 56 à 88%, dans le SAE 5, les poids sont en moyenne le double des poids théoriques attendus. Ces deux SAE ont un nombre de logements plus important que dans les autre SAE. Sur ce groupe de rotation au moins, ces deux SAE semblent moins bien équilibrés (fig. 19).

Table 12 – Écart absolu relatif en % par SAE

SAE	1	2	3	4	5	6	7	8	9	10	11	12	13
Moy.	21.5	23.4	10.8	1.4	107.5	25.0	8.7	16.8	5.8	14.1	4.4	15.6	10.2
Min.	21.5	14.8	5.6	0.7	89.6	23.5	2.2	13.9	3.7	4.8	4.4	6.2	4.0
Q1	21.5	24.2	5.6	0.7	89.6	23.5	2.2	13.9	3.7	15.4	4.4	16.6	4.0
Méd.	21.5	24.2	5.6	0.7	89.6	23.5	13.0	13.9	7.8	15.4	4.4	16.6	4.0
Q3	21.5	24.2	16.1	0.7	152.8	23.5	13.0	13.9	7.8	15.4	4.4	16.6	17.0
Max.	21.5	24.2	16.1	13.1	152.8	38.9	13.0	28.1	7.8	15.4	4.4	16.6	17.0

SAE	14	15	16	17	18	19	20	21	22	23	24	25	26
Moy.	4.5	15.3	67.8	17.1	11.4	6.8	6.2	4.2	1.4	5.3	5.3	13.2	19.4
Min.	3.2	14.4	56.2	8.7	2.6	1.3	4.6	3.1	1.4	0.0	4.1	3.4	11.3
Q1	3.2	14.4	56.2	8.7	13.4	1.3	4.6	3.1	1.4	0.0	4.1	14.1	21.2
Méd.	3.2	14.4	56.2	22.2	13.4	1.3	7.1	3.1	1.4	0.0	4.1	14.1	21.2
Q3	3.2	14.4	87.5	22.2	13.4	13.9	7.1	3.1	1.4	12.5	4.1	14.1	21.2
Max.	13.9	23.9	87.5	22.2	13.4	13.9	7.1	9.0	1.4	12.5	17.2	14.1	21.2

Note de lecture : Sur les 10 000 tirages effectués, les pondérations des monologements et des logements en petites adresses du SAE 1 s'écartent en moyenne de 21,5% de la pondération théorique du plan autopondéré. Écarts calculés sur 10 000 simulations à partir de la base de sondage des enquêtes ménages 2021 (enquête cartographique 2020).

Bilan

Les SAE sont pour la plupart d'entre eux construits de telle sorte que le scénario retenu se rapproche d'un plan autopondéré. Néanmoins, les SAE 5 et 16, du fait de leur grand nombre de logements et de leur grand nombre de logements en grandes adresses, s'écartent sensiblement du niveau théorique requis. Une modification légère de ces SAE, par exemple en redistribuant des îlots de monologements dans les SAE voisins, pourrait améliorer sensiblement la situation et réduire encore la dispersion globale des poids de sondage.

