Statistiques de mobilité et traces numériques de déplacements

Vincent Aguiléra[*], François Combes[*], Vincent Benezech[*], Chloé Million[+], Sylvain Allio[+]
[*]UMR LVMT, École des Ponts ParisTech, Champs-sur-Marne.
[+]Orange Labs, Belfort.

Résumé – *L'usage au quotidien des technologies de l'information et de la communication (TIC) au cours des déplacements — non seulement de personnes, mais également de marchandises — engendre des quantités importantes de données à la fois identifiées, datées et localisées. C'est le cas notamment des systèmes de billettique dans les transports en commun, des données de péages autoroutiers, ou encore des données de localisation nécessaires au fonctionnement des réseaux de téléphonie cellulaire. L'exploitation de ces traces numériques de déplacements offre des potentialités extrêmement intéressantes pour l'étude du fonctionnement des réseaux de transport et des territoires. Cet article présente une synthèse des premiers résultats obtenus au sein du Laboratoire Ville Mobilité Transport sur ce thème de recherches novateur, en partenariat avec Orange Labs et le Syndicat des Transports d'Île-de-France (STIF).*

Introduction

Le point initiateur des travaux présentés dans cet article (Aguiléra *et al.* 2012, 2013, 2014) est le constat que l'application de modèles de transport dynamiques nécessite une connaissance fine de la demande, dans l'espace et dans le temps, et que cette connaissance manque aujourd'hui. Il serait souhaitable de disposer d'un instrument de mesure robuste, partout disponible, et pouvant répondre à un large éventail des questions que posent le fonctionnement et l'exploitation des réseaux de transport, indépendamment du mode de transport considéré. Précisons que les traces numériques issues de systèmes de positionnement satellitaire comme le GPS ne sont pas considérées ici. En effet, si les données GPS sont bien adaptées aux transports de surface[1], elles sont indisponibles pour les systèmes de transport souterrains. Or ce mode de transport est bien souvent prépondérant en volume dans les grandes métropoles.

Cet article comprend deux sections suivies d'une conclusion. La première section présente des éléments relatifs au diagnostic de fonctionnement d'un territoire à partir de données de téléphonie mobile, en exposant la construction d'indicateurs de mesure de la mobilité sur le territoire de la Seine-et-Marne. La deuxième section traite de la mesure d'indicateurs de qualité de service dans un système de transport en commun : le RER A en Île-de-France. Deux sources de traces numériques sont utilisées et comparées : données de téléphonie mobile d'une part, et données de billettique d'autre part.

Ces travaux ont pu être réalisés en grande partie grâce au soutien du Syndicat des Transports d'Île-de-France (STIF) et de Orange Labs.

1. Le sujet a fait l'objet de nombreux travaux de recherche (voir par exemple Lind & Lindkvist 2006, Hofleitner *et al.* 2012, Chen & Bierlaire 2014). Par ailleurs aujourd'hui — pour le mode routier en particulier — des solutions opérationnelles sont proposées sur le marché par les fournisseurs de Floating Car Data (FCD).

1. Diagnostic de fonctionnement d'un territoire

Les travaux présentés dans cette section partent de l'idée que l'activité d'un réseau de téléphonie mobile capture une part substantielle des activités humaines sur un territoire. L'idée en soi n'est pas nouvelle : un bref état de l'art permet de situer l'originalité des travaux présentés ici par rapport à la littérature. Cet état de l'art, en sous-section 1.2, est précédé en sous-section 1.1 d'un exposé rapide du fonctionnement des réseaux de téléphonie. Cet exposé décrit comment, au cours de son fonctionnement quotidien, un réseau de téléphonie mobile et les mobiles qui y sont connectés échangent des messages dits de *signalisation*.

Alors que la plupart des travaux que nous avons identifiés dans la littérature portent uniquement sur l'analyse de traces d'appels, nous avons pu, dans le cadre d'un partenariat avec l'opérateur de téléphonie Orange, accéder à un ensemble de données beaucoup plus riche que les seules traces d'appels, et ce à l'échelle d'un département. Cette zone géographique d'intérêt, le département de Seine-et-Marne, est présentée en sous-section 1.3. La sous-section 1.4 présente la volumétrie du jeu de données, ainsi qu'un ensemble de statistiques permettant d'en appréhender le contenu. Enfin, la sous-section 1.5 propose des exemples d'analyse de mobilité sur le territoire Seine-et-Marnais.

1.1 Fonctionnement des réseaux mobiles

Cette sous-section présente le principe de fonctionnement d'un réseau de téléphonie mobile, ainsi que le rôle clé que joue la signalisation dans la localisation des mobiles. Elle permet notamment d'introduire la terminologie du monde de la téléphonie sans fil. Les notions d'*antenne* et de *cellule* sont présentées dans un premier temps (§ 1.1.1) ; puis viennent celles de *signalisation* (§ 1.1.2) et de *localisation* (§ 1.1.3).

1.1.1 Antennes et cellules

Un téléphone cellulaire émet et reçoit des informations (voix, messages, données) par des signaux radio échangés entre le téléphone et la (ou les) antennes les plus proches. Chaque antenne est connectée à une station de base, qui assure la liaison avec la partie filaire du réseau de télécommunications. Par la suite l'acronyme BTS (pour *Base Transceiver Station*) sera utilisé pour désigner une station de base. Le plus souvent les antennes et leurs BTS sont installées sur des points hauts (toits d'immeubles, pylônes, etc.). La zone géographique couverte par une station de base est appelée *cellule*. Typiquement, l'ouverture d'une antenne est de 120°, et trois antennes sont nécessaires à une BTS pour couvrir 360°. Les cellules se recouvrent partiellement à leurs frontières communes pour garantir la continuité de la couverture radio. En première approximation, on peut se figurer un réseau cellulaire comme un réseau d'hexagones réguliers qui forment les cellules, avec au centre de chaque cellule une BTS et son triplet d'antennes.

La situation réelle est un peu plus complexe. Avec le temps, plusieurs technologies ont émergé (2G, 3G, 4G), chacune avec des variantes. Elles coexistent aujourd'hui, et se superposent en plusieurs couches. À chaque technologie correspond une couche de cellules et, d'une technologie à l'autre, ces couches ne se superposent pas nécessairement. De fait, pour une technologie donnée, la taille d'une cellule dépend de trois facteurs principaux : la portée, les conditions locales de propagation, et la capacité. La portée dépend de la fréquence : plus la fréquence est élevée, plus les débits possibles sont importants, mais plus la portée est faible. Pour assurer une même couverture géographique, un réseau à haut débit type 4G comprend plus d'antennes qu'un réseau type 2G.

Les conditions locales de propagation sont données par l'environnement : les signaux radios peuvent être atténués, absorbés ou réfléchis par des obstacles de différentes natures (ex. : arbres, collines, bâtiments). Ceci est d'autant plus vrai que la fréquence est élevée.

Enfin, la capacité dépend de la densité spatiale de mobiles susceptibles de se connecter simultanément au réseau. Une BTS ne peut gérer qu'un nombre limité de communications en même temps. De ce fait la taille des cellules est plus faible dans les zones à forte densité d'occupation humaine que dans les zones rurales. Comme ordre de grandeur de la taille d'une cellule, on retiendra : quelques centaines de mètres dans les zones urbaines ; quelques dizaines de kilomètres dans les zones rurales.

1.1.2 Signalisation

Un réseau de téléphonie mobile et les terminaux mobiles qui y sont connectés échangent des données dites de *signalisation*. Un message de signalisation est appelé un *événement*. Lorsqu'il est émis par un mobile, un événement de signalisation contient plusieurs informations, dont les plus importantes pour la suite sont :

- son type ;
- l'antenne, et par conséquent la BTS, qui a reçu l'événement ;
- un horodatage, à la précision d'une milliseconde ;
- un identifiant du mobile ayant émis l'événement.

Lorsqu'un mobile est allumé et qu'il se connecte au réseau, il émet un événement de type ATT1 (pour *attachment*). L'événement symétrique, de type ATT0, est émis à l'extinction du terminal. D'autres types d'événements sont utilisés pour la gestion des communications, comme :

- COM0 : lorsqu'une communication (voix ou SMS) est reçue ;
- COM1 : lorsqu'une communication (voix ou SMS) est émise ;
- GMM_SERVICE : durant les échanges de données (connections internet) ;
- HO (pour *Hand Over*) : lorsqu'un changement de cellule survient pendant une communication.

1.1.3 Localisation

Une fonction importante des événements de signalisation est la gestion de la *localisation*. En effet, pour pouvoir transmettre les communications à destination des mobiles, le réseau a nécessairement besoin de connaître de façon approximative la position de chacun des mobiles qui y sont connectés. À défaut d'un tel mécanisme de localisation, il serait nécessaire de diffuser, pour chaque appel, une notification d'appel sur l'ensemble des antennes, ce qui n'est pas, à l'évidence, concevable.

Pour gérer la fonction de localisation, le réseau maintient une base de donnée distribuée. Au niveau le plus haut, le GSMC (pour *Global Mobile Switching Center*) maintient une table centralisée d'appariement entre les identifiants de mobiles et les LA (pour *Location Area*), c'est-à-dire des zones de localisation. Une zone de localisation est un regroupement logique de cellules géographiquement proches. Typiquement, une zone de localisation regroupe entre 10 et 100 cellules. Les informations de localisation d'une zone de localisation sont gérées par un MSC (pour *Mobile Switching Center*). Lorsqu'un appel doit être transmis à un mobile, le réseau consulte le GSMC pour connaître sa zone de localisation, puis transmet l'appel au MSC de cette zone. Le MSC maintient quant à lui une table de correspondance entre identifiants de mobiles et

BTS. Lorsque l'appel arrive au MSC, il est donc ensuite transmis à la BTS idoine, qui active alors les antennes appropriées.

Pour gérer la mobilité, il faut pouvoir maintenir à jour cette base distribuée. Des événements de signalisation sont utilisés à cette fin. Lorsqu'un mobile est allumé, il émet — en communiquant avec l'antenne la plus proche — un événement d'attachement au réseau (ATT1), et stocke en mémoire la zone de localisation de l'antenne. L'événement ATT1 provoque une mise à jour du MSC et du GSMC. Ensuite, typiquement toutes les dix secondes, le mobile sonde le réseau. Si la zone de localisation de la BTS avec laquelle il communique est la même que celle qu'il a en mémoire, rien ne se passe. Sinon, le téléphone émet un événement de type LAUN (pour *Location Area Update Normal*). Sa localisation est alors prise en charge par le MSC correspondant à la nouvelle zone de localisation. Par ailleurs, de façon systématique, si aucun événement LAUN n'est intervenu au cours d'une période de trois heures, un événement de type LAUP est émis.

1.2 État de l'art

Du fait même de leur principe de fonctionnement, les réseaux de téléphonie mobile constituent un capteur de mobilité des personnes. Les premiers travaux ont porté sur l'équivalent virtuel de capteurs trafic. L'une des premières expériences est reportée par Linnartz (1994) dans la baie de San Francisco. En France, l'usage de données de téléphonie mobile pour le mesure des flux de trafic a été expérimenté par Ygnace, d'abord en simulation, puis par la suite lors de tests opérationnels le long du couloir Rhôdanien (Ygnace, 2001). Depuis le sujet a été régulièrement revisité par des chercheurs (ex. : Bar-Gera 2007, Caceres *et al.* 2010).

La plupart des opérateurs de téléphonie cellulaire proposent aujourd'hui des solutions commerciales pour la mesure des flux de trafic. La plupart des systèmes de ce type sont fondés sur l'utilisation d'événements de type HO (pour *Hand Over*), qui se produisent lorsqu'un téléphone en communication passe d'une cellule à une autre. En effet, pour assurer la continuité du service lors d'un tel passage, deux cellules connexes échangent des informations détaillées (puissance de signal, fréquences allouées, etc.) qui permettent d'estimer finement la localisation et la vitesse du mobile. Localement, il est donc possible d'utiliser une paire (ou un petit groupe) de cellules comme capteur de trafic virtuel, avec un niveau de précision comparable à celui d'un capteur trafic standard type boucle électromagnétique.

À une échelle plus large que celle d'un capteur local de mesure de flux, les données de téléphonie mobile ont récemment émergé comme instrument possible de mesure de la mobilité des personnes à l'échelle d'un territoire. La plupart des recherches menées jusqu'à présent ont utilisé des CDR (pour *Call Detail Records*). Les CDR sont des jeux de données préparés à des fins de facturation. Ils utilisent un sous-ensemble réduit des données de signalisation, en considérant essentiellement uniquement les événements de type COM1, dont nous verrons qu'ils ne constituent qu'une faible fraction de l'ensemble des événements.

Des CDR ont notamment été utilisés : pour observer la dynamique d'une aire urbaine, en cartographiant des densités d'appels (Ratti *et al.* 2006, Rubio *et al.* 2013) ; pour des études de fréquentation touristique (Ahas *et al.*, 2008) ; pour calibrer des modèles de mobilité des personnes (Gonzales *et al.*, 2008) ; pour étudier des corrélations entre usage des téléphones mobiles et mobilité des personnes (Kang *et al.*, 2012) ; pour identifier des motifs d'activités quotidiennes (Calabrese *et al.*, 2013).

1.3 Zone géographique d'intérêt

La zone géographique d'intérêt est la Seine-et-Marne, l'un des huit départements de la région Île-de-France (Fig. 1a). Une attention particulière est portée aux environs de DisneyLand Resort Paris, avec trois points d'intérêt en voisinage immédiat les uns des autres.

À côté du parc d'attraction sont situés le centre commercial Val d'Europe, ainsi que la gare d'interconnexion TGV de Marne-la-Vallée Chessy. La zone est connectée à Paris par l'autoroute A4 et par le RER A (Fig. 1b). La gare TGV de Chessy est reliée au réseau européen de trains à grande vitesse, incluant en particulier l'Eurostar (Londres, Bruxelles) et le Thalys (Bruxelles, Cologne, Amsterdam). Les autres infrastructures de transport principales en Seine-et-Marne sont l'autoroute A104 — troisième rocade de l'Île-de-France, après le boulevard périphérique et l'autoroute A86 —, ainsi que les autoroutes A5 et A6. La population du département est d'environ 1,33 millions d'habitants, pour 11,73 millions en Île-de-France, selon les chiffres du recensement 2009. La surface du département Seine-et-Marne représente la moitié de celle de l'Île-de-France. La densité de population est donc nettement plus faible que la moyenne régionale. La répartition géographique de la population est très inhomogène, avec une concentration élevée le long de la frontière ouest et deux zones résidentielles importantes : l'une correspond à la ville nouvelle de Marne-la-Vallée, autour de Bussy Saint-Georges, le long du RER A et de l'autoroute A4 ; l'autre le long de l'autoroute A6, en remontant vers le Nord à partir de Melun.

1.4 Jeu de données

L'opérateur de téléphonie mobile Orange a fourni des données provenant de 14 zones de localisation qui couvrent l'ensemble de la Seine-et-Marne. Ce jeu de données correspond aux événements de signalisation enregistrés par plus de 600 BTS, dont 395 sont situés à l'intérieur du périmètre géographique du département. Il est présenté plus en détail dans ce qui suit, en commençant par la volumétrie générale (§ 1.4.1), en poursuivant avec la répartition géographique des données conjointement avec la répartition géographique des populations (§ 1.4.2), et en terminant par des statistiques globales (§ 1.4.3).

