JMS Journées de méthodologie statistique de l'Insee 2018

Ségrégation urbaine: un éclairage par les données de téléphonie mobile

Lino Galiana *(*), Benjamin* Sakarovitch *(*), Zbigniew* Smoreda *(**)*

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

(**) Orange Labs, SENSE

lino.galiana@insee.fr ; benjamin.sakarovitch@insee.fr ;
zbigniew.smoreda@orange.com

Mots-clés. Big data, Ségrégation, Analyse urbaine, Economie spatiale

Résumé

La ségrégation urbaine a principalement été étudiée à travers le prisme de la ségrégation résidentielle. Néanmoins, il s'agit d'un phénomène plus général, à la fois répartition spatiale non homogène des groupes sociaux mais aussi absence d'interactions entre individus issus de groupes sociaux différents. L'utilisation des données de téléphonie mobile, où sont enregistrées de manière systématique des contacts individuels et leurs localisations, ouvre de nouvelles perspectives pour les études de ségrégation. Cet article utilise l'ensemble des communications des clients de l'opérateur Orange pendant le mois de Septembre 2007. A partir des propriétés spatiales des données de téléphonie, les données fiscales géolocalisées `Filosofi` sont utilisées pour estimer le revenu probable de chaque utilisateur de téléphonie mobile. Nous utilisons les interactions téléphoniques et les déplacements dans l'espace pour constuire des indices individuels d'exposition aux différents groupes sociaux. Ceci nous permet d'étudier le profil social et spatial de la ségrégation sociale, dimensions que l'approche résidentielle ne peut mesurer conjointement.

Abstract

Segregation is a multi-dimensional phenomenon, mostly studied through the glance of residential segregation. Interactions and localizations that are recorded in mobile phone data open new perspectives for segregation studies. This paper uses Orange customers' September 2007 communications (18 millions individuals). We use geolocalized fiscal data (INSEE's `Filosofi` database) to simulate phone users' likely income. Interactions and mobility through space are used to construct individual exposure measures to other social groups. It enables studying social and spatial pattern for social segregation, a complementary view to the residential segregation approach.

Introduction

La ségrégation est la traduction d'une inégale répartition d'individus provenant de groupes sociaux différents dans l'espace, conséquence d'un ensemble de comportements pouvant provoquer de l'entre-soi, qu'il s'agisse de comportements volontaires ou d'effets de tri spatial sans volonté explicite (effets à la Schelling, 1969). Il s'agit à la fois d'un processus de séparation sociale et spatiale. Parmi les principaux vecteurs de ségrégation, les plus documentés sont les choix de localisation résidentielle (Maurin, 2007), les effets spatiaux de la segmentation du marché du travail (L'Horty, 2015) ou encore l'absence de mixité sociale dans la pratique de loisirs ou la consommation de biens publics (dans l'optique du vote avec les pieds de Tiebout, 1956). La ségrégation est ainsi un phénomène protéiforme et multidimensionnel.

La ségrégation urbaine a principalement été étudiée à travers le prisme de la ségrégation résidentielle (Oberti and Préteceille, 2016). Floch (2017) a récemment montré la manière dont le revenu, créateur d'une hiérarchie dans l'espace social, structure également l'espace physique. Néanmoins, les données utilisées pour analyser la ségrégation résidentielle ne permettent de mesurer que la répartition spatiale des groupes sociaux à travers la question résidentielle, pas les interactions entre individus issus de groupes sociaux différents. Comme Chamboredon and Lemaire (1970) l'ont montré dans leur travail fondateur, la mixité sociale - la proximité spatiale de groupes sociaux différents - peut très bien être associée à une absence d'interaction entre ceux-ci (distance sociale). De même, des individus provenant de quartiers fortement ségrégués peuvent se rencontrer lors de leurs déplacements (Le Roux, Vallée, and Commenges, 2017). Si l'approche résidentielle offre une perspective intéressante sur la ségrégation, elle ne permet d'éclairer qu'une partie du phénomène. Mesurer des aspects complémentaires à la dimension résidentielle, devrait offrir une vision plus complète des phénomènes de ségrégation.

Les données de téléphonie mobile, parce qu'elles permettent un enregistrement systé-

matique des contacts et des localisations des individus, offrent un complément intéressant à l'approche traditionnelle. Blondel, Decuyper, and Krings (2015) présentent plusieurs champs où les données de téléphonie mobile ouvrent de nouvelles perspectives. Les plus prometteurs sont ceux liés à l'exploitation des propriétés géographiques des données mobiles. Les données de téléphonie offrent, par exemple, la possibilité de suivre de manière précise la distribution des populations sur le territoire (Deville et al., 2014). De manière générale, les données de téléphonie mobile éclairent d'un regard nouveau les phénomènes où l'exposition d'un utilisateur de téléphone mobile à d'autres individus ou à un événement est en jeu (Bengtsson et al., 2011). Interpréter les phénomènes de ségrégation comme une exposition faible de certains individus à une communauté identifiée ouvre ainsi la voie à l'analyse de la ségrégation à partir des données de téléphonie mobile. Les caractéristiques mesurées dans les données des opérateurs de télécommunications correspondent aux dimensions (mobilité et interactions sociales) qui échappent à l'analyse de la ségrégation par le prisme résidentiel.

Néanmoins, les données de téléphonie, aussi riches soient-elles sur la mobilité et les interactions individuelles, ne sont pas suffisantes en elles-mêmes pour étudier la ségrégation. La ségrégation est en effet une thématique pour laquelle une information sur l'échelle sociale telle que les données de revenu est essentielle. Cette variable, déterminante pour l'accès au marché du logement, affecte la localisation résidentielle tout en déterminant le statut social d'un individu. Enrichir les données de téléphonie par des données de revenu ouvre de nouvelles perspectives quant à notre compréhension des phénomènes de ségrégation.

La Section 1 présente la manière dont les données de téléphonie mobile peuvent éclairer notre compréhension de la ségrégation urbaine. La méthodologie développée pour croiser données mobiles et administratives est présentée dans la Section 2. Des résultats provisoires sont détaillés dans la Section 3. L'Annexe A est dédiée à la présentation plus approfondie de la grille spatiale utilisée.

1 Données mobiles et ségrégation

1.1 Utiliser des données mobiles pour mesurer la ségrégation

Historiquement, les études sur la ségrégation urbaine française se sont appuyées sur la mesure de la répartition spatiale des Professions et Catégories Socioprofessionnelles (PCS)[1]. Ce type de variable permet en effet de définir des groupes sociaux. Cependant,

[1]Pour une synthèse, voir Oberti and Préteceille, 2016

la hiérarchisation de l'espace social permis par les PCS n'est pas univoque[2]. L'approche par les revenus s'est fortement développée au cours des dernières années. La variable de revenu est un marqueur fort de la ségrégation résidentielle et est ainsi la principale responsable de la répartition spatiale des individus tout en constituant une variable déterminante dans la hiérarchie sociale (Tovar, 2009). L'utilisation de données agrégées à une échelle infra-communale (IRIS dans le cas français) ou de données fiscales géolocalisées (Filosofi[3] en France) offre une vision de la distribution spatiale des revenus permettant une étude approfondie de la ségrégation résidentielle (Floch, 2017). La principale conclusion qui se dégage des études sur le cas français[4], qu'elles soient fondées sur une variable de revenu ou de PCS, est que la ségrégation concerne davantage les individus les plus élevés dans la hiérarchie sociale (Floch, 2017). L'organisation spatiale des villes, en particulier l'organisation mono- ou polycentrique de celles-ci, peut également affecter l'ampleur de la ségrégation (Madoré, 2015).

Une telle approche ne permet néanmoins pas de mesurer des interactions sociales mais seulement la présence dans un même espace résidentiel d'individus issus de groupes sociaux différents. De plus, l'approche résidentielle n'offre qu'une vision statique de la ségrégation et n'informe pas sur le mélange des populations en pratique, via, par exemple, la fréquentation par des individus issus de groupes sociaux différents de lieux communs. Mesurer cette dimension de mixité sociale est possible à partir d'enquêtes déclaratives et offre une vision dynamique de la ségrégation, permettant de tenir compte de potentiels contacts avec les individus par la coprésence en un même lieu à la même heure. En utilisant les données déclaratives de l'Enquête Globale de Transport (EGT), Le Roux, Vallée, and Commenges (2017) ont ainsi montré que la ségrégation est moins forte en journée, les groupes sociaux se mélangeant plus lors des heures de travail. Cependant, ces enquêtes sont contraintes en termes de taille d'échantillon et soumises à des biais d'imprécision ou des risques d'oubli de la part de la personne enquêtée, qui peuvent affecter la mesure de la ségrégation. Si ces enquêtes constituent une mesure intéressante de la manière dont les groupes sociaux se rencontrent dans l'espace, elles ne permettent pas toutes de mesurer la récurrence d'interactions, seulement la présence dans un même espace de populations issues de groupes sociaux différents.

Les données de téléphonie mobile offrent un complément prometteur aux données

[2]Les PCS permettent une hiérarchisation globale de l'espace social. Cependant, au sein de chaque PCS, des situations diverses, en particulier en termes de revenu, existent et tendent à relativiser la hiérarchie sociale qui peut être déduite des PCS (Dabet and Floch, 2014)

[3]La base de données FiLoSoFi (Fichier Localisé Social et Fiscal) offre une vision exhaustive de la composition des revenus des ménages français, tout en permettant une localisation précise de ceux-ci à partir de coordonnées géographiques. Les données Filosofi sont constituées à partir des sources de la DGFiP et enrichies de données sur les prestation sociales reçues par le ménage (fournies, entre autres, par la CNAF ou la CNAV). Cette base de données a remplacé le dispositif Revenus Fiscaux Localisés (RFL), moins complet. Nous exploitons le premier millésime disponible des données Filosofi. Celui-ci présente les revenus perçus au niveau du ménage pendant l'année fiscale 2011.

[4]Reardon and Bischoff (2011) ont montré que la conclusion est similaire dans le cas américain

déclaratives pour mesurer la ségrégation. L'enregistrement automatique des appels ou messages texte dans les comptes rendus d'appels[5] (*call details record*, CDR) offre une représentation d'une richesse inédite des interactions entre les individus et de leur mobilité, indépendante des oublis ou des imprécisions de déclaration. Ces données permettent, en particulier, un suivi précis de la mobilité individuelle (Williams et al., 2013) et des interactions interpersonnelles (Calabrese et al., 2011). L'exploitation de ces deux dimensions - mobilité et interactions - à une granularité spatiale et temporelle fine permet de caractériser le profil de la ségrégation dans sa dimension sociale et spatiale.

L'anonymisation des données de téléphonie mobile ôte tout caractère personnel à celles-ci et ne permet pas, en particulier, de connaître le groupe social auquel appartient une personne. Cependant, la dimension spatiale des données de téléphonie mobile peut être utilisée pour associer celles-ci aux données administratives spatialisées. En particulier, le profil de communication d'un individu peut être utilisé pour identifier un lieu de vie probable dont la richesse peut être connue à partir de données fiscales administratives. L'absence de caractéristique individuelle des CDR peut ainsi être compensée par l'utilisation des propriétés géographiques de ceux-ci, permettant l'estimation de celles-ci.

1.2 Enjeux méthodologiques du choix de l'indice de ségrégation

Les indices de ségrégation permettent des comparaisons spatiales - entre plusieurs zones géographiques - et temporelles - pour une zone donnée, évolution dans le temps. Cependant, comme Massey and Denton (1988) l'ont souligné, la manière de définir la ségrégation affecte la mesure de l'objet, chaque indicateur synthétique renvoyant à une dimension particulière de la ségrégation. De la typologie initiale de Massey and Denton (1988) (égalité, exposition, concentration, regroupement et agrégation spatiale), Reardon and O'Sullivan (2004) retiennent deux dimensions principales permettant de comprendre la ségrégation: l'*exposition* (ampleur par laquelle des membres d'un groupe social rencontrent ceux d'un autre groupe) et la *concentration* (ampleur par laquelle un groupe est concentré dans l'espace). Chaque dimension est mesurée par des indices différents et le choix de ceux-ci n'est pas neutre sur la mesure du phénomène (Apparicio, 2000). Les indices de dissimilarité (Duncan and Duncan, 1955), de Gini et l'indice d'entropie de Theil (1972) fournissent, entre autres, une mesure de la *concentration* des groupes sociaux dans l'espace. L'indice de dissimilarité s'interprète, par exemple, comme la part des individus qu'il faudrait déplacer et réallouer dans l'espace pour atteindre une distribution spatiale uniforme. Les indices d'exposition ou d'isolement fournissent eux, comme leur nom l'indique, une mesure

[5]Données recueillies par les opérateurs téléphoniques pour tarification. Les CDR fournissent des informations sur les antennes-relais utilisées pour transmettre la télécommunication, l'heure de l'interaction, et l'identifiant des deux individus concernés par l'interaction.

de l'*exposition* ou de son contraire l'*isolement*.

