Estimer la population résidente à partir de données de téléphonie mobile, une première exploration

Estimating the Residential Population from Mobile Phone Data, an Initial Exploration

Benjamin Sakarovitch*, Marie-Pierre de Bellefon*, Pauline Givord et Maarten Vanhoof*****

Résumé – De nombreux travaux s'intéressent à l'utilisation des données issues de la téléphonie mobile pour construire des indicateurs statistiques. Ces données ont l'intérêt de fournir des informations à la fois à une résolution spatiale élevée et à une haute fréquence. Plusieurs applications proposent par exemple de mesurer la population présente à des niveaux spatiaux ou temporels fins. L'exploitation de ces données pour construire des indicateurs statistiques soulève néanmoins des difficultés : les données d'un seul opérateur ne sont pas représentatives de la population totale, et ces données anonymisées sont souvent pauvres en caractéristiques socio-démographiques ce qui limite la qualité des redressements. Cet article s'appuie sur un fichier issu des enregistrements des activités d'abonnés d'un grand opérateur français pour donner un premier aperçu du potentiel mais aussi des problèmes posés par de telles données, illustré par l'estimation d'indicateurs de populations résidentes inférés à partir du simple enregistrement des activités des personnes.

Abstract – *Many studies are focused on using data derived from mobile phones to construct statistical indicators. Mobile phone data have the advantage of providing information with both high spatial resolution and at high frequency, allowing applications such as measurements of the spatial or temporal details of population presence. Nonetheless, using mobile phone data to construct statistical indicators raises difficulties: data from a single operator are not representative of the whole population and they often lack sociodemographic detail, which limits their quality for many applications. This article is based on a database of mobile phone records from subscribers collected by a large French operator. It aims to offer a view on the potential, but also the problems posed by mobile phone data, specifically by illustrating how indicators of residential populations can or can not be estimated from them.*

Codes JEL / JEL Classification : C55, C81, R23
Mots-clés : téléphonie mobile, comptes-rendus d'appels (CRA), population présente
Keywords: mobile phones, call detail records (CDR), present population

Ce travail a été mené dans le cadre d'une convention tripartite entre Orange, Eurostat et la direction de la méthodogie et de la coordination statistique et internationale de l'Insee. Les auteurs tiennent particulièrement à remercier Zbigniew Zmoreda et Cezary Ziemlicki pour leur accueil au sein du laboratoire Sense et leur appui pour l'utilisation des données de téléphonie mobile, les participants de la Task Force Big Data d'Eurostat et spécialement Michail Skaliotis et Fernando Reis, ainsi qu'Elise Coudin et Vincent Loonis pour leurs conseils. Les auteurs restent seuls responsables des erreurs ou omissions qui pourraient demeurer dans l'article.

** Insee (marie-pierre.de-bellefon@insee.fr ; benjamin.sakarovitch@insee.fr)*
*** Insee, Crest (pauline.givord@oecd.org)*
**** Open Lab, Newcastle University / Orange Labs (m.vanhoof1@ncl.ac.uk)*

Reçu le 20 février 2018, accepté après révisions 5 octobre 2018

Pour citer cet article: Sakarovitch, B., Bellefon, M. (de), Givord, P. & Vanhoof, M. (2018). Estimating the Residential Population from Mobile Phone Data, an Initial Exploration. *Economie et Statistique / Economics and Statistics*, 505-506, 109–132. https://doi.org/10.24187/ecostat.2018.505d.1968

L'exploitation des Big Data, liée à des progrès rapides dans les capacités de stockage et d'analyse de données massives, a pris un essor important au cours de la dernière décennie. Le potentiel de ces données, créées par la multiplication des traces numériques générées par l'activité des individus ou des entreprises, est souvent étudié sous l'angle de l'analyse prédictive ou de l'aide à la décision. Elles peuvent également être une source d'observations intéressantes pour la construction d'indicateurs statistiques, ce qui explique l'intérêt des instituts de statistiques publiques pour ces données[1]. Les opportunités pressenties pour l'usage de ces données seraient de réduire les délais de publications en profitant d'un accès très rapide à de l'information utile (par exemple dans le domaine de l'analyse conjoncturelle), mais également de produire des statistiques à un niveau plus fin (au niveau géographique en particulier) que celui qui peut être autorisé par les données d'enquête par exemple, et enfin de réduire la charge de collecte de l'information auprès des personnes et des entreprises. Ainsi, la collecte automatique de prix (à partir de sites de e-commerce, ou des données de facturation des grandes enseignes commerciales) est utilisée par plusieurs instituts de statistique pour enrichir la construction des indices de prix à la consommation[2]. L'utilisation pour la production statistique de sources alternatives ou complémentaires aux données « classiques » fait également l'objet de plusieurs travaux exploratoires. Si cette problématique n'est pas nouvelle puisque la statistique publique s'appuie depuis plusieurs décennies, outre sur les enquêtes statistiques, sur les sources administratives (par exemple, l'Insee utilise depuis longtemps son suivi statistique des salaires sur les déclarations sociales des employeurs), les données Big Data soulèvent des questions spécifiques, en particulier techniques (par exemple pour les données très volumineuses ou de format non structuré).

Les données issues de la téléphonie mobile font partie des sources identifiées comme particulièrement prometteuses pour compléter l'information statistique. Ces données correspondent aux enregistrements réguliers de la localisation des téléphones des abonnés des opérateurs de téléphonie mobile, ou tout au moins de l'antenne à laquelle ce téléphone s'est connecté (ainsi que de la date et de l'heure). Elles permettent donc de disposer d'informations sur les personnes présentes en un lieu, avec des niveaux de précision géographiques et temporels très fins. Alors que la statistique publique produit des informations sur la population résidente (*via* le recensement en particulier), l'accès de ces données à un niveau très fin permettrait de détecter le nombre de personnes *présentes* à un moment donné (qui dépend par exemple de la fréquentation touristique, des comportements d'activité, etc., cf. Terrier, 2009), ainsi que les flux de personnes entre plusieurs points. La localisation régulière des abonnés permet de construire des cartographies de la population présente et de son évolution (Deville *et al*., 2014 ; Debusschere *et al*., 2016 ; Ricciato *et al*., 2015). L'exploitation de ces données peut par exemple permettre de mesurer la variabilité de la fréquentation de certains lieux au cours de la journée ou de l'année, d'améliorer la connaissance précise des temps de transports selon les différents modes (en particulier pour les « petits » déplacements quotidiens) et de définir des matrices de mobilités à un niveau fin (voir Aguiléra *et al*., 2014, dans le cas de l'évaluation des performances du réseau de transport d'Île-de-France ou encore Demissie *et al*., 2014, sur le Sénégal). Les profils de fréquentation d'une zone à différents moments dans le temps sont susceptibles d'aider à analyser les dynamiques territoriales. On s'attend en effet à ce que les profils de présence (ou d'activité) changent au cours de la journée selon la nature du lieu (lieu de résidence, d'activité ou de transit). Toole *et al*. (2012) arrivent ainsi à distinguer, selon les profils quotidiens de présence observés à un niveau fin, l'activité principale des zones (commerces, résidentielles, industrielles ou parking par exemple) sur la région de Boston. Pour la France, Vanhoof *et al*. (2017) appliquent cette démarche à une échelle un peu plus large, les communes, et mettent en évidence une corrélation élevée entre les profils d'activité agrégée observés au niveau des antennes de téléphonie mobile et la typologie de la commune, telle que définie par les zonages en aires urbaines de l'Insee. Ces informations peuvent également enrichir l'analyse des réseaux interpersonnels, par l'analyse de la force des communications entre les abonnés (Grauwin *et al*., 2017).

L'exploitation de ces données soulève cependant plusieurs questions. En premier lieu, il est nécessaire de garantir le respect de la vie privée des abonnés. Pouvoir reconstituer des trajectoires individuelles grâce aux traces laissées par les abonnés expose à un risque élevé

1. Voir par exemple le mémorandum de Scheveningen (2013).

2. En France, le projet « données de caisse » s'appuie ainsi sur les enregistrements des prix issus des données de facturation de plusieurs grandes enseignes (voir Leonard et al., *2017, et* Economie et Statistique / Economics and Statistics *N° 509 à paraître).*

de « ré-identification ». Même en supprimant toutes mentions directes sur leur identité, il est possible d'attribuer avec une grande probabilité une trajectoire observée à une personne unique (Montjoye *et al.*, 2013). Cela exige que l'exploitation de ces données se fasse sur l'information agrégée à un niveau suffisant (au risque de réduire sa pertinence), ou à travers des procédures ne permettant pas d'accéder directement aux données sensibles[3]. Sur le plan technique, les données correspondant aux millions de clients envoyant quotidiennement des dizaines de SMS ou passant plusieurs appels, représentent des volumes gigantesques qui nécessitent des infrastructures de stockage et de calcul adaptées.

Les instituts nationaux de statistique (INS) s'intéressent aux potentiels de ces données. Un rapport d'Eurostat (2014) étudie ainsi l'apport de ces données comme source complémentaire pour améliorer la précision des indicateurs actuels de tourisme. Plusieurs INS ont lancé de premières expérimentations d'exploitation de ces données, et un programme de coordination a été lancé en 2016 pour mutualiser les expertises sur ce sujet[4]. L'une des questions porte sur les modalités d'un accès des INS à ces données qui garantissent le respect de la vie privée pour les abonnés, ainsi que le secret des affaires pour les entreprises concernées. Pour la France, ces questions ont été abordées par un rapport du CNIS sur la réutilisation des données des entreprises par la statistique publique (2016), qui s'intéressait en particulier au cas des données de téléphonie mobile[5]. D'autres INS européens ont également lancé des négociations avec des opérateurs nationaux sur des projets expérimentaux (Debusschere *et al.*, 2016 pour la Belgique). Ces expérimentations sont nécessaires pour définir quelles sont les informations nécessaires pour construire des indicateurs statistiques pertinents (Vanhoof *et al.*, 2018). Il s'agit ainsi d'évaluer dans quelle mesure des données agrégées, à la fois moins sensibles et moins coûteuses à traiter, peuvent apporter une contribution suffisante. Ces expérimentations permettent également de se confronter aux données et aux multiples questions que pose leur utilisation pour des indicateurs statistiques.

En premier lieu, les données de téléphonie mobile posent souvent des questions classiques de représentativité. L'accès aux données d'un opérateur ne fournira des informations que sur ses abonnés, qui ne constituent qu'une partie de la population. Les redressements nécessitent d'avoir des informations annexes, sur les caractéristiques démographiques de ces abonnés par exemple, mais aussi, pour obtenir des statistiques à un niveau spatial fin – ce qui est *a priori* l'un des intérêts principaux de ces données – sur l'implantation locale de ces opérateurs. Le taux d'équipement peut être variable en fonction des caractéristiques de la population : certains peuvent ne pas être équipés – par exemple Wesolowski (2013) met en évidence les problèmes de l'inégale répartition des téléphones dans différents groupes sociaux au Kenya pour l'exploitation de ce type de données. À l'inverse, dans les pays développés, certaines personnes peuvent être multi-équipées.

Une deuxième difficulté de l'exploitation de données de téléphonie mobile pour des mesures localisées tient au maillage des antennes, qui ne coïncide pas en principe avec les maillages géographiques usuels (les découpages administratifs par exemple). Les antennes ne sont pas réparties de manière uniforme – elles sont plus nombreuses dans les zones densément peuplées, moins dans les zones rurales. Pour les utiliser sur des unités territoriales plus classiques, il est nécessaire de procéder à une projection géographique de cette grille d'antennes, ce qui introduit des approximations (Ricciato *et al*., 2015).

Enfin, il est indispensable de clarifier ce qui peut être mesuré à partir des données de téléphonie mobile. Ces données sont produites « naturellement » (on parle parfois d'*organic data*, par opposition aux *designed data*, fournies par des enquêtes construites spécifiquement pour mesurer l'objet d'études[6]), elles sont simplement le reflet des traces laissées par les abonnés sur le réseau de téléphonie mobile. Pour qu'un indicateur statistique soit intelligible par tous, il est indispensable de s'accorder préalablement sur une définition de ce qu'on souhaite mesurer. Par exemple, un touriste est en général défini comme une personne

3. Par exemple, le projet Opal (http://www.opalproject.org/about-us/) propose de mettre à disposition des chercheurs une plateforme permettant de faire tourner des algorithmes sur des données de téléphonie mobile, auxquelles le chercheur n'a pas directement accès : on parle d'Open Algorithm plutôt que d'Open data.

4. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP5_Mobile_phone_data.

5. Voir rapport Cnis-Insee Réutilisation par le Système Statistique Public des Informations des Entreprises, *https://www.cnis.fr/wp-content/uploads/2017/09/DC_2015_2e_reunion_COM_Entreprises_GT_BON_Cnis-Insee.pdf.*

6. Cette distinction a été en particulier proposée par le Census Bureau, voir : https://www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html.

enregistrée « en dehors de son environnement habituel ». Les mesures de fréquentation touristique d'un lieu demandent donc de distinguer parmi les personnes présentes en ce lieu celles qui n'y résident pas mais également celles qui n'y travaillent pas régulièrement. Mesurer ces informations à partir des enregistrements des déplacements des abonnés nécessite d'être capable *a minima* d'identifier la résidence des personnes, voire leur lieu de travail « habituel » (Janzen *et al*., 2018). Plusieurs études ont été menées à partir des données de téléphonie mobile sur cette question. Ahas *et al*. (2010), par exemple, montrent qu'il est possible à partir des traces laissées sur le réseau par un individu de reconstituer ses « points d'ancrage » (*anchor places*), c'est-à-dire les lieux importants pour lui, où il passe de manière récurrente – son domicile et son travail étant les plus évidents d'entre eux (Ahas *et al*., 2010). Comme souligné également par Song *et al*. (2010), le temps passé par chacun se concentre en général sur un nombre limité de lieux. Plusieurs algorithmes ont été proposés pour identifier, à partir des profils de déplacement observés, le domicile probable d'un abonné (Vanhoof *et al*., 2017 ; Bojic *et al*., 2015 ; Isaacman *et al*., 2011). Ce point est essentiel car il est un préalable à de nombreuses autres analyses (Blondel *et al*., 2015), qui dépassent le simple cadre du tourisme.