1.4.1 Volumétrie générale

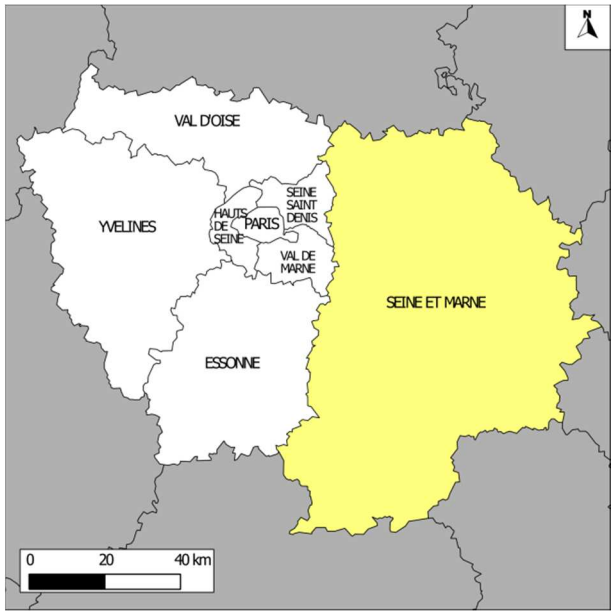
Les données utilisées par la suite dans cette section ont été collectées le samedi 22 décembre 2012, de 00h00 à 24h00. Le jour choisi étant très proche de Noël, on peut s'attendre à observer de nombreux déplacements pour motifs loisir ou achats. Les données concernent non seulement l'ensemble des abonnés de l'opérateur Orange, mais également les abonnées des opérateurs virtuels qui utilisent le réseau physique Orange, ainsi que les abonnés des opérateurs étrangers qui ont un accord d'itinérance en France avec Orange. Le jeu de données contient au total plus de 102 millions d'enregistrements, pour environ 1,85 millions de mobiles distincts. Si l'on compte uniquement les mobiles distincts vus par les BTS localisés en Seine-et-Marne[2], on arrive à un total d'environ 1,2 millions. Ce dernier chiffre correspond à environ 2,4 millions de personnes[3], à comparer à la population totale de Seine-et-Marne, d'environ 1,3 millions d'habitants. Ces chiffres sont repris dans le tableau 1.

2. C'est-à-dire en excluant les BTS localisées hors de Seine-et-Marne, mais présentes dans les données car appartenant à des zones de localisation à cheval sur la frontière départementale.

3. En considérant que la part de marché « réseau » de Orange — c'est-à-dire incluant le trafic de l'ensemble des opérateurs utilisant le réseau Orange — est de l'ordre de 50%, ce qui était en 2012 la moyenne nationale en France.

Figure 1. Cartographie de la Seine-et-Marne.

a) Localisation de la Seine-et-Marne au sein de la région Île-de-France.



b) Principales infrastructures de transport (routes et voies ferrées) en Seine-et-Marne, et localisation de la zone d'intérêt, à proximité immédiate du parc d'attraction DisneyLand Resort Paris.

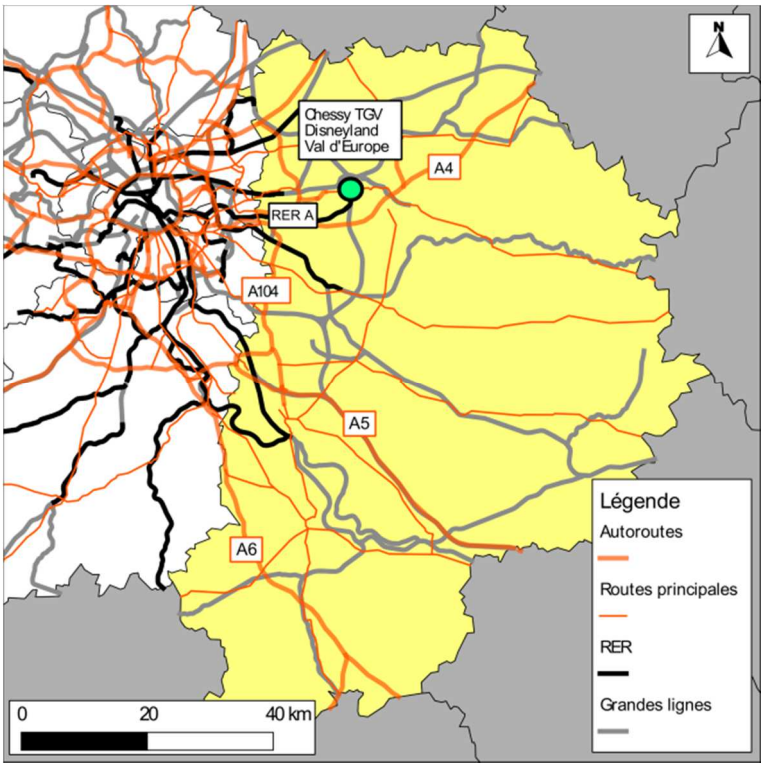


Tableau 1. Volumétrie générale des données.

	Nb. BTS	*Nb. évts*	*Nb. Tél.*	*Nb. évt/tél.*
Jeu de données complet	> 600	$102{,}01.10^6$	$1{,}85.10^6$	52,39
Seine-et-Marne seule	395	$68{,}58.10^6$	$1{,}20.10^6$	51,17

1.4.2 Répartition géographique

La répartition géographique des antennes de téléphonie suit de près celle de la population. Dans le diagramme de Voronoï de la carte des BTS en Seine et-et-Marne (Fig. 2b), le centre de chaque cellule correspond à une BTS. La comparaison entre les cartes de densité de la population au niveau communal (Fig. 2a) d'une part, et de densité des événements de téléphonie (Fig. 2b) d'autre part, est instructive. La Seine-et-Marne est divisée en 544 communes. La carte de la figure 2a représente la densité spatiale de population au niveau communal, exprimée en habitants par kilomètre carré, et calculée en divisant la population communale par la surface de la commune.

La carte de la figure 2b représente la densité spatiale de l'ensemble des événements du jeu de données. Les deux cartes illustrent la relation qui existe entre densité de population et densité du réseau téléphonique. En termes d'extension spatiale, on remarque clairement que la taille des cellules du réseau de téléphonie est d'autant plus petite que la densité de population est élevée. En termes d'activité du réseau, on remarque que la plupart des cellules dont la densité d'événements excède 100.000 événements par kilomètre carré couvrent une commune contenant une ville de plus de 15.000 habitants. Une exception notable est la zone autour de DisneyLand, où il n'existe pas à proprement parler une ville, mais où une intense activité téléphonique est observable.

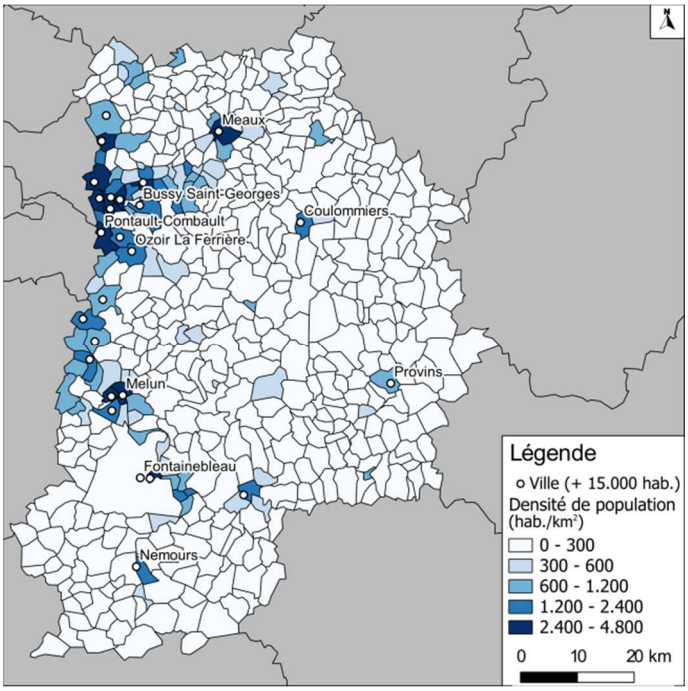
1.4.3 Statistiques

Ce paragraphe présente un ensemble de statistiques générales, dans l'objectif de mieux se figurer le contenu des données : distribution du nombre d'événements par type d'événement, distribution du nombre d'événements par mobile, et distribution des délais entre événements consécutifs.

Distribution des événements par type d'événement. La distribution des événements de signalisation par type d'événement est représentée par les camemberts de la figure 3. La signification des types d'événements principaux a été donnée précédemment, en sous-section 1.1. Pour les autres, préfixés par GMM_, il suffit de savoir qu'ils correspondent à des événements 3G de type *données*. Le camembert de gauche de la figure 3 donne la répartition des événements majoritaires. Trois types d'événements — GMM_SERVICE, COM0 et LAUN — représentent, à parts égales, environ 75% de l'ensemble des événements. GMM_SERVICE et COM0 correspondent à des communications vers les mobiles : soit de type *données* pour GMM_SERVICE ; soit de type *voix* (communication vocale ou SMS) pour COM0.

Figure 2. Cartes de densité en Seine-et-Marne.

a) Densité de population, à l'échelle de la commune.



b) Densité d'événements de signalisation, à l'échelle de la cellule.

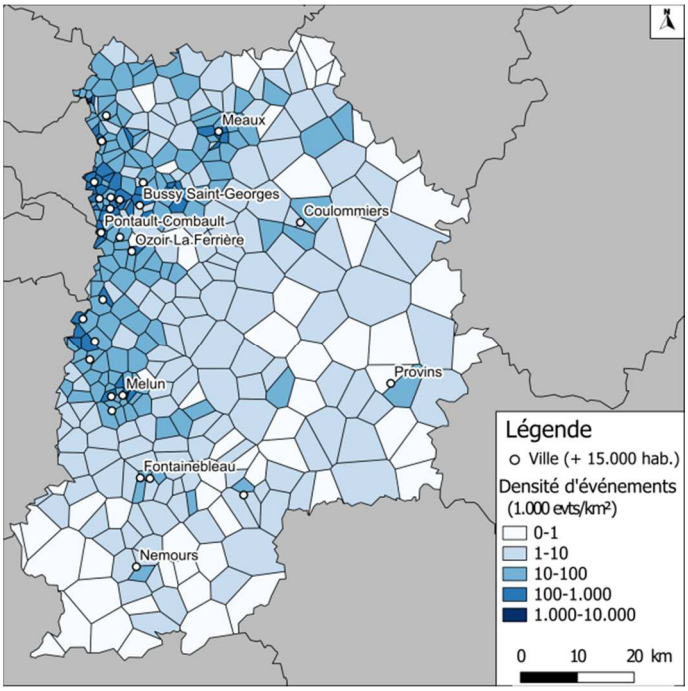
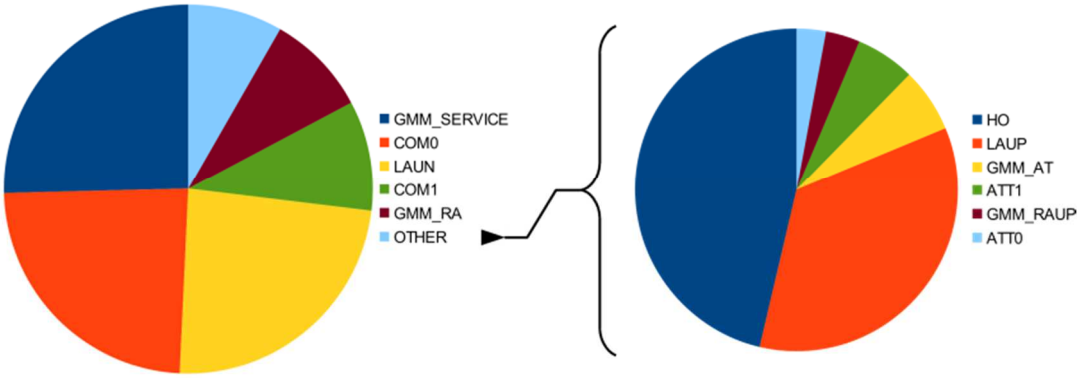


Figure 3. Distribution des événements par type d'événement.



Les événements LAUN sont des mises à jour de zones de localisation. Viennent ensuite, pour environ 10% chacun, les types COM1 (communications ou SMS émis) et GMM_RA (équivalent, pour les données, du type LAUN). Le camembert de droite de la figure 3 détaille la composition de la tranche OTHER. Les deux événements prépondérants sont ici HO et LAUP, à parts égales, environ 40% chacun.

Distribution du nombre d'événements par mobile. Indépendamment du type d'événement, observons maintenant la distribution des mobiles en fonction du nombre d'événements par mobile. Cette distribution est tracée figure 4, en fréquence et en fréquence cumulée. Dans l'ensemble, cette distribution décroît rapidement. Près de 9% des mobiles n'ont émis qu'un seul événement dans la journée sur la zone observée. Environ 15% ont émis entre 2 et 4 événements. La médiane est atteinte pour 14 événements par mobile. La distribution est assez régulière, sauf pour 8 événements, où l'on observe un pic qui sort nettement de la distribution. Un peu plus de 5% des mobiles observés ont émis exactement ce nombre d'événements durant la journée. Ce pic correspond pour l'essentiel à des mobiles ayant émis des événements de type LAUP. Si l'on se souvient qu'un événement de type LAUP est émis dans le cas où aucun événement de signalisation n'a été émis durant les trois dernières heures, le pic s'explique aisément. Il correspond pour l'essentiel à une population de mobiles connectés au réseau toute la journée, mais non utilisés. Enfin, il convient de noter que la queue de distribution est épaisse : près de 15% de l'ensemble des mobiles sont au-delà de 100 événements par jour.

Intervalle entre événements consécutifs. La distribution des intervalles de temps entre deux événements consécutifs pour un même mobile est tracée figure 5. Cette distribution a une forme assez régulière dans l'intervalle [1,170] minutes. On peut observer des pics pour des valeurs particulières, à 15, 30, 60 et 120 minutes, lesquels correspondent très certainement — comme c'est le cas du pic LAUP dans la distribution du nombre d'événements par mobile — à une particularité de fonctionnement déterministe, particularité que nous n'avons pas cherché à déterminer. Autour de 180 minutes, on observe la trace du « pic LAUP ». On notera que l'axe des ordonnées est logarithmique. Un modèle probabiliste expliquant la partie aléatoire de la distribution (i.e. hors les pics déterministes) ne serait certainement pas poissonien : dans ce cas, la loi des intervalles entre événements consécutifs serait exponentielle, alors qu'ici elle est super-exponentielle.

Figure 4. Distribution du nombre de mobiles, en fonction du nombre d'événements par mobile.