Les indices classiques sont généralement conçus pour des variables discrètes et ont été généralisés pour permettre des analyses de variables multimodales. Les variables continues, comme le revenu, impliquent d'utiliser d'autres indices ou d'être décomposées en variables à modalités discrètes (en déciles par exemple). Cependant, la distribution du revenu n'est pas homogène spatialement : les niveaux médians et la dispersion des revenus, par exemple, ne sont pas identiques entre les villes. Ainsi, la comparaison d'indices construits à partir de variables de revenu discrétisés est problématique car le choix des intervalles pour la discrétisation n'est pas le même pour toutes les villes. Pour résoudre cela, Reardon and O'Sullivan (2004) ont proposé un *rank ordered index* qui consiste à utiliser la variable de revenu pour ordonner les individus au sein d'une ville par niveaux de richesse puis à construire des indicateurs synthétiques à partir de la variable de rang. Le principal intérêt de cet indice est qu'il n'est pas sensible à la dispersion des revenus au sein des villes, permettant ainsi de comparer deux villes dont les niveaux ou la dispersion du revenu diffèrent.

Le choix de l'échelle d'analyse soulève également des enjeux méthodologiques. Le principal problème posé par la définition de celle-ci est le problème des aires modifiables (*modifiable area unit problem (MAUP)*, Openshaw and Openshaw, 1984) car l'agrégation spatiale induite par la définition d'aires d'analyse affecte la distribution observée au sein de l'unité spatiale [6]. Le découpage de l'espace, généralement administratif, provoque également des effets de rupture autour des frontières, allouant des populations potentiellement similaires à deux aires spatiales différentes (Griffith, 1980). Par exemple, choisir une échelle trop fine provoque une variance intra-classe géographique trop faible et une variance inter-classe trop forte. Dans le cas français, les découpages administratifs disponibles sont les notions d'aire et d'unité urbaine. La notion d'aire urbaine renvoie à une notion d'attraction des emplois alors que l'unité urbaine est définie par un critère de continuité du bâti. Ces notions, associées à un découpage infra-communal fin - IRIS ou données carroyées - permettent de prendre en compte des situations diverses, tout en étant un espace cohérent avec l'espace accessible aux personnes présentes dans la base de données de téléphonie mobile. L'utilisation de données fiscales géolocalisées à travers la base Filosofi offre une précision suffisante pour construire des distributions spatiales cohérentes avec le découpage choisi. Cependant, l'agrégation spatiale nécessaire pour construire des indicateurs de niveau de vie d'un quartier reste problématique lorsque le nombre de ménages fiscaux dans chaque carreaux est très hétérogène. L'agrégation spatiale, qui implique de réduire chaque cellule d'analyse à un nombre restreint de caractéristiques, pose problème si la variabilité du revenu au sein de chaque cellule (*within variability*) domine la variabilité entre les cellules (*between variability*). Cet arbitrage entre variabilités intra- et inter-cellules affecte la taille

[6] En d'autres termes, une situation de MAUP se matérialise lorsque deux découpages du territoire, qui induisent une agrégation au sein de chacun des carreaux d'analyse, d'une même réalité sociale risquent de provoquer des résultats différents

optimale de l'unité spatiale: des cellules de petite (resp. grande) taille tendront à démultiplier (resp. réduire) le nombre d'unités spatiales et réduire (resp. accroître) la variabilité infra-cellule au détriment de la variabilité inter-cellule.

2 Méthodologie

2.1 Définition de l'échelle d'analyse

Afin d'explorer la manière dont les données de téléphonie mobile permettent de comprendre les phénomènes de ségrégation urbaine, deux types de données sont utilisés: données de téléphonie mobile anonymisées et données fiscales administratives. Ces deux bases de données présentent une dimension spatiale commune à travers la géolocalisation, d'une part, des antennes-relais utilisées pour relayer l'interaction téléphonique et, d'autre part, du foyer fiscal du ménage. Les données de téléphonie proviennent des Comptes Rendus d'Appel (*Call Detail Records*, CDR) du mois de Septembre 2007[7] de l'opérateur Orange. Ceux-ci permettent d'identifier l'ensemble des interactions (appels et messages textes) de 18 millions de clients géolocalisés à partir des deux antennes-relais utilisées pour transmettre l'interaction. Afin d'enrichir les données individuelles de téléphonie mobile des données fiscales, les données de l'Insee présentant des attributs spatiaux peuvent être mobilisées. La base fiscale Filosofi 2011[8] qui permet de mesurer les revenus géolocalisés des ménages apparaît être la source d'information la plus pertinente à cet égard. Pour obtenir des données au niveau individuel, la mesure du revenu retenue est le revenu disponible du ménage par Unité de Consommation (UC), mesure la plus satisfaisante pour identifier le niveau de vie des individus composant le ménage[9]. L'enrichissement de la base téléphonique n'est pas à entendre au sens classique du terme puisque l'appariement n'est pas effectué en utilisant une caractéristique individuelle commune aux deux bases de données

[7] Les opérateurs de téléphonie mobile ont l'obligation légale de conserver les CDR douze mois puis de les effacer. Une autorisation de la CNIL permet d'utiliser cette base de données individuelle où chaque utilisateur se voit associer un identifiant unique minimisant le risque de ré-identification. Une version postérieure de cet article exploitera d'autres mois de l'année 2007. Une limite de l'utilisation de ces données provient du fait que les CDR utilisés ne permettent de capturer que les interactions entre clients Orange. Les communications de clients issus d'opérateurs concurrents ne sont ainsi pas présentes dans ces données. Si le profil social des clients Orange diffère de celui des opérateurs concurrents, ceci peut introduire un biais de sélection. Les CDR ne présentant pas d'information individuelle, il nous est impossible d'évaluer la représentativité des individus présents dans les CDR par rapport à la structure démographique et sociale française.

[8] Les données fiscales géolocalisées de l'année 2007 étaient trop incomplètes pour être exploitées. Pour cette raison, nous avons utilisé la première version de la base Filosofi disponible, à savoir 2011. Comme la hiérarchie des quartiers au sein des villes est supposée stable entre 2007 et 2011, l'écart entre les deux dates devrait avoir des effets mineurs.

[9] Dans un ménage, le premier adulte compte pour une unité de consommation. Chaque membre de plus de 14 ans compte pour 0,5 UC supplémentaire. Les enfants de moins de 14 ans comptent pour 0,3 UC.

mais en exploitant la caractéristique géographique commune aux deux bases.

Le choix de l'unité spatiale est fondamental. La littérature utilisant des données de téléphonie mobile privilégie un découpage de l'espace issu d'une tesselation par polygone de Voronoï (Voronoi, 1908; Aurenhammer, 1991). Nous utiliserons, par simplicité, la dénomination *voronoï* pour désigner de tels polygones. Ces polygones associent à chaque point de l'espace l'antenne la plus proche et entraînent un espace sans aire d'intersection entre polygones. Les *voronoï* peuvent être interprétés comme des aires d'influence de chaque antenne, c'est-à-dire, l'ensemble des points desservis par l'antenne la plus proche [10]. Les antennes constituent l'information spatiale la plus fine disponible dans les données de téléphonie mobile. Ainsi, les *voronoï* forment une échelle géographique fondamentale de l'analyse dont il est difficile de s'abstraire. Pour cette raison, il est nécessaire de disposer d'un réseau suffisamment dense de *voronoï* pour limiter les erreurs d'allocation de points dans l'espace produits par une décomposition spatiale par plus proches voisins alors qu'un phénomène plus complexe d'allocation d'antennes a lieu. Pour cette raison, seules les villes de Paris, Lyon et Marseille, où la densité de *voronoï* est importante, ont été étudiées.

Cependant, les *voronoï* n'offrent pas un découpage de l'espace satisfaisant. Ils produisent des espaces dont la taille et la densité de population sont très hétérogènes comme cela est présenté plus en détail dans l'Annexe A. En ne retenant que les *voronoï* situés dans les limites de l'unité urbaine des trois villes retenues (construites à partir de la base des unités urbaines de l'Insee)[11], la Table 1 souligne la forte hétérogénéité des polygones de *Voronoï* avec une distribution de la population par *voronoï* très dispersée, dans les trois villes. Les Figures 12 et 13 montrent les espaces urbains obtenus pour les trois métropoles. Le centre ville des trois agglomérations, où population, emplois et réseaux de transport sont les plus denses, est bien plus densément fourni en antennes-relais que le reste de l'espace urbain. De plus, la métropole parisienne dispose d'un équipement très dense en antennes-relais, avec presque 3000 antennes, tandis que les deux autres villes en possèdent chacune un peu plus de 400 (Table 2). Parmi ces antennes, certaines sont probablement des antennes secondaires, utilisées en cas de congestion de l'antenne principale pour lesquelles la constitution d'une aire d'influence propre par *voronoï* peut être fallacieuse. Cette parti-

[10] Dans la réalité, un appel peut être dirigé vers une antenne plus éloignée que l'antenne la plus proche si cette dernière est, par exemple, saturée. Nous n'avons pas de possibilité de savoir, pour un appel donné, quelle était l'antenne la plus proche, mais seulement celle qui a servi à transmettre l'interaction. L'utilisation de *voronoï* implique donc de faire l'hypothèse selon laquelle l'antenne qui a transmis l'appel était l'antenne la plus proche.

[11] L'unité urbaine est un ensemble de communes qui comporte sur son territoire une zone bâtie où vivent 2 000 habitants et où aucune habitation n'est séparée de la plus proche habitation de plus de 200 mètres. En outre, pour chaque commune concernée, plus de la moitié de la population doit résider dans cette zone bâtie. Les frontières de notre espace métropolitain peuvent différer de celles de l'unité urbaine car l'intersection entre *voronoï* et communes du pôle urbain n'est pas parfaite: certaines antennes localisées dans l'unité urbaine peuvent desservir une zone extérieure à celle-ci alors que des antennes extérieures peuvent avoir une emprise sur des points à l'intérieur de l'unité urbaine. Ces effets à la bordure de la zone métropolitaine ne joueront qu'à proximité de la frontière de l'unité urbaine.

tion de l'espace par tesselation de *Voronoï* ne suit donc pas un principe d'échantillonnage de la population selon des caractéristiques spécifiques mais est contrainte par des aspects techniques. En effet, il y a plus de *voronoï* dans les zones où les usages du téléphone sont importants, ce qui concerne les lieux de vie mais aussi les zones d'emploi ou réseaux de transports.

L'hétérogénéité des *voronoï* apparaît particulièrement problématique lors du croisement avec les données fiscales. L'existence d'unités spatiales de taille diverse crée des problèmes lors de la phase d'agrégation spatiale, nécessaire à la combinaison des bases. L'utilisation d'une telle partition de l'espace risque de favoriser le problème MAUP. Un autre problème que la Table 1 permet d'entrevoir est que la mesure des événements au niveau du *voronoï* tend à produire une population, dans les données de téléphonie mobile, trop homogènement dispersée sur le territoire. En effet, la dispersion de la population après allocation du domicile au niveau du voronoï, est très faible, à une échelle non cohérente avec la dispersion réelle de la population dans les données fiscales. Autrement dit, une allocation du domicile au niveau du *voronoï* implique de donner trop de poids à des antennes situées dans des régions peu denses, ce qui est particulièrement problématique lors du croisement des données.

Pour s'abstraire partiellement de la partition par *voronoï*, il apparaît préférable d'avoir comme unité spatiale une grille dont la dimension est fixe[12]. Le fait de considérer des présences au niveau des carreaux conditionnelle à une observation au niveau du *voronoï* permet de probabiliser la localisation d'un événement. A la place de mesurer avec certitude un événement au niveau du *voronoï* j, un événement se voit maintenant affecter au carreau i avec probabilité p_i^j où $\forall j,\ \sum_i p_i^j = 1$. Les probabilités sont définies à partir des aires d'intersection entre les deux partitionnements de l'espace[13]. L'enjeu du changement de maille consiste à partir d'observations au niveau d'une collection de *voronoï* $(v_1, ..., v_J)$ de projeter des événements au niveau de carreaux $(c_1, ..., c_I)$.

En notant $\mathcal{S}(v_j)$ la surface du *voronoï* j et $\mathcal{S}\big(c_i \cap v_j\big)$ celle de l'intersection entre le carreau i et le *voronoï* j, la probabilité d'observer un événement dans le carreau i sachant que celui-ci est mesuré dans le *voronoï* j est donnée par

$$p_i^j := \mathbb{P}\big(c_i|v_j\big) = \frac{\mathcal{S}\big(c_i \cap v_j\big)}{\mathcal{S}(v_j)} \tag{1}$$

Ces probabilités conditionnelles permettent ainsi de partiellement s'abstraire des problèmes de confiance dans la localisation des événements téléphoniques dans les es-

[12]Une autre manière d'harmoniser les espaces d'analyse serait d'utiliser des espaces à population identique. Les IRIS, découpage infra-communal de 2000 habitants, pourraient être utilisées. Cependant, les formes distordues des IRIS risquent de provoquer des allocations de domicile arbitraires dans la base téléphonique où il est préférable d'avoir des formes convexes.