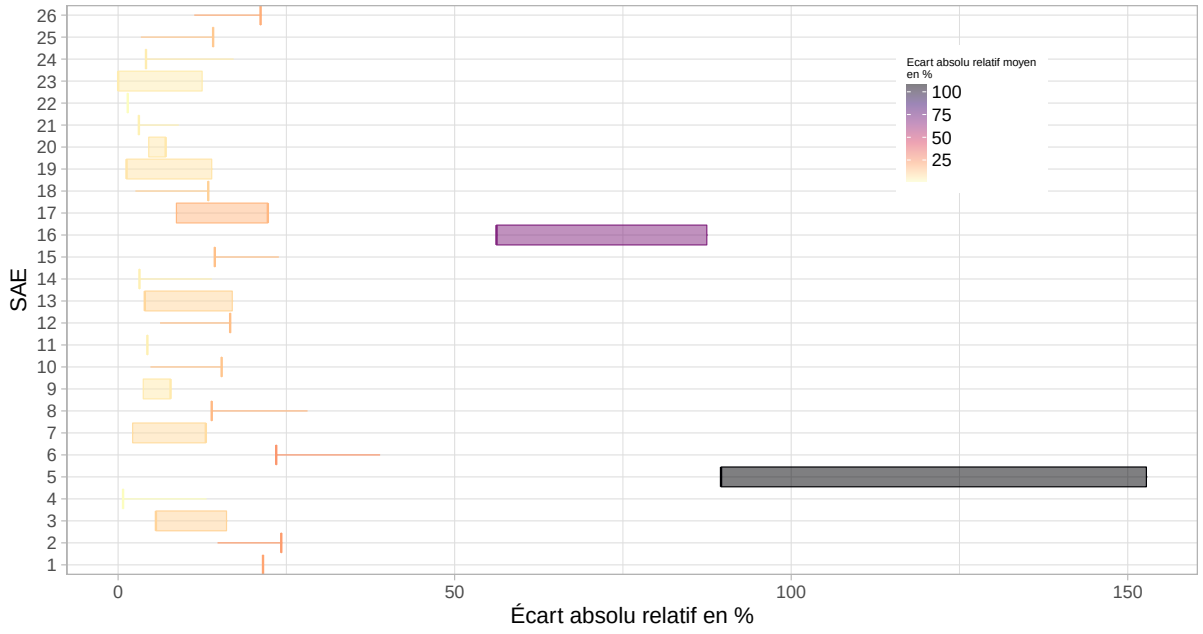


FIGURE 18 – Distribution des écarts à l'autopondération selon le SAE

Écarts calculés sur 10 000 simulations

Source : Base de sondage des enquêtes ménages 2021 - enquête cartographique 2020

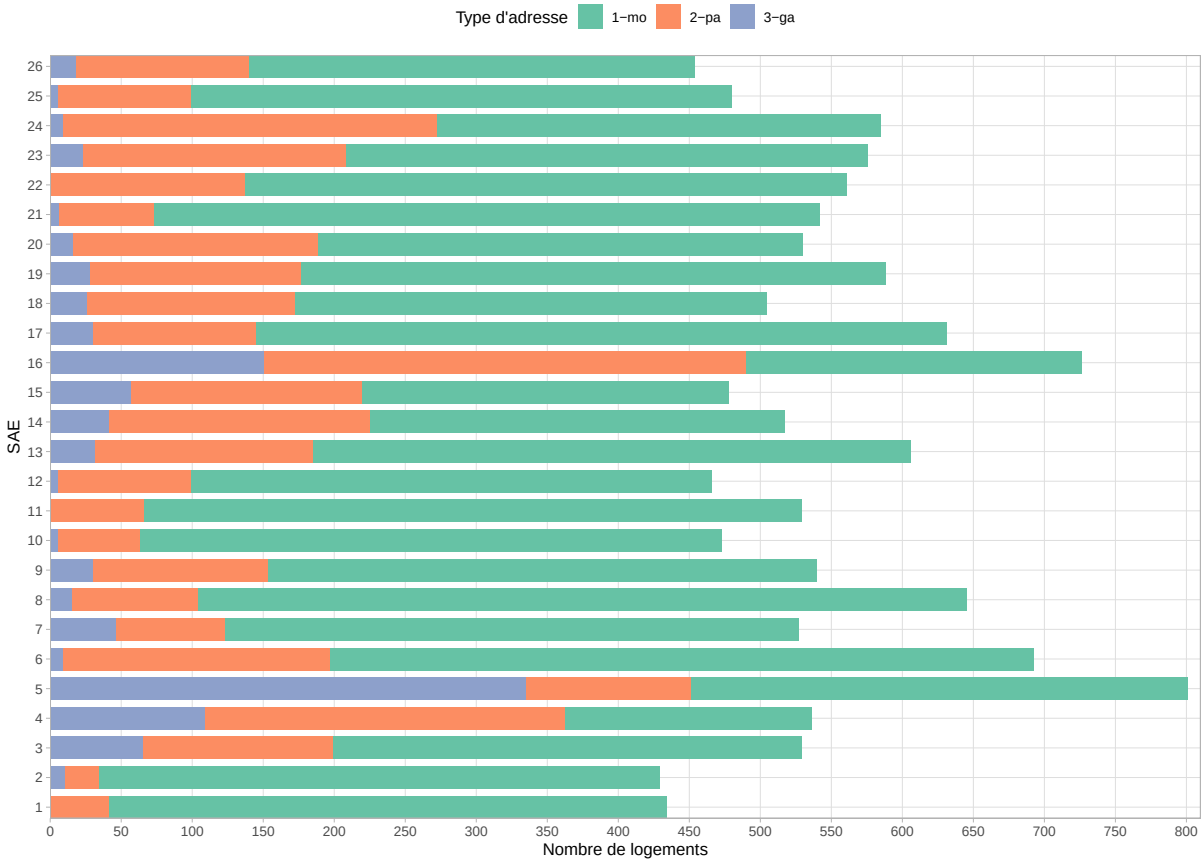


FIGURE 19 – Nombre de logements par type d'adresse et par SAE

Source : Base de sondage des enquêtes ménages 2021 - enquête cartographique 2020

Références

[1] A. Fleuret and J. Torterat, "Quinze ans d'enquêtes auprès des ménages dans les départements d'Outre-Mer," in *Journées de Méthodologie Statistique*, vol. 13, p. 1–86, 2015.

[2] Insee-DSDS, "Compte-rendu de la 1ère réunion du GT Mayotte du 10 juillet 2018." `https://www.agora.insee.fr/files/live/users/jj/dc/ji/DEQB4I/files/2018_14295_DG75-F201.pdf`, 2018.

[3] Insee-Mayotte, "Synthèse démographique, sociale et économique." `https://www.insee.fr/fr/statistiques/fichier/2018177/tiTEM.pdf`, 2020.

[4] Insee-Division-Emploi, "Enquête emploi, enquête sur l'emploi, le chômage et l'inactivité. méthodologie.." `https://www.insee.fr/fr/metadonnees/source/operation/s2022/documentation-methodologique`, 2021.

[5] Insee-DSDS, "Compte-rendu de la réunion du GT EEC Mayotte du 11 mars 2021." `https://www.agora.insee.fr/files/live/sites/dg-dsds/files/shared/F201_DERA/F201_PEEE/M%c3%a9lop%c3%a9e/Projet/GT%20Mayotte/2021_8733_DG75-F201.pdf`, 2021.

