Méthodologie de filtrage et de traitement de données de signalisation de la téléphonie mobile pour la construction de matrice origine-destination

Mariem Fekih, Patrick Bonnel, Zbigniew Smoreda, Tom Bellemans, Angelo Furno, Stéphane Galland

les Cahiers Scientifiques du Transport
N° 75/2019 - Pages 81-111

M. Fekih, P. Bonnel, Zb. Smoreda, T. Bellemans, A. Furno, St. Galland
Méthodologie de filtrage et de traitement de données de signalisation de la téléphonie mobile pour la construction de matrices origine-destination. Application à la Région Rhône-Alpes
JEL : C81, C83, R41, R22

MÉTHODOLOGIE DE FILTRAGE ET DE TRAITEMENT DE DONNÉES DE SIGNALISATION DE LA TÉLÉPHONIE MOBILE POUR LA CONSTRUCTION DE MATRICES ORIGINE-DESTINATION. APPLICATION À LA RÉGION RHÔNE-ALPES

MARIEM FEKIH IMOB, HASSELT UNIV., SENSE, ORANGE LABS
PATRICK BONNEL LAET, ENTPE
ZBIGNIEW SMOREDA SENSE, ORANGE LABS
TOM BELLEMANS IMOB, HASSELT UNIV.
ANGELO FURNO IFSTTAR, ENTPE, LICIT_UMR-T9401
STÉPHANE GALLAND CIAD, UNIV. BOURGOGNE FRANCHE-COMTÉ, UTBM

INTRODUCTION

Le développement d'enquêtes déplacements, comme les enquêtes ménages déplacements, est de plus en plus complexe avec la difficulté de constituer ou d'accéder à des bases de sondage, l'accroissement des aires d'enquêtes qui renchérit les coûts pour des enquêtes en face-à-face ou requiert l'usage

de plusieurs média d'interrogation, la diminution des taux de réponses (Arentze et alii, 2000 ; Atrostic, Burt, 1999 ; Ampt, 1997; Bonnel, 2003 ; Bonnel et alii, 2018b ; Stopher, Greaves, 2007 ; Zmud, 2003). Ces enquêtes sont pourtant historiquement la principale source de données pour comprendre et analyser la mobilité urbaine. Elles sont également utilisées pour construire et estimer les modèles de planification comme les modèles à quatre étapes (Bonnel, 2004 ; Ortuzar, Willumsen, 2011). Leur qualité est donc importante pour que les résultats de modélisation puissent être mobilisés dans les processus d'aides à la décision en termes d'investissement ou de politiques de transport. Si les enquêtes déplacements fournissent des données extrêmement utiles pour formaliser et estimer les modèles de choix de comportement (choix de destination, choix de mode par exemple), elles répondent moins précisément aux besoins de construction de matrices origine-destination (O-D) du fait d'effectifs insuffisants dans de nombreuses cases des matrices. Les enquêtes déplacements peuvent produire également des images tronquées de la mobilité urbaine du fait des déplacements non déclarés (Wolf et alii, 2003).

La disponibilité de grandes masses de données produites de manière automatique et passive comme les données de billettique (Munizagua et alii, 2010 ; Pelletier et alii, 2011 ; Zhao et alii, 2018)), les données de cartes bancaires ou les données de la téléphonie mobile, permet d'identifier la présence des individus dans l'espace et dans le temps. Plus particulièrement, les données de téléphonie mobile contiennent de plus en plus d'informations qui sont géolocalisées et horodatées. Les protocoles d'exploitation des réseaux permettent aussi de réduire l'intervalle entre deux observations en cas d'inactivité des mobiles. Nous avions, lors de précédents travaux, développé une méthode de construction de matrices origine-destination appliquée en Île-de-France (Bonnel et alii, 2015) que nous avions comparées à l'enquête globale transport conduite dans cette même région. Les travaux se sont depuis multipliés comme nous l'illustrons dans la Section 1. Toutefois, peu de travaux abordent la question de la validation des données produites à partir de la téléphonie mobile. Comme dans notre précédent article, nos travaux se situent dans ce cadre, mais avec une nouvelle application au cas de la Région Rhône-Alpes et avec des données de signalisation 2G et 3G. La méthodologie de traitement des données a été enrichie par le filtrage préalable de ces données pour supprimer les données non pertinentes pour la construction des matrices. Nous développons également une méthode d'expansion des données s'appuyant sur l'identification du domicile des possesseurs de mobile.

L'objectif de cet article est de tester cette nouvelle méthodologie pour produire des matrices origine-destination en comparant les résultats avec les données de l'enquête déplacements régionale conduite sur l'ensemble du territoire de la Région Rhône-Alpes.

Nous proposons tout d'abord une revue de la littérature (Section 1) avant de présenter les données utilisées (Section 2), la méthodologie de traitement des données afin de produire les matrices origine-destination (Section 3), permettant la comparaison avec les données externes (Section 4). Enfin, nous proposons les principaux enseignements de cette recherche ainsi que des pistes de développement (Section 5).

1. Revue de la littérature

Après plus de deux décennies d'existence des réseaux téléphoniques cellulaires, le téléphone mobile possède un fort taux de pénétration : 75,8 millions de cartes SIM (hors objets connectés) étaient activées en France début 2019, pour une population de 67 millions d'habitants (Arcep, 2019). Les opérateurs de téléphonie mobile disposent ainsi d'une grande masse de données, associée à la facturation ou à l'exploitation des réseaux. Chaque fois qu'un téléphone mobile est utilisé pour effectuer un appel, envoyer un SMS ou transférer des données, l'opérateur enregistre des données (ou CDR de l'anglais *Call Detail Record*) contenant l'horodatage, l'identifiant anonymisé du mobile, l'identifiant de l'antenne assurant la transmission. Il est ainsi possible de suivre le cheminement spatio-temporel du mobile.

1.1. Exploration des données mobiles pour l'analyse de la mobilité

Ces données sont largement utilisées pour analyser les comportements de mobilité, cartographier la présence humaine ou étudier la diffusion d'épidémies, la densité de population… (Frias-Martinez et alii, 2010 ; Hoteit et alii, 2014 ; Noulas et alii, 2013 ; Picornell et alii, 2015 ; Ratti et alii, 2006 ; Ricciato et alii, 2016 ; Sevtsuk, Ratti, 2010 ; Tizzoni et alii, 2014 ; Yue et alii, 2014). Les données sont également utilisées par les opérateurs pour optimiser le fonctionnement des réseaux de téléphonie (Zhang, Bolot, 2007). Blondel et alii (2015) et Wang et alii (2017) proposent des synthèses bibliographiques des travaux conduits à partir des données de téléphonie mobile.

Le potentiel de ces données massives a rapidement intéressé les chercheurs en mobilité (Asgari et alii, 2013 ; Becker et alii, 2013 ; Calabrese et alii, 2013). Gonzalez et alii (2008) ont été parmi les premiers à étudier la mobilité des utilisateurs à grande échelle, sur un échantillon de plus de 100 000 personnes. Cho et alii (2011) ont associé des données des réseaux sociaux des individus afin d'étudier les liens entre mobilité et réseau social. Ils en concluent que les trajets courts (inférieurs à 100 km) ont dans la plupart des cas une nature périodique, tandis que les trajets longs sont beaucoup plus influencés par le réseau social de l'individu (*i.e.* présence des amis).

La disponibilité de ces données sur de vastes territoires et sur de longues durées permet de conduire des analyses que ne permettent généralement pas les enquêtes classiques. Isaacman et alii (2010 ; 2011) ont ainsi comparé les

villes de New York et Los Angeles, ainsi que les différences saisonnières. La nature des données se prête également à l'analyse de l'attraction touristique temporaire ou permanente (Calabrese et alii, 2010). Afin d'enrichir ces données, certains auteurs ont proposé de croiser les données de téléphonie avec des analyses territoriales. Widhalm et alii (2015) ont ainsi développé une méthodologie permettant de créer une typologie de schéma d'activités qui est croisée avec des typologies spatiales. L'analyse appliquée sur plusieurs terrains comme Vienne en Autriche et Boston aux États-Unis d'Amérique met en évidence des similitudes en termes de profil de schéma d'activités, mais aussi des spécificités locales. Xu et alii (2015) ont étudié l'incidence de l'espace du domicile sur la mobilité. À partir d'un algorithme d'identification du domicile dans les données mobiles et d'une catégorisation spatiale de l'agglomération de Shenzhen en Chine, ils ont construit des profils de dispersion spatiale pour chaque cellule afin de différencier les profils de mobilité en fonction des caractéristiques des espaces d'origine. Huang et alii (2018) ont combiné les données de téléphonie mobile avec des données des réseaux de transport pour fournir des informations en temps réel. Jiang et alii (2017) ont utilisé les données de téléphonie pour modéliser les schémas d'activités à grande échelle dans l'agglomération de Singapour.

Les travaux à partir des données mobiles ont également porté sur la détermination de trajectoires de mobilité. Schlaich et alii (2010) ont ainsi développé une méthode pour déterminer la route utilisée entre les villes de Karlsruhe et Stuttgart en Allemagne. Jiang et alii (2011) approfondissent ces travaux pour affecter les usagers sur le réseau de la ville de Lisbonne. Le choix d'itinéraire entre le RER et le métro sur le réseau de transport collectif francilien a également été étudié à l'aide des données de la téléphonie mobile par Milion (2015) et Aguilera et alii (2014). La prise en compte de la congestion dans les véhicules leur a permis de dériver des valeurs monétaires de l'inconfort avec des montants proches de ceux issus d'une enquête de préférences déclarées.

Les données de la téléphonie mobile peuvent aussi être utilisées pour étudier les vitesses et les temps de parcours moyens. Ygnace (2001) a conduit une des premières études dans le sud de la France sur une autoroute traversant des zones rurales et urbaines. Les résultats apparaissent nettement meilleurs en zone rurale. À l'inverse en zone urbaine, des écarts importants apparaissent lorsque les données sont comparées à d'autres sources. Calabrese et alii (2011 ; 2013) ont étudié la vitesse moyenne et la longueur du trajet moyen dans l'agglomération de Boston. Cette analyse a été conduite tout au long de la journée, permettant de présenter des distributions horaires.