[13]Des méthodes fondées sur l'approche bayésienne inférant la localisation, conditionnelle à l'observation au niveau du carreau, existent également Tennekes (2018).

paces où la densité d'antenne est faible. Les événements mesurés pour un individu donné, qui constituent sa trace téléphonique $(v_{j_1}, ..., v_{j_J})$, deviennent des événements $\Big((c_{i_1}, ..., c_{i_I}), ..., (c_{i_1}, ..., c_{i_I})\Big)$ avec probabilités $\Big((p_1^{j_1}, ..., p_J^{j_1}), ..., (p_1^{j_J}, ..., p_J^{j_J})\Big)$[14].

La grille retenue est une couverture de l'ensemble du territoire de France métropolitaine par des carreaux de 500 mètres de côté. Cette grille se distingue du maillage des données carroyées de l'INSEE ou de la grille d'Eurostat (grille `INSPIRE`) pour deux raisons. D'abord, ces dernières sont construites à partir de carreaux de 200m de côté avec, éventuellement, des carreaux plus larges pour les zones où les enjeux de confidentialité sont problématiques. Notre grille est moins fine, ce qui se justifie à cause des enjeux de volume des données[15] mais induit une variabilité infra-cellule plus importante qu'avec des carreaux de taille réduite. La deuxième divergence entre la grille adoptée et la grille traditionnelle de l'INSEE provient du fait que cette dernière ne couvre pas l'ensemble du territoire métropolitain mais l'ensemble du territoire dévolu à l'usage résidentiel.

A l'échelle française, cette probabilisation de la localisation des événements induit un passage du nombre d'unités spatiales d'environ 18 000 *voronoï* (dont la surface médiane est 11km^2 et la surface moyenne 30km^2) à plus de 2 millions de grilles (dont la surface est fixée à 500m^2). Les limites des villes sont définies à partir des carreaux en intersection avec les communes constituant l'unité urbaine considérée. Les Figures 1 et 2 permettent de visualiser le quadrillage des trois unités urbaines considérées, à opposer au partionnement de l'espace par *voronoï* des Figures 12 et 13.

La Table 1 montre que la probabilisation des événements au niveau d'une grille apparaît plus satisfaisante que la localisation au niveau des polygones de Voronoï. D'abord, la distribution de la population dans les données fiscales apparaît plus homogène: la variance de la distribution de la population est réduite (Table 1) et la répartition des tailles d'unité spatiale apparaît plus équilibrée (Table 2). Ensuite, après détection du domicile depuis les données de téléphonie mobile, la population apparaît moins homogènement distribuée entre les territoires (Table 1), ce qui est cohérent avec la carte résidentielle où la population est concentrée dans un nombre restreint d'unités spatiales. Pour ces deux raisons, il apparaît préférable de s'abstraire d'une tesselation de Voronoï lors de l'analyse des propriétés spatiales des données de téléphonie mobile.

[14]Par exemple, si le *voronoï* j est intersecté par deux carreaux ($i = \{1, 2\}$), le premier couvrant 60% de sa surface et le deuxième 40%, les événements mesurés dans ce voronoi deviendront des 2-uplets (c_1, c_2) dont la probabilité associée est $(p_1^j = 0.6, p_2^j = 0.4)$.

[15]Pour le mois de Septembre, 3.9 milliards d'événements sont mesurés au niveau des antennes, un volume déjà conséquent. Le passage à une approche probabiliste avec des carreaux démultiplie la dimension de la base de données. Cet enjeu du volume des données impose ainsi de limiter la démultiplication induite par la probabilisation des événements, donc de limiter la finesse spatiale de la maille fondamentale.

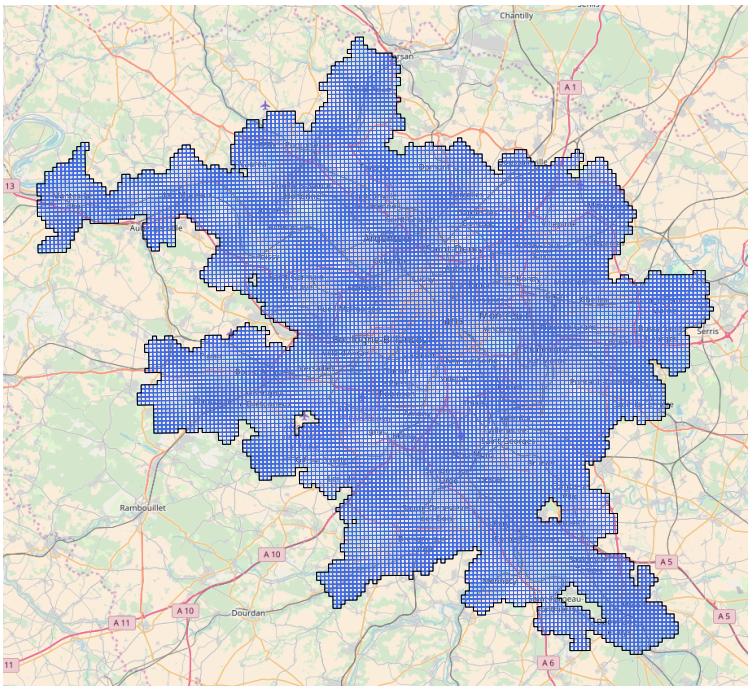
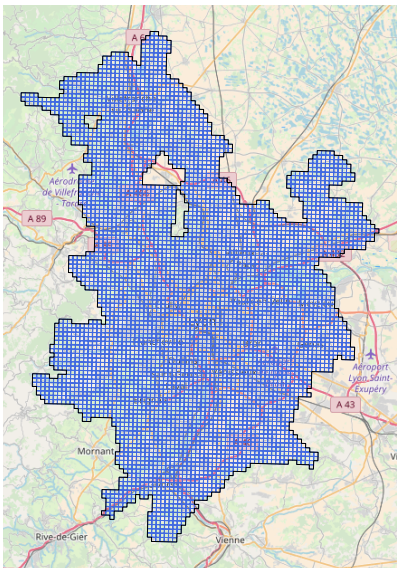


Figure 1 – Quadrillage de l'unité urbaine parisienne par des carreaux de 500 mètres de côté

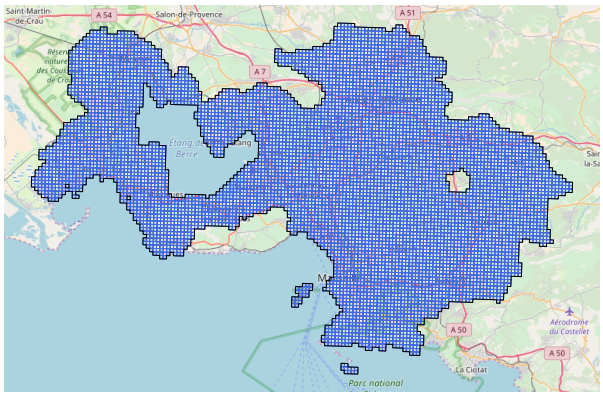
Table 1 – Ecart type de la population par unité spatiale selon la granularité spatiale adoptée

	Marseille		Lyon		Paris	
	Grille	Voronoï	Grille	Voronoï	Grille	Voronoi
Données Filosofi	3768	12 058	3423	10 861	6803	12 728
Données téléphonie après détection domicile	1189	684	1782	818	2450	1297
Ratio des écarts-types (%)	32	6	52	8	36	10

Exemple: A Paris, l'écart-type de la population par cellule (grille 500x500) est, dans les données fiscales, de 6803 personnes contre 12 728 avec une décomposition de l'espace par tesselation de *voronoï*.
L'écart type de la distribution de la population par unité spatiale dans les données de téléphonie mobile, après allocation du domicile, est de 2450 personnes avec une grille carroyée. Cela représente 36% de l'écart type dans les données fiscales avec la même granularité. Sur les données de téléphonie, l'écart type est de 1297 avec des cellules de voronoï, seulement 10% de la dispersion résidentielle observée dans les données fiscales.



(a) Lyon



(b) Marseille

Figure 2 – Quadrillage des unités urbaines de Lyon et Marseille par des carreaux de 500 mètres de côté

2.2 Appariement des bases fiscales et téléphoniques

L'historique d'appel et de messages texte d'un individu est utilisé pour déterminer un lieu de vie probable de l'utilisateur de téléphone. Afin de n'exclure aucun événement téléphonique, la recherche du lieu de vie n'est pas conduite à l'échelle des trois villes considérées mais au niveau de la France métropolitaine. Vanhoof et al. (2016) proposent, au niveau du *voronoï*, plusieurs heuristiques pour localiser les résidences des utilisateurs de téléphone. Nous avons adapté l'une d'elle - l'heuristique *distinct days* - à notre maille carroyée. Le lieu de vie d'un individu est identifié en sélectionnant la grille dans laquelle l'individu est localisé le plus fréquemment pendant la nuit, définie au sens large entre 19h et 9h du matin (si plusieurs grilles répondent à ce critère, un tirage aléatoire parmi ces grilles permet de sélectionner le lieu de résidence.). Autrement dit, un individu est supposé vivre dans le carreau où il a été localisé le plus de nuit distinctes.

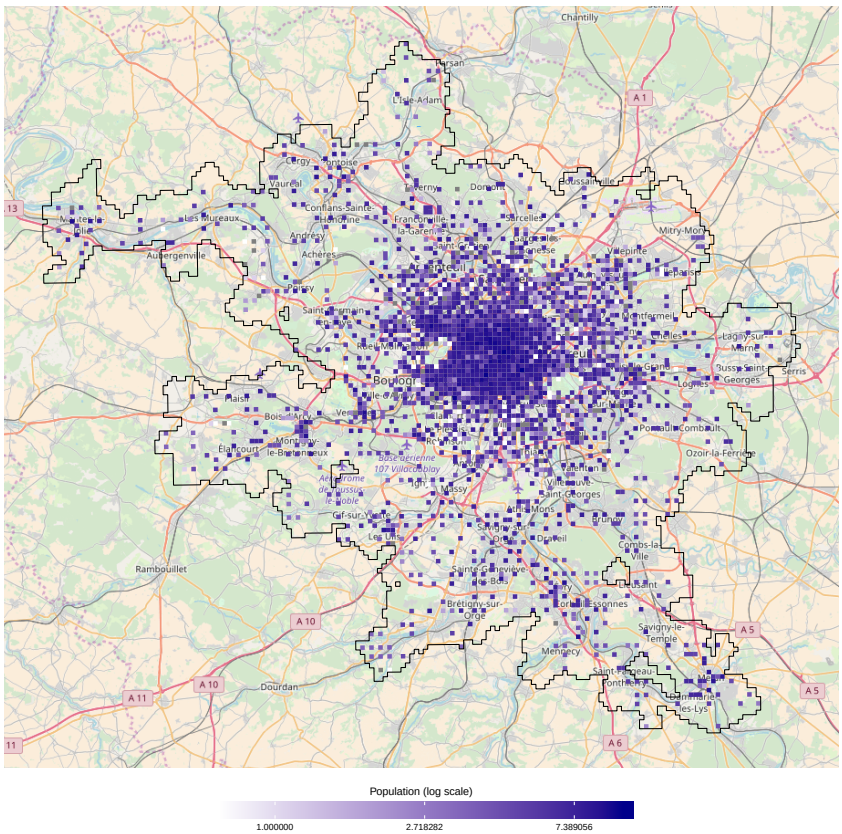
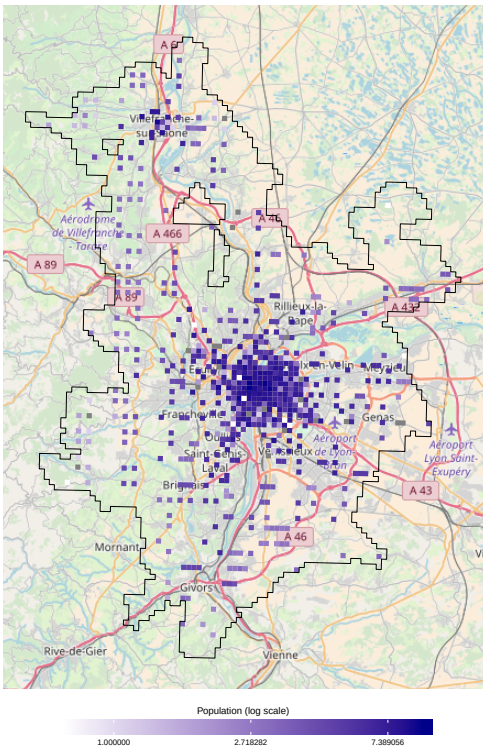