Cette étude se propose d'illustrer les questions empiriques soulevées par l'utilisation des données de téléphonie mobile à partir d'un exemple concret. Elle s'appuie sur les enregistrements exhaustifs des abonnés de l'opérateur historique de téléphonie français pendant cinq mois de 2007, et propose d'estimer des populations résidentielles, de les comparer aux estimations de la statistique publique prises comme données de référence et d'analyser les sources d'écart. Cette méthodologie permet de tester la pertinence de plusieurs algorithmes de détection de résidence et de plusieurs techniques de redressement des données : deux questions essentielles lorsque l'on souhaite utiliser des données mobiles pour produire des comptages.

La suite de l'article est organisée comme suit. Une première partie décrit les différents types d'enregistrement de téléphonie mobile et détaille les étapes pour passer de ces enregistrements à une mesure de comptage localisée. Une seconde partie développe les questions liées aux approximations faites pour modéliser la couverture des antennes. La partie suivante présente les différentes méthodes utilisées pour estimer des populations résidentes. Les questions de représentativité et des solutions qui peuvent être apportées pour les résoudre, ainsi que des comparaisons avec des sources de référence y sont détaillées. Enfin, la dernière partie propose des extensions d'utilisation de ces données pour caractériser les dynamiques présentielles de population.

Des enregistrements aux données

Un réseau de téléphonie mobile permet la communication *via* la transmission d'ondes radios entre les appareils, les antennes-relais et les commutateurs centralisés de l'opérateur qui dirigent vers d'autres antennes-relais la liaison pour le correspondant. Ces réseaux ont une structure cellulaire, c'est-à-dire que chaque antenne couvre une certaine zone et qu'un téléphone peut changer de cellule sans que la communication ne se coupe.

Principe des enregistrements de téléphonie mobile

Les données utilisées ici correspondent aux enregistrements par les antennes-relais du réseau de téléphonie cellulaire, qui signalent la présence des téléphones cellulaires des abonnés à proximité de ces antennes. Elles sont disposées sur des tours dont on peut connaître les coordonnées. Il est donc en principe possible de construire des indicateurs sur la fréquentation de certains lieux, ou les comportements de mobilité à des niveaux géographiques et temporels très fins. La fréquence et la régularité de ces enregistrements, et donc le niveau de finesse (la granularité) auquel on pourra construire des indicateurs, dépend du type de données. Celles-ci sont de plusieurs types.

Les CDR (*Call Detail Records*) ou CRA (comptes-rendus d'appels) correspondent à l'émission ou la réception d'un appel ou d'un SMS, soit une action volontaire de l'abonné : on parle de données actives. Ces données servent en général à la facturation, et les opérateurs les enregistrent donc « par défaut ». En France, ils ont l'obligation de conserver ces données pendant six mois. Outre des indications sur la localisation des abonnés, ces données peuvent être mobilisées par exemple pour des études sur les comportements des utilisateurs (fréquence des appels, préférence pour les SMS, etc.).

Les données de signalisation (*signaling data*), également appelées données passives, sont générées à partir des réseaux de télécommunication et internet (2G, 3G, 4G), par le fait que tous les téléphones mobiles se connectent régulièrement aux antennes les plus proches (avec une périodicité variable, pouvant s'étendre entre trois heures et une dizaine de minutes) sans nécessairement que cela provienne d'une action de l'usager sur le mobile. Elles apportent donc des informations bien plus complètes que les CDR si on souhaite par exemple mesurer la fréquentation d'un lieu à un moment précis, ou suivre les déplacements des personnes. En revanche, elles sont plus coûteuses à traiter. Par défaut, ces « événements » ne sont pas enregistrés par les opérateurs : le faire nécessite des volumes de stockage très importants.

En termes de couverture de la population, les données enregistrées par un opérateur, qu'elles soient actives ou passives, ne correspondent en principe qu'à celles de ses abonnés. Cependant, des accords d'itinérance (*roaming* en anglais) peuvent exister qui permettent aux abonnés d'un opérateur d'utiliser le réseau de ses concurrents quand il se trouve en dehors de la zone de couverture de son opérateur. En France, il y a peu d'accords d'itinérance entre les opérateurs nationaux, et cette situation « d'itinérance » correspond essentiellement à des abonnés étrangers. Cela signifie en particulier qu'il est possible d'identifier les personnes seulement de passage en France, à condition qu'ils utilisent leurs téléphones (pour les données CDR) ou tout au moins qu'ils le laissent allumé (pour les *signaling data*). La carte SIM permet en effet d'identifier le pays d'implantation de l'opérateur téléphonique, on peut alors inférer la nationalité probable de l'abonné du téléphone[7] (encadré 1).

La démarche pour passer des enregistrements à des comptages de population

Pour tirer des données enregistrées par le réseau mobile, des informations d'intérêt pour la statistique publique une série de traitements est nécessaire (schéma).

La première étape correspond à la cartographie des événements enregistrés sur le réseau

7. Avant juin 2017, ces frais d'itinérance à l'étranger étaient facturés par les opérateurs. Depuis cette date, la Commission européenne a imposé la fin de ces facturations. Il est possible que cela aboutisse à terme à créer un marché plus concurrentiel à l'échelle de l'Europe, des ressortissants d'un pays pouvant plus facilement recourir à un opérateur étranger, et qu'il soit donc plus difficile de repérer ces déplacements.

Encadré 1 – **Description des bases de téléphonie mobile utilisées**

L'étude porte sur l'exploitation d'un fichier anonymisé de données CDR (*Call Detail Records*), correspondant à l'enregistrement exhaustif des activités des abonnés de l'opérateur Orange sur le territoire français métropolitain pour une période de 5 mois, de mi-mai à mi-octobre 2007[(a)]. Elles correspondent à environ 18 millions de cartes SIM et plus de 20 milliards d'observations. Ces données ne contiennent aucune information directe sur le nom de l'abonné, ni sur son adresse. Il a cependant été possible pour l'étude de les compléter par certaines informations issues d'un fichier dit *Customer Relationship Management* (CRM), désigné dans l'article par « fichier client ». Ce fichier indique pour 12.4 millions de cartes SIM également présentes dans les CDR (soit environ deux tiers) le département de résidence déclaré par l'abonné. L'abonné (celui qui est identifié dans le fichier client) n'est pas nécessairement l'utilisateur du téléphone. C'est le cas des sociétés, mais aussi des parents finançant l'abonnement de téléphones portables utilisés par leurs enfants. Par ailleurs, les informations du fichier client peuvent « se périmer », par exemple en cas de déménagement, les mises à jour n'étant pas systématiques.

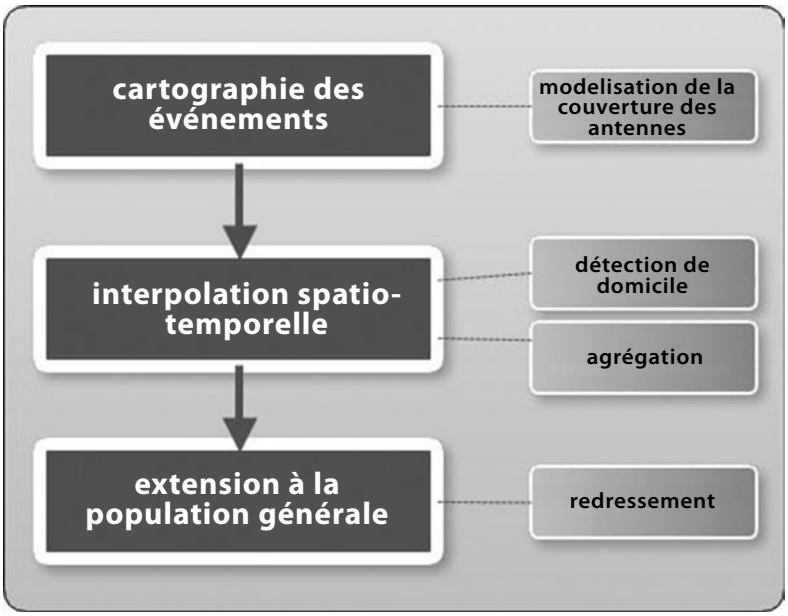
(a) Ces données anonymisées sont à disposition du laboratoire SENSE d'Orange Labs, pour des projets à vocation de recherche.

Tableau A
Structure des comptes rendus d'appels et des variables essentielles

Carte SIM émettrice	Carte SIM réceptrice	Type d'événement	Antenne émission	Antenne réception	Horodatage	Durée
SIM-1	SIM-2	Appel	A-1	A-2	13/06/2007 14:26:03	7m32s
SIM-1	SIM-3	SMS	A-3	-	25/08/2007 12:04:58	-

Note : dans le cas des SMS, on ne connaît pas l'antenne par laquelle passe la réception du message.

Schéma
Processus de traitement des données de téléphonie mobile pour la statistique publique



(appels ou SMS). Ces derniers ne sont localisés qu'indirectement, *via* l'antenne qui a transmis le signal. La localisation de l'évènement est inférée à partir des informations disponibles sur l'antenne. Pour cela, il est nécessaire d'une part de définir une grille spatiale sur laquelle on souhaite repérer les différents événements, et d'autre part de modéliser la zone de couverture de l'antenne (en particulier en fonction des caractéristiques techniques des antennes si elles sont disponibles, voir Ricciato *et al*., 2017). L'événement enregistré sera alors repéré sur la grille spatiale choisie, à partir de la prédiction fournie par le modèle de couverture de l'antenne sur lequel il a été modélisé. Comme détaillé dans la section « Un maillage très hétérogène du territoire », compte tenu des informations disponibles dans les données utilisées ici (les coordonnées des mâts supportant les antennes), on procède ici par tessellation de Voronoï (voir encadré 2), ce qui repose sur l'hypothèse la plus simple pour la modélisation des zones de couverture des antennes.

La seconde étape consiste à procéder à une interpolation spatio-temporelle pour passer de l'enregistrement des événements à un agrégat correspondant à une définition préétablie. Il s'agit de définir des unités d'agrégation (à la fois temporelles et spatiales) pour la production des indicateurs statistiques. Par exemple, on peut souhaiter construire des indicateurs de population présente en des lieux suivant un découpage administratif classique (au niveau de l'iris, la commune, etc.) à des moments précis de la journée, ou tout au moins sur des plages horaires fixes. La grille qui permet de passer des antennes à des lieux, liée aux caractéristiques techniques de ces dernières, ne se superpose pas naturellement au découpage territorial conventionnel. Comme décrit dans la section « La réallocation des cellules de Voronoï à une autre grille », il est donc nécessaire de procéder à une interpolation spatiale. Cette interpolation spatiale doit se doubler dans certains cas d'une interpolation temporelle car les enregistrements des cartes SIM n'ont pas de fréquence définie ni régulière : on pourra par exemple disposer grâce aux appels d'une localisation d'un même téléphone à 7h47 puis à 8h12 mais en revanche la localisation à 8h de ce même téléphone n'est pas directement connue. Si l'objet est de mesurer la population sur des heures précises il sera nécessaire de reconstituer le lieu de localisation probable à 8h à partir de ces données disponibles. Enfin, pour estimer des indicateurs de populations résidentes étudiées dans cet article, il faut par exemple être capable d'inférer le lieu de résidence probable, en fonction des heures et de la localisation des appels. La définition d'algorithmes de détection de résidence est détaillée dans la partie « Caractérisation du domicile :

"dis-moi quand tu téléphones, je te dirai où tu habites" ».

Dans la dernière étape, on cherche à obtenir des estimations correspondant à la population de référence, en se basant sur ces agrégats constitués à partir des enregistrements disponibles uniquement pour les abonnés d'un opérateur de téléphonie mobile. Ces redressements se font en fonction de sources externes (par exemple sur les parts de marchés des opérateurs). La section « Redresser les données pour obtenir des estimateurs de population résidente » présente plusieurs estimations possibles, en fonction de la richesse des informations annexes disponibles, en insistant sur les hypothèses sous-jacentes. Ces résultats sont comparés aux statistiques de référence (les populations résidentes, telles que mesurées par les sources fiscales retraitées par l'Insee).

L'approximation de la couverture des antennes : une simulation à partir des données fiscales

Un maillage très hétérogène du territoire

La couverture spatiale est inégale sur le territoire. Pour chaque opérateur de téléphonie mobile, les tours d'antenne relais, qui fournissent l'information principale sur la localisation, sont implantées de manière irrégulière sur le territoire. Comme le montre la figure I, en 2007, la distribution des antennes de l'opérateur Orange était très dense en zone urbaine, mais beaucoup moins dans les zones rurales. Par ailleurs, des infrastructures mobiles peuvent venir renforcer localement le réseau pour éviter qu'il ne se trouve saturé lors d'événements particuliers réunissant des foules importantes – compétition sportive, concerts, manifestations. De manière plus structurelle l'évolution des technologies (apparitions successives des 2G, 3G, 4G, etc.) entraîne un renouvellement du réseau et donc des modifications de la localisation des antennes.

En pratique, on peut inférer la position probable d'un téléphone en fonction des antennes auxquelles il s'est connecté. La solution la plus simple est de supposer qu'il s'est connecté à l'antenne la plus proche[8]. On peut définir une partition du territoire à partir d'une tessellation de Voronoï (encadré 2), qui fait correspondre à chaque antenne l'ensemble des points de l'espace dont elle est l'antenne la plus proche. Ce découpage est une approximation de celui produit par la couverture réelle des antennes. Il ne rend pas compte du fait que les véritables aires de couverture se superposent et que la charge des téléphones présents dans une zone est

8. Il s'agit d'une approximation qui repose sur l'hypothèse que les antennes émettent toutes avec la même puissance et dans toutes les directions. En réalité, une même tour peut abriter plusieurs antennes émettant dans des directions d'émission (azimut) et des portées différentes. Scholus (2015) ou Tennekes (2015) construisent un modèle d'inférence de la position du mobile qui repose sur l'observation fine des propriétés des antennes, ainsi que de la connaissance de la distance entre le téléphone et l'antenne ayant retransmis le signal. Ces informations (propriétés des antennes, distance au téléphone) ne sont cependant pas toujours disponibles dans les données. Par ailleurs, disposer d'informations très fréquentes peut permettre d'opérer des triangulations qui permettent une connaissance fine de la position d'un mobile. Dans le cas idéal où les distances à plusieurs antennes (au moins 3) sont rapportées il est possible de procéder par triangulation et déduire la position exacte du téléphone.