1.2. Exploration des données mobiles pour la construction de matrices origine-destination

Les travaux sur la construction de matrices origine-destination à partir des

données mobiles ont débuté dans les années 2000. Bolla et Davoli (2000) ont été les premiers à développer une application en Italie qui a été testée sur un petit échantillon dans (White, Wells, 2002). Les premiers travaux se sont concentrés sur les déplacements routiers et le plus souvent sur une portion restreinte du réseau routier comme un axe autoroutier ou un secteur bien défini. Akin et Sisiopiku (2002) ont travaillé sur la ville de Birmingham aux États-Unis d'Amérique mais avec un très petit échantillon ne contenant que 500 personnes. Leur algorithme génère des origines et des destinations en segmentant la journée en trois périodes :

. de minuit à 8h, où l'individu se trouve théoriquement chez lui ;
. de 8h à 16h où la personne se trouve théoriquement au travail ;
. de 16h à minuit où l'individu fait théoriquement des activités.

Les travaux se sont progressivement enrichis pour dépasser ces premières méthodes assez frustes et augmenter les tailles d'échantillon retenues. Bar-Gera, en 2007, a pour la première fois utilisé les données sans échantillonnage afin d'estimer le trafic sur une route d'Israël. Les travaux visaient à produire une matrice sur une route de 14 km avec 10 échangeurs tout en fournissant des données de vitesse. Calabrese et alii (2011) ont été les premiers à produire des matrices origine-destination sur un grand territoire dans la région de Boston aux USA. Bekhor et alii (2013) ont élargi l'analyse en travaillant à la construction d'une matrice origine-destination de déplacements de longue distance à l'échelle d'un pays (Israël).

Les premiers travaux ont principalement utilisé les données de facturation (CDR). Le volume de données de chaque mobile est dépendant de son niveau de consommation (appel, SMS, échange de données). Certains utilisateurs sont ainsi souvent exclus des bases de données car leurs données sont trop éparses. Les données de signalisation sont une solution pour réduire ce problème car elles contiennent des données ne dépendant pas du niveau de consommation du mobile comme les données de changement de zones de localisation qui sont automatiquement générées chaque fois qu'un individu change de zones de localisation (une zone de localisation est une zone regroupant un nombre plus ou moins important d'antennes pouvant aller jusqu'à quelques centaines en zone urbaine). Les données LAU (*location area update*) sont générées toutes les 3 heures afin de connaître en permanence la localisation du mobile. Bonnel et alii (2013 ; 2015) ont travaillé avec les données de signalisation 2G pour construire une matrice origine-destination concernant l'ensemble de la Région Île-de-France. Ils sont partis de la définition d'un déplacement qui constitue un mouvement entre deux activités stationnaires. Leur algorithme s'appuie sur la définition d'activités stationnaires à l'échelle de zones de localisation pour identifier des déplacements chaque fois qu'un mobile change de zone de localisation avec deux activités stationnaires à chaque extrémité.

1.3. « Validation » des données issues de la téléphonie mobile

Malgré la quantité de données mobilisées dans la plupart de ces travaux, la question de la représentativité des matrices ainsi produites se pose. Une partie des utilisateurs des réseaux est souvent exclue du fait d'un manque d'information pour les qualifier. Les utilisateurs d'un réseau ne sont pas forcément représentatifs de l'ensemble de la population même pour les opérateurs dominants. Malgré les taux de pénétration de la téléphonie mobile, des individus ne sont pas équipés ou ne se connectent pas tout le temps à leur mobile, d'autres sont multi-équipés... Les choix méthodologiques de traitement des données et d'extraction de matrices influent sur les résultats. L'utilisation de ces données pour décrire la mobilité de la population d'une région ou pour alimenter des modèles de simulation de la demande suppose pourtant que ces données soient représentatives de la population concernée. À notre connaissance, la plupart des travaux sur les données passives de la téléphonie mobile ne comportent pas de phase de confrontation des données produites avec des données externes à des fins de comparaison ou de « validation ».

En Angleterre, White et Wells (2002) ont été les premiers à tester la faisabilité, sur le comté du Kent, de la création d'une matrice origine-destination à partir des données de facturation (CDR). Les résultats obtenus ont été comparés avec une matrice origine-destination réalisée par enquête. Ils concluent que les données de facturation ne sont pas assez précises pour obtenir une matrice origine-destination fiable.

Caceres et alii (2007) ont calculé une matrice origine-destination sur une route d'Espagne reliant les villes de Huelva et de Séville. Quatre origines-destinations ont été considérées pour l'analyse. Les résultats obtenus ont été comparés à des comptages routiers réalisés sur le même territoire. Les différences entre les données de comptage et les données de téléphonie ne dépassent pas 4 %. Les auteurs en concluent que les résultats sont très prometteurs.

Mellegard (2011) a construit une matrice origine-destination sur une partie de la Suède. Il s'est appuyé sur une méthode développée par Kang et alii (2004) qui extrait des matrices à partir de données spatio-temporelles permettant d'identifier des zones de stationnarité et des zones de mouvement. La comparaison des données produites a toutefois été limitée à quelques origines-destinations.

Wang et alii (2012) ont construit des matrices horaires à San Francisco et Boston afin d'analyser le degré de saturation du réseau aux heures de pointe du matin. Les matrices produites ont été comparées en fonction de la quantité de données disponibles pour chaque mobile. Trois groupes ont ainsi été construit isolant ceux générant de grand volume de données et ceux utilisant peu leur mobile et enfin le groupe intermédiaire. Les matrices ont ensuite été

comparées à des données de comptage routier. Les résultats sont présentés comme très satisfaisants.

Calabrese et alii (2013) ont travaillé avec les données de Boston. Deux comparaisons ont été produites. Ils ont tout d'abord confronté le nombre de déplacements obtenus à celui de l'enquête nationale transport. La mobilité est apparue plus importante dans les données de la téléphonie mobile. Les auteurs expliquent les écarts principalement du fait des territoires d'analyse différents des deux sources de données et par une sous-estimation de la mobilité mise en évidence aux États-Unis d'Amérique en comparant les données de l'enquête avec des données GPS (Wolf et alii, 2003). La seconde comparaison concerne les distances parcourues. Ils ont utilisé les données issues des compteurs kilométriques des véhicules qui sont lus annuellement aux États-Unis dans le cadre des contrôles de sécurité des véhicules. Si des différences de niveau importantes sont mises en évidence, les auteurs soulignent que les structures de distance sont assez semblables.

Chen et alii (2014) ont produit un échantillon de données de téléphonie mobile simulées afin de disposer d'une base « exacte » complètement connue. Ils mettent en œuvre ensuite leur algorithme de détermination des lieux de domicile et de travail. Ils obtiennent des résultats tout à fait satisfaisants avec un bon niveau de reproduction des données sources. Les résultats sont en revanche moins bons pour la fréquentation des autres lieux d'activités.

Toole et alii (2015) ont développé et affiné des algorithmes de traitement des données de facturation (CDR). Ces algorithmes sont issus de travaux antérieurs (Jiang et alii, 2013 ; Zheng, Xie, 2011 ; Alexander et alii, 2015 ; Colak et alii, 2015 ; Wang et alii, 2012 ; Iqbal et alii, 2014). Les algorithmes permettent un filtrage des données afin d'extraire certaines données non pertinentes. Ils utilisent également des méthodes de fusion de données afin d'imputer des informations comme l'heure de début des déplacements ou le mode de transport. Les autres sources de données sont le recensement, les enquêtes déplacements, des données de congestion ou de comptages. Ces données externes sont également mobilisées pour déterminer les facteurs d'expansion des données de la téléphonie à l'ensemble de la population. Les données sont comparées aux enquêtes ménages déplacements de Boston et San Francisco (États-Unis d'Amérique), Rio de Janeiro (Brésil) et Lisbonne (Portugal). Les résultats sont présentés comme encourageants, même si des écarts parfois importants apparaissent. Dans la lignée de ces travaux d'autres auteurs proposent de combiner les données de la téléphonie avec d'autres sources de données comme des données GPS issues de taxi ou de voiture particulière (Huang et alii, 2018 ; Widhalm et alii, 2012), des données billettiques (Huang et alii, 2018), des données de modèles (Wismans et alii, 2018).

Dans les travaux cités précédemment de Bonnel et alii (2013 ; 2015), les

auteurs ont comparé les matrices produites aux données de l'enquête ménages déplacements produites également sur l'ensemble de l'Île-de-France (EGT, 2010). Ils ont obtenu des résultats très proches en termes de nombre de déplacements total. Les matrices sont également très proches (0-20 % d'écarts) pour les flux importants, tandis que les écarts peuvent être plus importants (jusque 70 %) pour les flux périphériques moins importants.

Les travaux de recherche sur la construction des matrices origine-destination sont nombreux et les travaux se poursuivent dans de nombreux centres de recherche. Ceux exploitant les données de signalisation sont toutefois moins nombreux que ceux issus des données de facturation malgré l'avantage des premières pour la prise en compte des utilisateurs du réseau générant peu de données de facturation (Fiadino et alii, 2017). Les travaux de comparaison avec des données censées représenter la réalité (« *ground truth* ») sont en revanche moins fréquents. La question de la représentativité statistique des données ainsi produites reste donc encore d'actualité d'autant plus que les données de téléphonie évoluent avec l'évolution des technologies et l'arrivée des nouveaux réseaux 4G, 5G… Les travaux de comparaison s'attachent plus souvent à comparer la structure des matrices que le nombre de déplacements car la détermination des facteurs d'expansion des données mobiles est souvent complexe (Graells-Garrido, Saez-Trumper, 2016 ; Wismans et alii, 2018). Cet article vise à apporter une contribution à l'état de l'art de l'exploitation des données mobiles pour la construction de matrices origine-destination. Il se situe dans la lignée de nos précédents travaux (Bonnel et alii, 2013 ; 2015 ; 2018a), en travaillant avec des données de signalisation collectées en 2017 à partir des réseaux 2G et 3G. Nous proposons des méthodes de filtrages des données afin d'éliminer certaines données provenant notamment d'automates. Nous appliquons également une méthode de détermination du domicile afin de produire des facteurs d'expansion spatialisés en mobilisant les données du recensement. Comme dans les précédents travaux, nous proposons une comparaison avec une enquête déplacements (EDR enquête déplacements régionale de la Région Rhône-Alpes).

2. Données utilisées : téléphonie mobile et enquête ménages déplacements

Nous présentons tout d'abord les données de téléphonie mobile utilisées avant de décrire l'enquête déplacements régionale utilisée comme source externe pour comparer les matrices origine-destination.