Figure 3 – Répartition population à Paris

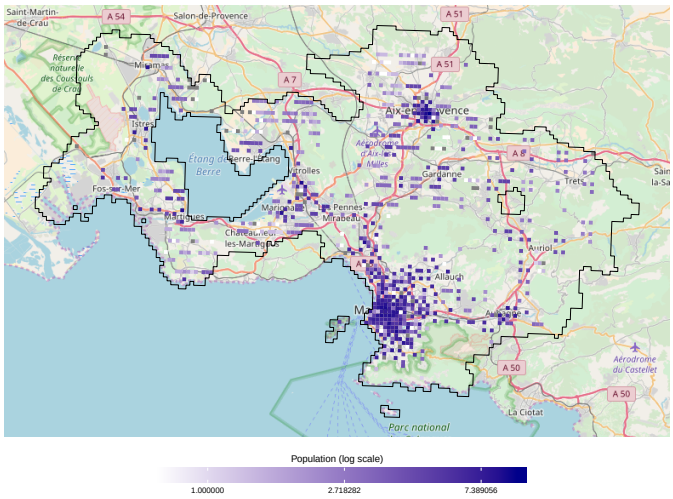
Le résultat de cette méthode d'attribution de domicile est présenté sur les Figures 3 et 4. La Figure 4 montre qu'il y a bien concentration des populations dans les espaces centraux les plus denses, caractéristique qui montre que cette méthode permet de globalement bien distribuer la population dans l'espace. Cette méthode d'allocation de la population permet également de bien représenter les espaces secondaires, comme Aix-en-Provence dans l'unité urbaine marseillaise. La Figure 3 montre également que les espaces vides au sein des espaces denses, comme par exemple les bois à la lisière de la petite couronne parisienne, sont également bien perçus grâce à cette méthode d'allocation de domicile. Cependant, cette allocation de domicile, malgré l'absence de filtre consistant à ne prendre en compte que les appels 9h et 19h, entraîne une concentration trop forte des populations autour des axes de communication[16]. Par exemple, dans le cas de Marseille (Figure 4), une concentration anormale de personnes autour de l'aéroport de Marignane se remarque. Cette tendance à la concentration des localisations de résidences autour des axes de communication est un effet induit par la concentration plus forte des antennes, et donc des événements téléphoniques mesurés, y compris le soir, autour de ces axes.

On estime ensuite le revenu d'un utilisateur de téléphone comme le revenu médian du

[16]La définition de la soirée est ici large car il est nécessaire, pour retrouver la trace probable du domicile, de disposer d'un nombre suffisant d'événements par jour. En 2007, le nombre d'appels ou messages textes était encore limité ce qui implique de garder une période de temps relativement large pour détecter le domicile.



(a) Lyon



(b) Marseille

Figure 4 – Quadrillage des unités urbaines de Lyon et Marseille par des carreaux de 500 mètres de côté

carreau alloué comme domicile. Les règles de confidentialité du secret statistique imposent de ne révéler aucune information pour des agrégats spatiaux construits sur des populations de moins de 11 ménages. Pour cette raison, les carreaux peu peuplés entrant dans cette catégorie sont d'abord agrégés avec leur voisin et, si, après cette étape, le seuil de 11 ménages est franchi, le revenu médian du carreau est construit à partir des ménages vivant dans ce carreau et ces voisins. Si, à l'inverse, le seuil de 11 ménages n'est pas dépassé, aucune information ne sera révélée sur le carreau ; le carreau est censuré. En pratique,

cela représente une perte mineure d'information car il s'agit de zones peu peuplées (Table 2).

2.3 Indices de ségrégation

Les données de téléphonie mobile constituent une mesure privilégiée de l'*exposition*[17] au sens de Reardon and O'Sullivan (2004) car elles mesurent la manière dont les individus issus de groupes sociaux différents interagissent. Les *rank-ordered index* créés par Reardon and Bischoff (2011) pour calculer des indices de ségrégation à partir d'une variable continue, comme le revenu, peuvent être adaptés à la structure des données de téléphonie mobile. En particulier, ces indices peuvent être adaptés pour mesurer les deux dimensions de la ségrégation que les données de téléphonie mobile permettent d'appréhender: contacts entre individus (ségrégation dans l'espace social) et déplacement dans l'espace (ségrégation dans l'espace physique). Ces deux dimensions requièrent deux indices différents[18].

En utilisant les revenus assignés à partir de Filosofi, on peut classer les individus d'une ville donnée du plus pauvre ($n = 1$) au plus riche ($n = N$)[19]. Ces rangs sont utilisés pour classer les individus et évitent la définition préalable de groupe sociaux plus ou moins arbitraires avant même le calcul d'un indice de ségrégation. Surtout, l'utilisation de rangs permet de contruire des indices de ségrégation ne dépendant pas de la dispersion des revenus entre les villes.

A partir de la variable de rang, une mesure de distance sociale entre les individus peut être définie. Celle-ci vise à objectiver une inégalité perçue entre deux individus à partir du critère de revenu. Introduire une distance dissymétrique, i.e. une mesure telle que la distance $d_{x \to y}$ diffère de $d_{y \to x}$, permet de rendre compte du fait que selon le groupe social auquel on appartient, on ne se représente pas de la même manière la distance aux autres groupes. Par exemple, un phénomène de peur du déclassement (Maurin, 2009) peut induire une perception des distances sociales telle que, pour un individu donné, les personnes ayant un revenu plus faible que lui apparaîtront plus distances que celles avec un revenu supérieur.

La fonction de distance sociale suivante entre x et y permet d'obtenir cette propriété de

[17] Rappelons que l'*exposition* mesure la probabilité pour un individu d'un groupe donné de rencontrer un individu issu d'un autre groupe. La notion de *concentration* renvoie, quant à elle, à la manière dont les individus se répartissent dans l'espace. Une future version de cet article inclura un indice d'entropie permettant de mesurer la concentration

[18] Ces indices ont été proposés par Yang Xu (*Hong Kong Polytechnic University*) dans le cadre d'un projet collaboratif que nous menons actuellement.

[19] Le principe adopté est celui du classement fractionnel. Les rangs égaux se voient associés le rang moyen qu'ils auraient eu sous un ordonnancement ordinal. Par exemple, si on dispose du vecteur de revenu suivant $(25, 35, 35, 40)$, les rangs associés seront $(1, 2.5, 2.5, 4)$

dissymétrie :

$$d_{x\to y} = \begin{cases} 0 & \text{si } j = x \\ \frac{|4x-4j|-1}{2(N-1)} & \text{si } j \leq 2x-1 \text{ et } j \neq x \\ \frac{j-1}{N-1} & \text{si } j > 2x-N \end{cases} \quad \text{si } x \leq \frac{N}{2} \tag{2}$$

$$d_{x\to y} = \begin{cases} 0 \text{ si } j = x \\ \frac{|4x-4j|-1}{2(N-1)} \text{ si } j \geq 2x-N \text{ et } j \neq x \\ \frac{N-j}{N-1} \text{ si } j < 2x-N \end{cases} \quad \text{si } x > \frac{N}{2} \tag{3}$$

avec N la population dans la ville considérée. Cette distance a des valeurs comprises entre 0 (deux individus qui ont le même niveau de richesse) et 1 (deux individus qui constituent les deux extrêmes de la distribution du revenu). Elle est asymétrique car un individu x peut avoir une perception de sa distance avec y différente de celle de y pour x ce qui se traduit par la possibilité que $d_{x\to y}$ et $d_{y\to x}$ diffèrent. Enfin, notons $s_{x\to y}$ l'indice de similarité sociale défini comme

$$s_{x\to y} = 1 - d_{x\to y} \tag{4}$$

A partir de cette mesure de distance sociale, deux indices mesurant l'exposition d'un individu à des personnes issues d'autres groupes sociaux peuvent être calculés.

2.3.1 Indice de ségrégation sociale

La première dimension de ségrégation que les données de téléphonie mobile permettent de mesurer correspond à une exposition dans l'espace social. Il s'agit de mesurer si les appels d'un individu sont principalement dirigés vers son groupe social d'origine où se dirigent uniformément vers l'ensemble des groupes sociaux.

Pour un individu x donné, sachant la liste de ses contacts $(y_1, ..., y_m)$ et le nombre d'interactions avec ces contacts $(f_1, ..., f_m)$, il est possible de construire l'indice de ségrégation sociale (*Social Segregation Index*, SSI) suivant, prenant valeurs dans $[0, 1]$,

$$SSI_x = \frac{\sum_{j=1}^{m} f_j s_{x\to y_j}}{\sum_{j=1}^{m} f_j} \tag{5}$$

où $s_{x\to y}$ est donné par eq. (4). Cet indice permet de mesurer, pour un individu x donné, une moyenne pondérée (par le nombre de contacts) de la similarité entre lui et ses contacts. Une valeur proche de 0 signifie que l'individu interagit principalement avec des individus éloignés de lui dans la hiérarchie des revenus. A l'inverse, un SSI proche de 1 signifie que la personne interagit principalement avec des membres de son groupe social. La valeur $1/2$ peut servir de référence égalitaire, situation où l'individu interagit avec des individus

issus de groupes sociaux différents de manière uniforme[20].

2.3.2 Indice de ségrégation physique

Les données mobiles permettent un suivi précis des déplacements d'un individu. L'exposition est cette fois spatiale car il est possible d'étudier la manière dont des individus issus de groupes sociaux différents peuvent être présents dans un même espace à un même instant. Les données de téléphonie mobile permettent de mesurer, pour un individu x donné, une suite de coordonnées géographiques et d'horodatage de la forme $\{(l_1, t_1), (l_2, t_2), ..., (l_n, t_n)\}$ avec l_i la localisation de l'utilisateur au moment t_i.

Sachant le nombre de fois où l'individu est mesuré à la localisation l dans un intervalle de temps T, notée $m_x(l, T)$, il est possible de calculer la probabilité que celui-ci se trouve à l'endroit l pendant la période T comme étant

$$\mathbb{P}_x(l, T) = \frac{n_x(l, T)}{n_x(T)} \tag{6}$$

où $n_x(T)$ est le nombre total d'appels passés par x sur la période de temps T. Le passage de l'échelle d'observation qu'est le *voronoï* à l'échelle d'analyse du carreau implique une redéfinition des probabilités. En notant $(v_l)_l$ les événements observés au niveau du *voronoï*, on a d'après eq. (6), en omettant de manière provisoire la dépendance à la période d'observation T afin d'alléger les notations,

$$\forall v_l \in \mathcal{V}, \quad \mathbb{P}_x(v_l) = \frac{n_x(v_l, T)}{n_x(T)}$$

Après passage par les probabilités conditionnelles (10), la présence au niveau du carreau est donnée par

$$\forall c_j \in \mathcal{C}, \quad \mathbb{P}_x(c_j) = \sum_{v_j \in \mathcal{V}} \mathbb{P}\big(c_j \big| v_j\big) \mathbb{P}_x(v_j) \tag{7}$$

Ainsi, la présence est déterminée au niveau du carreau pour être cohérent avec l'échelle spatiale choisie dans l'allocation de domicile. Elle dépend néanmoins implicitement de l'heure d'observation puisque deux individus ne se croisant pas à la même heure dans une localisation donnée ne seront pas considérés comme se croisant.

L'indice de ségrégation physique, au niveau individuel x et à un lieu donné c_j, consiste à mesurer la similarité de x avec les individus présents dans le même lieu au même instant (en la pondérant par la probabilité réciproque de se croiser). Autrement dit, en

[20] Sous l'hypothèse qu'un individu a N interactions réparties uniformément dans l'espace social (autrement dit le processus générateur des fréquences d'appels $(f_1, ..., f_N)$ suit une loi uniforme dans $[\![1, N]\!]$), alors $\forall x$, $\mathbb{E}(SSI_x) = \frac{1}{2}$ ce qui montre que la valeur $1/2$ peut servir de référence comme profil de communication uniforme indépendamment de la place de x dans la distribution du revenu

notant $U(c_j, T)$ l'ensemble des individus dont la probabilité de croiser x est non nulle (i.e. l'ensemble des y tels que $\mathbb{P}_y(c_j, T) > 0$)

$$\begin{aligned} PSI_x(c_j, T) &= \frac{\sum_{y \in U(c_j,T)} \mathbb{P}_x(c_j, T)\mathbb{P}_y(c_j, T)s_{x \to y}}{\sum_{y \in U(l,T)} \mathbb{P}_x(c_j, T)\mathbb{P}_y(c_j, T)} \\ &= \frac{\sum_{y \in U(c_j,T)} \mathbb{P}_y(c_j, T)s_{x \to y}}{\sum_{y \in U(c_j,T)} \mathbb{P}_y(c_j, T)} \end{aligned} \tag{8}$$

Les localisations $(c_1, ..., c_n)$ sont, dans notre cas, les événements mesurés dans les carreaux constituant le maillage urbain[21]. Pour déduire de $PSI_x(c_j, T)$ une mesure individuelle, sur une période de temps T, il suffit de faire la moyenne des $(PSI_x(c_j, T))_{c_j}$ pour un individu x sur l'ensemble des localisations $(c_1, ..., c_n) \in \mathcal{C}$ explorées pendant l'intervalle de temps T (moyenne pondérée par la probabilité de présence à l'endroit c_j pour tenir compte de la surreprésentation de certains lieux dans le parcours d'un individu). Autrement dit,

$$PSI_x(T) = \sum_{c_j \in \mathcal{C}} \mathbb{P}_x(c_j, T) \tag{9}$$

Cet indice de ségrégation physique est dans $[0, 1]$. Plus il s'approche de 1, plus l'individu tend à fréquenter des lieux fréquentés par des individus de son groupe social. L'intervalle de temps retenu dans cette étude est le suivant: on découpe la semaine en 48 fenêtres, 24 périodes de temps pour la semaine et 24 autres pour le weekend.