Encadré 2 – **Partitionner l'espace, la tessellation de Voronoï**

La tessellation de Voronoï est une partition de l'espace qui s'appuie sur un ensemble de points donnés : les graines. Chaque point du plan est alloué à la graine dont il est le plus proche. Les frontières entre les différentes régions du plan forment les côtés de polygones contenant exactement une graine.

Ce découpage du plan est utile pour le traitement des données mobiles lorsqu'on ne connaît que les localisations des différentes tours d'antennes (qui constituent donc ces graines). On fait alors l'hypothèse qu'un appel est émis par l'antenne la plus proche, ce qui signifie donc que le téléphone se trouve dans le polygone de Voronoï associé à cette antenne.

Figure A
Exemple d'une tessellation par polygones de Voronoï à partir de 7 points

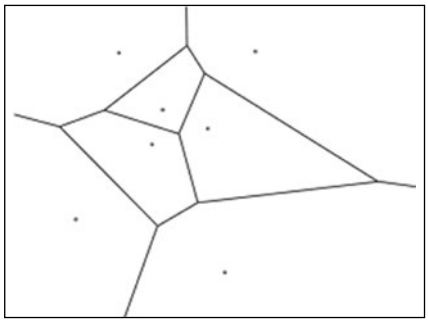
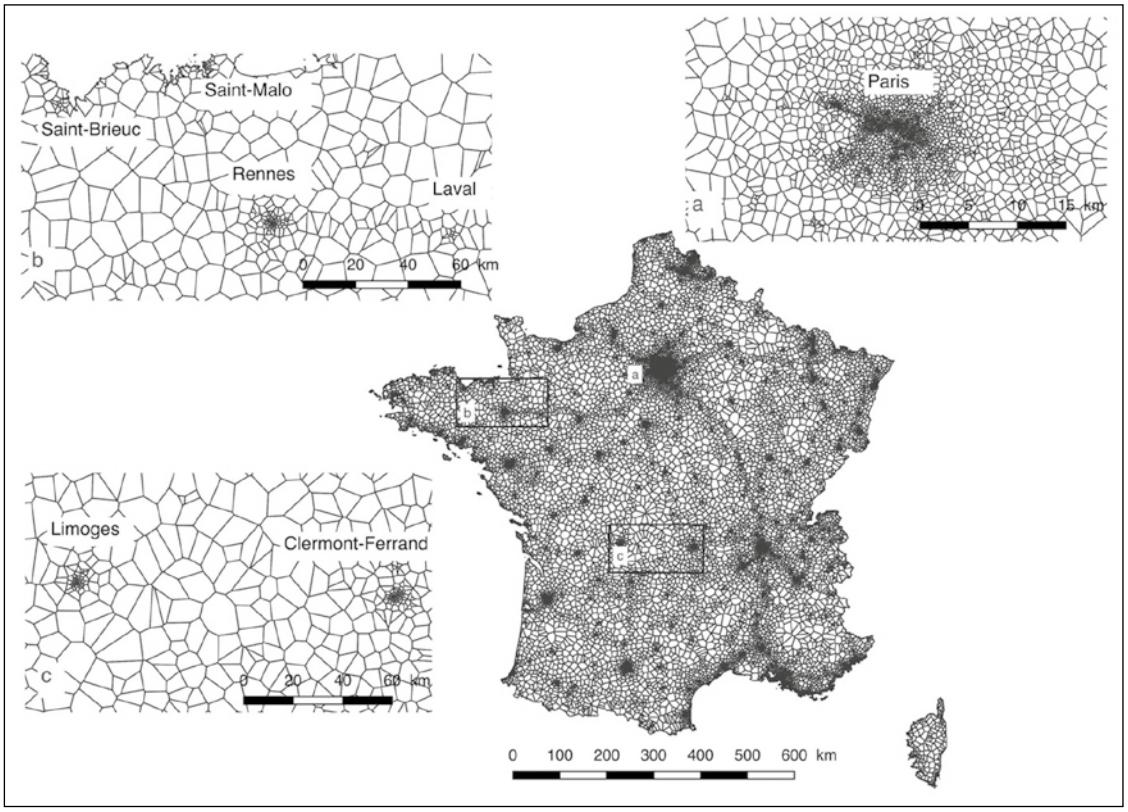
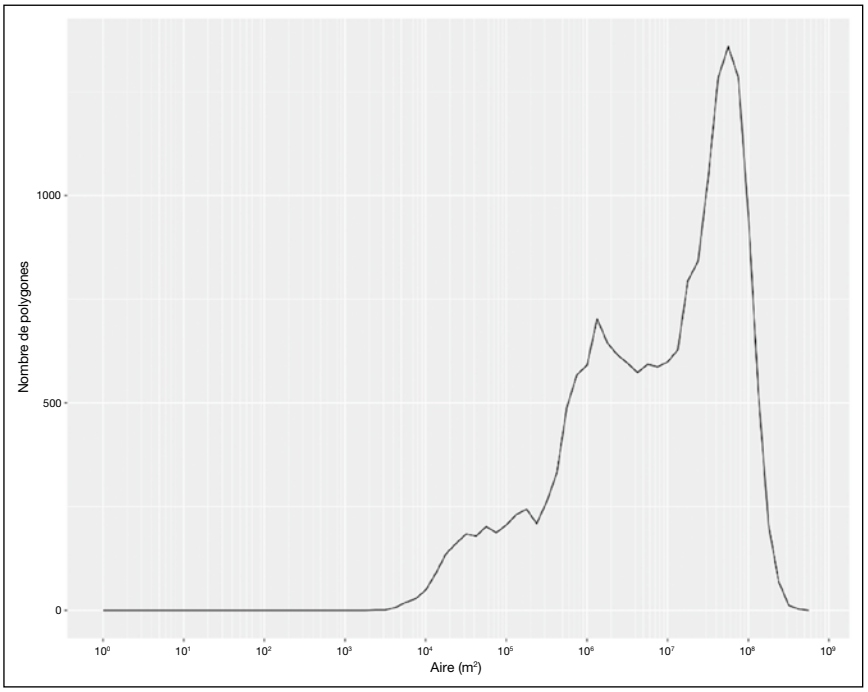


Figure I
Polygones de Voronoï associés aux antennes de l'opérateur Orange, France métropolitaine



Source : CDR Orange.

Figure II
Distribution des aires (en m²) des polygones de Voronoï associés aux antennes de l'opérateur Orange



Lecture : les modes de la distribution se trouvent à 10^6 et 10^8 m^2 (l'échelle du graphique est logarithmique), il n'y a pas de polygones d'aire inférieure à 10^3 m^2.

répartie entre les différentes antennes la couvrant. L'unité spatiale qu'on considère ensuite est donc le polygone de Voronoï, chacun étant construit autour d'une tour d'antennes. Du fait de la répartition inégale des antennes sur le territoire, les aires de ces polygones sont de tailles très variables (figure I). La figure II montre la distribution de leurs aires. On constate que si de nombreuses cellules de Voronoï ont une aire assez petite (quelques hectares) la diversité des aires couvertes est très importante et va jusqu'à plus d'une dizaine de milliers d'hectares pour quelques cellules. Ces aires importantes ne correspondent pas à la couverture effective des antennes mais proviennent de la tessellation de Voronoï dans les régions où les antennes sont très éloignées les unes des autres et peuvent même comprendre des zones en réalité « blanches » ou aucun signal n'est reçu. Les cellules de Voronoï les plus petites se trouvent dans les zones les plus densément peuplées.

La réallocation des cellules de Voronoï à une autre grille

La partition géométrique de l'espace, purement technique (liée à la position des antennes), ne coïncide évidemment pas avec les découpages du territoire utilisés pour la diffusion de données statistiques territorialisées. Les contours des polygones correspondant à la répartition des antennes n'ont aucune raison de se superposer aux limites administratives des communes ou départements, et ne sont pas non plus imbriqués dans les maillages plus fins utilisés par la statistique publique, comme les IRIS (les briques de base pour la diffusion d'information infra communale, imbriquées dans la géographie communale et constituant des unités de taille homogène en termes de population[9]). Estimer des statistiques territorialisées à partir des données de téléphonie mobile demande de revenir à une grille administrative classique. C'est notamment la condition pour la mise en relation avec d'autres informations fournies à l'échelle de cette grille administrative. On souhaite donc procéder à une interpolation. Dans la suite et à défaut d'informations plus pertinentes, cette interpolation sera faite simplement en fonction de la surface des unités spatiales considérées. La grille administrative de base choisie est la grille communale, divisée en arrondissements pour Paris, Lyon et Marseille. L'estimation de la population présente dans une unité géographique correspondra à la somme des populations estimées dans les polygones entièrement compris dans cette unité et d'un prorata de ces populations estimées (proportionnel à la part de la surface du polygone recouvrant cette unité géographique) pour les polygones à cheval sur plusieurs unités (selon l'équation 1).

$$N_c = \sum_{V_j} \frac{A_{V_j \cap C}}{A_{V_j}} N_{V_j} \quad (1)$$

Où N_C représente le nombre de résidents estimé dans l'unité géographique C, N_{V_j} le nombre de résidents détectés dans le polygone de Voronoï V_j, A_{V_j} l'aire de ce polygone de Voronoï, et $A_{V_j \cap C}$ l'aire de l'intersection entre l'unité géographique et le polygone de Voronoï.

Il s'agit donc d'une approximation, qui repose sur l'hypothèse que la densité de population présente est homogène sur l'ensemble du polygone. Cette hypothèse est évidemment contestable, en particulier dans les zones rurales où les habitations ne sont pas réparties de manière régulière (alors même que les antennes y sont moins nombreuses avec comme résultat des « mailles » plus grandes). Pour évaluer l'ampleur des approximations induites, nous reproduisons ces différentes étapes de modélisation sur des données classiques de statistique publique exhaustives et géolocalisées : les fichiers fiscaux.

Simuler la démarche sur les données fiscales pour évaluer l'ampleur de l'approximation

L'Insee dispose d'informations exhaustives à l'échelle du territoire sur la population résidente. Le Fichier Localisé Social et Fiscal (Filosofi), qui remplace et complète les fichiers Revenus fiscaux localisés (RFL) est constitué à partir des fichiers exhaustifs des déclarations de revenus des personnes physiques et de la taxe d'habitation. Ces informations sont disponibles à un niveau encore plus fin que les données de téléphonie mobile, puisqu'elles sont géolocalisées[10]. En revanche, la précision temporelle est bien moindre puisqu'elles sont produites annuellement. Par ailleurs, ces fichiers fiscaux ne renseignent que sur la résidence des personnes, et non sur leur présence effective dans certains lieux (qui peut varier au

9. https://www.insee.fr/fr/metadonnees/definition/c1523
10. https://www.insee.fr/fr/statistiques/fichier/2520034/donnee-carroyees-documentation-generale.pdf

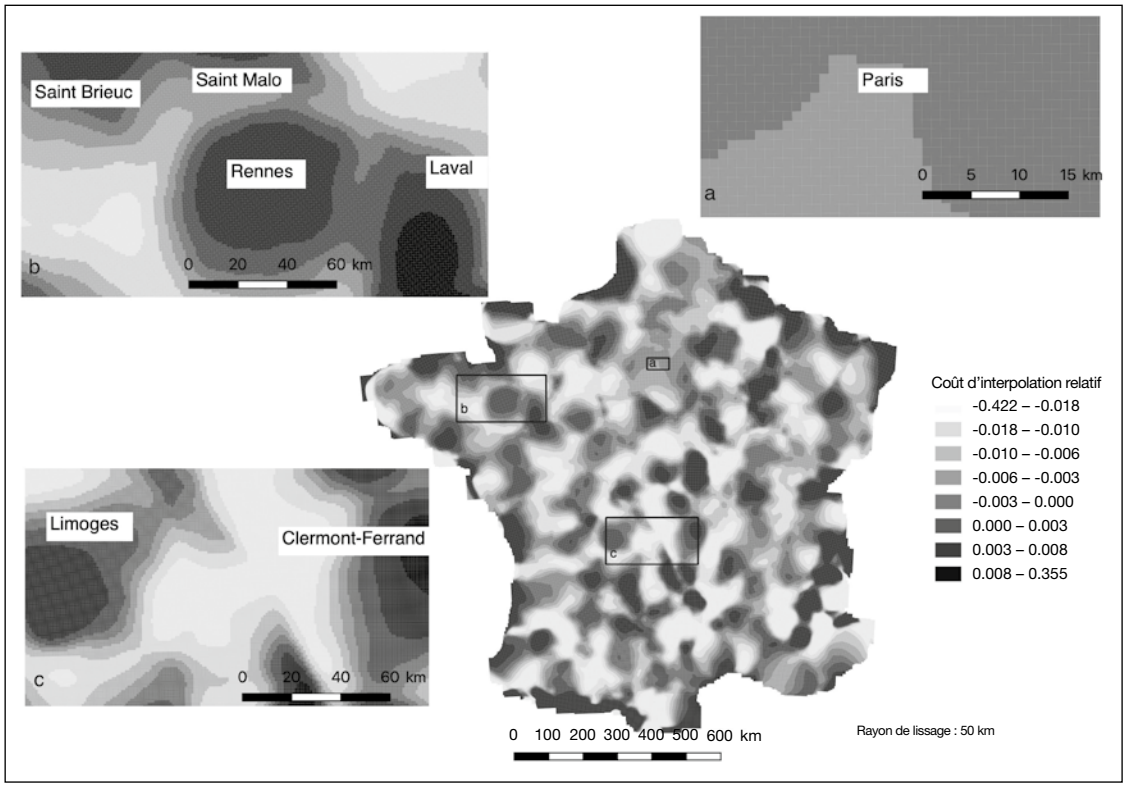
cours de la journée). Elles peuvent néanmoins constituer une source intéressante de comparaison pour évaluer la pertinence des données de téléphonie mobile pour reconstruire des indicateurs statistiques classiques comme les densités de population.