2.1. Données de signalisation d'Orange

Le réseau de téléphonie mobile est composé d'antennes qui assurent le fonctionnement de l'ensemble du réseau. Chaque antenne possède sa propre zone de couverture. Dans la pratique, la zone de couverture n'est pas uni-

forme car elle dépend de la densité d'utilisateurs du réseau. Les zones sont ainsi petites dans les zones denses et beaucoup plus étendues dans les zones peu denses. Le contour de la couverture spatiale des antennes n'est pas totalement fixe car elle dépend de l'activité de chacune des antennes, une antenne chargée pouvant se décharger sur ses voisines. Elle dépend également de la topographie et des conditions météorologiques. Par conséquent, la couverture ne peut être connue de manière précise, ce qui conduit le plus souvent à représenter la couverture des antennes sous forme de polygones de Voronoï[1] (Bonnel et alii, 2013). Les antennes sont ensuite regroupées en zones de localisation (LA-*Location Area*) pour les besoins de gestion des réseaux de téléphonie permettant un repérage plus rapide des téléphones en cas de communication (voix, SMS, données). La localisation d'un téléphone au sein d'une LA est connue en permanence, tandis qu'au niveau des antennes la position n'est connue qu'en cas de communication.

Les données de téléphones mobiles se présentent sous deux formes (Smoreda et alii, 2013) :

. Les données de facturation (CDR), qui enregistrent, à chaque fois qu'une personne reçoit ou émet un appel, un SMS, un échange de données ou une connexion internet, l'antenne par laquelle l'information est transmise, ainsi que l'heure correspondante ;

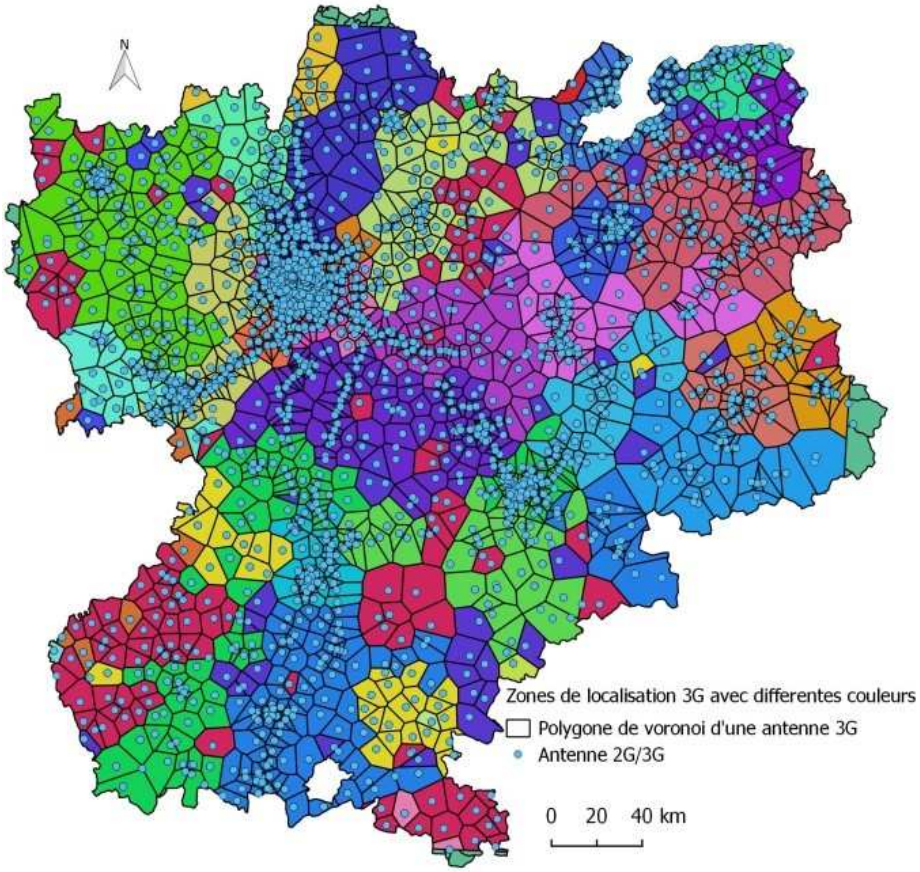
. Les données de signalisation, qui sont l'ensemble des données transitant par les antennes. Elles contiennent, en plus des données de facturation, des *handovers* (changement d'antennes durant la communication), un événement lorsqu'une personne change de LA et un rafraîchissement de la position du téléphone au moins toutes les 3 heures en cas d'inactivité (LAU-*location area update*) et enfin l'information que le téléphone est éteint ou allumé. Ces données sont récoltées grâce à des sondes dans le réseau.

Dans le cadre de cette recherche, Orange a mis à disposition des auteurs des données de signalisation issues du réseau de télécommunication en Rhône-Alpes. Il s'agit de données du réseau GSM (*Global System for Mobile communications*) qui fournit le service 2G et du réseau UMTS (*Universal Mobile Telecommunications System*) qui fournit le service 3G. Ces données concernent 2 millions d'utilisateurs en juin 2017 (à peu près 300 millions d'évènements horodatés et localisés à l'antenne). Les données sont exploitées sur une seule journée car le protocole d'anonymisation des données conduit à recoder toutes les 24h00 l'identifiant SIM. La journée disponible correspond au 1er juin 2017. Plus précisément afin d'harmoniser les données avec celle de l'enquête déplacements régionale, les données sont collectées du 1er juin 4h00 du matin au 2 juin à 4h00 du matin. La Figure 1a présente la couverture des antennes sur la Région Rhône-Alpes et son agrégation en

[1] Les polygones de Voronoï sont créés de sorte que chaque point dans un polygone est plus proche de l'antenne présente dans le polygone que des autres antennes.

zones de localisation dans le réseau 3G.

Figure 1a : Répartition des antennes et agrégation en zone de localisation du réseau 3G

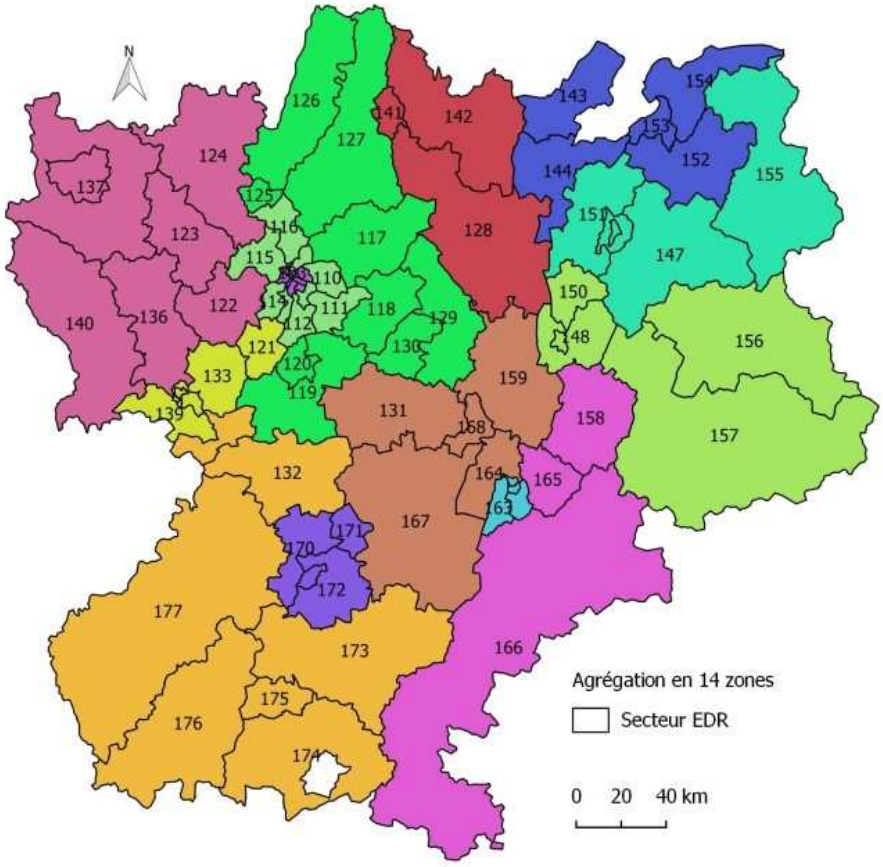


2.2. Données de « validation » : Enquête Déplacements Régionale (EDR)

L'Enquête déplacements régionale (EDR) constitue la principale source de connaissances concernant l'ensemble des déplacements des Rhônalpins depuis 2015. Il s'agit d'une enquête déplacements téléphoniques conduite avec une méthodologie proche du standard CEREMA (CERTU, 2008). Elle a été conduite en trois vagues entre l'automne et le début du printemps durant la période allant de 2012 à 2015. 37 450 individus âgés de 11 ans et plus ont décrit 143 000 déplacements réalisés un jour de semaine. L'enquête a été réalisée auprès de la population résidant en Région Rhône-Alpes. L'échantillon a été construit de manière aléatoire après stratification géographique en 77 secteurs (Figure 1b). La base contient des données sociodémographiques sur le ménage et sur l'individu enquêté qui décrit

l'ensemble des déplacements de la veille du jour d'enquête (de 4h00 du matin à 4h00 du matin le lendemain). Pour chaque déplacement, les principales caractéristiques du déplacement sont disponibles, notamment les heures de début et de fin et la zone d'origine et de destination.

Figure 1b : Agrégation en 14 zones des secteurs de l'EDR au sein de la Région Rhône-Alpes



Les données de l'EDR portent sur la totalité des motifs de déplacements, mais ne concernent que les résidents de la région Rhône-Alpes. Par conséquent, elles ne correspondent donc pas complètement aux données de la téléphonie mobile qui portent sur l'ensemble des personnes utilisant le réseau Orange qui sont présentes sur le territoire de la Région Rhône-Alpes, et cela indépendamment de leur lieu de résidence. Ces différences devront être prises en compte dans la comparaison des matrices origine-destination issues de chacune des bases.

La Région a une population de 5,2 millions d'habitants âgés de 11 ans et plus selon le recensement de 2015 sur un territoire de 43 700 km^2.