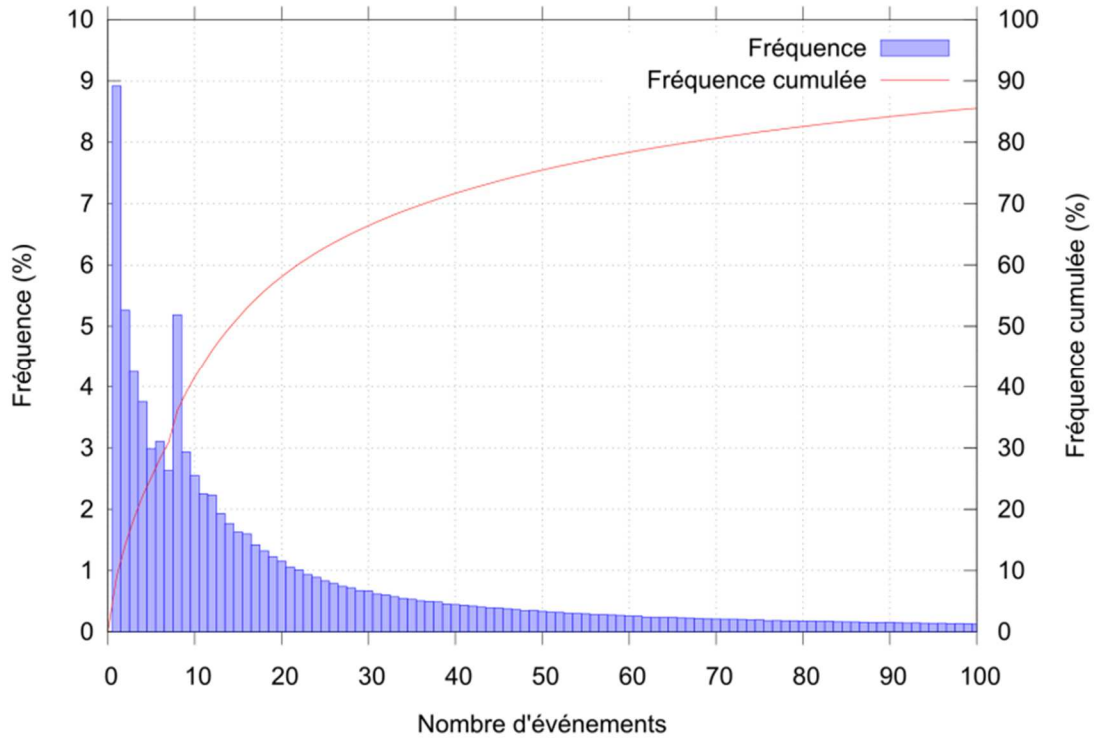
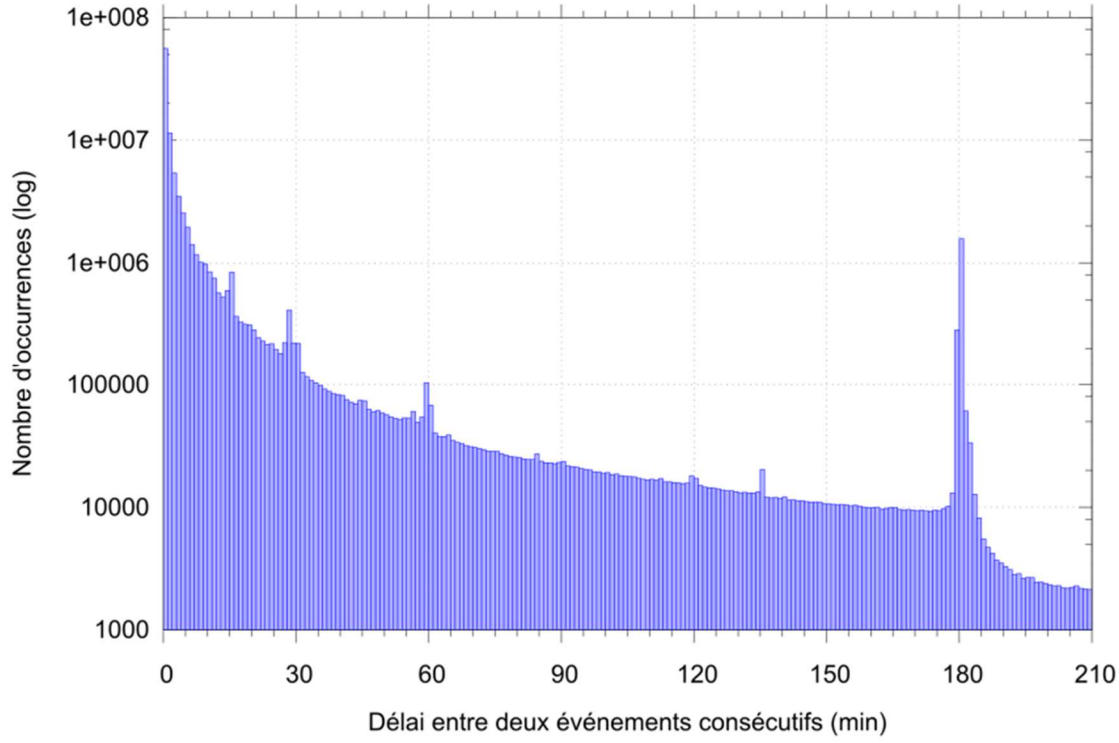


Figure 5. Distribution de l'intervalle de temps entre deux événements consécutifs pour un même mobile.



1.5 Analyse de mobilité

Fondamentalement, les données utilisées se résument à un ensemble E de triplets (m,s,h) ; avec :

- m un identifiant de mobile ;
- s un identifiant de BTS, ce qui permet de localiser géographiquement l'événement, à la précision des cellules de Voronoï de la figure 2b ;
- h l'instant auquel l'événement a été émis, avec une précision de 1 milliseconde.

Des exemples de traitements effectués sur le jeu de données E sont donnés ci-après. Ces exemples sont fondés sur la notion de trajectoire. Dans ce qui suit, on convient d'appeler trajectoire d'un mobile m la séquence — ordonnée par h croissant et notée tm — des couples (s,h) tels que (m,s,h) est dans E. Le § 1.5.1 analyse l'existence d'une corrélation entre nombre d'appels émis et comportement de mobilité, en classant les mobiles observés suivant le nombre d'appels émis et la longueur de leurs trajectoires. On s'intéresse ensuite aux mobiles peu actifs. En effet, il a été montré en section précédente qu'une fraction relativement importante de mobiles est caractérisée par une très faible activité. Les seuls événements observés sont des événements de type LAUP. Le § 1.5.2 analyse la répartition spatiale de ces mobiles. Enfin, le § 1.5.3 analyse les flux de mobiles dont la trajectoire a intersecté le centre commercial Val d'Europe.

1.5.1 Nombres d'appels et mobilité

On se pose ici la question de savoir si le comportement d'appel est corrélé ou non au comportement de mobilité. Convenons des définitions suivantes. Soit m un mobile, et t_m sa trajectoire, supposée non réduite à un point. Une paire $((s,h), (s',h'))$ formée de deux couples consécutifs de t_m est appelée un *segment*. La longueur d'un segment est la distance à vol d'oiseau séparant s de s'. La *longueur d'une trajectoire* non réduite à un point est la somme des longueurs de ses segments. Avec cette définition de la longueur d'une trajectoire, on associe à chaque mobile m : d'une part le nombre d'événements de type COM1 (appel émis ou SMS envoyé) dans la journée ; d'autre part la longueur de sa trajectoire.

De façon arbitraire, on décide de classer les mobiles dans des silos de largeur 10 dans le nombre d'événements COM1. Ainsi le silo [0;9] contient les mobiles ayant émis moins de 10 événements COM1 dans la journée ; le silo [10 ;19] ceux ayant émis au moins 10 et moins de 20 événements COM1 ; etc. Pour chacun de ces silos, on calcule la distribution de la longueur des trajectoires. Le résultat est illustré figure 6.

Les barres rouges indiquent le nombre de mobiles dans chaque silo (l'axe des ordonnées correspondant est l'axe de droite, en échelle logarithmique). Les distributions des longueurs de trajectoires sont représentées par les boîtes à moustaches en bleu (l'axe des ordonnées correspondant est l'axe de gauche). D'un silo à l'autre, les caractéristiques statistiques de ces distributions sont remarquablement stables. À l'exception du silo [0;9], dont les valeurs sont légèrement inférieures à celles des autres silos, les valeurs médianes, les écarts inter-quartiles et les valeurs extrêmes sont remarquablement proches. On n'observe pas ici, à l'échelle départementale, de corrélation manifeste entre comportement d'appel et comportement de mobilité.

Figure 6. Nombre de mobiles observés et distribution des longueurs de trajectoires, en fonction du nombre d'événements de type COM1 (appel ou SMS émis) par mobile.

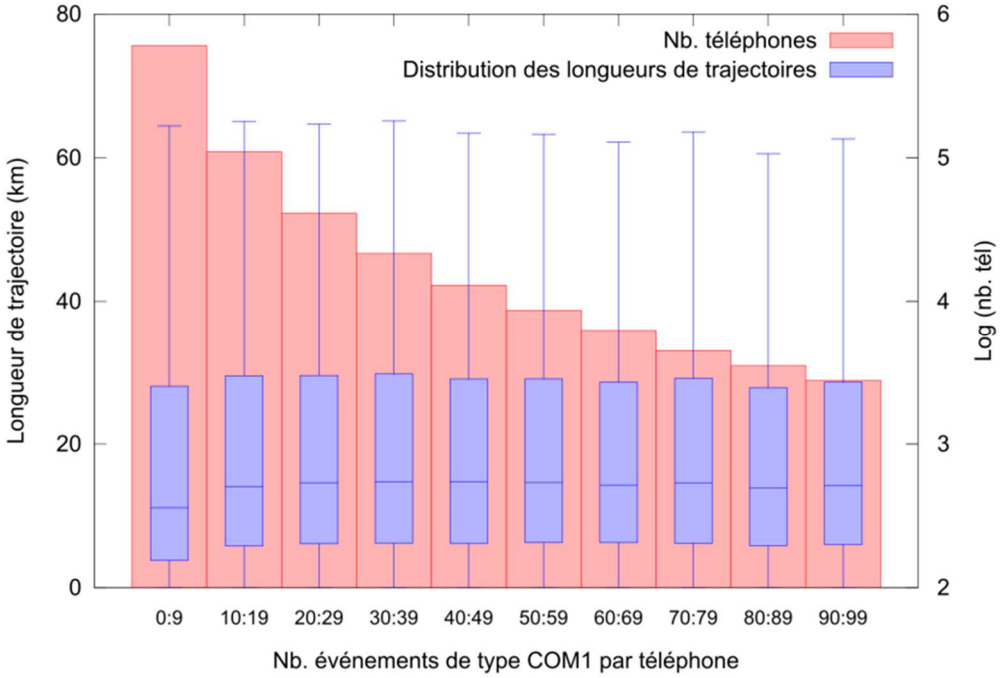
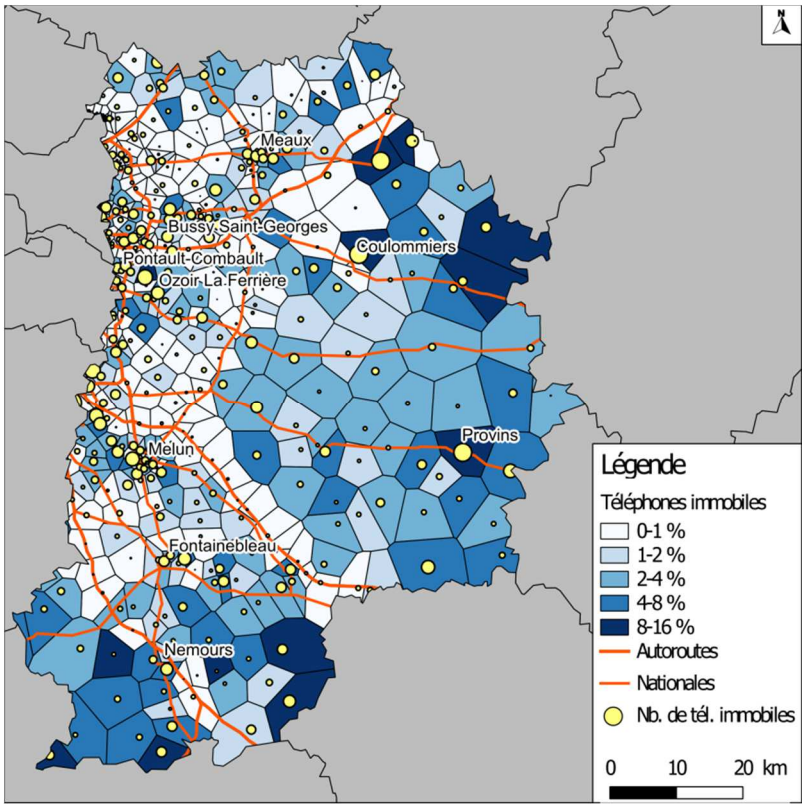


Figure 7. Cartographie de l'index d'immobilité et nombre de « mobiles immobiles » par BTS.



1.5.2 Immobilité

Les observations statistiques du § 1.4.3 montrent qu'une fraction non négligeable des mobiles sont très peu — voire ne sont pas du tout — utilisés, alors bien qu'allumés et connectés au réseau. Convenons de définir un mobile m comme *immobile* si sa trajectoire t_m vérifie les conditions suivantes :

- t_m contient au moins un segment ;
- tous les points de t_m partagent la même valeur de s, c'est-à-dire que le mobile m a toujours été vu par la même BTS ;
- un intervalle d'au moins deux heures sépare le premier point de t_m du dernier.

Parmi l'ensemble des mobiles observés, près de 90% ont une trajectoire non réduite à un point (Fig. 6). Parmi ceux-ci, 11% sont immobiles au sens défini ci-avant. En considérant maintenant une BTS s, on définit son *index d'immobilité* comme le rapport entre le nombre de mobiles immobiles en s et le nombre de mobiles distincts observés en s. Cet index d'immobilité est cartographié figure 7. Est également représenté sur cette même carte le nombre de « mobiles immobiles » par BTS. Sans surprise, les axes de transport majeurs creusent des sillons de faible immobilité. Les cellules qui présentent les plus fortes valeurs d'index d'immobilité, entre 8 et 16%, sont quant à elles situées près de villes en zone rurale (ex. : Coulommiers, Provins, La Ferté-sous-Jouarre, à l'Est de Meaux et au Nord de Coulommiers).

A priori, cet index d'immobilité est un marqueur intéressant du fonctionnement d'un territoire, qu'il conviendrait bien entendu d'interpréter en lien avec des analyses socio-démographiques plus subtiles.

1.5.3 Analyse de flux

Le troisième et dernier exemple d'utilisation des données est plus classique. L'objectif est d'étudier le flux de visiteurs du centre commercial Val d'Europe. De nombreuses analyses sont possibles à partir des données de signalisation, et susceptibles d'intéresser les acteurs du commerce de détail : durée de présence sur site ; origine géographique des visiteurs ; cheminement depuis le domicile jusqu'au centre commercial, et depuis le centre commercial jusqu'au retour au domicile, etc. Trois sont présentées ici, à savoir :

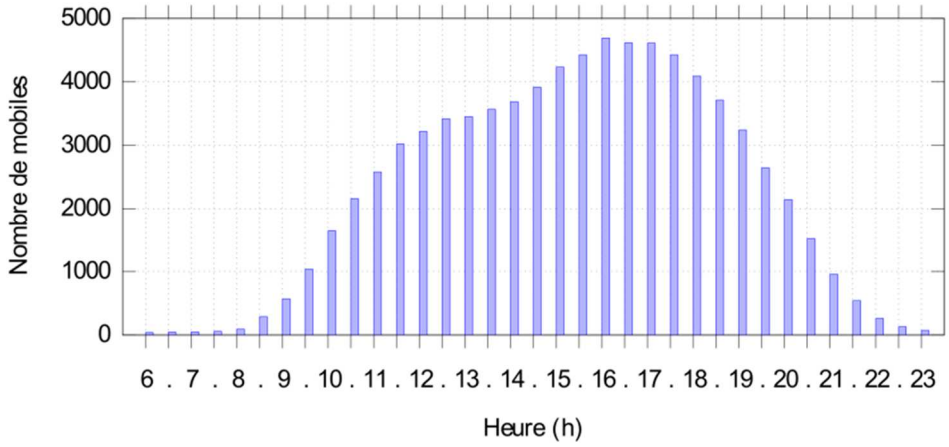
- la fréquentation du centre commercial au cours de la journée ;
- les débits de visiteurs, entrants et sortants ;
- la distribution des durées de présence sur site.

Le centre commercial Val d'Europe est un lieu public clos. Comme dans la plupart des lieux publics clos, la fourniture du service de téléphonie mobile nécessite l'utilisation d'antennes spécialisées dites « in-door ». Les antennes « in-door » de Val d'Europe sont connectées à une même BTS. Connaissant cette BTS, on considère l'ensemble des trajectoires de mobiles l'ayant intersecté au moins une fois dans la journée. À partir de cet ensemble de trajectoires, nous avons pu estimer le nombre de mobiles présents sur site par tranches d'une demi-heure. Le diagramme en bâtons de la fréquentation est donné figure 8a.

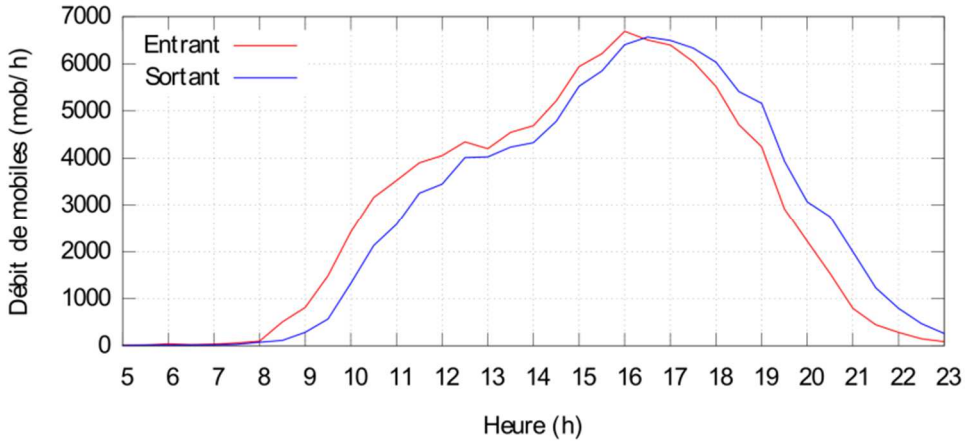
Le créneau horaire le plus chargé est 16h00-17h00, ce qui, à dire d'expert, correspond effectivement à la période de pointe classique d'un samedi. Les débits estimés, entrants et sortants, sont représentés figure 8b. Les débits les plus importants sont également observés sur le créneau 16h00-17h00. Enfin, la distribution des durées de présence estimées est tracée figure 8c, pour des tranches de largeur une demi-heure. Les durées de présence les plus fréquemment observées sont dans les tranches 1/2h à 1h et 1h à 1h30.