[6] Insee-Criem, "Note de tirage des groupes de rotation des îlots des petites communes mahoraises," 2020.

[7] Insee-DMTR, "Note de tirage des groupes de rotation des îlots des grandes communes mahoraises," 2020.

[8] P. Thibault, "Les villages de Mayotte en 2017," *Insee Analyses Mayotte*, Août 2019.

[9] P. Ardilly, *Les techniques de sondage*. Editions Technip, 2e édition ed., 2006.

[10] V. Loonis and M.-P. de Bellefon, *Manuel d'Analyse Spatiale*. Insee Méthodes, Insee, 2018.

[11] Y. Tillé, *Théorie des sondages : Échantillonnage et estimation en population finie*. Dunod, 2001.

[12] D. Haziza and E. Lesage, "A discussion of weighting procedures for unit nonresponse," *Journal of Official Statistics*, vol. 32, p. 129–145, Mars 2016.

[13] T. Deroyon, "Comment constituer des groupes de réponse homogène ? une comparaison de quelques méthodes appliquées aux enquêtes sectorielles annuelles en France," Octobre 2016.

[14] D. Haziza and J.-F. Beaumont, "On the construction of imputation classes in surveys," *International Statistical Review*, vol. 75, no. 1, p. 25–43, 2007.

[15] A. Rebecq, "Icarus : un package R pour le calage sur marges et ses variantes," 2016.