Note technique de l'Observatoire de la mobilité de la Région de Bruxelles-Capitale

L'apport des *big data* pour l'étude de la mobilité en Région de Bruxelles-Capitale : enjeux, opportunités et défis

Par Thomas Ermans, Céline Brandeleer et Michel Hubert

BRUXELLES MOBILITÉ
SERVICE PUBLIC RÉGIONAL DE BRUXELLES

Les auteurs

Thomas Ermans est géographe (ULB) et titulaire d'un master complémentaire en analyse de données statistiques (UGent). Chercheur au Centre d'Études Sociologiques (CES) de l'Université Saint-Louis – Bruxelles depuis 2014, il travaille principalement sur la mobilité urbaine au sein notamment de l'Observatoire de la Mobilité. Il a rejoint l'Institut Bruxellois de Statistiques et d'Analyses (IBSA – perspective.brussels) en 2019.
Contact : termans@perspective.brussels

Céline Brandeleer est politologue (USL-B/UCL). Chercheuse au Centre d'Études sociologiques (CES) de l'Université Saint-Louis – Bruxelles, depuis 2014, elle travaille principalement sur la mobilité urbaine, l'analyse de l'action publique et les inégalités sociales de mobilité, notamment au travers des *Cahiers* de l'Observatoire de la Mobilité de la RBC. Elle a rejoint l'Institut Bruxellois de Statistiques et d'Analyse (IBSA – perspective.brussels) en 2019.
Contact : cbrandeleer@perspective.brussels

Michel Hubert est docteur en sociologie, professeur ordinaire à l'Université Saint-Louis – Bruxelles, où il préside l'Institut de recherches interdisciplinaires sur Bruxelles (IRIB) et professeur visiteur au centre METICES de l'Université libre de Bruxelles (ULB). Il dirige aussi, depuis sa création en 2006, la revue *Brussels Studies* et est vice-président du *Brussels Studies Institute (BSI)*. Dans le cadre de ses recherches, il étudie notamment les pratiques de mobilité, ainsi que l'histoire et la structure des réseaux de transport et leur impact sur la ville et ses usagers. Michel Hubert coordonne depuis leur création les *Cahiers* de l'Observatoire de la mobilité.
Contact : michel.hubert@usaintlouis.be

La rédaction de la note technique "L'apport des *big data* pour l'étude de la mobilité en Région de Bruxelles-Capitale : enjeux, opportunités et défis" s'est clôturée en septembre 2018.
C'est pourquoi les sources mentionnées dans le document s'arrêtent à une date largement antérieure à sa date de publication.
Toutefois, les réflexions qui s'y trouvent restent valables.

Sommaire

Introduction

L'explosion des *traces* numériques dans une société toujours plus connectée a présidé au développement des *big data* (ou données massives). Ces dernières sont porteuses d'opportunités mais imposent de surmonter de nombreux défis techniques et de mobiliser, pour en faire ressortir l'information utile, des méthodes qui tranchent avec l'outillage "classique" des chercheurs, des administrations, bureaux d'études et, plus largement, des citoyens intéressés.

Cette nouvelle donne appelle un (re)positionnement de ces acteurs d'autant plus important que la montée sur le devant de la scène des *big data* répond à un engouement très vif dans le cadre de l'émergence de l'imaginaire des *smart cities* (ou villes intelligentes) qui lui confère un rapport de force avantageux vis-à-vis des méthodes et données plus classiques. Certaines questions sont ainsi posées : Faut-il encore réaliser des enquêtes de mobilité dans une ville *hyperconnectée* qui enregistre la moindre *trace* laissée par les personnes en mouvement ? Quelle place pour les études stratégiques dans la *smart city* ?, etc.

Cette note a donc pour objectif de fournir des pistes de réflexions quant aux positionnements possibles vis-à-vis des *big data*. En particulier, nous souhaitons investiguer la place qu'on peut leur attribuer dans l'outillage méthodologique des "experts", dans un sens très large, et les modalités de leur insertion dans les processus d'action publique.

Dans le détail, nous désirons dans cette note (1) synthétiser les éléments qui distinguent la production de connaissance dans le cas des *big data* – (2) étayer les rapports entre *big data* et *smart city* et ce qu'ils impliquent, non seulement en termes de production de savoir mais aussi d'action publique, en ce qui concerne particulièrement les modalités de passage entre le savoir produit et son *action* sur le réel – (3) approfondir les enjeux et défis qui se posent aux administrations publiques et aux chercheurs pour l'étude de la mobilité, notamment en matière de représentativité et de contextualisation des données, de propriété des données, de compétences techniques et de respect de la vie privée – et enfin (4) évoquer ces enjeux et défis de façon plus concrète autour de trois sources de données : les *floating mobile data (FMD)*, les *floating car data* (*FCD*) et les données de validation automatique des billets dans les transports publics.