Figure 8. Analyse du flux de visiteurs du centre commercial Val d'Europe.

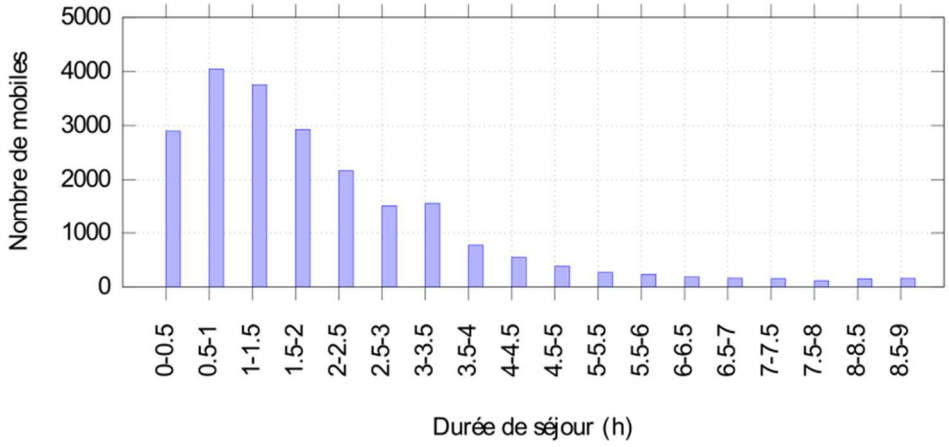
a) Nombre de mobiles présents, par 1/2 h.



b) Flux entrants et sortants.



c) Distribution des durées de présence.



Conclusion

Après avoir exposé le principe de fonctionnement des réseaux de téléphonie mobile, et en particulier des mécanismes de signalisation et de localisation, cette section a illustré, par différents exemples, le potentiel d'utilisation des données de signalisation pour l'étude du fonctionnement d'un territoire. Le département de Seine-et-Marne et le centre commercial Val d'Europe ont été pris comme zones d'études. Des éléments de conclusion complémentaires seront donnés en fin d'article, qui se poursuit par une section consacrée à la mesure d'indicateurs de qualité de service d'un système de transport en commun, le RER A.

2. Qualité de service du RER A

Les travaux présentés dans cette section sont relatifs à la mesure de la qualité de service dans un système de transports en commun, le RER A, à partir de traces numériques, billettique et téléphonie mobile. Si l'usage de données de billettique pour mesurer la qualité de service n'est pas une nouveauté, l'utilisation de données de signalisation de téléphonie mobile pour le « monitorage[4] » d'un système de transports en commun est, à notre connaissance, originale.

La suite est organisée comme suit. La sous-section 2.1 rappelle les enjeux de la mesure de la qualité de service et brosse un rapide état de l'art de l'usage de traces numériques dans ce domaine. La sous-section 2.2 présente le système étudié : structure du réseau RER A, analyse des effets de congestion, fonctionnement du système billettique et fonctionnement particulier du réseau de téléphonie mobile en milieu souterrain. La mesure de la densité de passagers par train, entre chaque station, est l'objet de la sous-section 2.3. Enfin, en sous-section 2.4, les deux sources de traces numériques, billettique et téléphonie, sont comparées pour la mesure de temps de parcours et de matrices origine-destination.

2.1 Enjeux et état de l'art

En transports en commun de voyageurs, la qualité de service est un indicateur qui intéresse plusieurs acteurs : les usagers, les opérateurs et les autorités organisatrices. Cependant, ces acteurs ne l'évaluent — et donc ne la mesurent — pas nécessairement de la même façon. Sur un trajet donné, les usagers sont concernés en particulier par le confort et la ponctualité du service. Sur le long terme, ils sont également sensibles à des indicateurs plus agrégés, comme la régularité ou la fiabilité du service. Les opérateurs, notamment dans le domaine ferroviaire, sont particulièrement attachés au respect de la grille horaire. Pour un opérateur, 90% de trains arrivant à l'heure prévue au terminus pourra être considéré comme un bon indicateur de la qualité du service. Cependant, si 90% des usagers utilisent les 10% de trains en retard, il est clair que la perception de la qualité du service par les usagers, pour ce qui concerne la ponctualité, va être sensiblement différente de celle évaluée par l'opérateur. En conséquence, les autorités organisatrices sont à la recherche d'indicateurs de qualité de service qui reflètent mieux le quotidien vécu par les usagers. Par exemple, le *Transit Capacity and Quality of Service Manual* américain définit comme indicateurs composant la qualité de service le confort, le respect des horaires, la praticité (*convenience*) ou encore la fiabilité (*reliability*).

Le confort — ou l'inconfort — ressenti par les usagers dépend du taux d'occupation du véhicule au sein duquel ils ont embarqué, ainsi que du temps passé dans une situation d'inconfort (ex. : position debout dans un véhicule saturé). Ce temps dépend lui du taux d'occupation initial au moment de l'embarquement. Plus le taux d'occupation initial est important, plus les chances d'obtenir une place assise sont faibles. Puis, le long du trajet, au fur et à mesure que des passagers embarquent et débarquent en station, les probabilités d'obtenir

4. Le terme est habituellement employé en médecine, notamment en réanimation chez les malades atteintes de troubles sévères, pour surveiller des paramètres physiologiques importants. Il paraît approprié pour évoquer le cas du RER A.

une place assise augmentent. Ainsi, une perception réaliste de la qualité de service d'un système de transports en commun, telle que perçue par les usagers, dépend en particulier :

- du taux d'occupation de chaque véhicule en chaque interstation ;
- du temps de parcours entre deux paires de stations consécutives, en fonction de l'heure de départ ;
- d'une connaissance fine des flux origine-destination et des choix d'itinéraires, notamment durant les périodes de pointe.

Les indicateurs de qualité de service sont traditionnellement établis en utilisant des méthodes de recueil de données souvent coûteuses et parfois peu fiables : observation des horaires réalisés, comptages manuels de passagers en station, enquêtes origine-destination. Si le développement de systèmes de billettique (AFC, pour *Automated Fare Collection*) et de localisation des véhicules (AVL, pour *Automated Vehicle Location*) ont permis d'améliorer la mesure de la qualité de service, l'usage de ces systèmes à cette fin connaît quelques limitations. La mesure de flux origine-destination peut en particulier être difficile.

Bertini & El-Geneidy (2003) listent de nombreux indicateurs de qualité de service pouvant être établis à partir de données AFC ou AVL, du point de vue de l'offre (vision opérateur) comme du point de vue de la demande (vision usager). Le rapport de Furth *et al.* (2006) fournit des recommandations exhaustives pour l'usage de ces données pour construire des indicateurs de qualité de service et aider à la bonne gestion d'un système de transport en commun. Les données de type AVL sont largement utilisées aujourd'hui dans les réseaux de bus. El Geneidy *et al.* (2011) montrent comment en déduire des mesures de qualité de service. Feng & Figliozzi (2011) donnent un exemple d'adaptation temps réel de l'offre à la demande à partir de données AVL.

Concernant les données de billettique, Reddy *et al.* (2009) ainsi que Nassir *et al.* (2011) décrivent des méthodes pour reconstruire des matrices origine-destination dans le cas — en pratique fréquent, car les validations ne sont souvent obligatoires qu'à l'embarquement — où seules les origines des déplacements sont connues. Un cas particulier intéressant est décrit par Frumin (2010) sur le réseau londonien, où la validation est obligatoire en sortie. Les données de téléphonie ont également été identifiées comme sources potentielles d'information pour des transports de surface. Friedrich *et al.* (2010) décrivent la construction de matrices origine-destination, sans distinction de modes. Wang *et al.* (2010) décrivent comment inférer des choix de modes en fonction de la vitesse des déplacements. Pour un état de l'art plus complet, voir Valerio (2009). Nous n'avons pas trouvé dans la littérature de précédent concernant le monitorage d'un système de transports en commun à l'aide de données de téléphonie mobile.

2.2 Le RER A

Cette sous-section présente les éléments techniques nécessaires à une bonne compréhension de la suite. Sont présentées tout d'abord, au § 2.2.1, les caractéristiques essentielles du réseau ferré souterrain de transports en commun parisien, ainsi que le fonctionnement de son système billettique. Ensuite, les effets de congestion dans le tronçon central du RER A sont analysés par l'examen statistique de tableaux horaires réalisés (§ 2.2.2). Les conditions particulières de fonctionnement du réseau de téléphonie mobile en milieu souterrain sont exposées (§ 2.2.3). Enfin, les débits horaires d'événements en station — mesurés à partir des données billettique d'une part, et des données de téléphonie d'autre part — sont comparés (§ 2.2.4).

Figure 9. Schéma du réseau ferré (métro et RER) de transports en commun, Paris et petite couronne. Le cercle en pointillé délimite approximativement la partie souterraine du réseau ferré.

2.2.1 Le réseau ferré souterrain et son système billettique

Le réseau ferré de transport en commun francilien comprend deux réseaux principaux. L'un est le métro, qui comprend 14 lignes, numérotées de 1 à 14. Le métro est pour l'essentiel souterrain et localisé dans Paris *intra-muros*. L'autre réseau est le RER (Réseau Express Régional) qui connecte la zone périurbaine à Paris par 5 lignes identifiées par des lettres, de A à E. Pour l'essentiel, le RER est souterrain dans Paris *intra-muros*, et aérien au-delà. Les deux réseaux sont représentés figure 9. Dans la suite, l'on s'intéresse à la partie souterraine des réseaux métro et RER, partie qui sera désignée par la suite *réseau ferré souterrain*. Le cercle en pointillés (Fig. 9) représente la limite approximative du réseau ferré souterrain. Toutes les stations du réseau ferré souterrain sont équipées de bornes billettique sans contact (Pass Navigo). D'après les statistiques du STIF (Syndicat des Transports d'Île-de-France), durant les périodes de pointe d'un jour ordinaire de semaine, près de 95% des déplacements sont effectués en utilisant des Pass Navigo. Du point de vue du fonctionnement du système billettique, le RER A est un système fermé. Un usager doit présenter son badge à l'entrée et à la sortie. Le métro est quant à lui un système semi-ouvert. Un usager doit présenter son badge à

l'entrée, mais pas à la sortie. Un usager du métro en correspondance avec le RER A doit présenter son badge trois fois : une première fois à l'entrée du métro ; une deuxième fois à l'entrée du RER ; et enfin une troisième fois à la sortie du RER. En cas de correspondance entre lignes de RER, par exemple à la station Châtelet pour les usagers en correspondance du RER B au RER A, le passage est libre. Du fait que les lignes A et B sont les plus chargées d'Île-de-France, une part importante des déplacements donne lieu à une paire d'événements billettique correspondant à une entrée le long de la ligne B et une sortie le long de la ligne A, ou inversement. Ces détails auront leur importance ultérieurement, en sous-section 2.4, lors de comparaisons de mesures de flux origine-destination en utilisant soit des données de billettique, soit des données de téléphonie.

2.2.2 Congestion dans le tronçon central du RER A

Le RER A est la ligne qui supporte le plus fort trafic en Île-de-France, avec en moyenne plus d'un million de trajets par jour ouvrable. La partie Est de la ligne comporte deux branches ; la partie Ouest trois. Le tronçon le plus chargé est situé entre les stations Châtelet et Auber (Fig. 9), avec 50.000 voyageurs par heure et par sens en heure de pointe du matin. Par la suite, nous désignerons par *tronçon central* la partie de la ligne située entre les stations Vincennes et Charles-de-Gaulle Étoile.

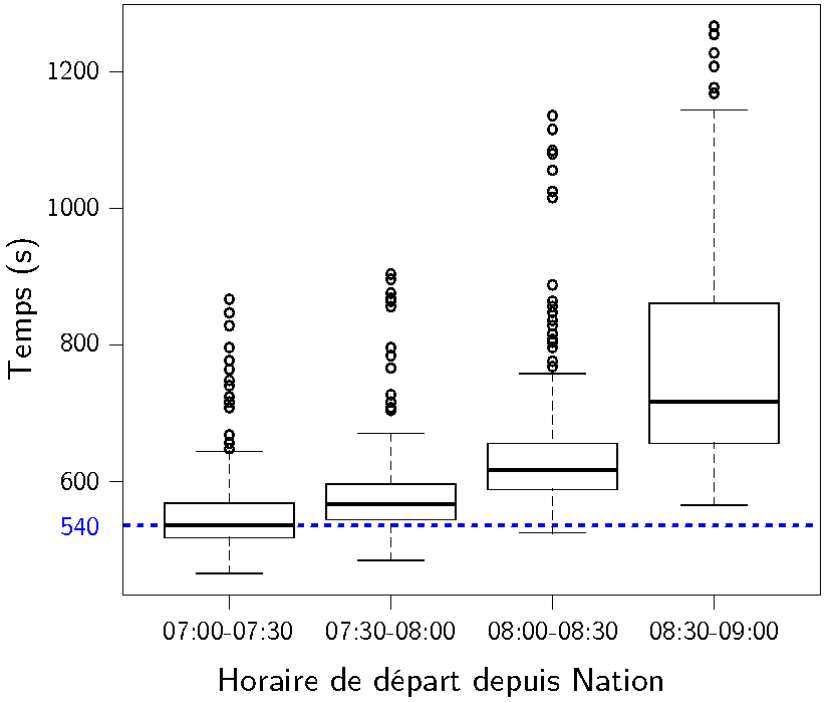
En heure de pointe, la fréquence théorique sur le tronçon central est d'un train toutes les deux minutes, soit 30 trains par heure, ce qui correspond à une capacité théorique de 60.000 voyageurs par heure. Cependant, les effets de congestion rendent cet objectif inatteignable. En pratique, le débit de trains observé durant les heures de pointe varie entre 25 et 27 trains par heure, soit une perte de capacité supérieure à 10%. Corrélativement, le RER A souffre d'irrégularité chronique dans les temps de parcours.

La qualité de service est perçue par les usagers comme très faible, du fait de l'incertitude sur les temps de parcours et du haut niveau d'inconfort dans les rames. Observons ces effets de congestion sur les temps de parcours et les temps de station à quai, en reprenant des résultats de Benezech (2013). Ce dernier a analysé — en utilisant les informations publiées en temps réel par l'opérateur du réseau, la RATP (Régie Autonome des Transports Parisiens) sur son site internet (http://www.ratp.fr) — les tables horaires du RER A, ce pour l'ensemble des jours de semaine des mois d'octobre et novembre 2011. Les résultats essentiels pour ce qui nous concerne sont représentés figure 10.