3 Résultats

3.1 Profils de ségrégation

Une version ultérieure de cet article présentera des résultats sur la ségrégation spatiale (indice présenté par l'eq. 9). La Figure 5 montre que les données mobiles permettent de mesurer une ségrégation sociale. Le mode des distributions de l'indice de ségrégation est proche de 0.6 pour les trois villes, légèrement inférieur à Marseille par rapport aux deux autres villes.

La Figure 6 permet de mesurer le profil de la ségrégation en fonction de l'appartenance à une classe de revenu (décile). Les trois villes exhibent un même profil en U suggérant que les groupes extrêmes sont les plus ségrégués. D'un point de vue global, ceci est cohérent avec les résultats de Floch (2017). Comme cela est mesuré par l'approche résidentielle,

[21] L'indice de ségrégation sociale (eq. 5) n'est dépendant de la décomposition de l'espace que parce qu'il découle de l'allocation de domicile et de revenu, tout deux dépendant de l'échelle d'analyse. L'indice de ségrégation physique introduit une nouvelle dépendance au maillage territorial puisque les événements étant probabilisés au niveau du carreau, les localisations ne sont plus des antennes mais des carreaux.

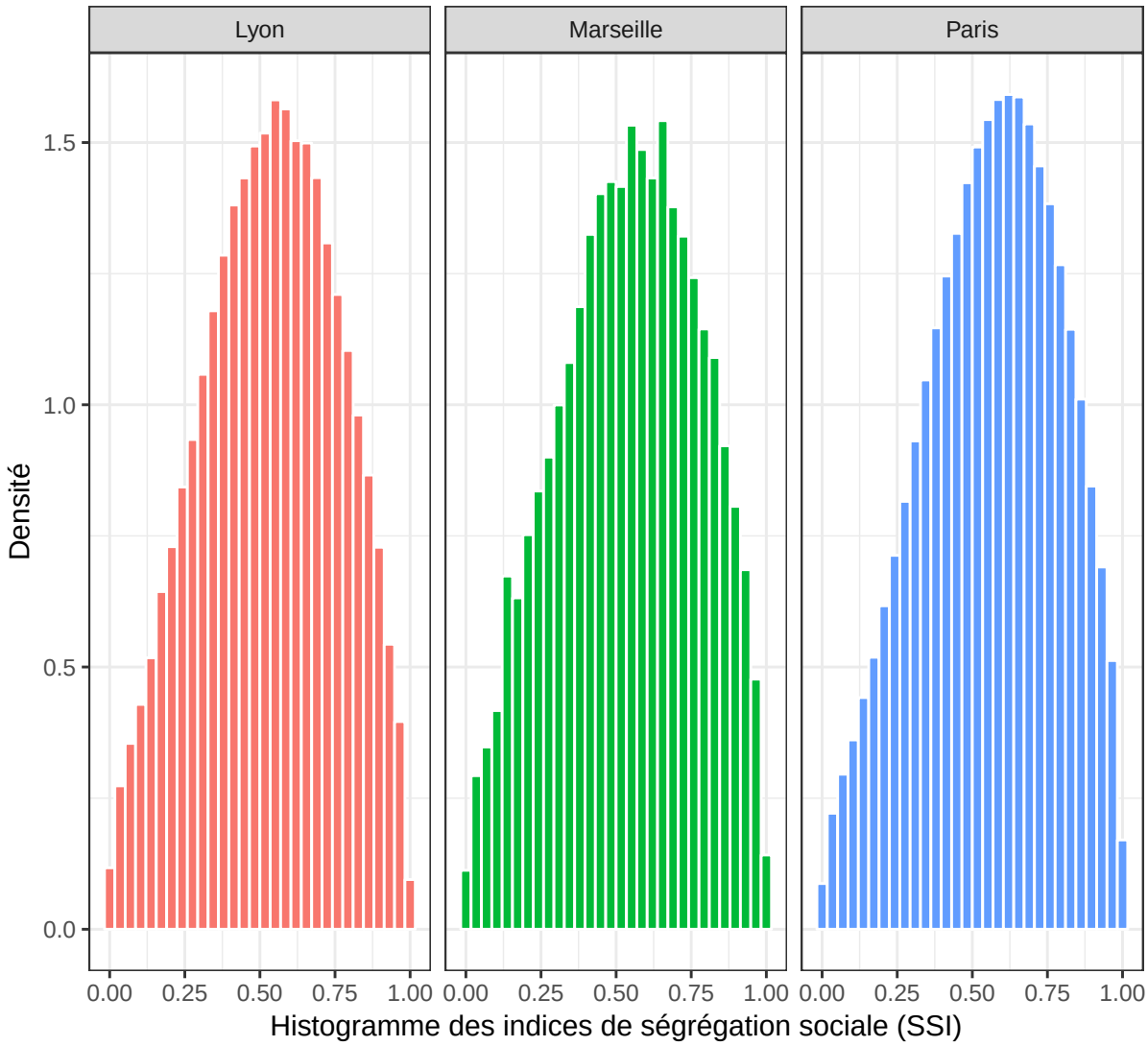


Figure 5 – Histogramme des indices de ségrégation individuels (eq. 5)

Paris est plus ségrégué que Lyon et Marseille dont les profils sont proches. Les données de téléphonie mobile apportent ainsi un éclairage, à l'échelle individuelle, du constat de Floch (2017) sur la ségrégation forte de l'espace parisien. Ces résultats suggèrent que la forte ségrégation de l'espace physique, mesurée par l'approche résidentielle, est également associée à une séparation des groupes sociaux dans l'espace social. La concentration des groupes sociaux dans l'espace affecte ainsi leurs liens, ce qui se retrouve dans les interactions téléphoniques[22]. A Lyon et Marseille, le fait que la ségrégation des groupes sociaux les plus riches soit relativement proche de celle des groupes les plus pauvres semble être en contradiction avec les résultats usuels. Les individus les plus riches tendent en effet à être les plus séparés dans l'espace (Oberti and Préteceille, 2016). Cette séparation spatiale pourrait intuitivement se retrouver dans l'espace social.

[22]Une version ultérieure offrira une vision plus fine du lien entre niveau individuel de ségrégation et revenu dans les déciles supérieurs.

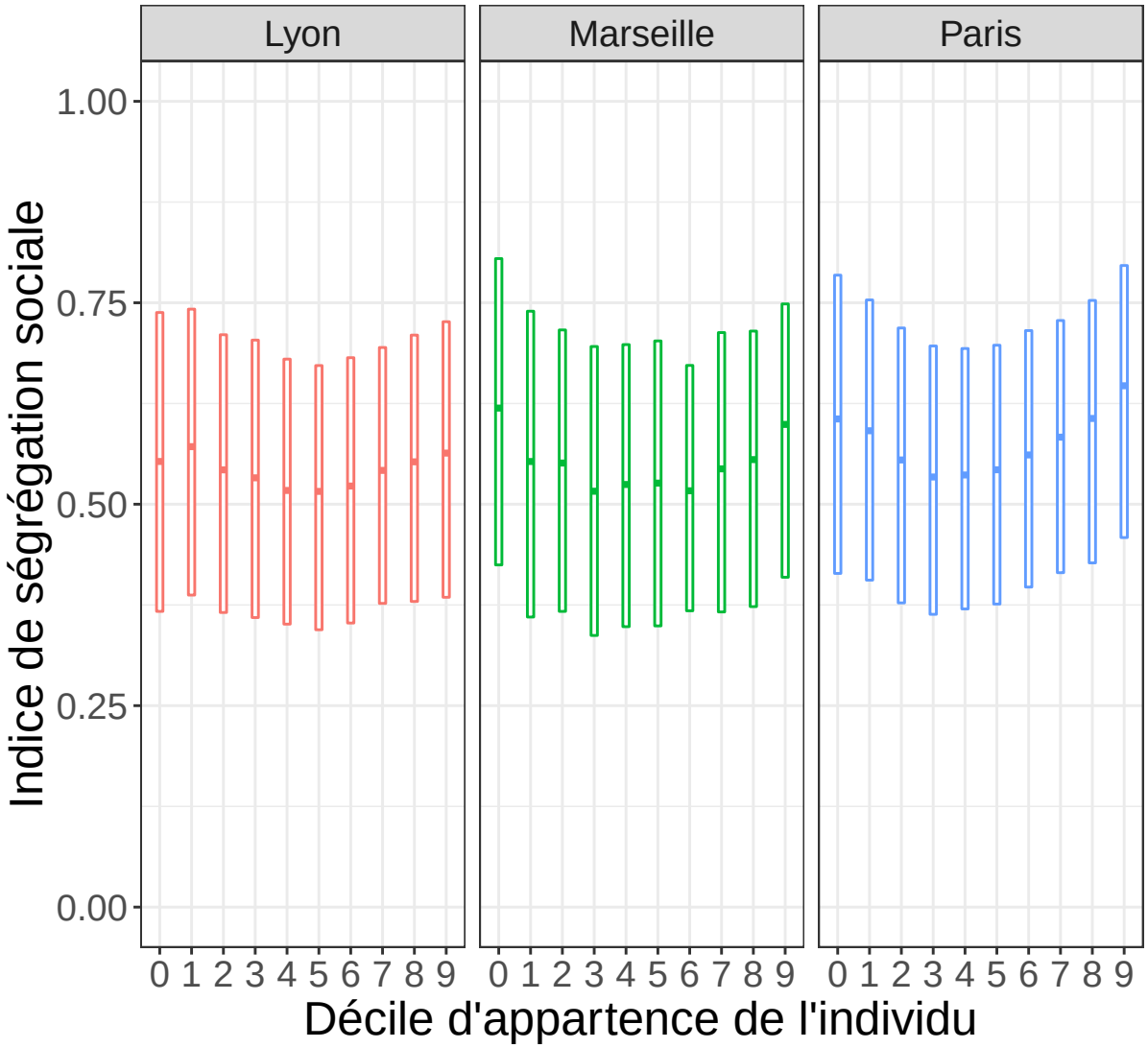


Figure 6 – Indice de ségrégation sociale (eq. 5) selon le groupe social de l'individu

Les Figures 7-9 représentent la valeur médiane de l'indice de ségrégation (eq. 5) dans chaque carreau où les utilisateurs de téléphonie mobile sont identifiés comme résidant. Les profils de ségrégation sont particulièrement marqués aux extrêmes de la distribution spatiale des revenus. A Paris (Figure 7), les quartiers où la ségrégation est la plus marquée sont situés au nord-est (quartiers populaires avec une surreprésentation des Quartiers Politique de la Ville) et dans l'ouest (quartiers plus riches). Au sein de Paris intra-muros, la ségrégation est particulièrement marquée dans le XVIe arrondissement. Dans l'agglomération lyonnaise, la concentration de zones de forte ségrégation sociale est moins marquée mais quelques îlots se remarquent dans des quartiers populaires du sud-est lyonnais ou à Villefranche-sur-Saône (Figure 8). La morphologie particulière de Marseille avec des quartiers populaires au sein du centre se retrouve avec une ségrégation plus marquée dans le centre de la ville. Dans cette ville, les plus pauvres sont d'ailleurs plus ségrégués que les plus riches (Figure 6). L'autre pôle de ségrégation (Aix-en-Provence) correspond lui à l'autre extrême de la distribution de revenu.

3.2 Limites et travail d'approfondissement

Mesurer la ségrégation sociale à partir de données de téléphonie mobile implique de réduire les relations sociales au champ plus étroit des interactions téléphoniques. On peut se demander dans quelle mesure les contacts téléphoniques sont représentatifs des interactions sociales en général. Afin d'assurer que les interactions mesurées correspondent bien à une relation sociale, les appels non réciproques, à l'échelle du mois, n'ont pas été considérés. Ce filtre permet d'éliminer une partie du bruit présent dans les données de téléphonie mobile pour mieux identifier les relations sociales. Malgré cela, identifier relations sociales et interactions téléphoniques comporte le risque d'exclure des individus de l'analyse qui, par exemple, n'utilisent pas les téléphones portables pour communiquer. La mesure de la ségrégation peut en être affectée si ces personnes tendent à appartenir principalement à un groupe social. L'usage généralisé du téléphone mobile permet d'espérer que l'assimilation entre ces deux ensembles ne biaise pas la représentation des interactions entre les groupes sociaux. Silm and Ahas (2014) ont ainsi montré qu'avec des données de téléphonie mobile, les résultats connus sur l'évolution dans le temps de la ségrégation spatiale, à savoir le fait que la ségrégation soit plus importante hors des heures de travail, peuvent être retrouvés.