À partir de ces données fiscales géolocalisées, nous estimons la répartition spatiale du nombre d'habitants aux niveaux des communes et des polygones de Voronoï. Nous interpolons la population communale à partir de cette partition en polygones et la comparons avec l'estimation directe au niveau communal afin de tester la procédure d'interpolation spatiale. On appelle « coût d'interpolation » l'erreur de mesure apportée par le fait d'utiliser une interpolation spatiale pour mesurer des agrégats au niveau de l'unité géographique administrative – plutôt que de la mesurer directement. En pratique, il s'agit de l'écart entre le nombre d'habitants mesuré directement au niveau de l'unité géographique et l'estimation obtenue à partir de l'interpolation spatiale (rapporté à la référence).

La figure III illustre la distribution sur le territoire de l'écart relatif de population communale introduit par l'interpolation spatiale à partir des polygones issus de la grille des antennes. Pour les communes situées en zone rurale, cette interpolation conduit en général à surestimer la population de la commune. L'interpolation spatiale repose en effet sur l'hypothèse que la densité de population est homogène sur l'ensemble d'un polygone. Dans les zones les moins densément peuplées, la grille d'antennes est moins serrée. Les polygones correspondants couvrent donc une plus grande surface, alors même que l'habitat est plus dispersé – ce qui rend l'hypothèse sous-jacente d'autant moins vraisemblable[11]. On retrouve ces écarts

11. Enfin, il faut noter qu'il s'agit ici d'écarts relatifs au nombre d'habitants de la commune – des écarts numériques peuvent être amplifiés pour les très petites communes.

Figure III
Écart relatif entre la population communale et la population estimée par interpolation spatiale (coût d'interpolation)



Lecture : le coût d'interpolation correspond à la différence entre la population communale obtenue directement dans la source fiscale, et celle estimée à partir de l'interpolation spatiale (à partir de l'équation (1)). Un coût d'interpolation négatif correspond à une sur-estimation de la population communale, un coût d'interpolation positif à une sous-estimation.
Source : Filosofi 2011 ; calculs des auteurs.

lorsque l'on estime les effets par taille de commune. Pour les communes de taille inférieure à 10 000 habitants, l'écart relatif lié à l'interpolation spatiale correspond à une surestimation de 53 % en moyenne (voir figure C1 du complément en ligne[12]). À l'inverse, pour les communes de plus de 10 000 habitants, l'interpolation spatiale a plutôt tendance à sous-estimer la population réelle de la commune – les écarts relatifs sont néanmoins plus réduits (sans être jamais négligeables) : ils sont de 10 % en moyenne.

Ces résultats suggèrent que le fait d'utiliser une grille qui ne se superpose pas directement aux découpages « classiques » est loin d'être anodin sur la qualité des estimations produites à partir de ces données. Une solution serait alors de s'abstraire du découpage administratif en considérant comme unité de base les polygones de Voronoï, mais elle a l'inconvénient de reposer sur une grille – celles des antennes – qui n'est ni stable dans le temps, ni homogène dans l'espace. Cette partition de l'espace repose également sur une approximation, qui n'est probablement pas sans conséquence sur la qualité des résultats obtenus : l'ensemble des antennes de la tour permettrait de couvrir uniformément ses alentours. En réalité, les antennes sont directionnelles, et ne couvrent que jusqu'à une certaine distance. Ceci explique d'ailleurs la présence de « zones blanches » cartographiées par l'ARCEP depuis 2017[13]. De plus leurs zones de couverture se superposent très fréquemment au contraire d'une tessellation. Disposer de ces informations sur les capacités techniques des antennes pourrait permettre d'affiner la partition réelle de l'espace correspondant. Par exemple, des travaux exploratoires du Bureau Central de la Statistique néerlandais (CBS) proposent d'utiliser une procédure d'inférence bayésienne pour attribuer un point de l'espace à l'une ou l'autre des antennes à proximité, en fonction de la puissance et de l'orientation de celles-ci[14]. Des travaux à venir pourront mettre en regard le gain obtenu en termes de précision et le coût en termes de complexité. Mais les données que nous utilisons ne contiennent pas les informations techniques nécessaires à cette exploration. Par ailleurs, comme discuté dans la suite, d'autres problèmes sont soulevés par l'utilisation de la téléphonie mobile, qui tiennent à la fois à la définition d'un concept (comment passer de l'enregistrement d'un appel téléphonique dans des données de gestion à un indicateur statistique ?[15]) et à celle de leur traitement statistique (comment obtenir des estimations représentatives de l'ensemble de la population à partir des abonnés d'un seul opérateur de téléphonie mobile ?).

Construire des indicateurs statistiques à partir des données

Caractérisation du domicile : « Dis-moi quand tu téléphones, je dirai où tu habites »

Les données dont on dispose correspondent aux traces laissées par les abonnés lors de leurs déplacements. La récurrence de ces passages signale *a priori* un usage des lieux spécifique à l'abonné. Il est ainsi possible d'inférer le lieu de domicile probable d'un abonné, ou son lieu de travail, ce qui est utile voire indispensable pour construire certains indicateurs statistiques, les indicateurs de temps de trajet domicile/travail ou de fréquentation touristique de certaines régions. Parmi les personnes dont la présence a été captée par des données de téléphonie mobile, il faut pouvoir identifier celles qui ne fréquentent pas le lieu de manière habituelle. Selon la définition « statistique » établie par l'Organisation mondiale du tourisme et la Commission statistique des Nations Unies, le tourisme correspond aux « activités déployées par les personnes au cours de leurs voyages et de leurs séjours dans les lieux situés en dehors de leur environnement habituel pour une période consécutive qui ne dépasse pas une année, à des fins de loisirs, pour affaires et autres motifs ». Si l'environnement habituel peut s'interpréter d'une manière plus ou moins extensive, il comprend *a minima* le domicile et le lieu de travail. Ces informations sont rarement disponibles dans les fichiers anonymisés auxquels les chercheurs ou statisticiens ont accès. Plusieurs algorithmes de détection de résidence ont donc été proposés. Leur principe général est de définir le domicile à partir de critères qui reposent sur la fréquence et/ou les horaires de présence (la nuit en général) dans ce lieu. Vanhoof *et al*. (2018) proposent une

12. Voir lien vers les compléments en ligne à la fin de l'article.

13. Le site https://www.monreseaumobile.fr/ permet d'observer les zones blanches en fonction du réseau et des opérateurs.

14. Ces travaux sont accessibles à travers le package R mobloc disponibles à l'adresse : https://github.com/MobilePhoneESSnetBigData/mobloc ; ils sont également décrits en néerlandais ici : https://circabc.europa.eu/webdav/CircaBC/ESTAT/ESTP/Library/2017%20ESTP%20PROGRAMME/46.%20Advanced%20Big%20Data%20Sources%20-%20Mobile%20phone%20and%20other%20sensors%2C%206%20%E2%80%93%209%20November%202017%20-%20Organiser_%20EXPERTISE%20FRANCE/Mobile_Phone2.pdf

15. Un indicateur statistique est entendu ici comme la quantification d'une réalité sociale (par exemple la population présente), suivant une convention à définir (pour Desrosières, 2008, « quantifier, c'est convenir puis mesurer »).

revue de ces différentes méthodes. On peut en distinguer cinq principales :

- activité maximale : le domicile est le lieu où la majorité des événements (émission, réception d'appels ou de SMS) se sont produits sur la période d'étude ;

- nombre de jours distincts : le domicile est le lieu où des activités ont été enregistrées pendant le plus grand nombre de jours distincts sur la période d'étude ;

- contrainte temporelle : le domicile est le lieu où la majorité des activités entre 19h et 9h ont été enregistrées sur la période d'étude ;

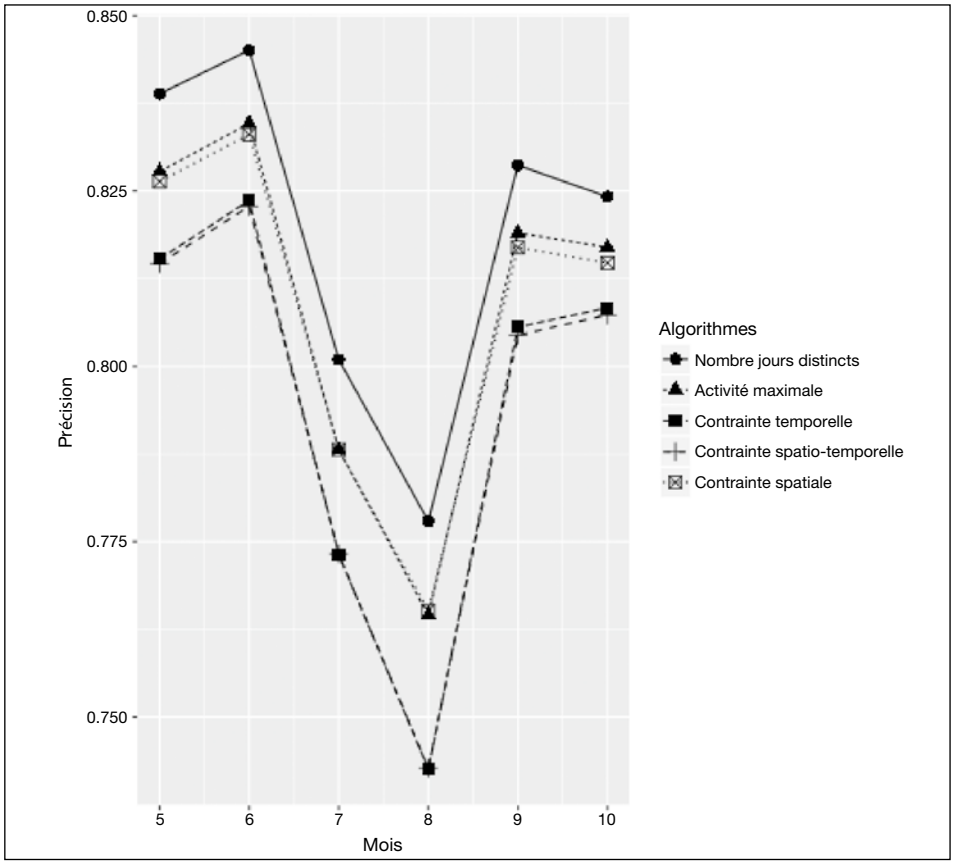
- contrainte spatiale : le domicile est le lieu où la majorité des activités ont été enregistrées dans un rayon de 1 km autour de l'antenne sur la période d'étude ;

- contrainte spatio-temporelle, qui correspond à la combinaison des deux précédentes.

Ces différents algorithmes correspondent tous à des intuitions raisonnables. Néanmoins, ils ont également tous leurs limites, et pour chacun d'eux, on peut aussi aisément penser à des situations où l'identification du domicile serait imparfaite. Pour un même abonné, des méthodes différentes peuvent identifier des lieux distincts comme domicile probable.

Pour évaluer la performance de ces différents algorithmes, nous disposons d'une information supplémentaire fournie par le fichier clients qui contient le code postal de la résidence de l'abonné. Cette information n'est disponible que pour les deux tiers des observations, mais il est néanmoins possible de rapprocher ce code postal de l'estimation du domicile fournie par les différents algorithmes. Par ailleurs, nous disposons également des données sur la population résidente fournies par les fichiers fiscaux localisés.

Figure IV
Précision au niveau départemental des algorithmes de détection de domicile d'après le fichier client



Note : la précision correspond à la proportion d'abonnés présents dans le fichier client pour lesquels l'algorithme de localisation détermine le même département que celui du fichier client.

La figure IV présente une comparaison de la précision des 5 algorithmes de détection de résidence proposés, estimée à partir des informations du fichier client. Les estimations sont menées pour chaque mois, et au niveau du département. La précision correspond à la proportion d'abonnés présents dans le fichier client pour lesquels on a correctement identifié le département dans lequel il réside (au sens où il correspond à celui du fichier client). Sur l'ensemble de la période d'étude, c'est l'algorithme correspondant au nombre de jours distincts (i.e. le lieu où des activités ont été enregistrées pendant le plus grand nombre de jours distincts) qui donne les meilleures performances. Même à ce niveau territorial assez agrégé[16], on observe que l'écart entre le département de résidence tel qu'identifié par les algorithmes d'une part et tel que déclaré dans le fichier client d'autre part reste élevé (il n'est jamais inférieur à 15 %). Ces divergences peuvent s'expliquer par la difficulté de ces méthodes heuristiques à identifier le domicile dans certains cas : par exemple, la baisse de la précision les mois d'été est notable et peut vraisemblablement s'expliquer par le fait qu'une part importante de la population est alors en congés et ne réside pas tout le mois dans son département habituel. Cet écart peut aussi être lié à un problème de qualité du fichier client. Même en négligeant les mois d'été, on observe une baisse de la précision sur l'ensemble de la période pour tous les algorithmes (les écarts observés en septembre-octobre sont plus élevés que ceux observés en mai-juin), ce qui peut être dû en partie à un effet de vieillissement du fichier client (par exemple un défaut de mise à jour en cas de déménagement). De plus, les données ne contiennent des enregistrements que pour la fin du mois de mai (18 jours) et le début du mois d'octobre (14 jours), ce qui peut également expliquer une moins bonne performance qu'en juin et septembre respectivement.

Il convient toutefois de noter qu'un usager est considéré comme détecté dans son département de résidence si l'antenne qui lui est attribuée par l'algorithme de détection de résidence est bien dans le département renseigné dans le fichier client. Il peut donc marginalement y avoir des effets de bord pour les antennes correspondant à des cellules de Voronoï à cheval sur plusieurs départements et donc des clients considérés comme détectés en dehors de leur département.

On pourra trouver en annexe des cartes représentant la répartition géographique de ces précisions pour les mois de juin et août.

Redresser les données pour obtenir des estimateurs de population résidente

Les données de téléphonie mobile dont nous disposons ne correspondent qu'aux abonnés d'un unique opérateur, qui ne représente qu'une fraction de la population des abonnés. Pour estimer des statistiques en niveau (par exemple, le nombre de personnes présentes en un lieu), il est donc nécessaire d'opérer des redressements.

Ces redressements doivent permettre de passer de la population des abonnés à la population totale, qui peuvent différer pour deux raisons. La première est que l'opérateur ne couvre qu'une partie des abonnés de téléphonie mobile. La part de marché de cet opérateur fournit une indication de l'ordre de grandeur de l'écart relatif qu'on s'attend à trouver entre la population réelle et les estimations « brutes » obtenues avec des données de téléphonie mobile. D'après l'Autorité de régulation des communications électroniques et des postes (ARCEP), la part de marché de l'opérateur Orange au niveau national était en 2007 de 46.7 %[17].