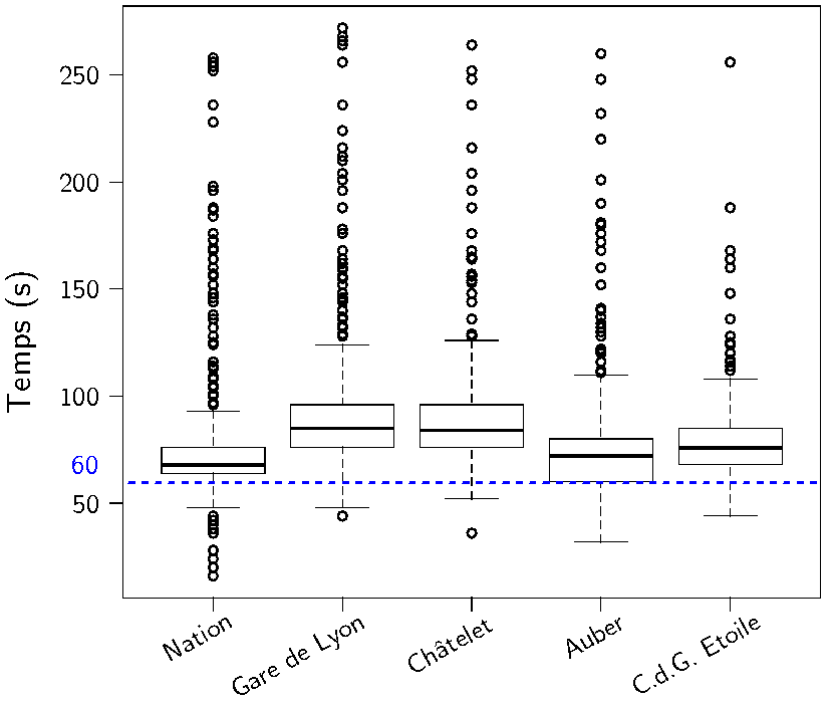
Le graphique en boîtes à moustaches de la figure 10a permet de se figurer l'évolution de la variabilité des temps de parcours entre les stations Nation et Aubert, pour quatre tranches de 30 minutes dans la période de pointe du matin, entre 07h00 et 09h00. Le temps nominal entre ces deux stations est de 9 minutes, soit 540 secondes. Ce temps nominal est représenté par la ligne bleue en pointillés (Fig. 10a). Entre 07h00 et 07h30, le temps nominal est respecté (la médiane est très proche de la valeur nominale), et la variabilité des temps de parcours est faible : les deux quartiles centraux sont proches de la médiane. Par contre, plus on avance dans la période de pointe, plus la médiane s'écarte de la valeur nominale, et plus la variabilité des temps de parcours augmente. Le maximum est ici atteint pour la tranche horaire 08h30-09h00, avec une augmentation de près de 30% de la valeur médiane et un quadruplement des écarts inter-quartiles par rapport à la tranche 07h00-07h30.

Figure 10. Congestion dans le RER A, au cours de la période de pointe du matin.

a) Distribution des temps de parcours entre les stations Nation et Aubert, par tranches de 1/2h au départ de Nation.



b) Distribution des temps à quai pour cinq stations du tronçon central.



Le graphique de la figure 10b permet d'analyser l'une des causes de ces augmentations. Le graphique représente les distributions des temps à quai observés durant les mois d'octobre et novembre 2011. La valeur nominale du temps à quai, là aussi représentée par la ligne bleue en pointillés, est de 60 secondes, dont une durée prévue de 40 secondes d'ouverture des portes. La cible est presque atteinte pour les stations Nation et Auber. Par contre, les stations Gare de Lyon et Châtelet se démarquent nettement, avec une augmentation du temps à quai de près de 30% par rapport à la cible. Ces deux stations étant d'importantes stations de correspondance — Gare de Lyon avec le réseau banlieue et TGV, Châtelet avec le RER B — il est vraisemblable d'expliquer l'augmentation des temps à quai par une augmentation des durées portes ouvertes, augmentation due aux flux importants et antagonistes de voyageurs montants et descendants. Au final, on retrouve ici le même cercle vicieux qu'en congestion routière : un excès structurel de la demande par rapport à l'offre entraîne une chute de capacité, ce qui amplifie, au cours de la période de pointe, l'écart entre l'offre et la demande.

2.2.3 Fonctionnement du réseau de téléphonie dans le réseau ferré souterrain

Le principe général de fonctionnement du réseau téléphonique en souterrain reste le même que celui exposé en section 1. Cependant, les conditions particulières de fonctionnement en souterrain imposent des adaptations. Tout d'abord, les antennes en surface ne peuvent communiquer avec les mobiles sous terre. De plus, les conditions de propagation des ondes électromagnétiques dans un milieu fermé souterrain diffèrent notablement de celles rencontrées en milieu ouvert aérien. En conséquence, le réseau téléphonique souterrain utilise des antennes dédiées, de deux types différents, suivant que l'on se trouve en station (sur le quai), ou en tunnel (entre deux stations). Les antennes en station sont des antennes d'intérieur classiques (antennes « in-door »). Les antennes en tunnel utilisent quant à elles des technologies particulières (câbles radiants ou guides d'ondes). Une particularité du réseau de téléphonie opéré par Orange dans le réseau ferré souterrain de Paris est qu'une très grande majorité des antennes souterraines, qu'elles soit en station ou en tunnel, appartiennent à une seule et même zone de localisation, qui sera désignée par la suite zone de localisation souterraine. Typiquement, un mobile qui emprunte le réseau ferré souterrain va émettre plusieurs événements de signalisation, dont des événements de mise à jour de zone de localisation (LAUN). En effet, peu de temps après l'entrée en zone souterraine, on s'attend à ce qu'un événement LAUN soit émis, puisque le mobile passe d'une zone de localisation aérienne à la zone de localisation souterraine. Ensuite, suivant l'activité téléphonique du mobile, zéro ou plusieurs événements de signalisation sont susceptibles d'être émis dans la zone de localisation souterraine. À l'issue du trajet, après que l'utilisateur soit remonté en surface, un nouvel événement de type LAUN est émis, puisque le mobile passe de la zone de localisation souterraine à une zone de localisation aérienne.

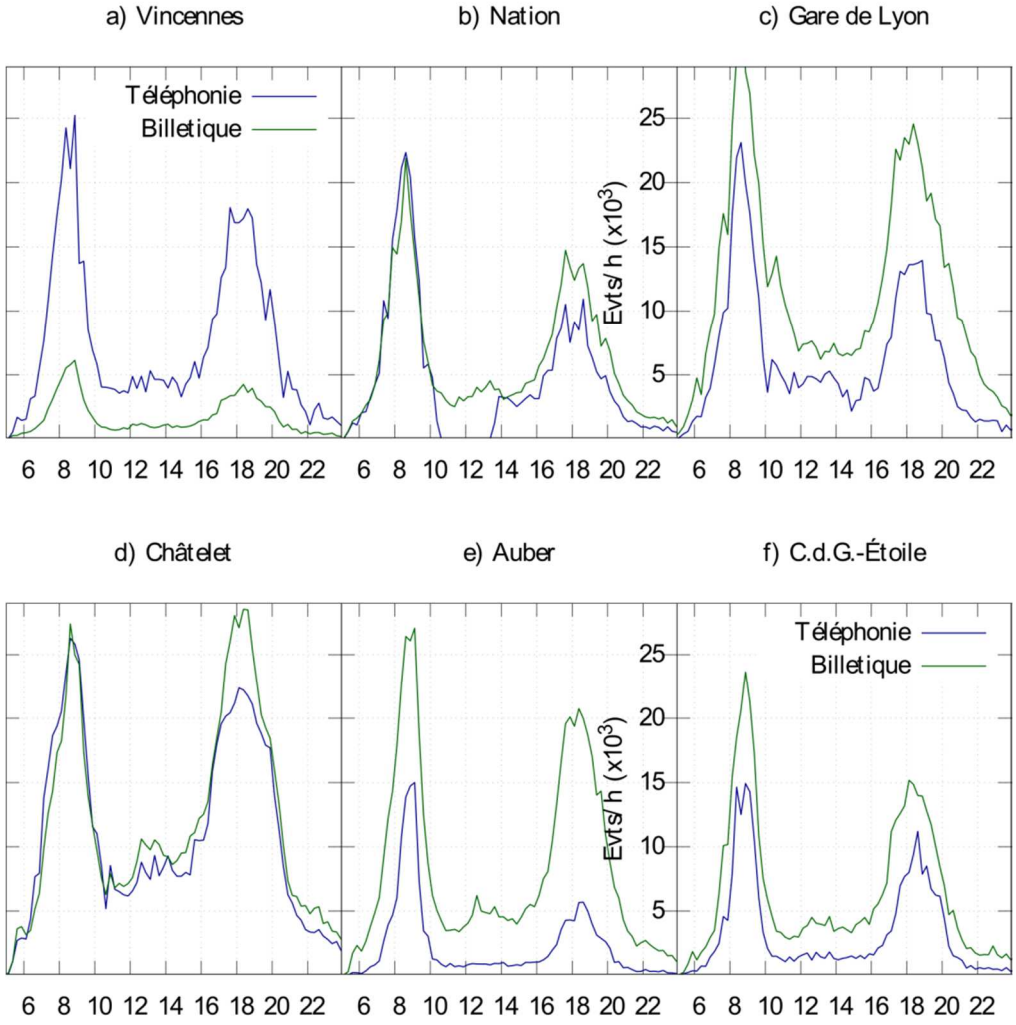
2.2.4 Comparaison entre données billettique et données de téléphonie

Les données de téléphonie et de billettique n'offrent pas les mêmes possibilités de mesures. Les données de billettique mesurent les entrées et, parfois, les sorties et les correspondances. Lorsque ces mesures existent, elles sont déterministes : au comportement de fraude et au dysfonctionnement des capteurs près, chaque badge laisse une trace. Les données de téléphonie mesurent quant à elles des trajectoires partielles, avec un fort caractère aléatoire. En théorie, les points d'entrée et de sortie de la zone de localisation souterraine sont identifiables par des événements LAUN. Entre ceux deux points terminaux, on peut trouver un nombre très variable de points intermédiaires, dépendant en particulier de l'activité téléphonique du mobile.

La figure 11 permet de comparer globalement ces deux sources de données. Les courbes de la figure représentent, pour six stations du tronçon central, les variations des débits horaires

d'événements de téléphonie et de billettique. Un événement de billettique correspond à une validation (entrée ou sortie) en station. Un événement de téléphonie correspond à l'émission par un mobile d'un événement de signalisation. Ces courbes correspondent aux données recueillies pour la journée du 13 octobre 2011. Pour chaque intervalle de 15 minutes, la courbe bleue représente le débit horaire du nombre d'événements vus par l'ensemble des antennes de la BTS en station. De même, pour chaque intervalle de 15 minutes, la courbe verte représente le débit horaire du nombre de validations sur l'ensemble des bornes billettique de la station.

Figure 11. Comparaison des flux d'événements de téléphonie (en bleu) et de billettique (en vert) sur six stations du tronçon central.



Les deux jeux de données ne fournissent pas la même information. D'un côté les données de téléphonie concernent les mobiles embarqués dans les trains, les mobiles à quai ou en tunnel piéton, entre l'entrée de la station et le quai. De l'autre côté, les données de billettique comptent le nombre de validation en entrée et en sortie, et ce depuis ou vers l'extérieur de la station (surface ou station de métro). Malgré ces différences, l'analyse comparée des courbes de la figure 11 est instructive. Les périodes de pointe du matin et du soir émergent clairement. À une constante multiplicative près — qui dépend de la station et de la période de pointe considérée —, les deux flux varient simultanément. Pour toutes les stations, à l'exception notable de Vincennes, le flux billettique est supérieur au flux téléphonie.

La différence observée à Vincennes (Fig. 11a) s'explique aisément, pour les deux raisons suivantes. Premièrement, cette station est située en tunnel, et la ligne est aérienne de part et d'autre : sur 1 km dans le sens Est-Ouest, vers Nation, et jusqu'au terminus dans le sens Ouest-Est. Ainsi, quel que soit le sens de circulation, tous les mobiles sont censés émettre un événement LAUN à la station Vincennes.

Deuxièmement, Vincennes n'est pas, en volume de passagers, une station majeure du réseau RER, ni une station de correspondance : le nombre de passagers qui embarquent ou débarquent à la station Vincennes est de façon générale faible devant le nombre de passagers d'un train. Il est donc normal d'observer un débit d'événements de téléphonie bien supérieur au débit d'événements billettique.

L'inverse est vrai pour la station Auber (Fig. 11e). Située loin des entrées de tunnel de chaque côté de la ligne, les antennes de téléphonie à Auber ne verront que peu d'événements LAUN de mise à jour de zone de localisation. Pour les stations Nation (Fig. 11b) et Châtelet (Fig. 11d), le fait que les deux courbes se superposent durant la matinée est pure coïncidence. Le creux autour de 12h00 dans les données de téléphonie à Nation est dû à un dysfonctionnement passager du réseau de téléphonie.

2.3 Densité de passagers

Le taux d'occupation est défini comme le ratio entre le nombre de passagers à bord d'un train et sa capacité. Il s'agit d'un déterminant essentiel de la qualité de service. L'objectif de cette section est de proposer une méthode de mesure du nombre de passagers à partir des données de téléphonie, et de vérifier son applicabilité. Deux jeux d'observations ont été utilisés. Le premier, décrit au § 2.3.1, est constitué d'observations terrain réalisées le 7 avril 2011 sur le tronçon central. Le second jeu de données contient les événements de signalisation enregistrés dans la zone de localisation souterraine le même jour. Il est décrit au § 2.3.2.

Les données de téléphonie n'enregistrent pas la trajectoire complète d'un téléphone : comme indiqué précédemment (§ 2.2.3), le nombre et la localisation des événements de signalisation émis par les mobiles peuvent varier significativement d'un mobile à l'autre. Ce point est examiné plus en détail, en comparant les trajectoires des trains reconstituées à partir des observations terrain aux trajectoires observées à partir des données de téléphonie (§ 2.3.3). Nous montrons ensuite que le nombre de mobiles à bord de chaque train peut être estimé à une constante multiplicative près (§ 2.3.4). Sous certaines hypothèses, cette constante — qui varie pour chaque inter-station —, peut être estimée (§ 2.3.5). Par suite, en utilisant l'information sur la capacité des trains provenant des observations terrain, il est possible d'estimer le taux d'occupation de chaque train, pour chaque inter-station. Ces estimations sont comparées aux observations terrain (§ 2.3.6).

2.3.1 Observations terrain

Trois types de trains circulent sur la ligne A du RER. Suivant le type de train, le nombre de places assises et la capacité totale varient. Pour chaque type de train, la capacité correspond à un taux d'occupation nominal de 4 passagers/m^2. Le tableau 2 détaille ces caractéristiques pour les types de matériels circulant sur la ligne.

Tableau 2. Caractéristiques des matériels circulant sur la ligne A du RER.

Type de train	*Places assises*	*Capacité totale*
MS61	432	1.760
MI84	600	1.900
MI2N	1.056	2.580

Entre 07h15 et 09h03, pour chaque rame circulant dans le sens Est-Ouest, et pour chaque station entre Vincennes et Charles-de-Gaulle Étoile, six enquêteurs du LVMT ont relevé à quai :

- le type de matériel ;
- une estimation du taux d'occupation de la voiture située en face d'eux, en utilisant l'échelle suivante :
  - **faible** : des sièges sont disponibles, les passagers sont quasiment tous assis ;
  - **normal** : quasiment aucun siège n'est disponible, quelques passagers sont debout ;
  - **élevé** : un grand nombre de passagers sont debout ;
  - **saturé** : un grand nombre de passagers sont debout et très serrés.
- les horaires correspondant aux événements suivants : arrivée du train en station, arrêt à quai, ouverture des portes, fermeture des portes, démarrage à quai, départ de la station.

2.3.2 Données de téléphonie

Orange a fourni l'ensemble des événements de signalisation enregistrés le 7 avril 2011 dans la zone de localisation souterraine. Dans ce qui suit, on considère uniquement les événements enregistrés par les BTS des stations Vincennes, Nation, Gare de Lyon, Châtelet, Auber et C.d.G.-Étoile. Ce jeu de données, noté par la suite L, contient essentiellement des triplets (m,h,s) avec :

- m un identifiant de mobile ;
- h un instant dans la journée, à la précision d'une milliseconde ;
- s une station du tronçon central.