A l'heure actuelle, les utilisateurs de téléphone mobile se voient assignés le revenu médian de la grille dans laquelle leur résidence est le plus probablement située. Les résultats ici représentés ne réflètent ainsi qu'une ségrégation reliée à la variance de revenu inter-unités spatiales. Pour tenir compte de la variance infra-grille, un travail de simulation plus complet de la distribution du revenu au sein des grilles est prévu. Cela devrait permettre de mieux représenter les inégalités de revenu dans leur dimension spatiale, à la fois entre les unités spatiales et au sein de celles-ci. Un autre avantage d'une approche par

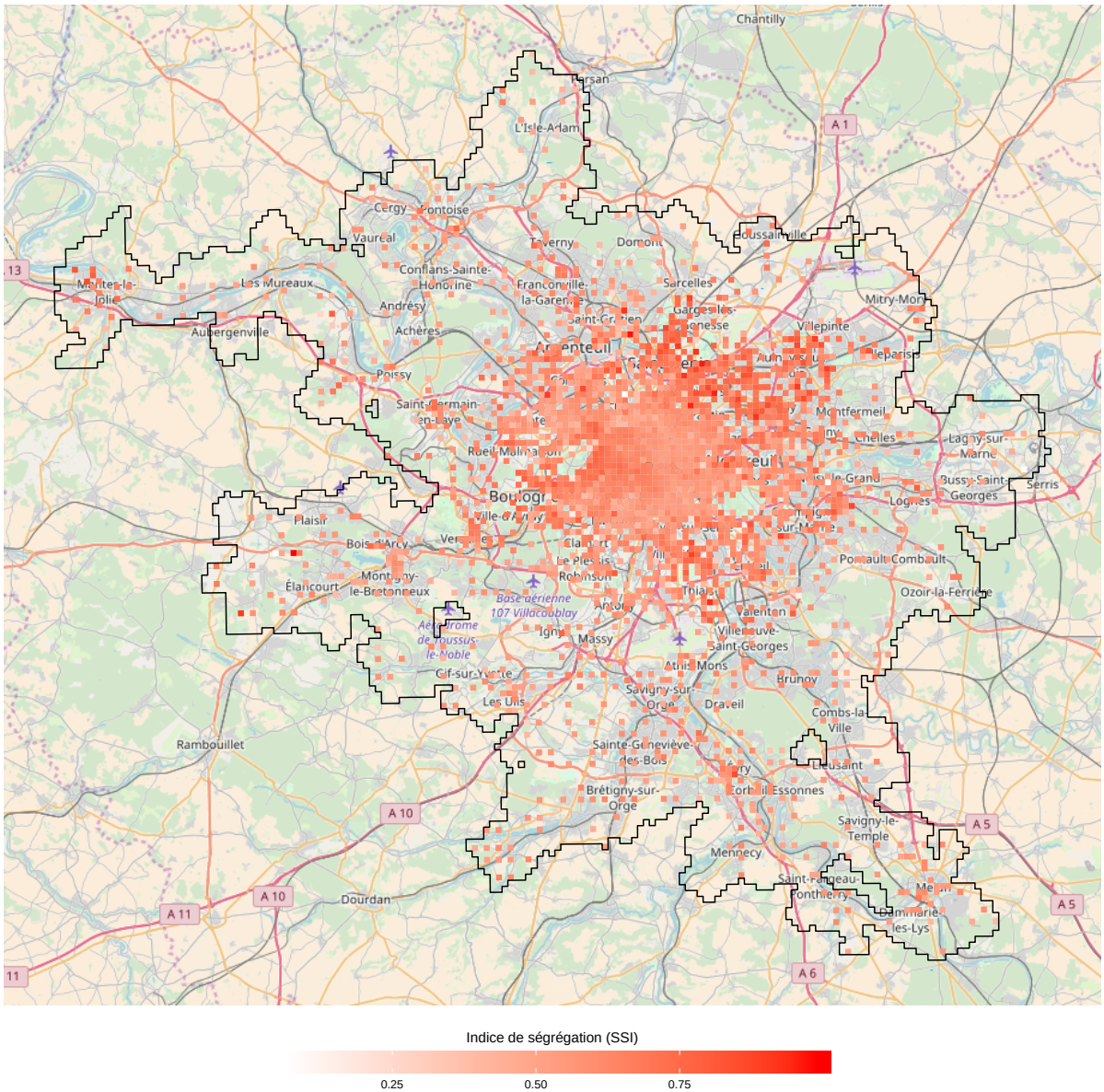


Figure 7 – Ségrégation sociale dans l'unité urbaine parisienne

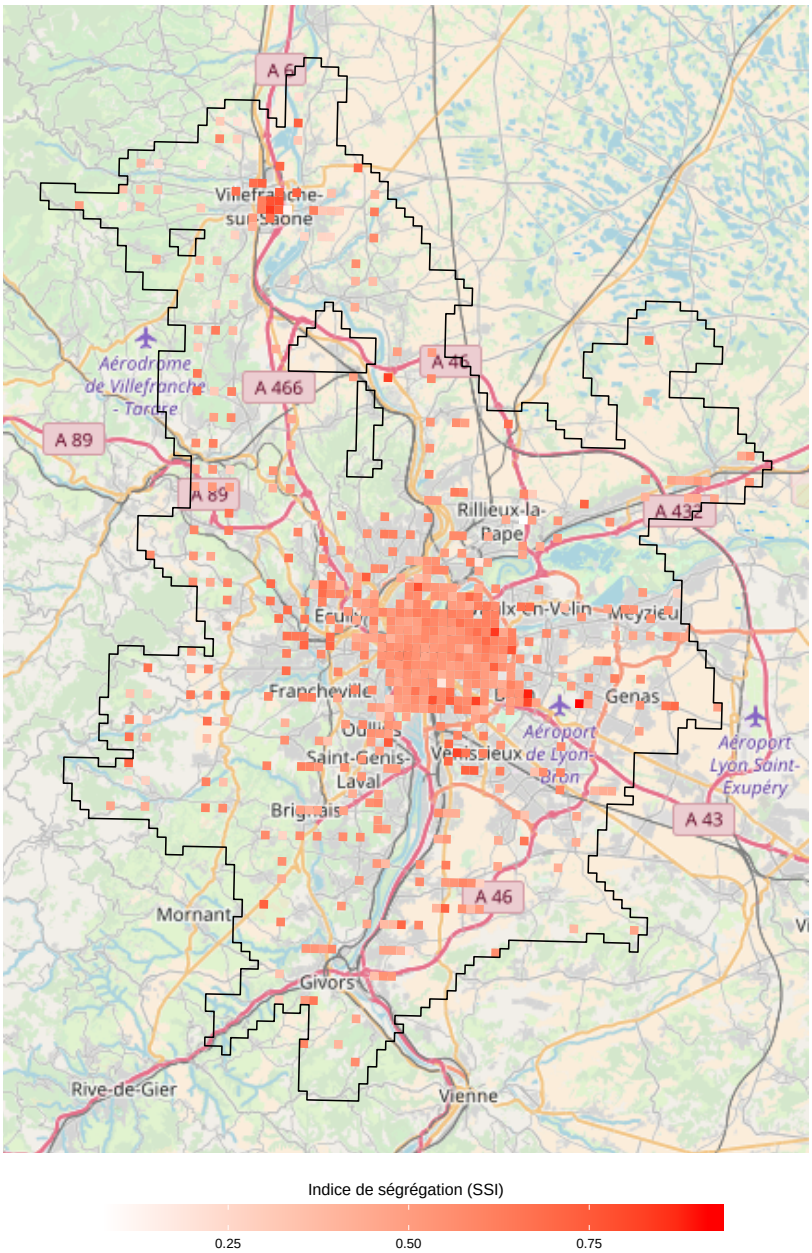


Figure 8 – Ségrégation sociale dans l'unité urbaine lyonnaise

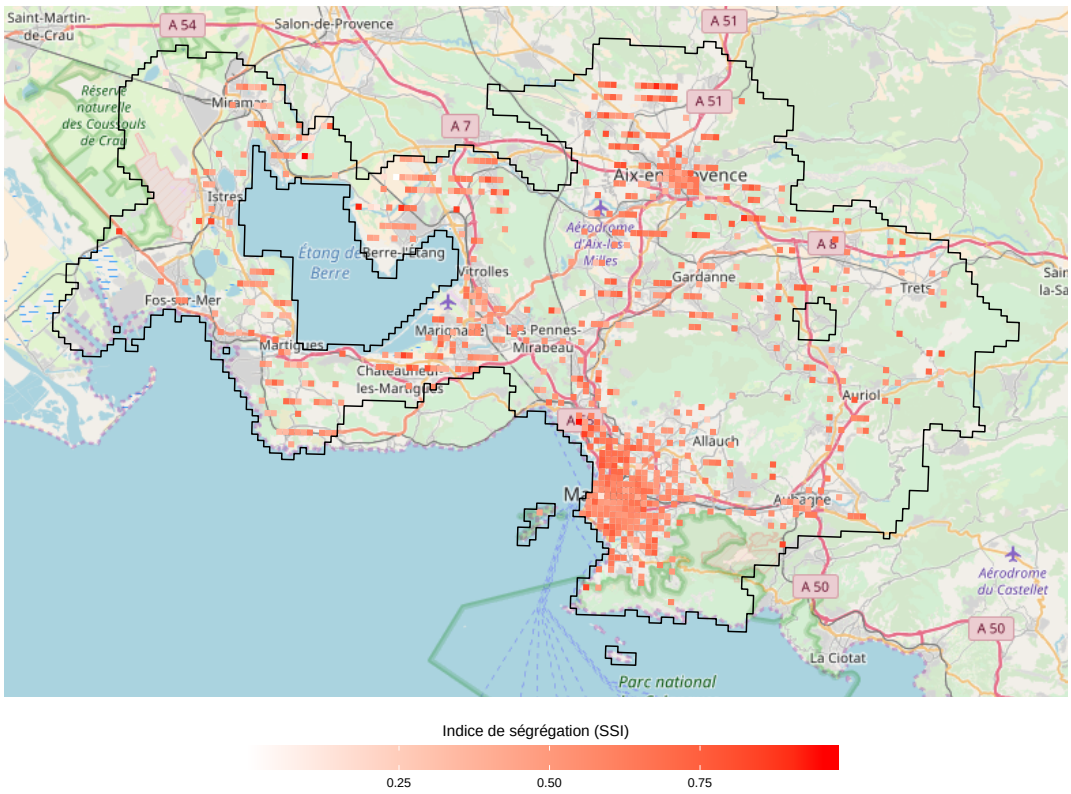


Figure 9 – Ségrégation sociale dans l'unité urbaine marseillaise

simulation est que celle-ci peut être reproduite pour mesurer la robustesse des résultats et déterminer des intervalles de confiance empirique des indicateurs de ségrégation.

Avant d'augmenter le nombre de simulations numériques pour obtenir l'imputation du revenu, il est nécessaire de s'assurer que chaque itération reproduit bien la manière dont les groupes sociaux s'organisent dans l'espace. La principale difficulté pour évaluer la robustesse de la méthodologie provient de la non connaissance du lieu de résidence effectif de l'individu. Un individu alloué de manière incorrecte à une unité spatiale se voit imputer aléatoirement un revenu d'une distribution qui ne le concerne pas. La cohérence de l'imputation de revenu dépend alors de deux facteurs: la précision dans la méthode d'allocation de domicile utilisée pour déterminer le lieu de vie et l'homogénéité sociale des unités spatiales voisines. Dans les quartiers relativement homogènes socialement, les unités voisines devraient avoir des distributions similaires. Dans ce cas, si l'erreur d'allocation est limitée (individu incorrectement assigné à une grille voisine), l'individu se verra aléatoirement imputer un revenu issu d'une distribution similaire à celle de son lieu de vie. Dans ce cas, l'erreur d'allocation de domicile n'est pas pénalisante. Cependant, si les unités voisines se situent dans des quartiers où les distribution de revenu sont très différentes, les quantiles des grilles voisines vont être distincts. Dans ce cas, l'erreur d'imputation du lieu de vie probable affecte l'imputation du revenu et peut entraîner un indice de ségrégation fallacieux. Si la variance inter-grilles domine la variance infra-grille, ce problème peut affecter la robustesse de l'analyse. Des indicateurs de stabil-

ité de l'allocation de domicile sont encore à construire.