3. Construction des matrices origine-destination

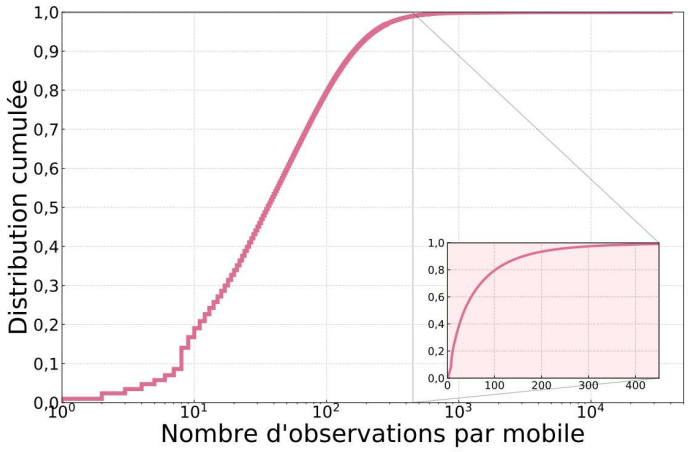
Lors de précédents travaux, nous avons développé une méthode simple de construction de matrice origine-destination à partir des données de signalisation (Bonnel et alii, 2013 ; 2015). Cette méthode a été appliquée sur les données parisiennes afin de comparer les matrices produites avec les données de l'Enquête Globale Transport. Les résultats furent très prometteurs, mais plusieurs problèmes ont été identifiés, notamment au niveau de la méthode d'expansion des données à l'ensemble de la population d'Île-de-France à partir des données Orange. Le présent article vise à répondre à certaines de ces limites. Nous proposons ainsi une méthodologie en 3 étapes :

. Analyse des données mobiles afin de filtrer les données non pertinentes (3.1) ;
. Construction des matrices origine-destination (3.2) ;
. Identification du domicile de la personne utilisant le mobile afin de déterminer un facteur d'expansion à l'échelle du zonage en secteurs de tirage de l'EDR (3.3).

3.1. Filtrage des données à partir de l'analyse des données mobile

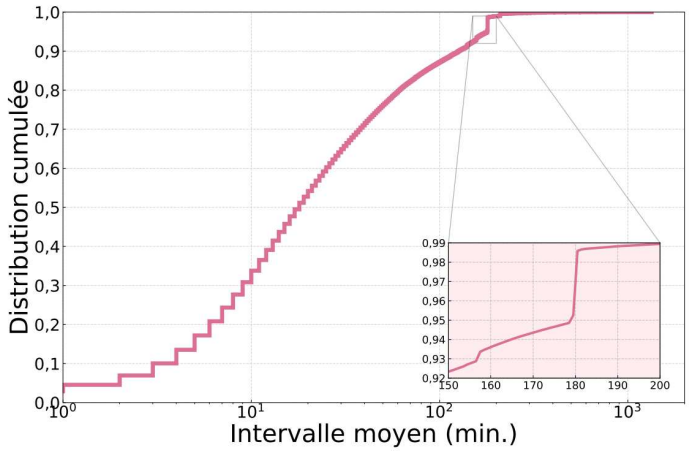
La méthode de construction des déplacements (Bonnel et alii, 2013 ; 2015. Section 3.2.) requiert au moins 4 observations par individu pour identifier un déplacement. De plus, en théorie chaque téléphone doit générer au moins 8 observations (LAU mise à jour de la localisation du téléphone toutes les 3 heures en cas d'inactivité du mobile). Afin de ne pas supprimer les téléphones déconnectés la nuit, nous supprimons seulement les mobiles générant 3 observations ou moins (Figure 2). Au sein de notre échantillon, 1 % des mobiles génèrent plus de 450 évènements. À l'opposé, moins de 5 % des mobiles génèrent 3 observations ou moins.

Figure 2 : Distribution cumulée du nombre d'observations par mobile



La procédure de LAU doit générer des données avec un intervalle de temps inférieur à 3h00. Toute durée supérieure à 3 heures entre deux évènements correspond à une absence du mobile du réseau de téléphonie rhônalpin. Cette absence peut être liée à une sortie du territoire ou à une déconnexion du téléphone pendant une durée supérieure à 3 heures (la connexion/déconnexion génère également une donnée dans la base de signalisation). Il n'est donc pas possible de suivre la localisation du mobile pendant cette absence. Nous décidons de filtrer les téléphones absents du territoire pendant plus de 3 heures à partir de l'analyse de la distribution du temps entre deux observations. La Figure 3a présente le temps moyen entre deux observations pour chaque mobile. Pour 95 % des mobiles, (*cf.* encart Figure 3a) ce temps est strictement inférieur à 3 heures. Seuls à peine plus de 1 % des intervalles sont supérieurs à 3 heures, désignant une absence du mobile pendant une période de temps indéfinie mais supérieure à 3 heures. Afin de ne pas éliminer les téléphones éteints la nuit, nous analysons la distribution du temps maximum entre deux évènements pour chaque mobile pendant la période 7h00 du matin-10h00 du soir qui correspond à la période d'activité pour la majorité des individus (Figure 3b). Un quart des mobiles ont un temps maximum entre deux évènements supérieur à 3 heures. Ces téléphones sont donc filtrés pour être éliminés de la base car nous ne pouvons pas les suivre tout au long de la journée.

Figure 3a : Distribution cumulée de l'intervalle moyen entre deux observations pour chaque mobile



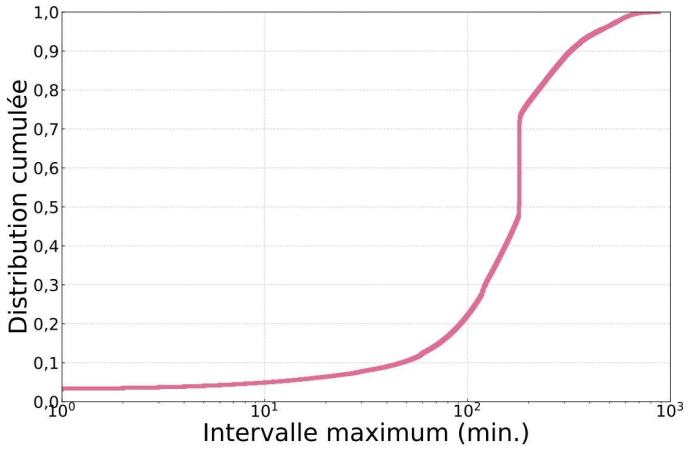
Le développement des réseaux de téléphonie mobile conduit à un usage du réseau de communication qui n'est pas limitée à l'utilisation des mobiles. De ce fait, les données collectées contiennent également des transactions générées par des machines (notamment internet des objets) qui correspondent à de gros volumes de données. Il est donc nécessaire de sélectionner uniquement les données utiles pour générer les déplacements (Wang, Chen, 2018).

La Figure 3b permet de constater que moins de 1 % des mobiles ont une durée maximum entre deux évènements inférieure à 1 minute voire quelques secondes. Il ne s'agit très probablement pas de téléphone avec ce rythme d'activité tout au long de la journée. Nous proposons d'utiliser la notion d'entropie pour identifier ces terminaux et les filtrer :

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log(p(x_i)) \quad (1)$$

avec X la distribution du nombre d'évènements au cours des 24 heures d'un terminal et $p(x_i)$ la proportion des évènements dans l'intervalle de temps d'une heure x_i. Les valeurs d'entropie ont été normalisées pour être incluses dans l'intervalle (0-1).

Figure 3b : Distribution cumulée de l'intervalle de temps maximum entre deux évènements (7h00-22h00) pour chaque mobile



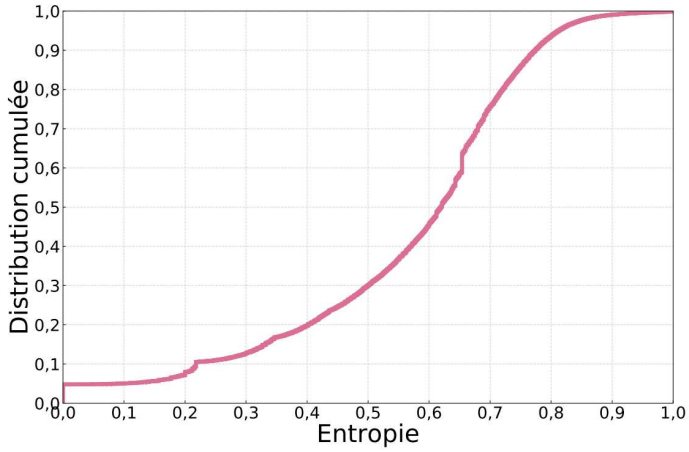
L'entropie permet de mesurer l'uniformité de la distribution des évènements associés à chaque terminal. La Figure 4 montre que seul un très petit nombre de terminaux (moins de 1 %) ont une valeur d'entropie supérieur à 0,9 ce qui correspond à un intervalle de temps entre chaque évènement extrêmement régulier et donc très probablement à une machine. Nous filtrons les terminaux avec une entropie supérieure à 0,9.

3.2. Construction des matrices origine-destination

La définition d'un déplacement dans l'EDR est fournie par le CEREMA (Certu, 2008) : un « *déplacement est le mouvement d'une personne, effectué pour un certain motif, sur une voie publique, entre une origine et une destination, selon une heure de départ et une heure d'arrivée à l'aide d'un ou plusieurs moyens de transports* ». Le déplacement est défini par une origine et une destination qui correspondent à un motif et par conséquent à une activité stationnaire pour reprendre la définition du CEREMA. Il est donc

indispensable de pouvoir identifier deux évènements successifs dans la même zone avec un temps suffisant entre ces ceux-ci afin de pouvoir associer ce temps de présence à la réalisation d'une activité stationnaire.

Figure 4 : Distribution cumulée de l'entropie (équation 1)



Le zonage défini par les polygones de Voronoï pour les données mobiles est différent de celui de l'EDR en 77 secteurs. Afin d'assurer une correspondance entre les deux découpages, nous associons chaque antenne au secteur dans lequel elle se situe.

La réalisation d'une activité stationnaire au sein d'un secteur suppose de rester suffisamment longtemps à l'intérieur du même secteur afin de distinguer les activités des déplacements. Comme dans l'étude précédente en Île-de-France, nous considérons différents seuils de durée pour imputer les activités (stationnaires) (Bonnel et alii, 2013 ; 2015). Compte tenu de la taille des secteurs, la plupart des déplacements repérés dans la matrice sont réalisés à l'aide de modes motorisés. Nous considérons donc que la durée de traversée de la plupart des secteurs est incluse dans un intervalle entre 30 min et 1 heure. Par ailleurs, la durée moyenne entre deux observations est de 54 minutes dans la base de données filtrées. De plus, 50 % des mobiles ont une valeur de la durée moyenne entre deux évènements de plus de 30 minutes. Il nous semble donc peu réaliste de retenir des valeurs en dessous de 30 minutes pour le seuil. Nous considérons que cela risquerait d'induire un trop grand nombre de fausses activités. Inversement, une valeur supérieure à 1 heure réduirait significativement le nombre d'activités.