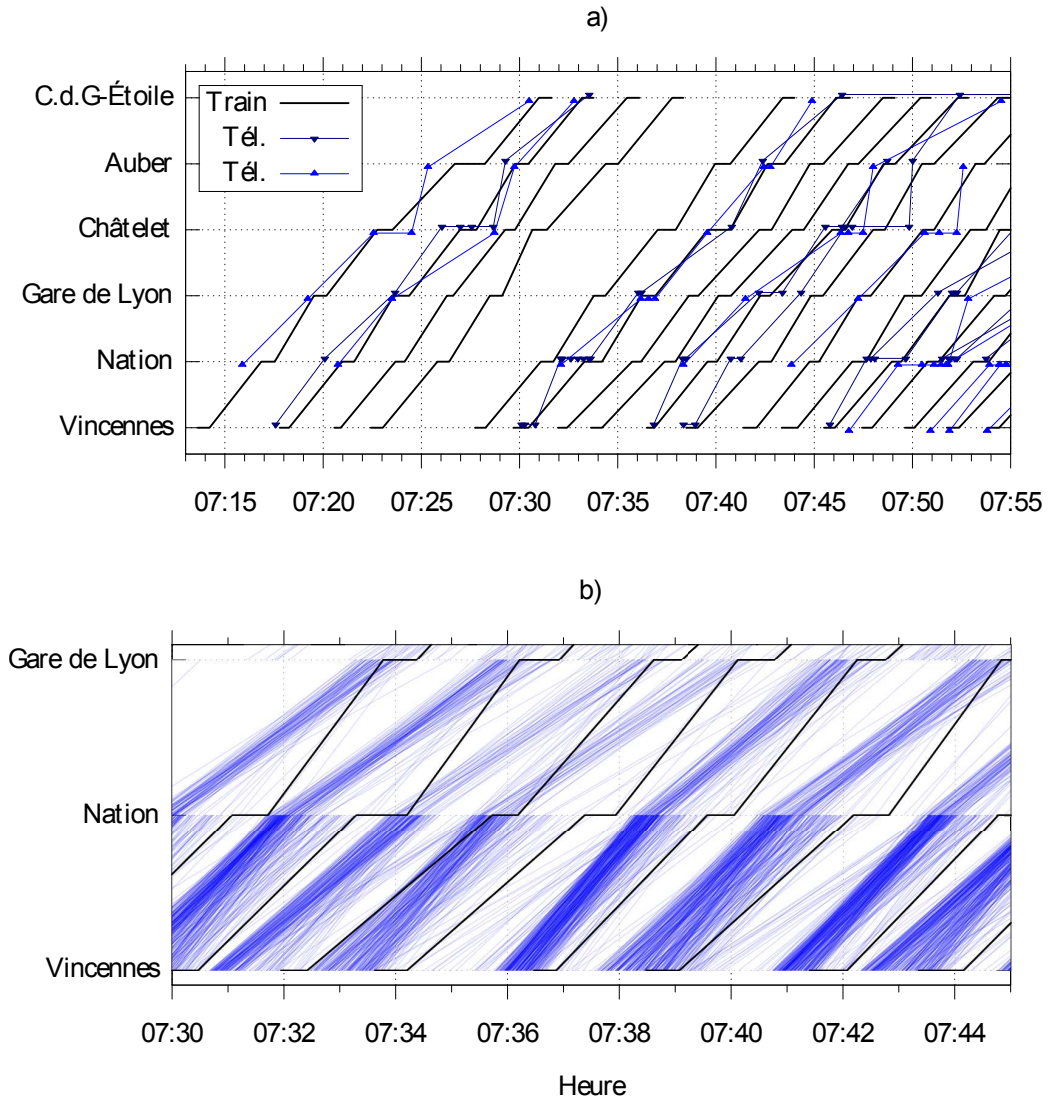
L contient environ 1 million d'enregistrements, pour environ 293.000 identifiants de mobiles distincts. En moyenne, chaque mobile apparaît donc 3,65 fois dans L. 32% des mobiles n'apparaissent qu'une fois. 63% apparaissent trois fois ou moins. 4% apparaissent dix fois ou plus.

2.3.3 Trajectoires des trains

Les trajectoires des trains observées par les enquêteurs à quai sont tracées dans le diagramme temps-espace de la figure 12a, pour les horaires de départ depuis Vincennes compris entre 07h15 et 07h55, ainsi que quelques-unes des *5-trajectoires* contenues dans L.

Une *n-trajectoire* est un sous-ensemble T de L tel que tous les triplets de T partagent une même valeur de m, et tel que, en ordonnant les triplets de T par h croissant, n stations du tronçon central apparaissent successivement dans le bon ordre. Comme l'on peut s'y attendre à partir des statistiques du § 2.3.2, les 2-trajectoires sont beaucoup plus nombreuses que les 5-trajectoires. La figure 12b montre les 2-trajectoires observées entre les stations Vincennes et Nation d'une part, et entre les stations Nation et Gare de Lyon d'autre part, pour des horaires de départ depuis Vincennes compris entre 07h30 et 07h44. On peut visuellement constater que l'arrivée de chaque train en station est corrélée avec un pic dans le débit de téléphones observés. Entre les stations Vincennes et Nation les rafales de 2-trajectoires précèdent l'arrivée du train en station parce que la plupart des mobiles émettent leur LAUN dans le tunnel précédent l'arrivée en station, avant que le train n'arrive en station. De façon similaire, les LAUN observés avant Nation sont émis dans le tunnel entre Vincennes et Nation, avant que le train ne s'arrête à Nation.

Figure 12. Trajectoires des trains et des mobiles. a) ensemble des trajectoires des trains sur le tronçon central, et exemples de 5-trajectoires de mobiles. b) exemples de 2-trajectoires.



Il faut remarquer aussi que les trajectoires de mobiles ne se superposent pas nécessairement aux trajectoires de trains. Certaines 2-trajectoires semblent correspondre à des sauts d'un train à l'un des trains suivants. Il ne s'agit pas d'erreur de mesures, mais bien de comportements possibles de la part des usagers. En effet, dans le sens Est-Ouest, les RER ne circulent pas entre toutes les combinaisons possibles entre les deux branches Est et les trois branches Ouest. Ainsi les usagers souhaitant effectuer un trajet depuis le Nord-Est vers le Nord-Ouest doivent nécessairement débarquer dans l'une des stations du tronçon central pour effectuer une correspondance.

2.3.4 Nombre de 2-trajectoires par train

La figure 13 est une représentation alternative des 2-trajectoires de la figure 12b. Figure 13, une 2-trajectoire entre Vincennes et Nation est représentée par un point dans le plan. Ce point a pour abscisse l'instant auquel le téléphone a été détecté à Vincennes, et pour ordonnée un instant, s'il existe, auquel le téléphone a été détecté à Nation après avoir été détecté à

Vincennes. Formellement, un point apparaît dans ce plan ssi il existe un couple (l_1,l_2) de L^2, avec $l_1=(m_1,h_1,s_1)$ et $l_2=(m_2,h_2,s_2)$, et tel que :

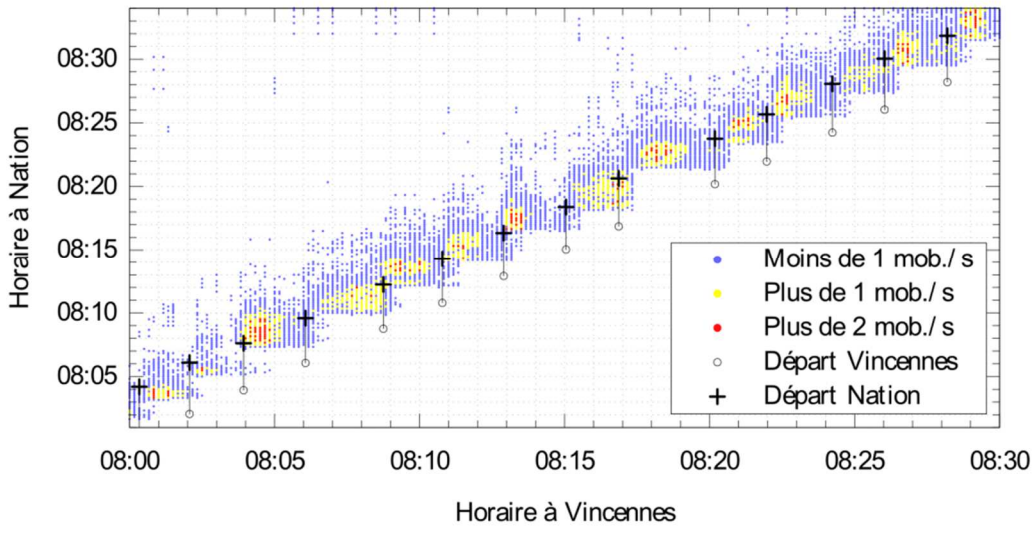
$$(m_1 = m_2) \wedge (h_1 < h_2) \wedge (s_1 = \text{Vincennes}) \wedge (s_2 = \text{Nation})$$

Pour un mobile m, si plus d'une paire (l_1,l_2) vérifie cette condition, un seul point est sélectionné arbitrairement, de sorte que ce point minimise le temps de parcours $h_2 - h_1$. Le plan est discrétisé par une grille régulière, au pas de 10 secondes sur chaque axe, et on compte le nombre de points par cellule de cette grille. Cette grille est représentée figure 13, en associant une couleur à chaque cellule, en fonction du nombre de points qu'elle contient. Le bleu correspond à un débit maximal de 1 mob./s ; le jaune à un débit de mobiles compris entre 1 et 2 mob./s ; le rouge à un débit supérieur à 2 mob./s.

Quant aux trajectoires des trains — telles que mesurées à partir des observations terrain — elles sont représentées figure 13 par des segments de droites. Les cercles sur la diagonale correspondent aux instants de départ depuis la station Vincennes. La croix à l'autre extrémité de chaque segment indique l'horaire de départ depuis la station Nation.

Des techniques standard de segmentation (détection de contour et croissance de région) permettent de partitionner cette carte de densité des 2-trajectoires de sorte à associer à chaque 2-trajectoire un train, et, par conséquent, à associer à chaque train, pour chaque inter-station, un nombre de mobiles. Pour un même effectif de mobiles, le nombre ainsi obtenu dépend de la paire de stations successives considérées, et plus précisément de la probabilité pour un mobile d'être détecté en deux stations consécutives. Pour les raisons suggérées au § 2.2.4 cette probabilité varie avec la paire de stations successives considérée, ainsi qu'avec l'heure de la journée.

Figure 13. Densité de 2-trajectoires entre Vincennes et Nation.



2.3.5 Probabilité de détection d'un téléphone

On note :

- $P(s)$ la probabilité pour un mobile d'être observé à la station s ;
- s^- (resp. s^+) la station en aval (resp. en amont) de s ;

et on suppose que $P(s^-)$, $P(s)$ et $P(s^+)$ sont indépendantes. Alors, le nombre de mobiles $N_{t,s}$ à bord d'un train t à la station s pour lequel le nombre $n_{t,s}$ de 2-trajectoires a été observé est :

$$N_{t,s} = n_{t,s} \,/\, P(s)P(s^-)$$

Reste à estimer $P(s)$. C'est l'objet de ce paragraphe. Écrivons que :

$$P(s) = P(s \mid s^+ \cap s^-) = P(s^+ \cap s \cap s^-) \;/\; P(s^+ \cap s^-)$$

En d'autre termes : la probabilité $P(s)$ pour qu'un mobile soit observé à la station s, sachant qu'il a été observé en s^+ (la station en amont de s) et en s^- (la station en aval de s), est le ratio entre (i) la probabilité que ce mobile ait été observé dans les trois stations s^+, s et s^-, et (ii) la probabilité qu'il ait été observé en s^+ et en s^-.

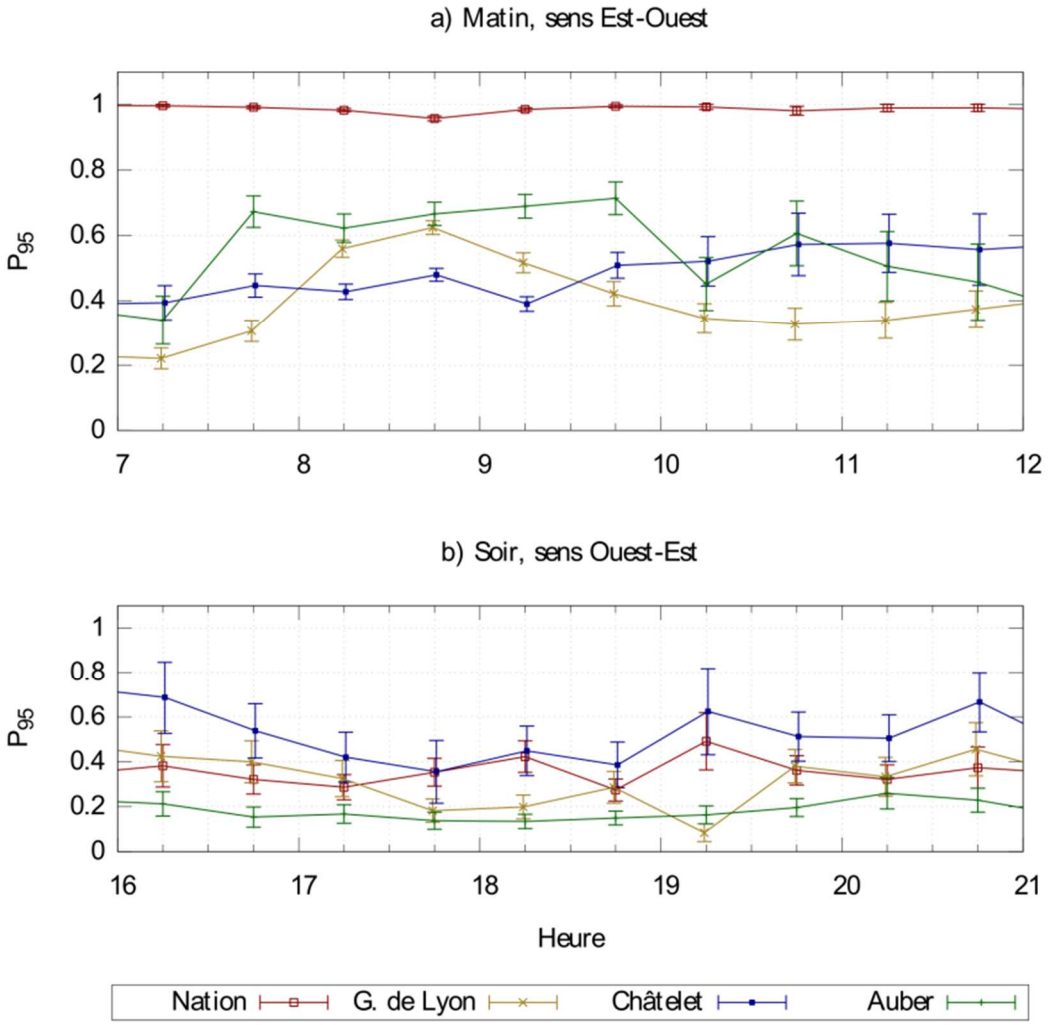
On peut ainsi construire un estimateur simple de $P(s)$. En notant $n_{s+ \cap s \cap s-}$ le nombre de mobiles observés dans les trois stations s^+, s et s^-, et $n_{s+ \cap s-}$ le nombre de mobiles observés aux deux stations s^+ et s^-, on obtient que :

$$f = n_{s+ \cap s \cap s-} / \, n_{s+ \cap s-}$$

est un estimateur de la probabilité $P(s)$ pour un mobile d'être observé en s. L'intervalle de confiance à 95% est donné par :

$$P_{95}(s) = f \pm 1.96 \, (f(1-f) \,/\, n_{s+ \cap s-})^{1/2}$$

Figure 14. Variations de l'estimateur P_{95}.



L'estimateur P_{95} dépend de s, de la direction, ainsi que de l'heure dans la journée. Ses variations sont représentées graphiquement figure 14a, pour tous les intervalles consécutifs de longueur 30 minutes entre 07h00 et 12h00 pour les quatre stations Nation, Gare de Lyon, Châtelet et Auber, dans la direction Est-Ouest. À la station Nation, P_{95} est quasiment égal à 1 pendant toute la matinée, et l'intervalle de confiance est étroit. Le matin, la plupart des mobiles ayant été observés à Vincennes et à Gare de Lyon sont également observés à Nation. À la station Châtelet, l'indicateur P_{95} augmente lentement durant la matinée, pour passer de 0,4 à 0,5. À la station Gare de Lyon, l'indicateur P_{95} est contenu dans l'intervalle [0,4 ; 0,5] durant la période de pointe du matin, et dans l'intervalle [0,2 ; 0,4] en dehors de la période de pointe du matin. La station Auber se comporte de façon quasi-identique à la station Gare de Lyon, avec un plateau remarquable durant la pointe du matin.

2.3.6 Taux d'occupation des trains

Comme montré au § 2.3.4, on peut, à partir des 2-trajectoires, associer à chaque train t et pour chaque paire de stations consécutives (s,s^+) un nombre $n_{t,s}$ de mobiles. À partir de l'estimateur P_{95} défini ci-avant (§ 2.3.5), le taux d'occupation du train t entre les stations s et s^+, noté $O_{t,s}$, est estimé par :

$$O_{t,s} = n_{t,s} / (a\, K_t\, P_{95}(s)\, P_{95}(s^+))$$

avec K_t la capacité du train t (cf. Tab. 2) et a la part de marché réseau[5].