La seconde raison est qu'il n'y a pas de correspondance simple entre la population des personnes physiques et celle des cartes SIM. Toutes les personnes physiques ne possèdent pas de téléphone (comme les très jeunes enfants), et à l'inverse certaines en possèdent plusieurs (pour des raisons professionnelles en particulier). Il faut donc tenir compte du taux de pénétration, c'est-à-dire le ratio du nombre de téléphones sur la population de référence (la population au 1er janvier de l'année $N-1$ publiée par l'Insee). En 2007, par exemple, le nombre de téléphone portable par habitants estimé par l'ARCEP était de 85.6 % sur l'ensemble du territoire métropolitain. Il était de 81.6 % pour la région Rhône-Alpes mais de seulement 66.0 % en Franche-Comté. Dans deux régions, l'Île-de-France et la région PACA, ces taux étaient même supérieurs à 100 (respectivement de 122.3 % et 104.3 %)[18]. Une partie de ces écarts peut être mise en lien avec les caractéristiques des populations. Le

16. Le niveau plus agrégé est a priori *moins intéressant, l'intérêt éveillé par les sources issues de la téléphonie mobile étant justement d'obtenir des estimateurs avec une granularité spatiale fine.*
17. Voir l'Avis n° 07-0706 de l'ARCEP en date du 6 septembre 2007, https://www.arcep.fr/uploads/tx_gsavis/07-0706.pdf
18. ARCEP, Le Suivi des Indicateurs Mobiles – les chiffres au 31 décembre 2007 : https://www.arcep.fr/index.php?id=9545 « Répartition géographique des clients métropolitains ».

baromètre numérique du CREDOC montre par exemple de fortes disparités selon l'âge en 2007 : la presque totalité des 18-24 ans était équipée d'un téléphone alors que ce n'était le cas que d'un tiers des plus de 70 ans[19].

Formellement, le passage du nombre d'abonnés N_{HD_i} identifiés comme résidant dans une unité spatiale donnée *i* (à partir de l'algorithme de détection de résidence – HD pour *home detection* – correspondant au nombre de jours distincts, le plus efficace d'après les résultats plus haut) à la population résidente dans cette unité est fournie par l'opération comptable suivante :

$$\widehat{N_i} == \tau_i^{-1} \cdot \alpha_i^{-1} \cdot N_{HD_i} \quad (2)$$

où α correspond à la part de marché locale de l'opérateur Orange, τ au taux de pénétration. Ces deux paramètres sont susceptibles de varier sur l'ensemble du territoire, à la fois pour des questions de couverture mais aussi de composition de la population résidente. Pour obtenir des estimations finement localisées on souhaiterait donc disposer d'information précise sur les variables correspondant au redressement (au moins la part de l'opérateur et le taux de pénétration) à des niveaux géographiques fins. Cependant, ces dernières sont en général disponibles à un niveau assez agrégé (national ou régional). Les utiliser uniformément sur l'ensemble du territoire expose au risque de ne pouvoir distinguer entre des écarts réels de population et des couvertures (ou des parts de marchés) différentes entre ces unités.

Pour quantifier l'importance de ces différents effets, nous estimons, à partir des données de téléphonie mobile, des densités de population résidente – qui peuvent donc être comparées à celles observées à partir de la source fiscale – en utilisant des redressements s'appuyant sur un ensemble croissant d'information annexe. L'estimation « brute » consiste à simplement corriger d'un effet taille – en utilisant le ratio du nombre d'abonnés disponibles dans le fichier par la taille de la population résidente en France métropolitaine (soit 18 millions pour une population totale métropolitaine d'environ 62 millions en 2007). Cette estimation très frustre peut être affinée en utilisant le fait que nous disposons ici d'une information supplémentaire et rare correspondant au fichier des clients. Ce dernier permet d'avoir une estimation de la répartition territoriale des abonnés. En pratique, on utilise ce fichier pour reconstruire des redressements au niveau départemental. Ce niveau géographique apparaît à la fois suffisamment large pour réduire les problèmes d'approximation spatiale qui se posent à partir de l'utilisation de la grille fournie par les polygones de Voronoï, et suffisamment fin pour qu'on puisse négliger l'hétérogénéité spatiale des parts de marché de l'opérateur étudié et du taux de pénétration de la population. Le nombre d'abonnés résidant dans le département *k* est estimé à partir des adresses disponibles dans ce fichier. Ces dernières n'étant disponibles que pour une partie du fichier de cartes SIM dont nous disposons, nous redressons par la taille de ces fichiers (ce qui revient à supposer que le défaut de couverture du fichier client est homogène sur l'ensemble du territoire). La part de marché départementale correspond alors simplement au ratio de cette estimation du nombre d'abonnés résidant dans le département sur le nombre total d'habitants de ce département fourni par les sources fiscales.

$$\alpha_k \tau_k = \frac{Tot_{HD}}{Tot_{CRM}} \cdot \frac{N_{CRM_k}}{N_{Insee_k}} \quad (3)$$

Où *k* représente l'indice du département. Deville *et al*. (2014) proposent d'estimer les densités de population communales à partir de données mobiles équivalentes et d'un modèle prenant en compte « l'effet superlinéaire des zones densément peuplées sur les activités humaines ». Nous reprenons donc cette méthode à titre de comparaison avec les différents redressements que nous proposons[20].

La population est alors estimée par le modèle :

$$N_{Insee_C} = \alpha \cdot N_{HD_C}^{\beta} \quad (4)$$

où les paramètres α et β sont eux-mêmes estimés par régression linéaire généralisée et avec N_{Insee_C} le nombre de résidents d'après la source fiscale dans la commune et N_{HD_C} le nombre de personnes repérées comme résidentes dans la commune avec les données mobiles.

Ces redressements sont appliqués aux estimations obtenues pour la population résidente que nous pouvons comparer avec les statistiques fournies par les données fiscales agrégées à

19. Baromètre du numérique 2015 disponible à https://www.arcep.fr/uploads/tx_gspublication/CREDOC-Rapport-enquete-diffusion-TIC-France_CGE-ARCEP_nov2015.pdf (tableau 2, p. 24).

20. Le modèle proposé par Deville et al. *(2014) porte sur les densités de population. Le modèle est estimé par moindres carrés pondérés par la population des communes sur les logarithmes des densités. L'intérêt de la statistique publique est davantage tourné vers des comptages de population. Nous privilégions donc un modèle plus adapté aux comptages et nous estimons les paramètres par régression linéaire généralisée qui repose sur une famille de Poisson (équation 4), sur lequel est appliquée une fonction de lien logarithmique.*

ENCADRÉ 3 – **Similarité cosinus et coefficient de corrélation empirique**

Pour chaque niveau géographique, on peut définir les vecteurs des observations à partir des données issues de la source fiscale ($\vec{x}$) et des données de téléphonie mobile ($\vec{y}$). On appelle alors coefficient de corrélation empirique :

$$cor\left(\vec{x},\vec{y}\right)=\frac{\overrightarrow{(x-\bar{x})\cdot(y-\bar{y})}}{\left\|(x-\bar{x})\right\|\cdot\left\|(y-\bar{y})\right\|} \quad (5)$$

Où $\bar{x}$ et $\bar{y}$ correspondent aux moyennes empiriques sur l'échantillon. Il est aussi standard d'utiliser la similarité cosinus, qui permet de mesurer si ces deux vecteurs sont proches. Formellement, il s'agit du produit scalaire rapporté au produit des normes des deux vecteurs.

$$\mathrm{cosim}\left(\vec{x},\vec{y}\right)=\cos(\theta)=\frac{\vec{x}\cdot\vec{y}}{\left\|\vec{x}\right\|\cdot\left\|\vec{y}\right\|} \quad (6)$$

Cette mesure est donc indépendante de la norme de chaque vecteur. Elle est *a priori* plus indiquée pour mesurer des densités, tandis que le coefficient de corrélation renseigne plutôt sur les divergences en niveau.

des échelles spatiales plus ou moins fines. Ces redressements, certes frustres, permettent d'estimer les ordres de grandeur, et la répartition territoriale, des écarts entre les populations légales et celles qui peuvent être estimées par les données de téléphonie mobile.

Concrètement, nous comparons ici les comptages obtenus à partir de la source fiscale, qui a l'avantage d'être géolocalisée et que nous pouvons donc mobiliser à des échelles spatiales plus ou moins fines, avec ceux obtenus avec les données de téléphonie mobile pour un concept proche de population résidente. La corrélation des estimateurs fournies par ces deux sources est mesurée par le biais de deux indicateurs : la similarité cosinus et le coefficient de corrélation empirique (encadré 3). Ces indicateurs sont tous les deux indépendants de la taille de la population concernée. Il s'agit donc de vérifier que les estimations fournies par la source de téléphonie mobile donnent des densités de population résidente cohérentes avec celles données par la source fiscale. L'objectif *in fine* est de comparer les estimations de nombre de personnes en niveau. Cependant on peut simplement évaluer un ordre de grandeur des erreurs obtenues en utilisant les enregistrements de téléphonie mobile pour reconstituer le nombre de résidents d'une unité géographique.

Nous avons mesuré les écarts à plusieurs niveaux de granularité. En premier lieu, au niveau des polygones de Voronoï, qui est l'échelle spatiale la plus fine disponible avec les données de téléphonie mobile. Comme discuté plus haut, le découpage du territoire auquel il correspond ne se superpose pas naturellement à des découpages statistiques ou administratifs. On utilise donc également les découpages en IRIS (premier niveau infra-communal), en communes, en zones d'emploi puis en départements. La figure V représente la corrélation et la similarité cosinus (qui est indépendante de la taille des unités initiales mais compare la cohérence globale des estimations) entre l'estimation de population et la population issue des données fiscales géo-référencées, pour chacun de ces niveaux de granularité. Notons qu'il y a ici deux raisons de trouver des différences entre les résultats fournis par les deux sources. D'une part, les concepts de mesure de la résidence ne sont pas les mêmes (dans un cas, l'information est directement issue de la déclaration de résidence fiscale, dans l'autre elle n'est obtenue que de manière très indirecte à partir des comportements d'appels de l'abonné). D'autre part, l'une des sources est exhaustive quand l'autre nécessite des redressements – sachant que le nombre d'informations auxiliaires permettant ces redressements sont faibles.

Les résultats font apparaître des divergences importantes au niveau des estimations obtenues à des niveaux très fins : les divergences les plus grandes sont observées au niveau Iris – la corrélation empirique étant de 0.61. Au niveau des polygones de Voronoï, les observations sont plus proches (par rapport à la maille Iris, le fait de ne pas recourir à une interpolation enlève une source d'écart).

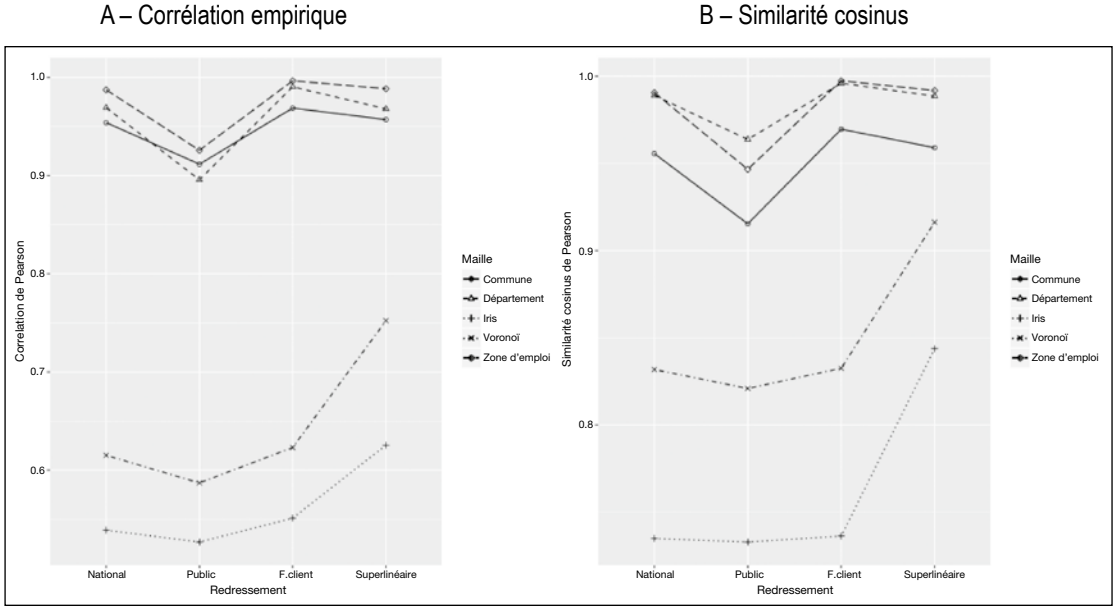
L'écart est plus faible aux niveaux plus agrégés. Il correspond essentiellement à la précision de l'algorithme de résidence, qui peut varier sur les différents départements (en particulier parce que la répartition des antennes n'est pas homogène sur le territoire). Les redressements s'appuient justement sur les données fournies par la source fiscale au niveau du département, et il n'est pas surprenant que les estimations obtenues soient très proches. En revanche, il était moins attendu d'observer que la perte de

précision au niveau communal soit faible par rapport au niveau départemental.

Nous avons également testé la qualité de nos estimations sur un zonage statistique *a priori* plus adapté à nos données : le découpage en zones d'emploi. Une zone d'emploi est un espace géographique à l'intérieur duquel la plupart des actifs résident et travaillent (Aliaga, 2015). Ce zonage est construit de façon itérative, avec pour objectif de maximiser le nombre d'actifs qui résident et travaillent sur la zone. En 2010, la France compte 322 zones d'emploi, qui forment une partition complète du territoire et sont de surfaces similaires, intermédiaires entre communes et départements. Les zones d'emploi sont toutes plus ou moins centrées sur une aire urbaine. Ce zonage est adapté à l'étude du marché du travail local. On peut considérer que la plupart des actifs qui résident dans une zone d'emploi effectueront l'ensemble de leurs appels téléphoniques dans cette même zone, du moins durant les jours ouvrés. S'il existe une imprécision sur la localisation précise du domicile d'un individu, il y a néanmoins de fortes chances pour que l'algorithme situe le domicile de l'individu dans la bonne zone d'emploi puisque celle-ci recouvre en principe l'ensemble des déplacements effectués par l'individu. Les zones d'emploi nous semblent donc être une échelle géographique pertinente pour analyser les estimations de population faites à partir des données de téléphonie mobile. C'est bien au niveau de ces zones d'emploi que les estimations sont les plus corrélés avec la population de référence, quel que soit le mode de redressement.