Nous définissons l'ensemble des activités associées à un mobile en analysant la succession des évènements générés par le mobile. Nous identifions l'horaire du premier et du dernier d'entre eux et inclus dans un même secteur EDR. Si la durée entre ces deux évènements est supérieure au seuil retenu, nous identifions une activité dans le secteur correspondant. Les déplace-

ments sont générés à partir de la succession des activités ainsi définies.

3.3. Identification du domicile

Lors de nos précédents travaux, nous avons défini un facteur d'expansion unique pour l'ensemble du territoire considéré en prenant le rapport entre la population du territoire et le nombre de mobiles identifiés sur le même territoire. Cette définition est problématique. En effet, le taux de pénétration d'Orange est variable selon les zones. Pour résoudre ce problème, nous avons construit un algorithme de définition du domicile. Les données disponibles sont très limitées du fait de l'obligation de changer l'identifiant associé au mobile toutes les 24h00. Nos données couvrant la période 4h00 du matin jour J à 4h00 jour J+1, nous faisons l'hypothèse que nous disposons de deux nuits partielles : jour J de 4h00 du matin à 7h00 du matin et jour J+1 de 22h00 (jour J) à 4h00 du matin (jour J+1). Nous identifions ensuite les antennes associées à chacun des évènements stationnaires (les *handovers* qui correspondent à des changements d'antennes et les LAU sont exclus) pendant les deux périodes considérées. Le domicile est défini comme étant localisé dans la zone de l'antenne contenant le plus de connexions. L'antenne est ensuite associée au secteur de l'EDR selon la méthodologie définie en 3.2.

Cette méthode permet d'identifier 1,27 millions de domicile au sein de la Région Rhône-Alpes, soit près de 25 % de la population de la région. Ils représentent 62 % de l'ensemble des terminaux présents dans notre base de données au cours de la journée étudiée.

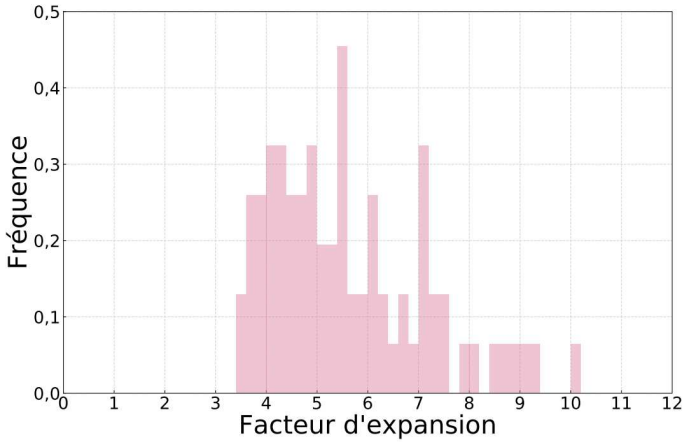
3.4. Définition des facteurs d'expansion par secteur

Le filtrage des données défini à la Section 3.1, permet d'exclure de la base de données les terminaux ne correspondant probablement pas à des téléphones mobiles. Nous excluons également les mobiles dont la présence au sein du territoire rhône-alpin ne peut pas être assurée ou pour lesquels le nombre d'observations est insuffisant pour appliquer l'algorithme de définition des déplacements. L'application de ces filtres conduit à conserver un échantillon de 985 000 mobiles, ce qui correspond à 77,3 % des mobiles pour lesquels un domicile a été identifié et approximativement à la moitié des terminaux présents dans la base initiale.

Nous disposons maintenant d'une base permettant de définir les facteurs d'expansion à l'échelle des secteurs de tirage de l'EDR. Le facteur d'expansion d'une zone est défini comme étant le rapport entre la population du secteur selon le dernier recensement disponible (2015) et le nombre de mobiles ayant le domicile identifié dans le secteur. La définition des secteurs de tirage de l'EDR étant basée sur le découpage en IRIS de l'INSEE, il n'est pas nécessaire de faire des hypothèses complémentaires pour obtenir la population de chaque secteur de tirage.

La Figure 5a présente la distribution des facteurs d'expansion parmi l'ensemble des 77 secteurs. Elle souligne l'importance de la définition spatialisée des facteurs d'expansion avec un rapport de 3 entre le facteur le plus grand et le plus petit. La Figure 5b présente les résultats spatialisés des facteurs d'expansion. Les agglomérations des plus grandes villes de Rhône-Alpes ont en général des facteurs d'expansion plus élevés. Il est difficile d'identifier des raisons précises compte tenu de la méthodologie de traitement des données pouvant générer plusieurs biais. Il est probable que l'absence des données 4G au sein de notre base puisse expliquer ce résultat, les utilisateurs du réseau 4G étant notoirement urbains.

Figure 5a : Distribution des facteurs d'expansion au sein des 77 secteurs de l'EDR



4. Résultats et comparaison entre les matrices de déplacements de la téléphonie mobile et celles issues de l'enquête déplacements

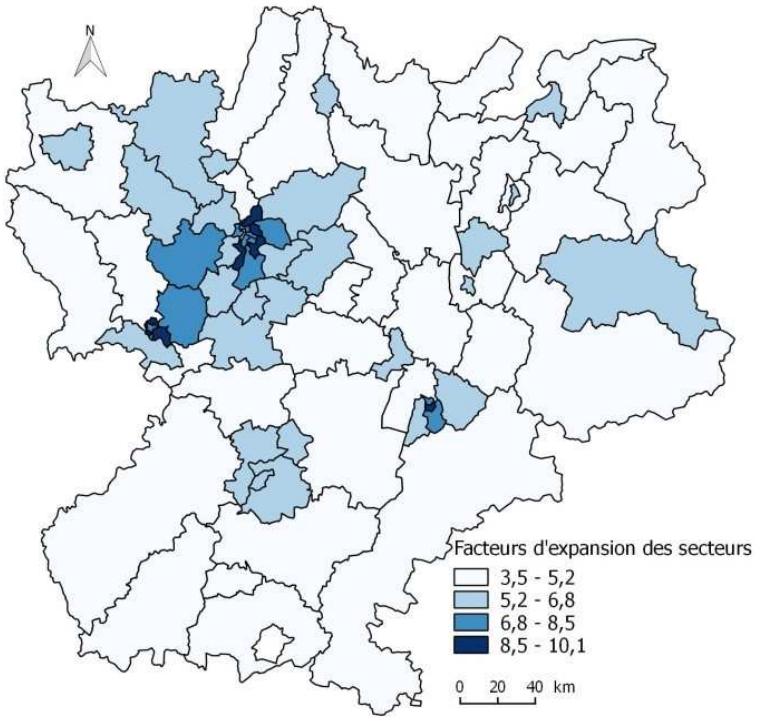
Les matrices origine-destination de déplacements de l'EDR sont obtenues en appliquant les mêmes règles que celles définies pour les données mobiles en termes de durée d'activités et de zonage.

À l'échelle des 77 secteurs de tirage, la matrice contient 5 929 cases. 143 000 déplacements ont été enquêtés. Les effectifs enquêtés sont donc faibles voire très faibles dans la très grande majorité des O-D, ce qui génère des intervalles de confiance très larges et donc de fortes incertitudes sur le nombre de déplacements de la matrice O-D de l'EDR (Armoogum, Madre, 1998).

Nous avons analysé plus finement les matrices de déplacements. Seules 40 % des cellules de la matrice O-D de l'enquête déplacements régionale sont non nulles, alors que ce pourcentage est de l'ordre de 95 %, quel que soit le seuil de 30 min à 1 heure pour les matrices issues de la téléphonie

mobile. Ce constat illustre les limites des enquêtes déplacements pour produire des matrices origine-destination sur les zonages habituellement requis. Nous procédons à une agrégation supplémentaire de la matrice, afin de ne conserver que 14 zones permettant d'avoir une précision suffisante pour la plus grande partie des paires origine-destination au niveau de la matrice EDR (Cf. Figure 1b).

Figure 5b : Distribution spatiale des facteurs d'expansion au sein de Rhône-Alpes



Le nombre de déplacements des matrices EDR/mobile est évidemment dépendant du choix du seuil (Tableau 1). Les seuils de 30 et 40 minutes fournissent les résultats les plus proches pour les deux matrices. Ces durées ne sont pas incompatibles avec les durées de traversée de la majorité des secteurs de tirage, à l'exception des secteurs des zones montagneuses dont la surface est beaucoup plus importante. Ces derniers secteurs génèrent toutefois moins de déplacements que les autres. Nous privilégions ces deux seuils pour la suite des analyses.

Tableau 1 : Nombre de déplacements des matrices EDR et mobile en fonction du choix de durée du seuil de définition des activités

Seuil de durée pour la définition des activités	60 minutes	50 minutes	40 minutes	30 minutes
EDR (en milliers)	2 211	2 260	2 344	2 448
Téléphone mobile (en milliers)	1 607	1 743	1 905	2 108

Afin d'étudier la distribution des flux pour les seuils de 30 et 40 minutes pour les deux bases de données, la corrélation de rang de SPEARMAN[2] a été calculée et est évalué à 0,95. Elle correspond à une très forte corrélation dans la hiérarchie des origines-destinations des deux matrices.

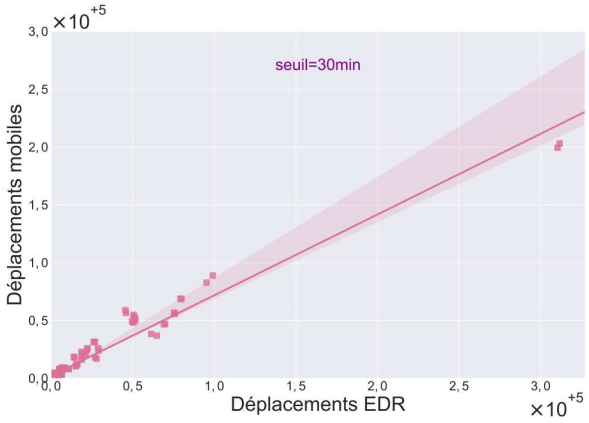
Nous poursuivons notre analyse en faisant une régression entre les déplacements des deux matrices pour chacun des seuils de 30 et 40 minutes. y_{ij} représente le nombre de déplacements de la matrice issue des données de signalisation et x_{ij} celui de la matrice EDR :

$y_{ij} = 0{,}70.x_{ij} + 2\ 193$ avec un $R^2 = 0{,}96$ (Figure 6a, seuil de 30 minutes) ;

$y_{ij} = 0{,}66.x_{ij} + 1\ 964$ avec un $R^2 = 0{,}96$ (Figure 6b, seuil de 40 minutes).