Pour s'assurer de la robustesse de l'appariement construit, des tests formels sont nécessaires. Par exemple, un test d'adéquation entre la distribution empirique simulée et la distribution empirique observée dans Filosofi peuvent constituer une première étape. Ces tests d'adéquation peuvent être conduits grille par grille. L'inconvénient d'une telle approche est qu'elle ne rend pas compte des phénomènes d'autocorrélation spatiale de chacune des séries étudiées (Filosofi et données simulées), i.e. de la dépendance des observations aux valeurs observées dans les unités voisines. Produire un test prenant en compte les phénomènes de corrélation spatiale (Lee, 2001) apparaît nécessaire pour s'assurer que l'imputation sur les données mobiles est cohérent avec les données observées. Pour tester la robustesse de l'enrichissement des données, la comparaison des résultats sur la distribution spatiale du revenu avec des indices classiques de ségrégation (par exemple indice d'entropie de Theil, 1972) apparaît également importante. Une fois la pertinence de la méthodologie d'imputation du revenu contrôlée, la répétition de cette méthode un certain nombre de fois (itérations successives) devrait permettre d'obtenir des intervalles de confiance pour les indices de ségrégation et d'analyser la stabilité des résultats.

Les premiers résultats sur la ségrégation offrent des pistes intéressantes sur la mesure de la ségrégation sociale. Cette dimension, absente des approches fondées sur la ségrégation résidentielle, est pourtant fondamentale. Elle offre cependant une vision complémentaire aux résultats de Madoré (2015) ou Floch (2017). Les données de téléphonie mobile devraient également permettre de mesurer l'évolution dans le temps de la ségrégation à à travers le calcul de l'indice de ségrégation physique (eq. 9). Les résultats de cette approche devraient offrir un complément intéressant aux résultats mesurés à partir d'une perspective résidentielle. Cela devrait permettre de comprendre la manière dont les groupes sociaux se rencontrent à différents moments de la journée. Cela permettrait de mettre en regard des plages temporelles où la ségrégation est particulièrement marquée (le soir) et d'autres avec plus de mixité sociale.

4 Conclusion

Les données de téléphonie mobile apparaissent être une source intéressante pour mesurer l'ampleur du phénomène de ségrégation dans les grandes métropoles françaises. Sur cette question, déjà très documentée, la vision fine des interactions sociales et de la mobilité individuelle qu'offrent les données de téléphonie permet d'étudier des dimensions complémentaires à la question de la ségrégation résidentielle. L'utilisation de la base fiscale Filosofi pour enrichir les données de téléphonie mobile permet de mesurer l'ampleur des phénomènes de ségrégation. Sans aucune connaissance *a priori* des caractéristiques individuelles des utilisateurs de téléphone, il est possible de déduire leur lieu de vie probable et

leur imputer un revenu à partir de la distribution dans le voisinage de celui-ci. Les indices de ségrégation, construits pour tirer profit des interactions sociales et de la mobilité individuelle mesurés dans les comptes rendus d'appels, offrent des résultats complémentaires à l'approche résidentielle. Par une approche orientée autour des interations téléphoniques, il est possible de retrouver des résultats proches de ceux mesurés par une approche résidentielle (Madoré, 2015; Floch, 2017). Il reste encore à construire une analyse fine de la dynamique temporelle de la ségrégation. Le croisement des données téléphoniques et fiscales est une méthodologie innovante offrant des perspectives d'analyse fine de la ségrégation. Cette méthodologie implique néanmoins la construction d'indicateurs et de tests statistiques adaptés. Cet enjeu dépasse le sujet de la ségrégation (Ollion and Boelaert, 2015) et conditionne la qualité des études exploitant la richesse des données mobiles.

A Annexe: Passage du *voronoï* à la grille

A.1 Enjeux

Cette annexe offre plus de détails sur le changement d'unité spatiale. La Figure 10 montre, à l'échelle de la France métropolitaine, l'hétérogénéité de la distribution des antennes et ainsi celle du partionnement par *voronoï*. Dans les zones rurales, la densité des antennes est peu marquée, induisant des *voronoï* de grande superficie. Ceci est particulièrement remarquable pouvant être considérés comme des zones blanches, par exemple les Alpes du sud, la Corse ou encore la région champenoise.

La forme des polygones de Voronoï dépend de la distribution locale des antennes. Une tesselation fondée sur la distribution locale de la population, et non des antennes, devrait produire des unités spatiales très différentes. En particulier, la tesselation de Voronoï couvre l'ensemble du territoire alors que l'espace résidentiel se concentre dans certaines zones: 9% des sols français sont artificialisés (Fontes-Rousseau and Jean, 2015), une part moindre est dédiée à l'espace résidentiel. Les *voronoï* étant généralement de superficie large dans les espaces ruraux, une partie faible des *voronoï* ne rencontre pas de résidence dans la base fiscale Filosofi (Table 2). Ce nombre réduit d'unités spatiales à population faible n'est pas désirable puisque la population, en réalité, est concentrée dans un espace réduit, ce qui devrait se réfléter par un nombre plus important d'unités spatiales peu peuplées.

Le principal problème de la tesselation de Voronoï est qu'elle ne repose pas sur un échantillonnage équilibré de la population. Ceci pose ainsi des problèmes d'agrégation spatiale (MAUP, Openshaw and Openshaw, 1984) car cela amène à construire des statistiques descriptives sur des tailles de population très hétérogènes. Comme le montre la Table **??**, la variance de la population au sens de la résidence fiscale par unité spatiale est extrêmement forte avec une tesselation de Voronoï. Autrement dit, avec des *voronoï*, l'agrégation spatiale est construite sur un nombre d'observation très hétérogène. L'utilisation d'une telle partition de l'espace risque ainsi de favoriser le problème MAUP, comme anticipé.

Un autre problème que la Table 1 permet d'entrevoir est que la mesure des événements au niveau du *voronoï* tend à produire une population, dans les données de téléphonie mobile, trop homogènement dispersée sur le territoire. En effet, la dispersion de la population, après allocation du domicile au niveau du voronoï, est très faible, à une échelle non cohérente avec la dispersion réelle de la population dans les données fiscales. Autrement dit, une allocation du domicile au niveau du *voronoï* implique de donner trop de poids à des antennes situées dans des régions peu denses, ce qui est, encore une fois, problématique lors du croisement des données.

La Table 1 montre également que la probabilisation des événements au niveau d'une grille apparaît plus satisfaisante que la localisation au niveau des polygones de Voronoï.

D'abord, la distribution de la population dans les données fiscales apparaît plus homogène: la variance de la distribution de la population est réduite (Table 1) et la répartition des tailles d'unité spatiale apparaît plus équilibrée (Table 2). Ensuite, après détection du domicile depuis les données de téléphonie mobile, la population apparaît moins homogènement distribuée entre les territoires avec un ratio des écarts-types plus proche de la part de marché d'Orange (Table 1). Pour ces deux raisons, il apparaît préférable de s'abstraire d'une tesselation de Voronoï lors de l'analyse des propriétés spatiales des données de téléphonie mobile.

Principe

Le passage à une grille se fait à partir des aires d'intersection avec les *voronoï*. Soit $\mathcal{S}(v_j)$ la surface du *voronoï* j et $\mathcal{S}\big(c_i \cap v_j\big)$ celle de l'intersection entre le carreau i et le *voronoï* j. La probabilité d'observer un événement dans le carreau i sachant que celui-ci est mesuré dans le *voronoï* j est donnée par

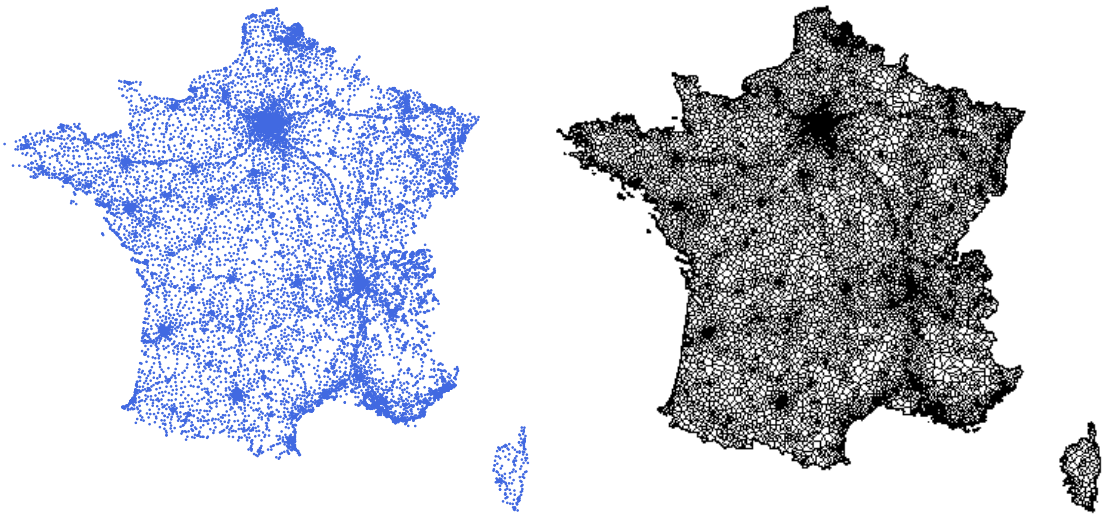
$$p_i^j := \mathbb{P}\big(c_i|v_j\big) = \frac{\mathcal{S}\big(c_i \cap v_j\big)}{\mathcal{S}(v_j)} \tag{10}$$

Le conditionnement par v_j provient du fait que les observations sont effectuées au niveau du *voronoï*. Le passage à une probabilité non conditionnelle, nécessaire pour la détection de domicile ou le calcul de l'indice de ségrégation physique (eq. 9) est donné, pour l'individu x, par

$$\forall c_j \in \mathcal{C}, \quad \mathbb{P}_x(c_j) = \sum_{v_j \in \mathcal{V}} \mathbb{P}\big(c_j\big|v_j\big)\mathbb{P}_x(v_j)$$

où $\mathbb{P}_x(v_l)$ représente la probabilité d'être observé dans le *voronoï* v_l (part des appels de x dans ce *voronoï*).

Le passage de probabilités au niveau du *voronoï* $(p(v_j))_j$ à des probabilités au niveau des carreaux $(p(c_i))_i$ est illustré dans un cas simple avec la Figure 11 où seuls deux *voronoï* sont fréquentés.



(a) Répartition des antennes (b) Tesselation de Voronoï

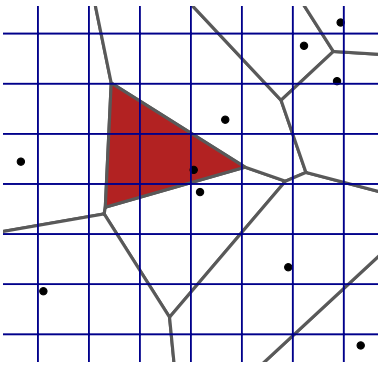
Figure 10 – Distribution des antennes et *voronoï* à l'échelle française

Table 2 – Statistiques descriptives sur la population (dans Filosofi) par unité spatiale

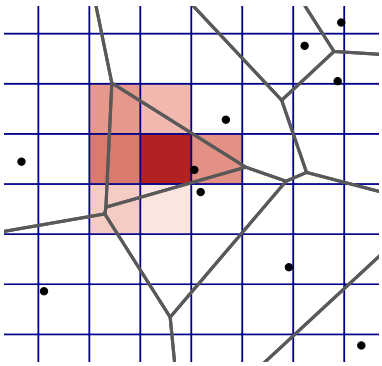
	Marseille		Lyon		Paris	
Taille cellule	Grille	Voronoï	Grille	Voronoï	Grille	Voronoï
Cellules vides	0 (44.41)	0 (0.45)	0 (24.06)	0 (0)	0 (29.46)	0 (5.2)
Moins de 11 ménages	1.04 (21.15)	0.01 (1.36)	0.9 (26.87)	0.01 (2.1)	0.14 (11.98)	0.01 (2.33)
Entre 11 ménages et 50 individus	0.95 (5.57)	0.01 (0.45)	0.79 (6.82)	0 (0.47)	0.11 (2.4)	0.01 (1.2)
Entre 50 et 200 individus	5.66 (11.88)	0.13 (4.77)	5.51 (16.39)	0.08 (3.04)	1.17 (9.04)	0.19 (6.07)
Entre 200 et 1000 individus	25.81 (11.64)	1.88 (13.18)	26.35 (18.27)	2.27 (15.19)	13.21 (20.89)	2.93 (19.41)
Plus de 1000 individus	66.55 (5.35)	97.97 (79.77)	66.46 (7.59)	97.64 (79.21)	85.37 (26.23)	96.85 (65.78)
Total	1 758 986 (8091)	1 758 986 (440)	1 740 388 (5412)	1 740 388 (428)	10 990 852 (12 617)	10 990 852 (2998)

Pour chaque catégorie de taille, la première ligne correspond à la part totale de population concernée et la seconde ligne (entres parenthèses) correspond à la part d'unités spatiales concernées. Dans la catégorie *"Total"* est renvoyée la population et le nombre total d'unités spatiales.