Les figures V-A et V-B permettent aussi de comparer les écarts obtenus selon les informations annexes disponibles : simple ratio du nombre d'abonnés, utilisation des « données publiques » (part de marché nationale de l'opérateur et taux de pénétration régionaux), utilisation du fichier client qui permet de redresser sur la population observée au niveau du département et par l'estimation du modèle superlinéaire proposé par Deville *et al*. (2014). Les meilleures estimations sont obtenues à partir des informations du fichier clients. En revanche, l'utilisation d'information annexe comme les taux de pénétrations a plutôt tendance à détériorer les estimations par rapport à une simple règle de trois sur le volume des abonnés rapportés à la population métropolitaine. L'utilisation des taux de pénétrations régionaux, qui peut masquer des comportements très hétérogènes au niveau infra-communal, apporte ici plus de bruits qu'une amélioration

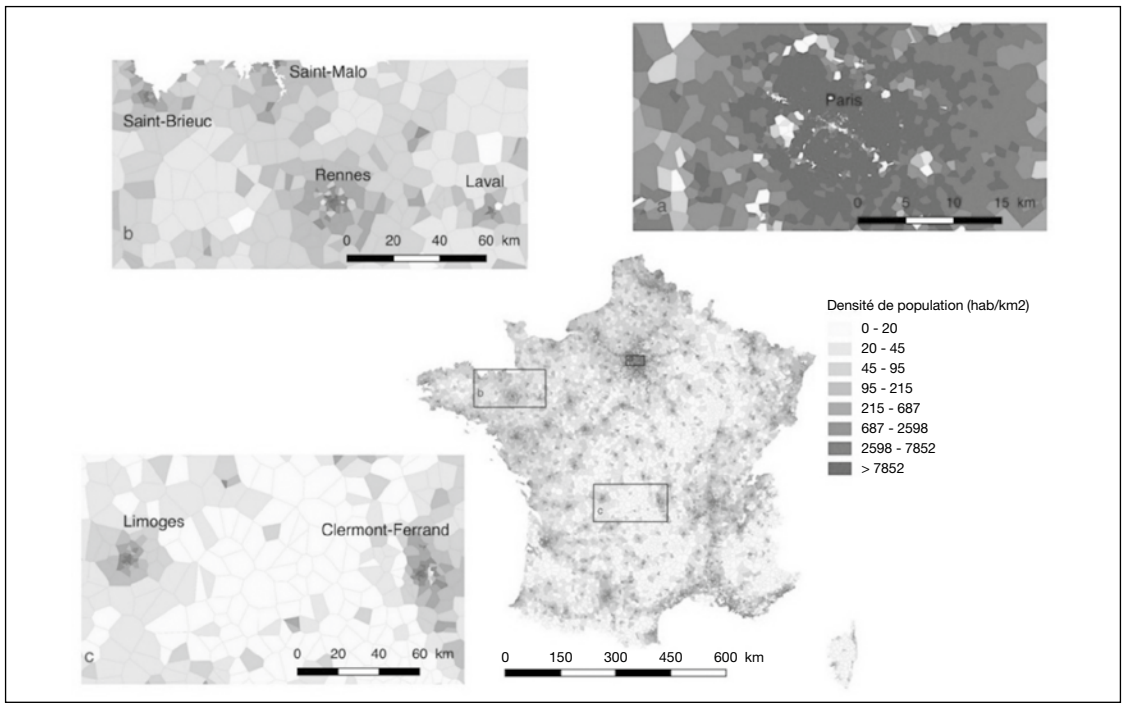
Figure V
Corrélation empirique et similarité cosinus entre les estimations de population résidente et la source fiscale en fonction du mode de redressement et de la maille d'agrégation



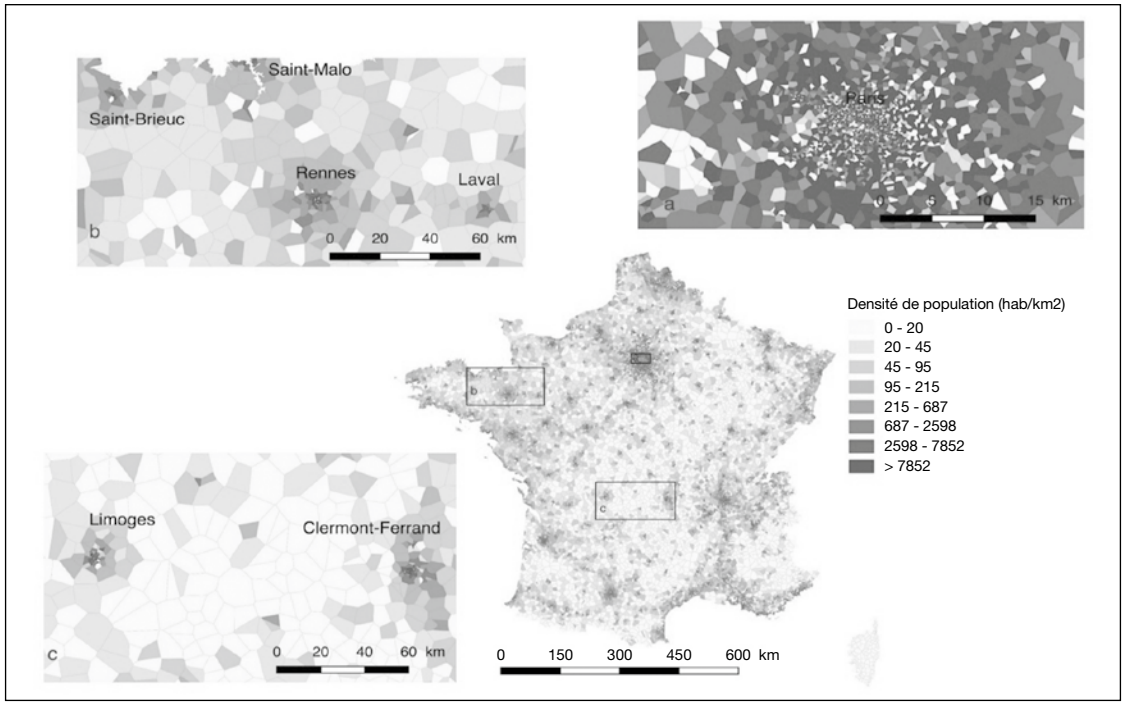
Lecture : au niveau des zones d'emploi, en redressant les estimations à partir du fichier client on trouve une corrélation de 0.99 entre la population estimée à partir des données mobiles et la population résidente fiscale.
Sources : CDR, fichier client pour le redressement « f.client », données Arcep 2007 pour le redressement « public » et Filosofi 2011 ; calculs des auteurs.

Figure VI
Densités de population par polygone de Voronoï calculées à partir des données fiscales (A) et de mobiles (B)

A – Données fiscales



B – Données de mobiles



Note : les estimations sont redressées au niveau départemental à l'aide du fichier client.
Source : A, Filosofi ; B, CDR, fichier client et Filosofi ; calculs des auteurs.

de la précision de l'estimation. Par ailleurs, le modèle superlinéaire estimé au niveau national n'apporte pas de meilleurs résultats au sens de la corrélation empirique ou de la similarité cosinus qu'en redressant par les parts de marché départementales. C'est en prenant en compte une information sur la représentativité des clients de l'opérateur à un niveau géographique intermédiaire (le département) qu'on obtient les meilleurs résultats, même sans tenir compte d'éventuels effets non linéaires mais avec un redressement local simple.

Les figures VI et VII fournissent une représentation cartographique – au-delà d'indicateurs agrégés nationalement – pour comparer les différences entre densités de population estimées par les données fiscales et mobiles (avec le redressement départemental par les parts de marché). D'autre part, la comparaison de ces deux paires de cartes permet d'illustrer combien les estimations au niveau communal sont plus proches de la référence qu'à l'échelle des polygones de Voronoï. Sur les zooms, particulièrement autour de Paris, il est clair que le changement de maille et l'agrégation par commune ou arrondissement apporte une information plus proche des références disponibles.

La carte de la figure VIII présente les écarts relatifs entre les prédictions réalisées au niveau communal et redressées au niveau départemental à l'aide du fichier client avec les populations communales obtenues à partir du fichier fiscal. Les zones où l'écart est le plus élevé correspondent sensiblement aux parties du territoire où la procédure d'interpolation spatiale crée les approximations les plus fortes (comme illustré dans la figure III). On reste donc essentiellement dépendant de la maille que représentent les cellules de Voronoï pour produire une estimation communale. L'imprécision est d'autant plus importante que l'hypothèse d'uniforme répartition de la population dans le Voronoï a moins de chance d'être vérifiée (dans les zones d'habitats non uniformes sur le territoire de la commune). Les écarts entre estimation et références sont parfois très importants. Dans certaines zones, la population de la commune est sous-estimée de près de la moitié la population de la commune tandis que dans d'autres elle est surestimée de plus du double (figure VIII). Ces chiffres recouvrent les estimations de la section « Simuler la démarche sur les données fiscales pour évaluer l'ampleur de l'approximation » sur le coût de l'interpolation dans la source fiscale. Ce résultat est aussi confirmé par une analyse plus systématique des erreurs par une analyse statistique (voir complément en ligne C4).

Les indicateurs tels que le coefficient de corrélation ou la similarité cosinus ne tiennent pas compte de l'organisation spatiale des points mesurés. Il est cependant vraisemblable que les écarts entre les variables observées et prédites soient spatialement corrélés, comme illustré par les figures III et VIII. On peut supposer par exemple des phénomènes de compensation entre des communes proches, qui sont en partie couvertes par les mêmes antennes et donc par les mêmes polygones de Voronoï. Les estimations de population par Voronoï seront réparties entre ces communes, ce qui créera une corrélation entre les valeurs estimées sur ces communes. Par ailleurs, l'erreur liée à l'utilisation d'une interpolation spatiale étant corrélé à la densité

ENCADRÉ 4 – ***I* de Moran**

Les indices d'autocorrélation spatiale permettent de mesurer la dépendance spatiale entre les valeurs d'une même variable en différents endroits de l'espace. Plus les valeurs des observations sont influencées par les valeurs des observations qui leur sont géographiquement proches, plus l'autocorrélation spatiale est élevée.

- L'autocorrélation spatiale est positive lorsque des valeurs similaires de la variable à étudier se regroupent géographiquement.

- L'autocorrélation spatiale est négative lorsque des valeurs dissemblables de la variable à étudier se regroupent géographiquement : des lieux proches sont plus différents que des lieux éloignés.

- En l'absence d'autocorrélation spatiale, on peut considérer que la répartition spatiale des observations est aléatoire.

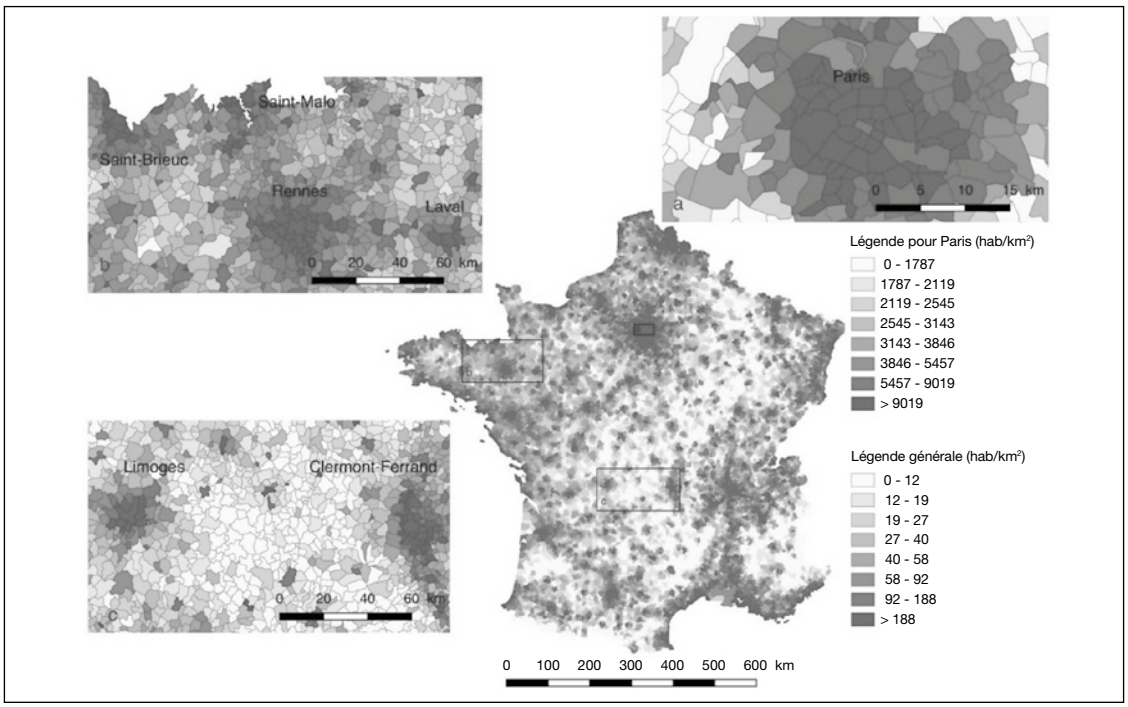
L'indice de Moran permet de comparer la façon dont les observations voisines co-varient, à la covariance de l'ensemble des observations. La notion de voisinage est introduite grâce aux poids w_{ij} qui valent 1 si les observations y_i et y_j sont voisines, et 0 sinon. L'hypothèse nulle est une absence d'autocorrélation spatiale.