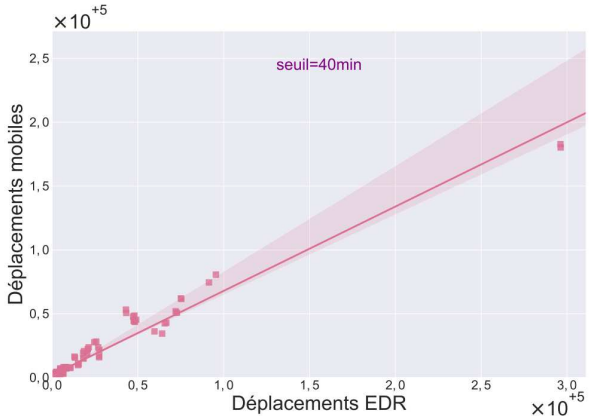
Figure 6 : Régression entre les matrices de déplacements issues des données de signalisation et de l'EDR (zonage en 14 zones) pour un seuil de

(a) 30 minutes



[2] Le Coefficient de corrélation de rang de SPEARMAN est calculé sur les numéros d'ordre des valeurs de deux variables ordinales, ici l'ordre des flux O-D au sein de chacune des deux matrices.

(b) 40 minutes



Nous obtenons une très forte corrélation avec un R^2 de 0,96 pour les deux seuils. En revanche, les pentes ne sont pas proches de 1. Cela signifie que les nombres de déplacements restent significativement différents, même si les structures sont proches. L'analyse des graphiques de régression permet d'identifier deux observations fortement éloignées du nuage de points qui risquent d'influencer fortement la pente de la droite. Ces deux observations correspondent aux flux entre les deux zones de l'agglomération lyonnaise (centre et périphérie) qui représentent une part importante de l'ensemble des déplacements de la Région. Ces déplacements semblent sous-estimés en comparaison des données de l'EDR. Il est possible que la durée retenue pour la définition des activités soit trop élevée pour ces deux zones dont la surface est nettement plus faible que les autres en raison de leur densité.

Nous reprenons l'analyse de la régression en enlevant ces deux points qui s'écartent fortement du nuage de points :

$y_{ij} = 0{,}85.x_{ij} + 877$ avec un $R^2 = 0{,}95$ (Figure 7a, seuil de 30 minutes)

$y_{ij} = 0{,}80.x_{ij} + 788$ avec un $R^2 = 0{,}95$ (Figure 7b, seuil de 40 minutes)

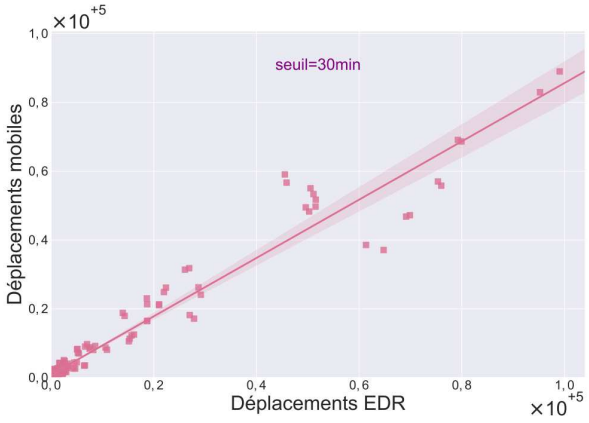
Le coefficient de détermination reste excellent. La pente se rapproche de 1, signifiant un nombre de déplacements proche pour chaque origine-destination. La constante est faible au regard du nombre moyen de déplacements par origine-destination (11 500)[3]. L'analyse plus fine des origines-destinations montre une sous-estimation fréquente pour les zones adjacentes, tandis que les flux sont plus proches pour les zones plus distantes. Il est possible que le choix du seuil amène à omettre des sorties du domicile liées à des activités et des déplacements de courte durée entre zones adjacentes. L'ana-

[3] La pente des deux régressions est toutefois statistiquement différente de 1 au seuil de 95 %. Il en de même pour les constantes qui sont statistiquement différentes de 0.

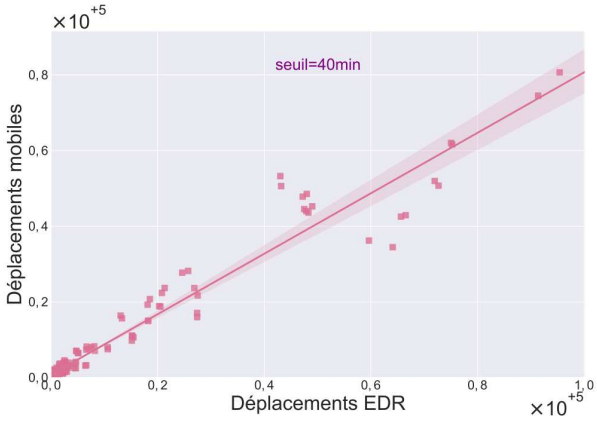
lyse en termes de pourcentage de variation entre les origines-destinations des deux matrices, et non pas en termes d'effectif, identifie certaines origines-destinations avec des écarts importants. Il s'agit principalement des origines-destinations ayant des flux faibles au niveau de l'EDR pour lesquels les intervalles de confiance sont très larges.

Figure 7 : Régression entre les matrices de déplacements issues des données de signalisation et de l'EDR avec élimination des flux entre les deux zones de Lyon pour un seuil de

(a) 30 minutes



(b) 40 minutes



5. Discussion et conclusion

Comme de nombreux auteurs l'ont déjà montré, les données de la téléphonie mobile permettent de construire des matrices origine-destination. Cependant

les travaux sur la validation de ces matrices ne sont pas très fréquents. Dans le cadre de cet article, nous proposons une nouvelle méthode d'estimation de matrices origine-destination à partir de données de signalisation du réseau de téléphonie mobile en comparant les résultats avec ceux de la dernière enquête déplacements régionale de Rhône-Alpes. Les résultats sont encourageants et plusieurs enseignements peuvent en être tirés.

L'utilisation des données de signalisation plutôt que des données de facturation permet de prendre en compte les personnes ayant une utilisation plus limitée de leur mobile grâce aux changements de zones de localisation et aux mises à jour périodiques de la localisation des mobiles.

La méthodologie mise en place permet un filtrage des données afin d'exclure les mobiles ayant un volume d'activité très élevé avec un profil très régulier. Le recours à l'entropie permet ce filtrage avec un seuil fixé à 0,9 compte tenu du profil de la courbe (*cf.* Figure 4). L'analyse de la durée maximale entre deux observations permet d'exclure les mobiles qui sont soit déconnectés du réseau, soit sortis du territoire. L'utilisation des données de signalisation à la différence des données de facturation permet de disposer des mises à jour périodiques de la localisation des mobiles connectés (LAU périodiques). La durée entre deux évènements ne doit pas dépasser 3 heures, seuil que nous avons retenu dans cette recherche. Enfin, la définition des activités stationnaires requiert au moins 2 évènements, et au moins 4 évènements pour identifier un déplacement, qui est par définition le mouvement dans l'espace entre deux activités successives. Sur cette base, et considérant que les LAU périodiques génèrent théoriquement 8 évènements par jour, nous avons exclu les mobiles ayant moins de 4 observations et qui sont soit sortis du territoire d'études, soit ont été déconnectés une partie de la journée. La méthodologie de filtrage des données que nous proposons contribue à la qualité des données mobilisées pour la construction des matrices.

Lors de précédents travaux (Bonnel et alii, 2013 ; 2015), nous avions estimé un unique facteur d'expansion pour l'ensemble du territoire étudié. Notre analyse illustre l'importance d'un traitement aussi fin que possible de l'espace afin de déterminer un facteur d'expansion par zone. La méthodologie proposée s'appuie sur un algorithme de détermination du domicile du porteur du mobile. La contrainte de modifier l'identifiant du mobile toutes les 24 heures pour des raisons de protection de la vie privée, nous a forcé à retenir une méthodologie assez simpliste. La disposition de données avec le même identifiant sur une plus longue période permettrait évidemment d'enrichir et d'améliorer l'algorithme afin d'accroître la qualité de cette détermination. Malgré tout, nous mettons en évidence des facteurs d'expansion allant de 3,5 à 10, soit un rapport de presque 3 entre certaines zones. L'analyse spatialisée met en évidence des facteurs globalement plus élevés pour les zones les plus denses des grandes agglomérations de Rhône-Alpes. Ce constat est probablement lié à l'absence des données 4G, mais peut être

aussi dû à des comportements spécifiques des habitants des grandes agglomérations comparativement à ceux dans le reste du territoire.

Ce traitement des données améliore significativement les résultats (Bonnel et alii, 2018a). Nous obtenons ainsi un nombre de déplacements proche de celui de l'EDR pour les choix de seuils de 30 et 40 minutes. La structure des deux matrices est très proche avec une corrélation de rang de Spearman évaluée à 0,95. Celle-ci montre que la hiérarchie des flux au sein des deux matrices est très proche. De plus, la régression des flux de chacune des origines-destinations mobiles par ceux de l'EDR fournit un très bon coefficient de détermination de 0,95. La pente n'est pas égale à 1 mais n'en est pas trop éloignée. La constante est également faible au regard du nombre moyen de déplacements par origine-destination. L'analyse plus fine au niveau des origines-destinations confirme la qualité des résultats sur les flux les plus importants entre zones non adjacentes. Pour les plus petits flux, les résultats sont de qualité moindre, mais cela peut être dû à la précision des données de l'EDR qui est faible voire très faible pour les plus petits flux.