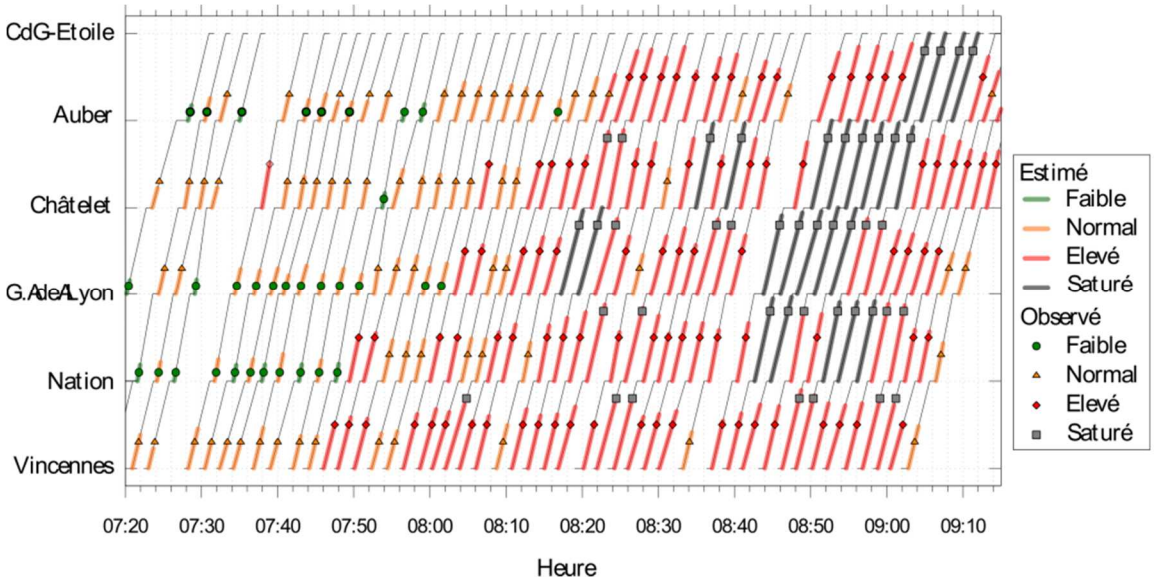
Les taux d'occupation, soit estimés à partir de l'équation ci-avant, soit observés sur le terrain (cf. § 2.3.1), sont représentés figure 15. Les taux d'occupation observés sont représentés par des symboles colorés (rond, triangle, losange, carré). Les taux d'occupation estimés sont représentés par des segments colorés dont la longueur, proportionnelle au taux d'occupation estimé, est tronquée à 100%. La couleur est fonction du taux d'occupation, sur la même échelle discrète que les observations. Dans l'ensemble, estimations et observations sont en bon accord : plus de 80% des valeurs estimées sont conformes aux valeurs observées. Lorsqu'ils existent, les écarts sont faibles et semblent pour l'essentiel explicables par des effets de seuil et les difficultés que peut avoir un observateur à quai à se représenter le taux d'occupation d'un train complet. Par exemple, pour la station Gare de Lyon entre 07h30 et 08h03, dix observations sont classées *faible* alors que les estimations sont classées *normal*.

Conclusion

Les expériences et résultats décrits dans cette sous-section 2.3 ont été réalisés à partir d'un jeu de données correspondant à la date du 7 avril 2011, et limité à la zone de localisation souterraine. À partir des événements de signalisation, nous avons pu estimer, en bon accord avec des observations terrain, les taux d'occupation des trains en chaque paire de stations consécutives sur le tronçon central. Chronologiquement, il s'agissait de fait de notre première expérience d'utilisation des données de téléphonie. Nous nous sommes volontairement limité à l'utilisation des seules données de la zone de localisation souterraine pour i) découvrir les contraintes expérimentales liées à l'acquisition et au traitement de ces données, relativement volumineuses et ii) être certains des conditions d'observation des déplacements observés, de sorte à ne pas être confrontés dans un premier temps aux difficultés d'interprétation des données de surface.

5. La part de marché réseau comprend le trafic induit par les opérateurs virtuels utilisant le réseau physique Orange. En 2001, d'après http://goo.gl/1HSI7, elle était égale à 46,6%.

Figure 15. Taux d'occupation des trains.



2.4 Temps de parcours et matrice origine-destination

Les données utilisées en sous-section 2.3 concernaient uniquement la zone de localisation souterraine, ce qui ne permet pas par exemple de capturer les destinations des déplacements souterrains. À cette fin, nous avons construit une deuxième expérience, à plus large échelle, et utilisant elle aussi deux jeux de données. Le premier est composé des événements de signalisation d'une cinquantaine de zones de localisation — dont, bien entendu, la zone de localisation souterraine — les autres couvrant en surface une large bande le long du RER A, ce afin de pouvoir capturer les événements LAUN de sortie de la zone souterraine. Le deuxième jeu de données est composé de données billettique.

L'objectif était de comparer ces deux sources de données entre elles, sur deux types de mesures : des temps de parcours et des matrices origine-destination. Cette sous-section présente les résultats obtenus lors de cette deuxième expérience. Les données billettique sont décrites dans un premier temps (§ 2.4.1) ; puis viennent les données de téléphonie (§ 2.4.2) ; ensuite la mesure de temps de parcours (§ 2.4.3) ; et enfin la mesure de matrice origine-destination (§ 2.4.4).

2.4.1 Données billetique

Le STIF a mis à disposition les moyens permettant le traitement de données de validation Navigo sur l'ensemble du RER A, pour l'ensemble de la journée du 13 octobre 2011, pour un total de $1{,}9.10^6$ enregistrements. Chaque enregistrement contient un identifiant pseudonymisé de Pass Navigo, un horaire, un identifiant de station, et un type d'événement (entrée ou sortie). Ce jeu de données contient 648.000 identifiants distincts, soit en moyenne 2,71 événements par badge. La distribution du nombre de badges, en fonction du nombre d'événements observés par badge, est tracée figure 16 (barres vertes). La plupart des badges ont été détectés soit deux fois dans la journée — ce qui peut correspondre à des trajets domicile-travail pendulaires en correspondance avec une autre ligne de RER — soit quatre fois dans la journée — déplacements pendulaires entre deux stations du RER A. Les nombres de badges ayant une ou trois validations dans la journée (Fig. 16) est relativement important, et en tout état de cause supérieur à ce que l'on peut *a priori* imaginer.

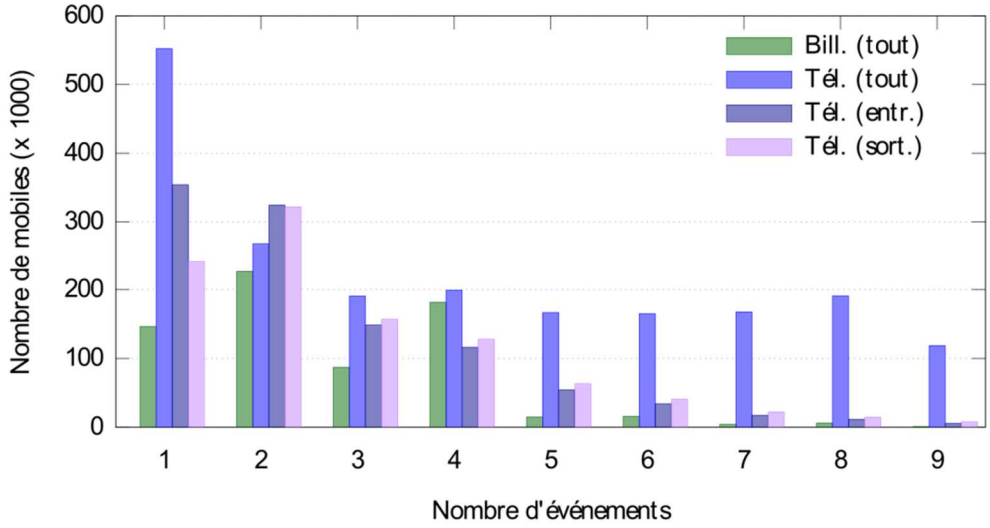
2.4.2 Données de téléphonie

Ce jeu de données contient des événements de signalisation intervenus le 13 octobre 2011 dans 49 zones de localisation. L'une d'entre elles est la zone de localisation souterraine. Les 48 autres sont des zones de localisation classiques, en surface, qui couvrent une large bande autour du tronçon central du RER A. Cette zone de couverture aérienne comprend plus de 8.000 antennes, et couvre en particulier un nombre important de stations de métro en connexion avec le tronçon central du RER A. Le jeu de données — noté E — contient 29.10^6 enregistrements pour un total de $2{,}9.10^6$ mobiles distincts, soit une moyenne de 10 événements par téléphone.

Chaque élément e de E est un 5-uplet $e=(m,c,h,l,l')$, avec :

- m un mobile ;
- c une BTS ;
- h un instant ;
- l la zone de localisation de c ;
- l' la zone de localisation précédente, dans le cas d'un événement de type LAU.

Figure 16. Distribution des badges Navigo (en vert) et des mobiles (en bleu) en fonction du nombre d'événements observés.

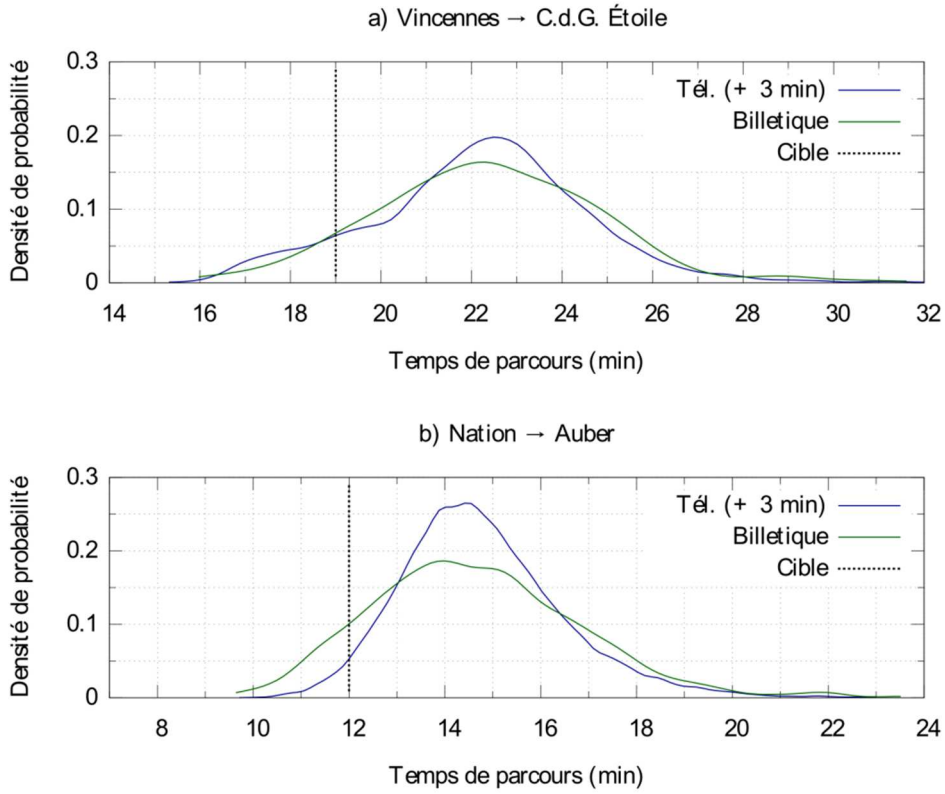


Événements entrants et sortants. Pour chaque 5-uplet $e=(m,c,h,l,l')$ de E, si l est l'identifiant de la zone de localisation souterraine et si l' est l'identifiant d'une zone de localisation de surface, e est qualifié d'événement *entrant*. À l'inverse, si l est l'identifiant d'une zone de localisation aérienne et si l' est l'identifiant de la zone de localisation souterraine, alors e est qualifié d'événement *sortant*.

La distribution du nombre de mobiles en fonction du nombre total d'événements durant la journée est représentée figure 16. Sont également tracées les distributions du nombre de mobiles en fonction du nombre d'événements entrants (resp. sortants). Pour un nombre d'événements (entrant ou sortant) supérieur ou égal à 2, ces distributions sont très comparables. Le jeu de données capture bien les entrées/sorties des mobiles qui sont entrés et sortis plus d'une fois dans la journée de la zone de localisation souterraine. Par contre, ces deux valeurs diffèrent pour les téléphones qui n'ont effectué qu'un événement entrant (resp. sortant) au cours de la journée. La raison en est que le jeu de données capture l'ensemble des

événements entrants, mais que les sorties qui ont lieu en dehors de la zone de couverture des 48 zones de localisation en surface ne sont pas détectées.

Figure 17. Comparaison entre les temps de parcours obtenus à partir des données billettique (en vert) et les données de téléphonie (en bleu), journée du 13/10/2011. a) de la station Vincennes vers la station C.d.G.-Étoile, dans la matinée. b) de la station Nation vers la station Auber, après-midi et soirée.



2.4.3 Temps de parcours

Les deux jeux de données — billettique et téléphonie — contiennent suffisamment d'enregistrements pour estimer les distributions de temps de parcours entre deux stations quelconques du tronçon central. Un temps de parcours mesuré à partir des données de billettique est de borne à borne : il comprend le temps de marche depuis la borne d'entrée jusqu'au quai, le temps d'attente à quai, et le temps de marche du quai jusqu'à la borne de sortie. Les temps de parcours mesurés à partir des données de téléphonie correspondent ici à des 2-trajectoires entre stations non nécessairement consécutives. La figure 17 représente des fonctions de densité de la distribution des temps de parcours : dans la matinée, depuis la station Vincennes vers la station C.d.G.-Étoile (Fig. 17a) ; dans l'après-midi et en soirée, depuis la station Nation vers la station Auber (Fig. 17b). Dans les deux cas, les fonctions de densité obtenues à partir des données de billettique d'une part, et des données de téléphonie d'autre part, sont comparables — à une translation près de +3 minutes pour les données de téléphonie. Cette différence est aisément explicable par le fait que les données de billettique comprennent des composantes — temps de parcours de borne à quai et temps d'attente à quai — absentes des données de téléphonie. Dans les deux figures 17a et 17b, le trait en pointillé illustre le temps de parcours nominal tel que prévu par l'opérateur. Plus de 80% des usagers subissent un temps de parcours supérieur au temps nominal. Plus de 40% subissent un temps

de parcours supérieur de 20% au temps nominal. Bien que non illustré ici, on vérifie également que l'évolution de la distribution des temps de parcours par tranches de 1/2h au cours de la période de pointe du matin est très comparable à celle observée durant deux mois d'observation (Fig. 10a).

2.4.4 Mesure de matrice origine-destination

L'objet est ici d'estimer des flux origine-destination depuis des stations du tronçon central vers la station La Défense. Le quartier de La Défense est le plus grand quartier d'affaires d'Île-de-France. Géographiquement, la station RER La Défense est la première, dans le sens Est-Ouest, après C.d.G. Étoile. Les flux estimés à partir des données billettique correspondent aux trajets avec une entrée dans le tronçon central, et une sortie à La Défense. Ainsi les usagers en correspondance de RER — en particulier en provenance du RER B à la station Châtelet — ne sont pas pris en compte. Par contre, les usagers en correspondance de métro — par exemple à la station Nation — le sont. Les flux estimés à partir des données billettique sont tracés figure 18 (bâtons de gauche). La période de pointe s'étend de 08h00 à 10h00, avec une hyper-pointe entre 8h30 et 9h30. La plupart des trajets ont pour origine la station Auber, avec près de 40% des trajets ayant cette origine durant l'hyper-pointe. Pour obtenir des valeurs comparables avec les données de téléphonie, des traitements supplémentaires sont nécessaires. Il faut en effet prendre en compte le rabattement métro vers RER présent dans les données de billettique. Voici comment nous avons procédé.