Ces trois sources de données ne sont évidemment pas les seules *big data* propres à l'étude de la mobilité mais elles comptent parmi les plus fréquemment utilisées. Elles suscitent par ailleurs un intérêt particulier sur la scène bruxelloise avec le développement de fournitures *big data* de la part de différents opérateurs privés (Proximus, Be Mobile, TomTom, etc.) mais aussi avec les débouchés potentiels des données issues de la carte Mobib (généralisée à tous les types de formule tarifaire de la STIB depuis 2016).

©STIB-MIVB

1. *Big data* : des données nouvelles pour des manières nouvelles de produire du savoir

1.1. Des données inattendues produites en *masse*

Dans une société qui se veut toujours plus *smart*, toujours plus connectée (à internet, au téléphone, au *GPS*, etc.), une gamme d'actions toujours plus importante laisse une trace numérique qui, enregistrée, génère un flux (et un stock) de données qui croît de manière exponentielle[1]. La production de données a longtemps été l'apanage d'institutions (administrations, entreprises, associations) et restreinte à des catégories spécifiques (date de naissance, catégorie professionnelle…). Elle est aujourd'hui de plus en plus le fait des individus eux-mêmes ou de machines procédant à une collecte automatisée et s'étend à des domaines non accessibles jusqu'alors avec une telle précision, tels que nos déplacements, nos goûts, nos relations, etc. (Cytermann, 2015). L'apparition parfois imprévue mais généralement souhaitée de banques de données massives a ainsi été accompagnée du développement de solutions en termes de stockage, d'organisation des flux de données, d'harmonisation des définitions et des formats, etc.

Cette apparition a donné lieu à la définition déjà ancienne mais encore d'actualité des *big data* (ou données massives) autour de la notion des 3 V (Laney, 2001), formule désignant les termes anglais de *volume*, *velocity* et *variety*, qui sont autant de caractéristiques qu'endosseraient les *big data*. Dans ce cadre, le volume (*volume*) souligne la masse de données particulièrement importante, la vitesse (*velocity*) le fait que le processus de génération de données est très rapide, souvent en continu, alors que la diversité (*variety*) fait référence aux sources et formats variés des données collectées, qui témoignent surtout de leur caractère peu ordonné, peu conforme à l'application directe d'un traitement analytique.

Il est important de souligner que cette définition naît en réponse à la difficulté technique alors éprouvée par des analystes de l'e-commerce de faire face à des flux de données mettant à rude épreuve les capacités de traitement d'alors. Sans trop savoir d'où elle vient, il est fréquent également d'entendre ou de lire la définition suivante : "[toute donnée qui ne rentre pas dans une feuille excel]" (Batty, 2013 : 274). Cette définition, en insistant sur le volume, l'aspect *big,* tend à résumer le phénomène *big data* à une problématique technique.

Cette formulation nous semble, de ce fait, peu satisfaisante. À partir de quelle taille un échantillon est-il suffisamment grand ? À partir de quelle vitesse de génération de données ? De quelle diversité ? Il n'y a pas de doute sur le fait que les *big data*, en leur capacité de données, développent certains traits caractéristiques mais il nous semble que ceux-ci sont moins primordiaux pour comprendre le phénomène que le mode particulier de production de savoir qu'ils induisent, que certains auteurs voient comme l'avènement d'un nouveau paradigme. Pour ne citer qu'eux, Graham et Shelton (2013 : 256) évoquent ainsi un "[paradigme informatique]" et Antoinette Rouvroy (2014 : 413) parle d'un "nouveau régime de vérité".

1.2. Un défi de taille : faire émerger le signal

Il est intéressant de souligner à ce stade les différences entre les caractéristiques des données d'enquêtes, les données administratives et les *big data*. Les données d'enquête sont généralement des données collectées dans le cadre d'un dispositif visant à répondre à des questions de recherche préalablement déterminées. L'échantillon est connu et, bien que l'ensemble des données collectées puisse être considérable, le volume des données reste généralement limité[2]. Les données administratives, au contraire, ne sont pas collectées dans un but de recherche et leur collecte est moins harmonisée au sein de la population (dont on connaît généralement l'échantillon). Il s'agit souvent de bases de données complexes, multidimensionnelles et nettement plus volumineuses que dans le cadre d'enquêtes, nécessitant un traitement spécifique (Connelly *et al.*, 2016).

De ce point de vue, deux caractéristiques plus fondamentales, ou peut-être seulement plus synthétiques, des *big data* nous semblent être la distance importante entre la donnée et le signal[3], qu'on retrouve pour partie dans le *variety* des 3 V, et le fait que les données ne sont pas produites, calibrées, a priori pour se prêter à l'analyse de phénomènes sociaux[4]. Pour aborder cet aspect, certains auteurs évoquent un 4e V, pour véracité (*veracity* en anglais) (voir notamment André De Palma, 2017 : 22) pour insister sur la qualité variable de la donnée, qui nécessite un retraitement avant usage.