De nombreuses hypothèses sont nécessaires pour construire les matrices de déplacements à partir des données de téléphonie mobile. Nous reprenons ici les plus importantes pour tenter de dégager des pistes d'approfondissements :

. Les données de la téléphonie portent sur tous les déplacements réalisés par les personnes présentes sur le territoire de Rhône-Alpes. En revanche, les données de l'EDR ne concernent que les résidents de Rhône-Alpes. La disposition de données conservant un identifiant unique sur une plus longue période et sur un territoire plus large que celui de la Région, permettrait d'identifier les domiciles se trouvant hors de Rhône-Alpes pour filtrer ces données dans une optique de comparaison avec des données de l'enquête déplacements ;

. Les données de l'EDR sont représentatives d'un jour moyen sur la période automne-printemps, tandis que les données de téléphonie ne portent que sur une seule journée de juin. Il est sûrement intéressant de pouvoir disposer de données de téléphonie sur une plus longue période pour étudier d'une part la variabilité des données collectées par l'opérateur mobile et d'autre part pouvoir construire un jour moyen similaire à celui de l'EDR. Nous revenons sur cette question dans les perspectives en fin de conclusion ;

. Nous avons retenu une hypothèse uniforme de stationnarité de 30 à 40 minutes pour l'ensemble des zones. Les résultats sont très sensibles à ce seuil (*cf.* Tableau 1). Le choix de la durée de 30 à 40 minutes est arbitraire, même si elle semble cohérente avec la taille des zones et les vitesses des déplacements motorisés. Il est possible d'affiner cette durée afin de l'adapter aux caractéristiques de chaque zone et aux vitesses pratiquées ;

. L'expansion des matrices issues des données de téléphonie mobile s'est appuyée sur une hypothèse très simple de détermination du domicile du fait du changement d'identifiant toutes les 24 heures. Il serait pertinent d'améliorer l'algorithme sous réserve de disposer de données sur une plus longue période. Même si nous avons pu déterminer un coefficient par zone, ce qui représente une amélioration de notre méthodologie, son application nécessite de supposer que la population des possesseurs de téléphone mobile pour lesquels nous avons pu reconstituer au moins un déplacement est représentative de l'ensemble de la population présente sur le territoire. S'il est peu probable que l'on puisse accéder aux données démographiques des utilisateurs du réseau Orange pour d'évidentes raisons de respect de la vie privée, il n'est pas exclu de tenter de collecter des informations via d'autres sources. Calabrese et alii (2011) et Bekhor et alii (2013) ont analysé la distribution spatiale des possesseurs de téléphone mobile en la comparant aux données de recensement. Bekhor et alii (2013) ont également utilisé des données d'enquêtes sur les déplacements contenant des questions sur la téléphonie. Ces données permettraient d'identifier les biais éventuels des échantillons de téléphonie pour redresser les données à l'aide des données de mobilité des enquêtes ménages déplacements ;
. Enfin, nous n'avons pas analysé les données des possesseurs de téléphone mobile pour lesquels nous disposons de moins de 4 évènements. Il serait pourtant utile d'identifier ceux qui ont connecté ou déconnecté leur téléphone pendant la journée étudiée pour les différencier de ceux qui sont entrés ou sortis de la zone d'étude au cours de la journée.

Au-delà de ces développements méthodologiques, il serait intéressant d'étudier la variabilité des matrices origine-destination produites avec les données de la téléphonie mobile. Les données d'enquêtes déplacements portent en général sur une seule journée et conduisent à postuler implicitement une reproduction des comportements individuels d'un jour sur l'autre. La disponibilité potentielle de ces données en continu permet donc d'envisager l'analyse de la variabilité quotidienne ou hebdomadaire des matrices ainsi produites. Il serait également possible d'étudier la saisonnalité des comportements de déplacements. Cette analyse de la variabilité serait d'ailleurs également utile pour la comparaison avec les données de l'enquête déplacements régionale pour étudier la dispersion des résultats comparativement à ceux de l'enquête régionale.

Les données de la téléphonie mobile apparaissent ainsi très prometteuses pour l'analyse de la mobilité spatiale, mais nécessitent encore de nombreuses recherches avant de pouvoir valider pleinement leur utilisation pour construire des matrices origine-destination à des fins de modélisation

transport ou de planification des déplacements, notamment pour celles utilisant des techniques de modélisation au niveau microscopique telles que les systèmes multiagents. Elles présentent un avantage important comparativement aux matrices issues des enquêtes déplacements, car elles permettent d'estimer des flux pour la plupart des origines-destinations alors qu'avec l'EDR 60 % des origines-destinations ont un flux nul malgré un zonage en seulement 77 zones, très loin de la finesse habituellement requise pour les travaux de modélisation.

Remerciements : Les auteurs tiennent à remercier d'une part Orange pour l'accès aux données de téléphonie mobile et d'autre part la Région Rhône-Alpes pour la mise à disposition des données de l'Enquête Déplacements Régionale. Seuls les auteurs sont toutefois responsables des propos tenus dans cet article.

Bibliographie

Aguilera 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**, Vol. 43, n° 2, pp. 198-211.

Akin D., Sisiopiku V. (2002) Estimating Origin-Destination Matrices Using Location Information from Cellular Phones. **Proc. NARSC RSAI**, Puerto Rico, USA.

Alexander L.P., Jiang S., Murga M., González M.C. (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data. **Transportation Research Part C**, Vol. 58, pp. 240-250.

Ampt E.S. (1997) Response Rates - Do they matter? In Bonnel P., Chapleau R., Lee-Gosselin M., Raux C. (eds.) **Les enquêtes de déplacements urbains: mesurer le présent, simuler le futur**. Programme Rhône-Alpes Recherches en Sciences Humaines, Lyon, pp 115-125.

Arcep (2019) **Autorité de Régulation des Communications Electroniques et des Postes, Observatoires réseaux et services mobiles**. https://www.arcep.fr/cartes-et-donnees/nos-publications-chiffrees/observatoire-services-mobiles/obs-mobile-t1-2019.html.

Arentze T., Timmermans H., Hofman F., Kalfs N. (2000) Data needs, data collection, and data quality requirements of activity-based transport demand models. In Transport surveys, raising the standard, **TRB transport circular E-C008**, pp. II-J/1-30.

Armoogum J., Madre J.-L. (1998) **Redressement de l'enquête transports pour l'estimation de matrices Origine-Destination**. Rapport INRETS n° 223.

ASGARI F., GAUTHIER V., BECKER M. (2013) **A survey on human mobility and its applications**. arXiv preprint arXiv:1307.0814.

ATROSTIC B.K., BURT G. (1999) **Household non-response: what we have learned and a framework for the future**. Statistical Policy working paper 28, Federal Committee on Statistical methodology, Office of Management and Budget, Washington, pp. 153-180.

BAR-GERA H. (2007) Evaluation of a cellular Phone-Based System for Measurement of Traffic Speeds and Travel Times: A Case Study from Israel. **Transportation Research Part C**, Vol. 15, n° 6, pp. 380-391.

BEKHOR S., COHEN Y., SOLOMON C. (2013) Evaluating Long-Distance Travel Patterns in Israel by Tracking Cellular Phone Positions. **Journal of Advanced Transportation**, Vol. 47, n° 4, pp. 435-446.

BLONDEL V.D., DECUYPER A., KRINGS G. (2015) A survey of results on mobile phone datasets analysis. **EPJ Data Sci. 4**, 10. https://doi.org/10.1140/epjds/s13688-015-0046-0

BOLLA R., DAVOLI F. (2000) Road Traffic Estimation from Location Tracking Data in the Mobile Cellular Network. **Proc. IEEE WCNC**, Chicago, USA.

BONNEL P. (2003) Postal, telephone and face-to-face surveys: how comparable are they? In STOPHER P.R., JONES P.M. (eds.) **Transport Survey Quality and Innovation**. London: Elsevier, pp 215-237.

BONNEL P. (2004) **Prévoir la demande de transport**. Presses de l'Ecole Nationale des Ponts et Chaussées, Paris, 425 p.

BONNEL P., FEKIH M., SMOREDA Z. (2018a), Origin-Destination estimation using mobile network probe data. **Transportation Research Procedia**, Transport Survey Methods in the era of big data: facing the challenges, Vol. 32, pp. 69-81. https://doi.org/10.1016/j.trpro.2018.10.013

BONNEL P., HOMBOURGER E., SMOREDA Z. (2013) **Quel potentiel des données de la téléphonie mobile pour la construction de matrices origine-destination de déplacement–application à l'Île-de-France**. Rapport de Recherche, Laboratoire d'Économie des Transports, Orange Labs, 133 p.

BONNEL P., HOMBOURGER E., OLTEANU-RAIMOND A.-M., SMOREDA Z. (2015) Passive mobile phone dataset to construct origin-destination matrix: potentials and limitations. **Transportation Research Procedia**, Vol. 11, pp. 381-398. doi: 10.1016/j.trpro.2015.12.032

BONNEL P., MUNIZAGA M., MORENCY C., TREPANIER M. (Eds) (2018b) Transport Survey Methods in the era of big data: facing the challenges. **Transportation Research Procedia**, Vol 32, 666 p. https://www.sciencedirect.com/journal/transportation-research-procedia/vol/32/suppl/C

CACERES N., WIEDEBERG J.P., BENITEZ F.G. (2007) Deriving Origin-Destination Data from a Mobile Phone Network. **IET Intelligent Transport System**, Vol. 1, n° 1, pp. 15-26.

CALABRESE F., DIAO M., DI LORENZO G., FERREIRA JR J., RATTI C. (2013) Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. **Transportation Research Part C**, Vol. 26, pp. 301-313.

CALABRESE F., DI LORENZO G., LIU L., RATTI C. (2011) Estimating Origin-Destination Flows using Mobile Phone Location Data. **IEEE Pervasive Computing**, Vol. 10, n° 4, pp. 36-44.

CALABRESE F., PEREIRA F., DI LORENZO G., LIU L., RATTI C. (2010) The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events. **Proc. Pervasive Computing**, Helsinki, Finlande.

CERTU (2008) **L'enquête ménages déplacements standard CERTU**. Lyon, éditions du CERTU, 203 p.

CHEN C., BIAN L., MAC J. (2014) From traces to trajectories: How well can we guess activity locations from mobile phone traces? **Transportation Research Part C**, Vol. 46, pp. 326-337.

CHO E., MYERS S.A., LESKOVEC J. (2011) Friendship and Mobility: User Movement in Location-based Social Networks. **Proc. ACM SIGKDD**, San Diego, USA.