Considérons un mobile m ayant effectué un événement entrant $e = (m,l_e,h_e)$ et un événement sortant $e = (m,l_s,h_s)$ dans la matinée, i.e. avec $0 < h_e < h_e < 12$h. Il faut tout d'abord pouvoir déterminer si l'événement sortant s a eu lieu à proximité de la Défense. Si tel est le cas, il faut ensuite pouvoir déterminer si le lieu l_e de l'événement entrant e est un lieu vraisemblable d'origine d'un déplacement en RER vers La Défense depuis une station du tronçon central. En effet, l'événement entrant peut avoir eu lieu dans une station du métro. Un modèle simple d'affectation, détaillé ci-après, a été utilisé à cette fin.

Le voisinage de surface de La Défense. Certains mobiles ont émis un événement souterrain à La Défense, puis un événement en surface. Parmi ceux-ci, plus de 95% l'ont fait en moins de trois minutes. On considère, arbitrairement, que l'ensemble des BTS de surface correspondantes définissent le *voisinage de surface de La Défense*. Par suite, parmi la population des trajectoires de mobiles de l'ensemble E, celles qui contiennent un événement sortant dans le voisinage de surface de La Défense sont considérées comme ayant pour destination La Défense.

Temps de parcours entre stations souterraines. Soit (u,v) une paire de stations souterraines distinctes. Notons

$$T_{u,v} = \{(o,d) \in E^2,\ o.c \in u,\ d.c \in v,\ o.m=d.m,\ o.h<d.h\ \}$$

l'ensemble des 2-trajectoires entre u et v. Si $T_{u,v}$ n'est pas vide, on définit le temps de parcours entre u et v pour un départ depuis la station u à l'heure h, noté $t_{u,v}(h)$, par :

$$t_{u,v}(h) = d.h - o.h\ ,\ (o,d) = \begin{cases} \text{argmin}_{t \in T'_{u,v}} \{t.d.h\} & \text{si } T'_{u,v} \text{ est non vide} \\ \text{argmax}_{t \in T_{u,v}} \{t.o.h\} & \text{sinon} \end{cases}$$

avec

$$T'_{u,v} = \{\ (o,d) \in T_{u,v}\ ,\ h < o.h\ \}$$

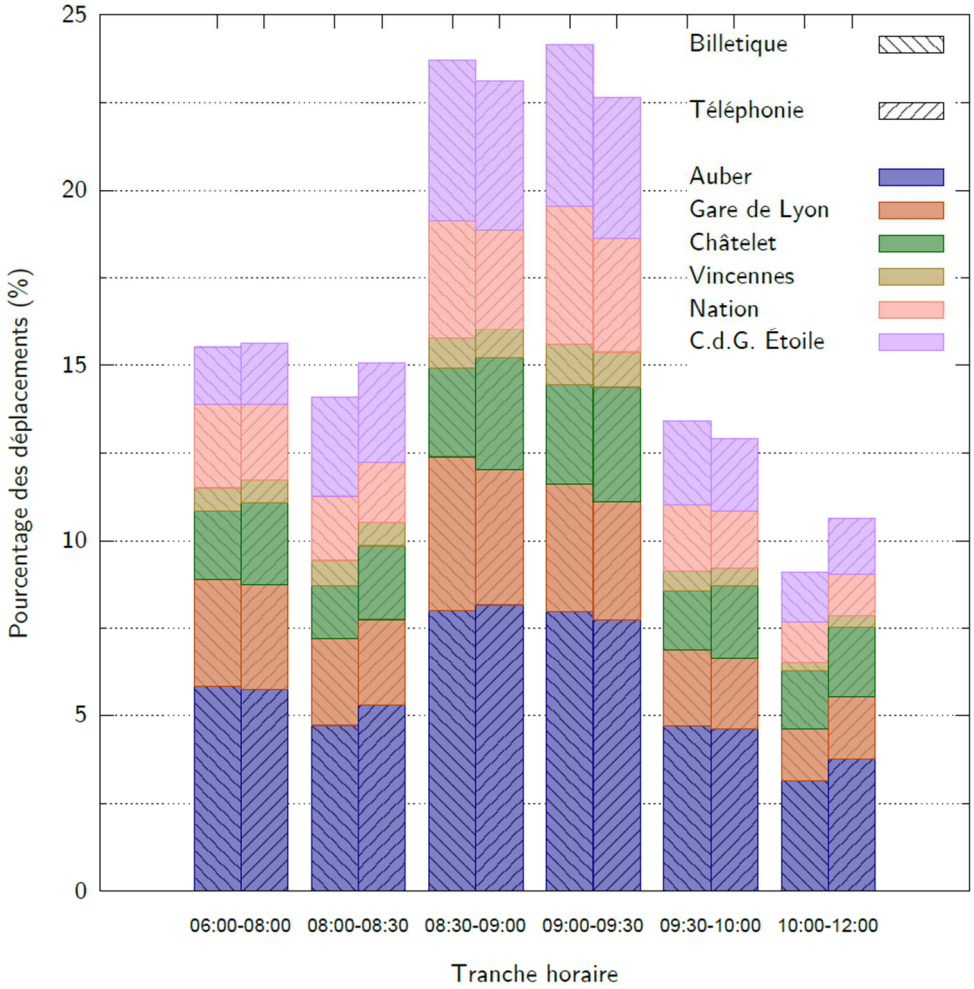
Modèle simple d'affectation. Soit un mobile m ayant (i) émis un événement sortant d avant 12h00 dans le voisinage immédiat de la Défense et (ii) ayant émis un événement entrant avant d. Soit o son dernier événement entrant précédant d. Le temps de parcours entre o et d est

$t_m = d.h - o.h$. En supposant que le trajet inclut une correspondance avec l'une des stations du tronçon central, alors il existe une station du tronçon central, notée s^*, qui minimise la différence entre t_m et le temps de parcours entre o et d en connectant à s^*, c'est-à-dire :

$$s^* = \text{argmin}_s \{ (t_m - (t_{o,s}(o.h) + t_{s,d}(o.h + t_{o,s}(o.h))))^2 \}$$

Dans une affectation « tout ou rien », on considérerait que si la différence entre t_m et le temps de parcours de la route passant par s^* excède un seuil arbitrairement fixé (ex. : 10%), l'hypothèse serait rejetée. Sinon, la trajectoire serait affectée à la paire (s^*,d). Sur ce principe, mais pour mieux distribuer les trajectoires sur les différentes stations, nous avons utilisé, plutôt qu'une affectation « tout ou rien », une affectation logit entre les alternatives viables. Les résultats sont illustrés par la figure 18 (bâtons de droite), pour différentes tranches horaires dans la matinée. Pour chacune de ces tranches, le nombre total de téléphones affectés à La Défense depuis l'une des stations centrales représente 50% ± 2% du nombre de badges Navigo. Pour chacune des tranches horaires considérées dans la matinée, la proportion de mobiles affectés à chaque station origine dans le tronçon central est très proche de celle issue des données billettique. La plupart des différences proviennent d'une sur-estimation de l'origine Châtelet dans les données de téléphonie. Il peut s'agir d'un biais systématique introduit par le modèle simple d'affectation mis en œuvre.

Figure 18. Comparaison entre les mesures de flux OD obtenus à partir des données billettique (bâtons de gauche) et les données de téléphonie (bâtons de droite) depuis les stations du tronçon central vers La Défense, pour plusieurs tranches horaires dans la matinée du 13/10/2011.



Conclusion

Cette section a montré que le traitement de données de téléphonie permet la mesure des quantités nécessaires à l'évaluation d'indicateurs de qualité de service : densité de passagers par train, temps de parcours, matrices origine-destination. Si des travaux complémentaires de validation et de calibration sont nécessaires, le potentiel révélé est intéressant à plusieurs titres.

Premièrement, les données existent et ne requièrent pas de déploiement particulier de capteurs spécifiques, tout en répondant à plusieurs besoins. Les risques, financiers et technologiques, d'un déploiement opérationnel sont donc *a priori* limités.

Deuxièmement, les indicateurs proposés ne nécessitent pas l'utilisation de données de la part de l'opérateur de transport. La disponibilité d'un tel instrument de mesure indépendant peut être intéressante lorsque les relations entre usagers, opérateurs de transport et autorité organisatrice des transports revêtent un caractère conflictuel.

Troisièmement, les données de téléphonie permettent d'étudier simultanément, et à des échelles spatiales et temporelles très fines, les interactions entre offre et demande de transport. Cela ouvre la voie à l'étude et à la compréhension de phénomènes difficiles à capturer avec des moyens classiques d'observation, qu'il s'agisse de phénomènes transitoires comme l'adaptation des comportements en situation de crise (ex : grèves, interruptions de service), ou de phénomènes récurrents comme les évolutions de choix de route ou de mode en fonction du taux d'occupation.

Conclusion

Les travaux présentés dans cet article ont illustré le potentiel d'utilisation de traces numériques de déplacements pour l'étude de la mobilité des personnes. Les données de signalisation de téléphonie mobile ont constitué la trame de fond, complétée en section 2 par des données de billettique et des données de « web mining ». La section 1 a présenté une application au diagnostic de fonctionnement d'un territoire, en proposant des exemples d'analyses sur le territoire de la Seine-et-Marne. La section 2 a été consacrée à la mesure d'indicateurs de qualité de service dans le réseau ferré souterrain parisien. Ces travaux, de nature exploratoire, partaient de l'objectif de disposer d'un instrument de mesure des déplacements robuste, partout disponible, et pouvant répondre à un large éventail des questions que posent le fonctionnement et l'exploitation des réseaux de transport, indépendamment du mode considéré. Revenons un instant sur la disponibilité et la robustesse.

D'un point de vue technique, le caractère de disponibilité universelle semble acquis. Les opérateurs de téléphonie ont déployé, et continuent de déployer, des réseaux ayant une très large couverture géographique — et en tout état de cause une couverture la plus complète possible des zones à forte densité d'occupation humaine, y compris dans les souterrains des systèmes de transport collectifs urbains. De ce fait, la garantie de disponibilité des données, là où se posent des questions transport, surpasse de loin celle d'autres types de traces numériques telles que les traces GPS. Se posent maintenant des questions de disponibilité opérationnelle, qui sortent du champ de la recherche. Il va s'agir, dans un futur proche, et pour différents acteurs (opérateurs téléphoniques, opérateurs de transport, usagers des différents réseaux, autorités organisatrices, ...) de définir des conditions d'exploitation de ces données, dans le respect de différentes contraintes, à commencer celles liées au respect de la vie privée.

Le caractère de robustesse semble également acquis : le taux d'équipement de la population et les volumes de données sont tels que les régularités statistiques constatées sont, sauf à remettre en cause les lois des grands nombres, difficilement réfutables. Ceci n'exclut pas, à l'évidence, l'existence de biais possibles. C'est très certainement dans la compréhension et le

redressement de ces biais que se trouveront des complémentarités entre méthodes d'enquêtes classiques et traitement de traces numériques.

Bibliographie

Aguiléra V., Allio S. *et al.* « Estimating the Quality of Service of Underground Transit Systems with Cellular Network Data ». *Procedia - Social and Behavioral Sciences*, 2012, 48, pp. 2262-2271.

Aguiléra V., Allio S. *et al.* « Using cell phone data to measure quality of service and passenger flows of Paris transit system ». *Transportation Research Part C : Emerging Technologies,* 2013.

Aguiléra V., Allio S., Milion C. « Territory analysis using cell-phone data », in Proc. of the 5th Transportation Research Arena, Paris, France, 2014.

Ahas R., Aasa A. *et al.* « Evaluating passive mobile positioning data for tourism surveys : An Estonian case study ». *Tourism Management*, 2008, 29(3), pp. 469-486.

Bar-Gera H. « Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times : A case study from Israel ». *Transportation Research Part C : Emerging Technologies*, 15(6), pp. 380-391, 2007.

Benezech V. *Travellers' experience of quality of service in urban transport networks : a stochastic approach*. Thèse de doctorat, École Doctorale Ville Transport Territoires, Univ. Paris Est, 2013.

Bertini R., El-Geneidy, A. « Generating transit performance measures with archived data ». *Transportation Research Record : Journal of the Transportation Research Board*, 2003, 1841, pp. 109-119.

Caceres N., Romero L. M. *et al.* « Traffic Flow Estimates Inferred from Mobile Phone Networks », in Proc. of the 12th World Conference on Transportation Research (WCTR), Lisbon, Portugal, 2010.

Calabrese F., Diao, M. *et al.* « Understanding individual mobility patterns from urban sensing data : A mobile phone trace example ». *Transportation Research Part C : Emerging Technologies*, 2013, 26, pp. 301-313.

Chen J., Bierlaire M. « Probabilistic multimodal map-matching with rich smartphone data ». *Journal of Intelligent Transportation Systems*, 2014.

El-Geneidy A., Horning J., Krizek K. « Analyzing transit service reliability using detailed data from automatic vehicular locator systems ». *Journal of Advanced Transportation*, 2011, 45(1), pp. 66-79.

Feng W., Figliozzi M. « Using Archived AVL/APC Bus Data to Identify Spatial-Temporal Causes of Bus Bunching », in Proc. of the 90th Transportation Research Board Annual Meeting, Washington D.C., U.S.A., 2011.

Friedrich M., Immisch K. *et al.* « Generating Origin-Destination Matrices from Mobile Phone Trajectories », *Transportation Research Record : Journal of the Transportation Research Board*, 2011, 2196, pp. 93-101.

Frumin M. *Automatic Data for Applied Railway Management : Passenger Demand, Service Quality Measurement, and Tactical Planning on the London Overground Network*. Ph.D. thesis, Massachusetts Institute of Technology, 2010.

Furth P. G., Hemily B. *et al.* (2006). « Using Archived AVL-APC Data to Improve Transit Performance and Management », TCRP REPORT 113, Transit Cooperative Research Program of the Transportation Research Board.

Gonzalez M. C., Hidalgo C. A., Barabasi, A. L. « Understanding individual human mobility patterns ». *Nature*, 2008, 453, pp. 779-782.

Hofleitner A., Herring R., Bayen A. « Arterial travel time forecast with streaming data : A hybrid approach of flow modeling and machine learning ». *Transportation Research Part B : Methodological*, 2012, 46(9), pp. 1097-1122.

Kang, C., Ma, X. *et al.* « Intra-urban human mobility patterns : An urban morphology perspective ». *Physica part A : Statistical Mechanics and its Applications*, 2012, 391(4), pp. 1702-1717.

Lind G., Lindkvist, A. « OPTIS - Optimised traffic in Sweden ». Rapport technique, Movea Traffic Consultancy Ltd, 2006.

Linnartz J. M. W., Hamerslag R. « Monitoring the Bay Area Freeway Network using Probe Vehicles and Random Access Radio Channels » California PATH research report ITS-PRR-94-23, U. C. Berkeley, California, U.S.A, 1994.

Nassir N., Khani A. *et al.* « Transit Stop-Level Origin-Destination Estimation Through Use of Transit Schedule and Automated Data Collection System ». *Transportation Research Record : Journal of the Transportation Research Board*, 2011, 2263, pp. 140-150.

Ratti C., Williams S. *et al.* « Mobile landscapes : using location data from cell phones for urban analysis ». *Environment and Planning part B : Planning and Design*, 2006, 33(5), pp. 727-748.

Reddy A., Lu A. *et al.* (2009). « Application of Entry-Only Automated Fare Collection (AFC) System Data to Infer Ridership, Rider Destinations, Unlinked Trips, and Passenger Miles ». *Transportation Research Record : Journal of the Transportation Research Board*, 2009, 2110, pp. 128-136.

Rubio A., Sanchez A., Frias-Martinez, E. « Adaptive non-parametric identifica-tion of dense areas using cell phone records for urban analysis ». *Engineering Applications of Artificial Intelligence*, 2013, 26(1), pp. 551-563.

Valerio D. (2009). *Road Traffic Information from Cellular Network Signaling*. Rapport technique, Forschungszentrum Telekommunikation Wien, 2009.

Wang H., Calabrese F. *et al.* « Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records », in Proc. of the 13th International IEEE Conference on Intelligent Transportation Systems, 2010, pp.318-323.

Ygnace J. « Travel time and speed estimates on the French Rhône corridor network using cellular phones as probes ». Rapport technique INRESTS/LESCOT, programme SERTI V, projet STRIP (System for Traffic Information and Positioning). Lyon, France, 2001.