$$I_w = \frac{n}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}$$

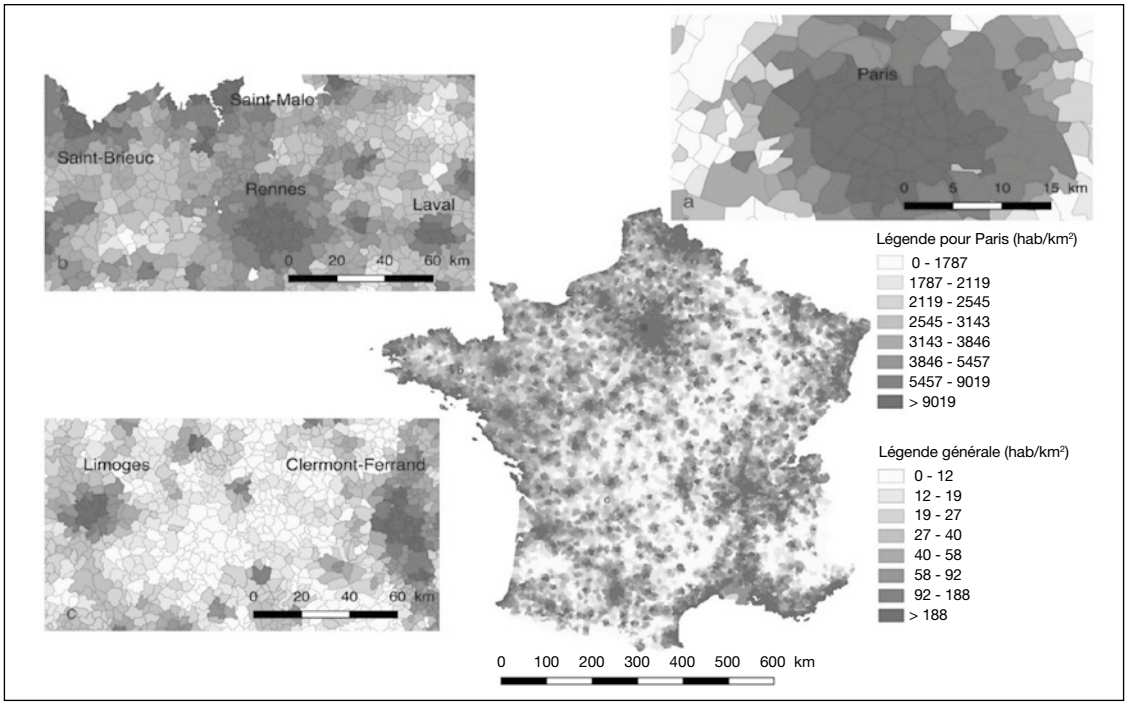
$I_w > 0$ si il y a une autocorrélation spatiale positive

Figure VII
Densités de population par commune calculées à partir des données fiscales et de mobiles

A – Données fiscales

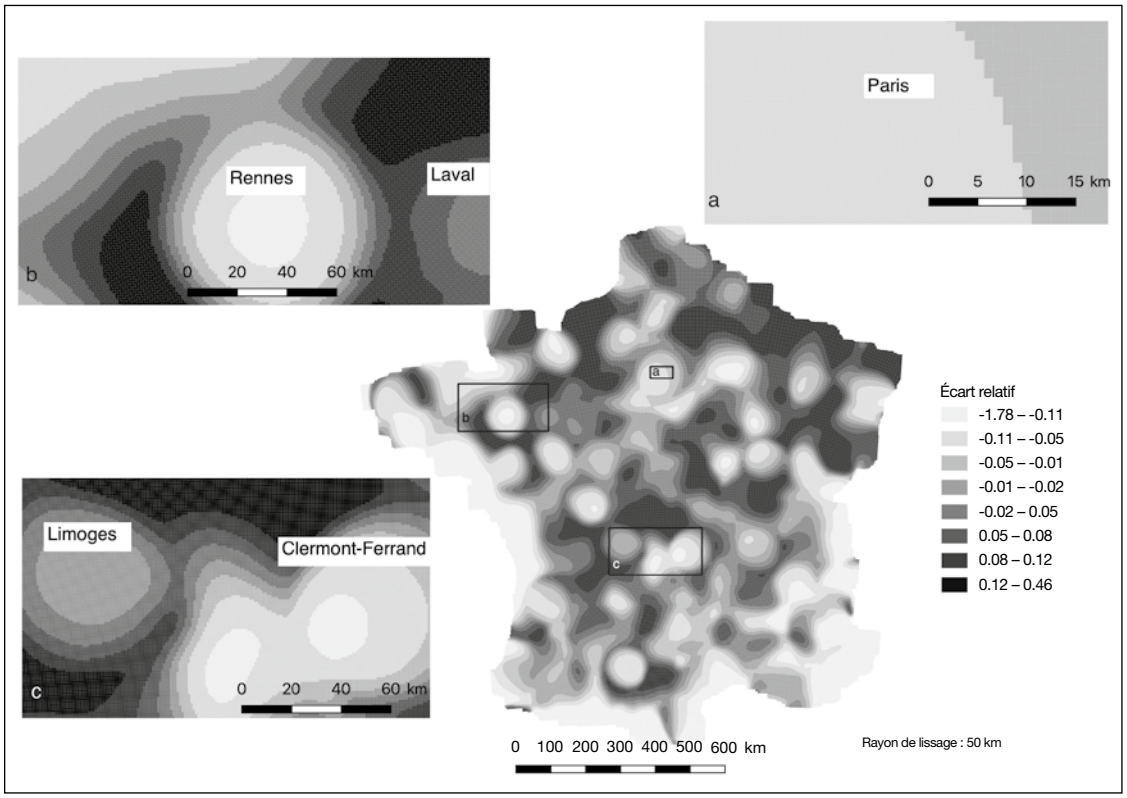


B – Données de mobiles



Note : les estimations sont redressées au niveau départemental à l'aide du fichier client.
Source : A, Filosofi ; B, CDR, fichier client et Filosofi ; calculs des auteurs.

Figure VIII
Carte de l'écart relatif par commune entre l'estimation de population résidente redressée par le fichier client et la source fiscale



Note : les écarts sont lissés spatialement pour la représentation.
Lecture : dans les zones les plus claires la population estimée est surestimée d'un facteur compris entre 0.11 et 1.78, dans les zones les plus foncées elle est sous-estimée d'un facteur compris entre 0.12 et 0.46.
Source : compte-rendu d'appels et fichier client de 2007 d'Orange et Filosofi 2011 ; calculs des auteurs.

de population, il est probable que les écarts pour des communes voisines soient proches. Les indicateurs d'autocorrélation spatiale tels que le *I* de Moran (encadré 4) sont un élément supplémentaire pour illustrer ces phénomènes.

Nous avons calculé la valeur du *I* de Moran pour quatre variables : le coût d'interpolation brut, le coût d'interpolation relatif (par rapport au nombre d'habitants de la commune), l'écart brut et l'écart relatif. Les quatre indices sont significatifs, ce qui confirme que ces variables ne sont pas réparties aléatoirement sur le territoire, et qu'il y a bien un phénomène spatial en jeu.

L'indice d'autocorrélation spatiale de Moran du coût d'interpolation brut est négatif – et non significatif. Ceci s'explique par le fait que lorsque le découpage en polygones de Voronoï conduit à surestimer la population d'une commune, la population des communes voisines est sous-estimée, puisque le total de population est constant. En revanche lorsque le coût d'interpolation est ramené au nombre d'habitants, cet indice devient positif – mais très faible même s'il est significatif (tableau 2). Diviser par la taille de la population estimée lisse en effet les différences puisque les zones surestimées voient leur poids diminuer relativement aux zones sous estimées.

Tableau 1
Autocorrélation spatiale des écarts et du coût de l'interpolation

Variables	Valeur *I* de Moran
Écart brut	0.14***
Écart relatif	0.13***
Coût d'interpolation brut	- 0.11
Coût d'interpolation relatif	0.009***

Note : *, **, *** indiquent la significativité aux seuils de 10, 5 et 1 %.

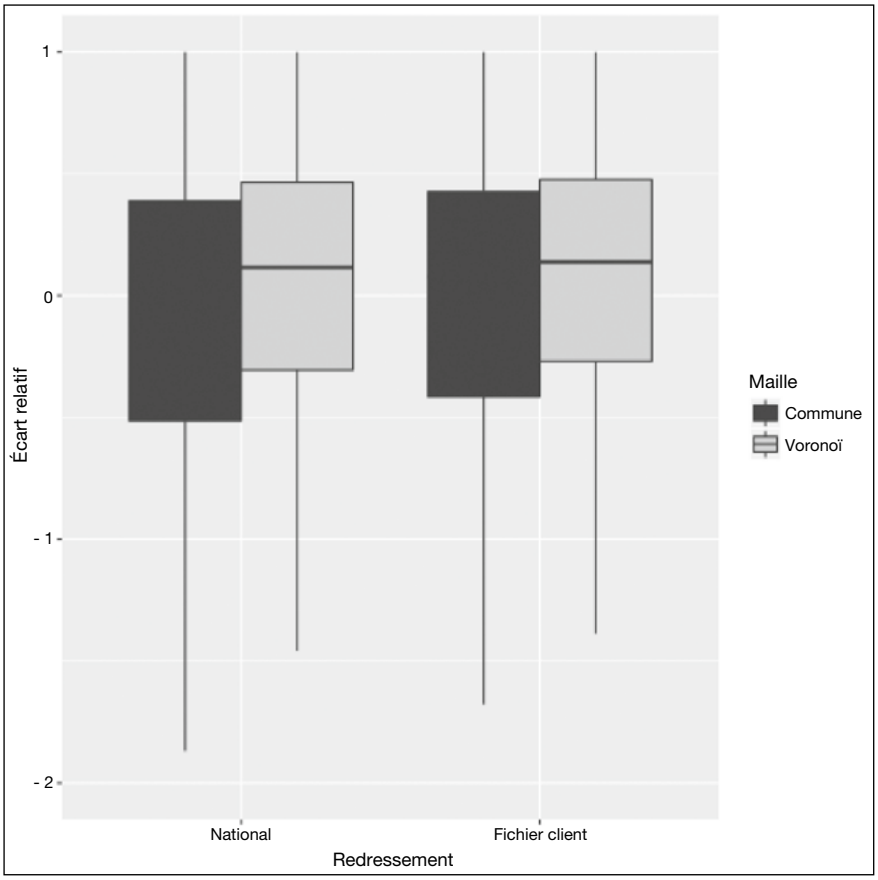
Les écarts bruts et écarts relatifs sont corrélés positivement dans l'espace, signe que certaines zones concentrent de façon significative les communes présentant des écarts plus élevés ou plus faibles que la moyenne.

Enfin, la distribution des écarts de population communale, comme représentée sur la figure VI, est plus resserrée lorsque l'on redresse à l'aide du fichier clients au niveau départemental. Cependant la médiane de cet écart relatif reste plus faible à la fois au niveau de la cellule de Voronoï et de la commune avec le redressement simple et uniforme (figure IX).

Utiliser la granularité temporelle : estimer les variations saisonnières

Un atout important des données de téléphonie mobile, outre la précision spatiale, est de disposer de données répétées avec une forte périodicité. On dispose en effet d'enregistrements en continu sur la présence de personnes utilisant le réseau. Cette dimension était utilisée indirectement dans les estimations précédentes pour identifier la résidence probable des abonnés, mais il s'agissait ensuite d'estimer des grandeurs statiques (la population). Exploiter plus directement les aspects dynamiques peut fournir des informations intéressantes sur la dynamique des territoires, en étudiant par exemple les variations saisonnières de fréquentation. Ces indicateurs pourraient compléter les indicateurs classiques de la statistique publique : ceux-ci renseignent sur les évolutions des populations sur le temps long (fournis par les recensements), ou à un niveau temporel plus fin sur la fréquentation touristique. Les enregistrements de téléphonie mobile peuvent permettre d'identifier, avec une forte précision géographique, les zones sur lesquelles on observe des écarts élevés au cours de l'année. S'intéresser aux variations plutôt qu'au niveau remédie en partie aux fragilités mises en lumière par les analyses précédentes. En particulier, disposer de la variabilité locale

Figure IX
Distribution des écarts entre la population estimée et la source fiscale

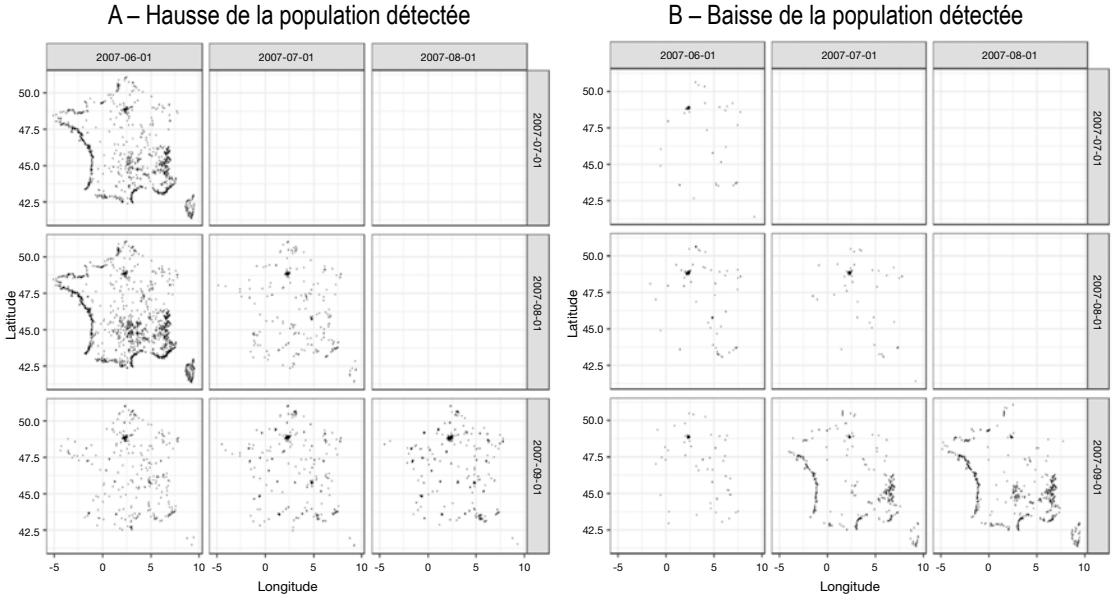


Note : pour une meilleure visualisation les points aberrants ne sont pas représentés. Toutefois ils représentent une part non nulle de la population : pour environ 250 cellules de Voronoï où aucun résident fiscal n'est réputé habiter un total de près de 60 000 usagers ont été estimés y vivre.

des parts de marchés de l'opérateur dont on utilise les données est moins primordial pour estimer des grandeurs relatives qu'en niveau.