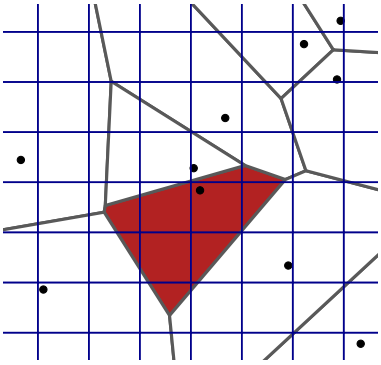
Exemple de lecture: A Paris, lorsqu'on adopte une granularité spatiale fondée sur la tesselation de Voronoï, 65.78% des polygones contiennent plus de 1000 personnes. 96.85% de la population parisienne vit dans ces polygones. Lorsqu'on adopte une grille carroyée, ce ne sont plus que 26.23% des cellules qui appartiennent à cette catégorie de taille, pour 85.37% de la population.



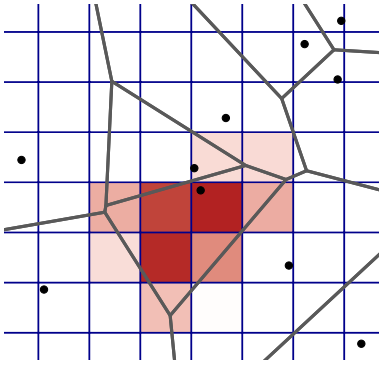
(a) $p(v_1) = 2/3$



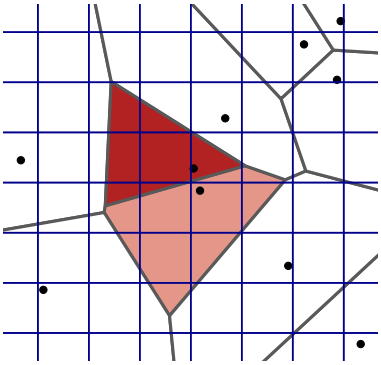
(b) $(p(c_i|v_1))_i$ pour $p(v_1) = 2/3$



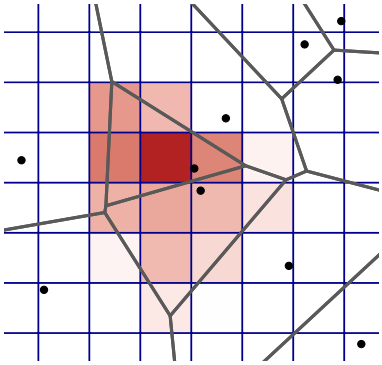
(c) $p(v_2) = 1/3$



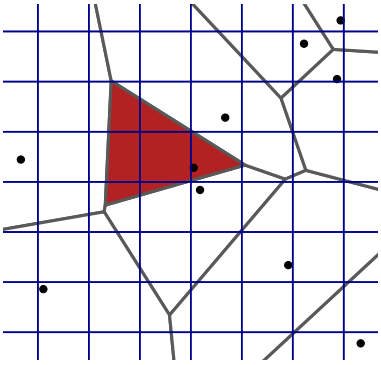
(d) $(p(c_i|v_2))_i$ pour $p(v_2) = 1/3$



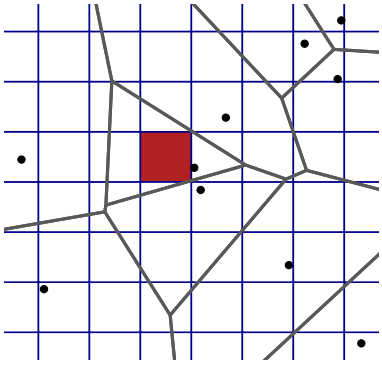
(e) $p(v_1) = 2/3$ et $p(v_2) = 1/3$



(f) $(p(c_i))_i$ pour $p(v_1) = 2/3$ et $p(v_2) = 1/3$



(g) Voronoi le plus fréquent



(h) Cellule la plus probable

Figure 11 – Tesselation de Voronoï et grille: un exemple

Lecture: sur une nuit, le *voronoï* 1 est localisé avec probabilité 2/3 (a) et le *voronoï* 3 avec probabilité 1/3 (c). Les probabilités sont attribuées au niveau des cellules à partir de la surface d'interaction (cf. (b) et (d)). On obtient ainsi une répartition globale des appels au niveau des *voronoï* (e) et des cellules (f). Le lieu de vie est ici assigné à la cellule (resp. *voronoï*) dont la probabilité est maximale (algorithme du *max action* de (Vanhoof et al., 2016)).

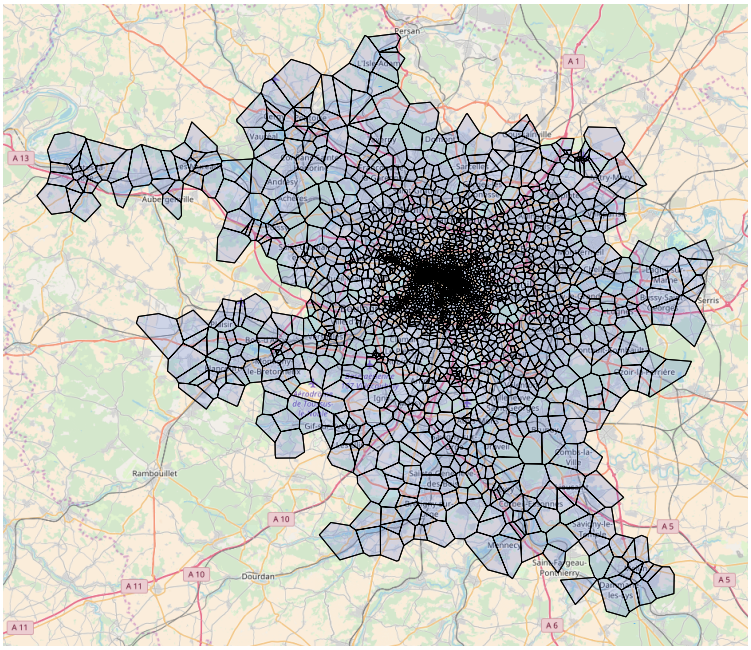
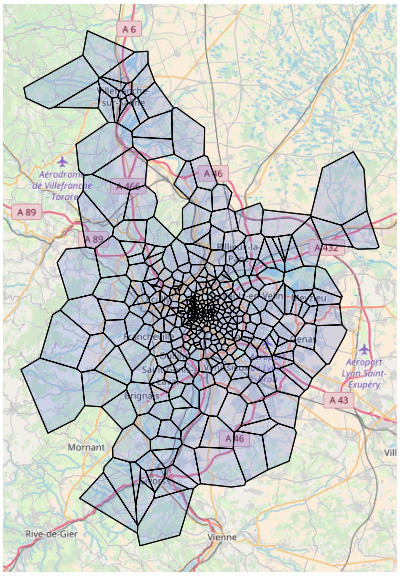
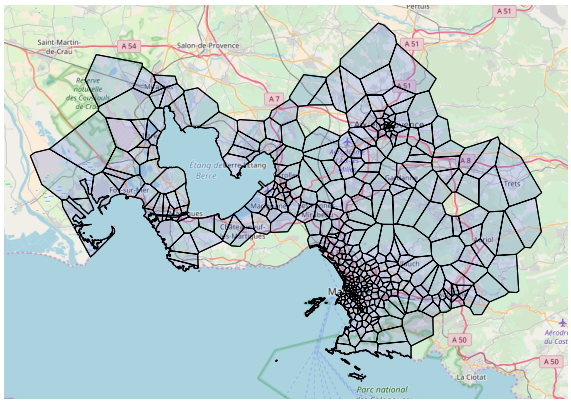


Figure 12 – Décomposition de l'unité urbaine parisienne avec une tesselation de *Voronoï*



(a) Lyon



(b) Marseille

Figure 13 – Décomposition des unités urbaines de Lyon et Marseille avec une tesselation de *Voronoï*

References

Apparicio, Philippe (2000). “Les indices de ségrégation résidentielle: un outil intégré dans un système d’information géographique”. In: *Cybergeo: european journal of geography*.

Aurenhammer, Franz (1991). “Voronoi diagrams—a survey of a fundamental geometric data structure”. In: *ACM Computing Surveys (CSUR)* 23.3, pp. 345–405.

Bengtsson, Linus et al. (2011). “Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti”. In: *PLoS medicine* 8.8, e1001083.

Blondel, Vincent D, Adeline Decuyper, and Gautier Krings (2015). “A survey of results on mobile phone datasets analysis”. In: *EPJ Data Science* 4.1, p. 10.

Calabrese, Francesco et al. (2011). “Interplay between telecommunications and face-to-face interactions: A study using mobile phone data”. In: *PloS one* 6.7, e20814.

Chamboredon, Jean-Claude and Madeleine Lemaire (1970). “Proximité spatiale et distance sociale. Les grands ensembles et leur peuplement”. In: *Revue française de sociologie*, pp. 3–33.

Dabet, Gaëlle and Jean-Michel Floch (2014). “La ségrégation spatiale dans les grandes unités urbaines de France métropolitaine: une approche par les revenus”. In: *Document de travail de la Direction de la Diffusion et de l’Action régionale*.

Deville, Pierre et al. (2014). “Dynamic population mapping using mobile phone data”. In: *Proceedings of the National Academy of Sciences* 111.45, pp. 15888–15893.

Duncan, Otis Dudley and Beverly Duncan (1955). “A methodological analysis of segregation indexes”. In: *American sociological review* 20.2, pp. 210–217.

Floch, Jean-Michel (2017). “Standards of living and segregation in twelve French metropolises”. In: *Economie et Statistique* 497.497, pp. 73–96.

Fontes-Rousseau, Camille and René Jean (2015). “L’artificialisation des terres de 2006 à 2014 : pour deux tiers sur des espaces agricoles”. In: *SSP Agreste Primeur*.

Griffith, Daniel A (1980). “Towards a theory of spatial statistics”. In: *Geographical Analysis* 12.4, pp. 325–339.

Le Roux, Guillaume, Julie Vallée, and Hadrien Commenges (2017). “Social segregation around the clock in the Paris region (France)”. In: *Journal of Transport Geography* 59, pp. 134–145.

Lee, Sang-Il (2001). “Developing a bivariate spatial association measure: an integration of Pearson’s r and Moran’s I”. In: *Journal of Geographical Systems* 3.4, pp. 369–385.

L’Horty, Yannick (2015). “Territoires, emploi et politiques publiques: présentation générale”. In: *Economie & prévision* 1, pp. I–X.

Madoré, François (2015). “Approche comparative de la ségrégation socio-spatiale dans les aires urbaines françaises”. In: *Annales de géographie*. 6. Armand Colin, pp. 653–680.

Massey, Douglas S and Nancy A Denton (1988). “The dimensions of residential segregation”. In: *Social forces* 67.2, pp. 281–315.

Maurin, Éric (2007). “31. La ségrégation urbaine, son intensité et ses causes”. In: *Repenser la solidarité*. Presses Universitaires de France, pp. 621–633.

Maurin, Eric et al. (2009). “La peur du déclassement”. In: *Seuil: La République des Idées*.

Oberti, Marco and Edmond Préteceille (2016). *La ségrégation urbaine*. La Découverte.

Ollion, Étienne and Julien Boelaert (2015). “Au-delà des big data. Les sciences sociales et la multiplication des données numériques”. In: *Sociologie* 3, vol. 6.

Openshaw, Stan and S Openshaw (1984). “The modifiable areal unit problem”. In: Geo Abstracts University of East Anglia.

Reardon, Sean F and Kendra Bischoff (2011). “Income inequality and income segregation”. In: *American Journal of Sociology* 116.4, pp. 1092–1153.

Reardon, Sean F and David O’Sullivan (2004). “Measures of spatial segregation”. In: *Sociological methodology* 34.1, pp. 121–162.

Schelling, Thomas C (1969). “Models of segregation”. In: *The American Economic Review* 59.2, pp. 488–493.

Silm, Siiri and Rein Ahas (2014). “The temporal variation of ethnic segregation in a city: Evidence from a mobile phone use dataset”. In: *Social Science Research* 47, pp. 30–43.

Tennekes, Martijn (2018). “Geographic Location of Events”. In: *Vignette for mobloc package*.

Theil, Henry (1972). *Statistical decomposition analysis; with applications in the social and administrative sciences*. Tech. rep.

Tiebout, Charles M (1956). “A pure theory of local expenditures”. In: *Journal of political economy* 64.5, pp. 416–424.

Tovar, Élisabeth (2009). “Mesurer la pauvreté et la ségrégation en Île-de-France: une approche capabiliste”. In: *Document de travail du Centre d’Études de l’Emploi* 116.

Vanhoof, Maarten et al. (2016). “Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics”. In: *Proceedings of the Conference of the Italian Statistical Society*.

Voronoi, Georges (1908). “Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites.” In: *Journal für die reine und angewandte Mathematik* 133, pp. 97–178.

Williams, Nathalie E et al. (2013). “Measurement of human mobility using cell phone data: developing big data for demographic science”. In: *Population Association of America Annual Meeting*.