COLAK S., ALEXANDER L.P., ALVIM B.G., MEHNDIRETTA S.R., GONZÁLEZ M.C. (2015) Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. **Transportation Research Board Annual Meeting, 94th**, Washington, DC, USA.

EGT (2010) **Enquête Globale Transport de l'Île-de-France**. http://www.stif.org/IMG/pdf/%20Enquete_globale_transport_BD-2.pdf

FIADINO P., PONCE-LOPEZ V., ANTONIO J., TORRENT-MORENO M., D'ALCONZO A. (2017) Call Detail Records for Human Mobility Studies: Taking Stock of the Situation in the "Always Connected Era. In **Proceedings of Big-DAMA**. ACM Press, pp. 43-48. https://doi.org/10.1145/3098593.3098601

FRIAS-MARTINEZ V., FRIAS-MARTINEZ E., OLIVER N. (2010) A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records. **Proc. AAAI AI-D**, Palo Alto, USA.

GONZALEZ M.C., HIDALGO C.A., BARABASI A.-L. (2008) Understanding Individual Human Mobility Patterns. **Nature**, Vol. 453 (7196), pp. 779-782.

GRAELLS-GARRIDO E., SAEZ-TRUMPER D. (2016) A Day of Your Days: Estimating Individual Daily Journeys Using Mobile Data to Understand Urban Flow. **2nd International Conference on IoT in Urban Space**, ACM Press, pp. 1-7. https://doi.org/10.1145/2962735.2962737

HOTEIT S., SECCI S., SOBOLEVSKY S., RATTI C., PUJOLLE G. (2014) Estimating human trajectories and hotspots through mobile phone data. **Computation Network**, Vol. 64, pp. 296-307.

HUANG Z., LING X., WANG P., ZHANG F., MAO Y., LIN T., WANG F.-Y. (2018) Modeling real-time human mobility based on mobile phone and transportation data fusion. **Transportation Research Part C: Emerging Technologies**, Vol. 96, pp. 251-269. https://doi.org/10.1016/j.trc.2018.09.016

IDATE (2009) Observatoire économique de la téléphonie mobile-faits et chiffres 2008. **Mobile et société**, n° 9, pp. 6-15. http://www.fftelecoms.org/sites/default/files/contenus_lies/mobile_et_societe_9.pdf

IQBAL M.S., CHOUDHURY C.F., WANG P., GONZÁLEZ M.C. (2014) Development of origin–destination matrices using mobile phone call data. **Transportation Research Part C**, Vol. 40, pp. 63-74.

ISAACMAN S., BECKER R., CACERES R., KOBOUROV S., ROWLAND J., VARSHAVSKY A. (2010) A Tale of Two Cities. **Proc. ACM HotMobile**, Annapolis, USA.

ISAACMAN S., BECKER R., CACERES R., KOBOUROV S., MARTONOSI M., ROWLAND J., VARSHAVSKY A. (2011) Ranges of Human Mobility in Los Angeles and New York. **Proc. IEEE PerCom Workshops**, Seattle, USA.

JIANG S., FIORE G.A., YANG Y., FERREIRA JR J., FRAZZOLI E., GONZÁLEZ M.C. (2013) A review of urban computing for mobile phone traces: current methods, challenges and opportunities. Proceedings of the **2nd ACM SIGKDD International Workshop on Urban Computing**. August 11-14, Chicago, Illinois,. ACM Press, pp. 1-9.

JIANG S., VINA-ARIAS L., FERREIRA J., ZEGRAS C., GONZÁLEZ M.C. (2011) Calling for Validation: Demonstrating the use of Mobile Phone data to Validate integrated land use Transportation models. **Proc. 7VCT**, Lisbon.

JIANG S., FERREIRA J., GONZÁLEZ M.C. (2017) Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. **IEEE Transactions on Big Data**, Vol. 3, pp. 208-219.

KANG J.H., WELBOURNE W., STEWART B., BORRIELLO G. (2004) Extracting Places from Traces of Locations. **Proc. ACM WMASH**, Philadelphia, USA.

MELLEGARD E. (2011) **Obtaining Origin/Destination-Matrices from Cellular Network Data**. Master's Thesis, Chalmers University of Technology.

MILION C. (2015) **Méthodes et modèles pour l'étude de la mobilité des personnes par l'exploitation de données de radiotéléphonie**. Thèse de doctorat de l'Université Paris-Est, Marne La Vallée.

MUNIZAGA M., PALMA C., MORA P. (2010) Public transport OD matrix estimation from smart card payment system data. **12th World Conference on Transport Research**, Lisbon, Paper n° 2988.

NOULAS A., MASCOLO C., FRIAS-MARTINEZ E. (2013) Exploiting Foursquare and Cellular Data to Infer User Activity in Urban Environments. **Proc. IEEE**, MDM, Milan, Italie.

OLTEANU-RAIMOND A.-M., BAHOKEN F., COURONNÉ T., SMOREDA Z. (2013) Proposition de matrices de flux temporelles issues de l'activité d'individus mobiles. Actes du **Colloque International de Géomatique et d'Analyse Spatiale (SAGEO 2013)**, Brest, France.

ORTÚZAR J. de D., BATES J. (2000) Workshop summary, Transport surveys, raising the standard. **TRB transport circular E-C008**, pp. II-J/31-35.

ORTÚZAR J. de D., WILLUMSEN L.G. (2011) **Modelling Transport**. Wiley (4th ed.).

PELLETIER M.P., TRÉPANIER M., MORENCY C. (2011) Smart card data use in public transit: A literature review. **Transportation Research Part C**, Vol. 19, pp. 557-568.

PICORNELL M., RUIZ T., LENORMAND M., RAMASCO J., DUBERNET T., FRIAS-MARTINEZ E. (2015) Exploring the potential of phone call to characterize the relationship between social network and travel behavior. **Transportation**, Vol. 42, pp. 647-668.

RATTI C., WILLIAMS S., FRENCHMAN D., PULSELLI R.M. (2006) Mobile landscapes: using location data from cell phones for urban analysis. **Environment and Planning B**, Vol. 33, n° 5, pp. 727-748.

SCHLAICH J., OTTERSTATTER T., FRIEDRICH M. (2010) Generating Trajectories from Mobile Phone Data. **Proc. TRB Annual Meeting**, Washington D.C, USA.

SEVTSUK A., RATTI C. (2010) Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. **Journal of Urban Technology**, Vol. 17, n° 1, pp. 41-60.

SMOREDA Z., OLTEANU-RAIMOND A.M., COURONNÉ T. (2013) Spatiotemporal Data from Mobile Phones for Personal Mobility Assessment. In **Transport survey methods: best practice for decision making**, Emerald Group Publishing Limited, pp. 745-768.

STOPHER P.R., GREAVES S.P. (2007) Household travel surveys: where are we going? **Transportation Research Part A**, Vol. 41, pp. 367-381.

TIZZONI M., BAJARDI P., DECUYPER A., KING G.K.K., SCHNEIDER C., BLONDEL V., SMOREDA Z., GONZALEZ M.C., COLIZZA V. (2014) On the Use of Human Mobility Proxies for Modeling Epidemics. **PLOS Computational Biology**, Vol. 10, n° 7. http://dx.doi.org/10.1371/journal.pcbi.1003716

TOOLE J.L., COLAK S., STURT B., ALEXANDER L.P., EVSUKOFF A., GONZALEZ M.C. (2015) The path most traveled: Travel demand estimation using big data resources. **Transportation Research Part C**, http://dx.doi.org/10.1016/j.trc.2015.04.022

WANG P., HUNTER T., BAYEN A., SCHECHTNER K., GONZALEZ M.C. (2012) Understanding Road Patterns in Urban Area. **Scientific Reports**, 2 (1001) pp. 1-6.

WANG Z., HE S.Y., LEUNG Y. (2017) Applying mobile phone data to travel behaviour research: A literature review. **Travel Behaviour and Society**. https://doi.org/10.1016/j.tbs.2017.02.005

WHITE J., WELLS I. (2002) Extracting Origin Destination Information from Mobile Phone Data. **Proc. IEEE RTIC**, London, UK.

WIDHALM P., YANG Y., ULM M., ATHAVALE S., GONZALEZ M.C. (2015) Discovering urban activity patterns in cell phone data. **Transportation**, Vol. 42, pp. 597-623.

WIDHALM P., NITSCHE P., BRÄNDIE N. (2012) Transport mode detection with realistic Smartphone sensor data. Proceedings of the **21st International Conference on Pattern Recognition (ICPR2012)**, pp. 573-576.

WISMANS L.J.J., FRISO K., RIJSDIJK J., DE GRAAF S.W., KEIJ J. (2018) Improving A Priori Demand Estimates Transport Models using Mobile Phone Data: A Rotterdam-Region Case. **Journal of Urban Technology**, Vol. 25, pp. 63-83. https://doi.org/10.1080/10630732.2018.1442075

WOLF J., OLIVEIRA M., THOMPSON M. (2003) Impact of underreporting on mileage and travel time estimate-results from Global Postionning System enhanced household travel survey. **Transportation research record**, n° 1854, pp. 189-198.

XU Y., SHAW S.-L., ZHAO Z., YIN L., FANG Z., LI Q. (2015) Understanding aggregate human mobility pattern using passive mobile phone location data: a home-based approach. **Transportation**, Vol. 42, pp. 625-646.

YGNACE J.-L. (2001) **Travel Time/Speed Estimates on the French Rhone Corridor Network using Cellular Phones as Probes**. INRETS STRIP Project Technical Report.

YUE Y., LAN T., YEH A.G.O., LI Q.-Q. (2014) Zooming into individuals to understand the collective: a review of trajectory-based travel behaviour studies. **Travel Behaviour Society**, Vol. 1, n° 2, pp. 69-78.

ZHAO Z., KOUTSOPOULOS H.N., ZHAO J. (2018) Individual mobility prediction using transit smart card data. **Transportation Research Part C: emerging technologies**, Vol. 89, pp. 19-34.

ZHANG H., BOLOT J. (2007) Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Networks. **Proc. ACM MobiCom**, Montréal, Canada.

ZHENG Y., XIE X. (2011) Learning travel recommendations from user-generated GPS traces. **ACM Transactions on Intelligent Systems and Technology**, Vol. 2, n° 1, article 2.

ZMUD J. (2003) Designing instruments to improve response: keeping the horse before the cart. In STOPHER P.R., JONES P.M. (Eds) **Transport Survey Quality and Innovation**. Oxford, Elsevier, Pergamon, pp 89-108.