À titre d'illustration, on se concentre sur les mois d'été, et on calcule pour chaque mois le nombre de personnes distinctes identifiées dans les enregistrements de l'opérateur sur une zone durant un mois donné, rapporté au nombre de résidents les mois précédents (en utilisant ici aussi l'algorithme le plus efficace lié au nombre de jours distincts de présence sur un mois). On raisonne directement sur la grille fournie par les polygones de Voronoï, pour s'affranchir des difficultés liées à la transposition au découpage administratif présentées plus haut. On dispose donc, pour chaque cellule de Voronoï, de 6 variables correspondant aux ratios pour les mois de juillet, août et septembre, rapportés à l'estimation des résidents pour les mois de juin, juillet et août. Sur l'ensemble des cellules de Voronoï, ces variables ont une distribution qui correspond approximativement à une loi log-normale centrée autour de 1 – qui correspond à une situation où les personnes présentes un mois donné sont identiques à celles identifiées comme résidentes le mois précédent. Ces écarts peuvent cependant être très élevés, ce qui se traduit par une queue très épaisse de la distribution. Pour mettre en évidence la répartition géographique de ces écarts, on représente sur la figure X le logarithme de ces variables, selon les différents mois. Pour mieux faire ressortir les fortes variations, on représente sur des cartes distinctes les zones où les évolutions sont les plus marquées, avec des évolutions de population d'un mois sur l'autre supérieures à 50 % (figure X-A) ou inférieures à 50 % (figure X-B). Les évolutions sont conformes à l'intuition. On observe sur les principales zones marquées par une forte concentration touristique (zones littorales ou de montagne en particulier) de fortes augmentations de population entre juin et juillet puis entre juillet et août, qui se résorbent en septembre pour revenir à une situation similaire à celle avant les deux mois de départ en vacances.

Figure X
Variation de la population présente par mois



Lecture : entre juin et août la population détectée comme habitant autour des antennes a plus que doublé, essentiellement sur le littoral et en montagne (figure partie A). En complément en ligne C4, les points bleus clairs montrent les antennes autour desquelles la population a diminué, de moins de la moitié.
Source : CDR ; calculs des auteurs.

Sur le reste du territoire, les évolutions sont moins marquées – on observe cependant aussi des évolutions saisonnières marquées, avec des flux positifs en dehors des grands pôles urbains au cours des mois d'été, qui s'inversent en septembre.

* *
*

Ces premières analyses suggèrent qu'il serait difficile avec des sources de téléphonie mobile de reproduire des statistiques précises de comptage de la population telles que celles produites par la statistique publique. Ce résultat n'est pas en soi surprenant, compte tenu des différences de concepts entre les deux sources (résidence fiscale déclarée versus résidence reconstituée par des analyses). On peut également mentionner les limites inhérentes au caractère « actif » des données utilisées, les localisations sont fréquentes en moyenne mais pas toujours très régulières. Les données de *signaling*, qui fournissent des informations sur la localisation à une fréquence systématique, peuvent par exemple permettre de mieux identifier les résidences. Même en se limitant aux données de CDR, la généralisation des forfaits illimités sur les textos (encore peu répandus en 2007) a multiplié leur usage – et donc également les possibilités de localiser plus régulièrement les abonnés. Par ailleurs la disponibilité de para-données sur les couvertures des antennes semble cruciale dans la mesure où une large partie des écarts trouvés semble provenir de l'approximation faite par la modélisation des zones de couverture par une tessellation de Voronoï.

Cette évolution rapide des usages liés à la téléphonie mobile pose une question majeure pour l'utilisation de ce type de données par la statistique publique. Les indicateurs produits par la statistique reposent sur des concepts clairs et partagés – une convention de mesure sur la grandeur qu'on souhaite mesurer. Pour les utiliser sur la durée, il est *a priori* nécessaire que des données (et ce à quoi elles correspondent) soient cohérentes temporellement. Une évolution constante des contenus, et des méthodes nécessaires pour les traiter, risque de compliquer l'interprétation des résultats. Il paraît donc encore prématuré de viser la publication d'indicateurs standardisés à partir des données de téléphonie mobile. Par ailleurs, l'utilisation de données d'un seul opérateur pose des questions importantes sur la possibilité d'accéder aux informations nécessaires pour le redressement, en particulier concernant les parts de marché locales, condition nécessaire pour redresser à un niveau fin. Enfin, la couverture inégale du territoire soulève des difficultés à reproduire des analyses précises, sur des maillages qui aient du sens.

Malgré ces limites, les enregistrements issus de la téléphonie mobile fournissent une riche matière première pour des études structurelles, car ils permettent d'éclairer des phénomènes territoriaux, en donnant des informations sur les comportements des individus ou d'autres variables utiles pour l'aménagement territorial. Pucci *et al*. (2015) présentent ainsi un exemple d'utilisation de ce type de données pour décrire les pratiques et usages de l'espace urbain (dans lequel le maillage des antennes de téléphonie mobile est suffisamment serré pour permettre des analyses précises), et Aguilera *et al*. (2014) les utilisent sur des mesures de performance des réseaux de transport urbain (temps de transport, occupation des trains, etc.). On peut supposer que ces variables soient moins sensibles au choix de l'opérateur de téléphonie mobile et donc que les questions de redressement se posent avec moins d'acuité. Galiana *et al*. (2018) s'intéressent quant à eux à l'étude de la ségrégation sociale et spatiale, dans les unités urbaines de Paris, Lyon et Marseille. En identifiant la résidence probable de l'abonné, et en caractérisant le quartier dans lequel il réside en fonction des caractéristiques socio-économiques fournies par l'Insee, on peut calculer des indicateurs de ségrégation sociale, quantifiant la propension des personnes à ne communiquer qu'avec des personnes résidant dans un quartier similaire au sien en termes de niveau de revenu, et à évaluer si ce comportement est plus ou moins marqué selon que l'on réside dans un quartier privilégié ou non. Cette étude propose également de mesurer la ségrégation dans l'espace et son évolution, qui correspond au fait de croiser au cours de la journée ou de la semaine des personnes provenant de quartiers variés, ou au contraire au fait de rester confiné dans un entourage similaire au sien. □

Lien vers les compléments en ligne : https://www.insee.fr/fr/statistiques/fichier/3706213?sommaire=3706255/505-506_Sakarovitch-de-Bellefon-Givord-Vanhoof_complement.pdf

BIBLIOGRAPHIE

Aguiléra, V., Allio, S., Benezech, V., Combes, F. & Milion, C. (2014). Using cell phone data to measure quality of service and passenger flows of Paris transit system. *Transportation Research Part C: Emerging Technologies*, 43(2), 198–211. https://doi.org/10.1016/j.trc.2013.11.007

Ahas, R., Silm, S., Järv, O., Saluveer, E. & Tiru, M. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1), 3–27. https://doi.org/10.1080/10630731003597306

Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre, J.-L, Nurmi, O., Potier, F., Schmücker, D., Sonntag, U. & Tiru, M. (2014). *Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics: Consolidated Report.* Luxembourg: Publications Office of the European Union. https://doi.org/10.2785/55051

Aliaga, C. (2015). Les zonages d'étude de l'Insee: une histoire des zonages supracommunaux définis à des fins statistiques. *Insee Méthodes*, 129. https://www.insee.fr/fr/information/2571258

ARCEP (2008). Le Suivi des Indicateurs Mobiles – les chiffres au 31 décembre 2007. https://archives.arcep.fr/index.php?id=9545&L=1

Blondel, V. D., Decuyper, A. & Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1), 1–55. https://doi.org/10.1140/epjds/s13688-015-0046-0

Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S. & Ratti, C. (2015). Choosing the Right Home Location Definition Method for the given Dataset. In: Liu, T.-Y., Scollon C., Zhu W. (Eds.) *Social Informatics. 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings*, pp. 194–208. Springer International Publishing. https://doi.org/10.1007/978-3-319-27433-1_14

Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., De Meersman, F., Seynaeve, G., Wirthmann, A., Demunter, C., Reis, F. & Reuter, H. I. (2016). Big data et statistiques : un recensement tous les quarts d'heure…, *Carrefour de l'Economie*, 2016/10. https://economie.fgov.be/fr/file/801/download?token=Juj2pHbV

Debusschere, M., Sonck, J. & Skaliotis, M. (2016). Official statistics and mobile network operator partner up in Belgium, *The OECD Statistics Newsletter* N° 65, 11–14. https://issuu.com/oecd-stat-newsletter/docs/oecd-statistics-newsletter-11-2016?e=19272659/40981228

Demissie, M. G., Phithakkitnukoon, S., Sukhvibul, T., Antunes, F., Gomes, R. & Bento, C. (2016). Inferring Passenger Travel Demand to Improve Urban Mobility in Developing Countries Using Cell Phone Data: A Case Study of Senegal. *IEEE Transactions on Intelligent Transportation Systems*, 17(9), 2466–2478. https://doi.org/10.1109/TITS.2016.2521830

Deville, P., Linard, C., Martine, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D. & Tatem, A. J. (2014). Dynamic population mapping using mobile phone data, 111(45), 15888–15893. https://doi.org/10.1073/pnas.1408439111

DGINS (2013). Scheveningen Memorandum on Big Data and Official Statistics. https://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13

Desrosières, A. (2008). *Pour une sociologie historique de la quantification : L'Argument statistique I.* Paris : Presses des Mines. https://doi.org/10.4000/books.pressesmines.901

Galiana, L., Sakarovitch, B. & Smoreda, Z. (2018). *Ségrégation urbaine un éclairage par les données de téléphonie mobile*. Journées de méthodologie statistique de l'Insee, 12-14 juin 2018. http://jms-insee.fr/wp-content/uploads/S25_2_ACTEv2_GALIANA_JMS2018.pdf

Grauwin, S., Szell, M., Sobolevsky, S., Hövel, P., Simini, F., Vanhoof, M., Smoreda, Z., Barabási, A.-L. & Ratti, C. (2017). Identifying and modeling the structural discontinuities of human interactions. *Scientific Reports*, 7. https://doi.org/10.1038/srep46677

Grégoir, S., & Dupont, F. (2016). La réutilisation par le système statistique public des informations des entreprises. *Rapport du groupe de travail Insee-Cnis*. https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2016_143_reutilisation_syst_stat_information_ets.pdf

Isaacman, S., Becker, R., Caceres, R., Kobourov, S., Martonosi, M., Rowland, J. & Varshavsky, A. (2011). Identifying Important Places in People's Lives from Cellular Network Data. In: Lyons, K., Hightower, J. & Huang, E. M. (Eds.), *Pervasive Computing*, vol. 6696, pp. 133–151. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21726-5_9

Janzen, M., Vanhoof, M., Smoreda, Z. & Axhausen, K. W. (2018). Closer to the Total? Long-Distance Travel of French Mobile Phone Users. *Travel Behaviour and Society*, 11, 31–42.
https://doi.org/10.1016/j.tbs.2017.12.001

Léonard, I., Sillard, P., Varlet, G. & Zoyem, J.-P. (2017). Données de caisses et ajustements qualité. Insee, *Document de travail* N° F1704.
https://www.insee.fr/fr/statistiques/fichier/2912650/F1704.pdf

Montjoye, Y. A. (de), Hidalgo, C.A., Verleysen, M. & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Science Report*, 3.
https://doi.org/10.1038/srep01376

Pucci, P., Manfredini, F. & Tagliolato, P. (2015). Mobile Phone Data to Describe Urban Practices: An Overview in the Literature. In: *Mapping Urban Practices Through Mobile Phone Data,* pp. 13–35. Springer, Cham.
https://doi.org/10.1007/978-3-319-14833-5_2

Ricciato, F., Widhalm, P., Craglia, M. & Pantisano, F. (2015). *Estimating Population Density Distribution from Network-based Mobile Phone Data*. Luxembourg: Publications Office.
https://doi.org/10.2788/863905

Ricciato, F., Widhalm, P., Pantisano, F. & Craglia, F. (2017). Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. *Pervasive and Mobile Computing*, 35, pp. 65–82.
https://doi.org/10.1016/j.pmcj.2016.04.009

Scholtus, S. (2015). Aantekeningen over het toewijzingsalgoritme voor Daytime Population. Statistics Netherlands, *Internal CBS note*.

Song, C., Qu, Z., Blumm, N. & Barabasi, A.-L., (2010). Limits of Predictability in Human Mobility. *Science* 327(5968), 1018–1021.
https://doi.org/10.1126/science.1177170

Tennekes, M. (2015). Uitvoering toewijzings algoritme. Statistics Netherlands, *Internal CBS note*.

Tennekes, M. (2019). *R package for mobile location algorithms and tools: MobilePhoneESSnetBigData/mobloc*. R, Mobile Phone ESSnet Big Data.
https://github.com/MobilePhoneESSnetBigData/mobloc (Original work published 2018)

Terrier, C. (2009). Distinguer la population présente de la population résidente. Insee, *Courrier des Statistiques* N° 128, 63–70.
https://www.epsilon.insee.fr/jspui/bitstream/1/8564/1/cs128k.pdf

Toole, J. L., Ulm, M., González, M. C. & Bauer, D. (2012). *Inferring land use from mobile phone activity*. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12* (p. 1). Beijing, China: ACM Press.
https://doi.org/10.1145/2346496.2346498

Vanhoof, M., Combes, S., & de Bellefon, M.-P. (2017). Mining mobile phone data to detect urban areas. In: *Proceedings of the Conference of the Italian Statistical Society*. Florence, Italy: Firenze University Press.
https://eprint.ncl.ac.uk/file_store/production/241585/32829DBE-235C-4902-A175-0A8A0BDCAFD4.pdf

Vanhoof, M., Plotz, T. & Smoreda, Z. (2017). Geographical veracity of indicators derived from mobile phone data. In: *Netmob 2017 Book of abstracts*.
https://arxiv.org/abs/1809.09912

Vanhoof, M., Reis, F., Ploetz, T., & Smoreda, Z. (2018). Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics. *Journal of Official Statistics*, 34(4), 935–960.
https://doi.org/10.2478/jos-2018-0046

Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W., & Buckee, C. O. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of The Royal Society Interface*, 10(81), 20120986–20120986.
https://doi.org/10.1098/rsif.2012.0986