À ce titre, le défi posé par les *big data* n'est pas comment absorber plus vite et mieux une masse de données en croissance permanente mais comment en faire émerger la valeur ajoutée (Floridi, 2012). Cet objectif s'assortit d'outils analytiques propres qui font pleinement partie de la définition du cadre *big data* et confèrent à celui-ci une position épistémologique particulière qui ne le substitue pas poste pour poste aux outils et analyses classiques.

Les techniques mobilisées dans le cadre *big data* relèvent essentiellement d'un processus d'extraction de connaissance des bases de données[5] (Miller, 2010 – Fayyad, 1996), une expression qui souligne cette distance entre donnée et signal évoquée plus haut. Nous prenons ici le temps d'en présenter schématiquement les principales étapes (telles que définies par Miller (2010 : 189)), d'une part parce qu'il fait peu de doute que seule une fraction

[1] A titre indicatif, si en 2010 le volume mondial de données numériques stockées s'élevait à 1,2 zettaoctets (1,2.1021 octets, soit 1.200 milliards de gigaoctets), le consultant IDC (2014) estime qu'il grimpera jusqu'à 44 zettaoctets en 2020, soit une multiplication d'un facteur 37 en 10 ans.

[2] L'Enquête socio-économique de 2001 dont la collecte ambitionne l'exhaustivité de la population visée reprend approximativement 10 millions de lignes pour une centaine d'attributs et constitue de ce point de vue un contre-exemple.

[3] Par *signal*, on entend ici l'information utile, véritablement informative par rapport à un objectif de recherche spécifique, en opposition au "bruit" contenu dans les données, qui parasite l'accès au signal.

[4] Une caractéristique partagée avec les données administratives.

[5] En anglais *knowledge discovery from databases*.

d'analystes très formés soit en mesure de se les représenter concrètement, mais également pour insister sur l'importance du facteur humain dans les décisions successives qui y sont prises, qui colore le processus d'une part évidente de subjectivité :

- Sélection des données : sélection du sous-ensemble qui sera utilisé pour l'analyse.
- Nettoyage des données (doublons, données manquantes, etc.) et enrichissement éventuel par l'apport de données auxiliaires.
- Réduction du nombre de dimensions et/ou projection des données dans des espaces de représentation plus efficaces.
- Application de techniques de *data mining*, visant à faire apparaître des récurrences, des corrélations (*patterns*), des singularités cachées dans les données.
- Interprétation des données et rapport de conclusions.

Au sein de ce processus, c'est l'ensemble des techniques rassemblées autour du *data mining* qui produisent sur les données le saut qualitatif décisif en termes d'information produite. Il s'agit d'outils adaptés qui visent généralement moins à expliquer qu'à décrire et à révéler les structures cachées au sein des données en recourant à différentes techniques telles que le *clustering* (groupement d'observations, d'objets similaires entre eux), la classification (affectation d'objets à des classes selon une règle éventuellement préétablie), l'association (sur base des relations entre objets, prédiction de valeurs futures, sur la base de régressions ou d'arbres de décision par exemple), l'étude des déviations (objets qui dévient singulièrement de la norme attendue) (Miller, 2010).

Le *machine learning* et le *deep learning* (voir encadré ci-dessous) constituent des outils de *data mining* particuliers qui, au travers d'apprentissages de règles de décisions se prêtent davantage à une valorisation dynamique du signal remonté, qui peut potentiellement déclencher une action en temps réel **(2.2. Smart city et big data : un nouvel imaginaire urbain)**.

Machine learning, deep learning et intelligence artificielle

Parmi les techniques de *data mining*, les concepts et outils que constituent le *machine learning*, le *deep learning* et l'intelligence artificielle tendent à prendre le devant de la scène, aussi bien dans les usages que dans les discours médiatiques.

De ces trois termes, le plus général est celui de *machine learning* (algorithme apprenant en français), qui caractérise des techniques de *data mining* et qui repose sur la mise en œuvre d'algorithmes déterminant un ensemble de règles sur base d'un premier jeu de données (phase d'entraînement, d'apprentissage), qui sont ensuite appliquées sur un second. Vayatis (2017 : 55) précise à ce sujet qu'il y a "[d]eux niveaux d'algorithme à considérer : un premier algorithme A qui réalise l'apprentissage proprement dit en se connectant à une base de données (entrée de l'algorithme A) optimisant un certain critère (une fonction d'utilité) qui représente le parti pris de l'algorithme sur ce qui est considéré comme une bonne décision, et qui calibre une règle de décision (sortie de l'algorithme A). Un deuxième algorithme P consiste à exécuter la règle de décision produite par A sur de nouvelles données pour produire une aide à la décision (selon les contextes, on parle parfois de prédiction)."

Le tournant des années 2000 a vu la découverte de nouvelles techniques algorithmiques applicables à la problématique du *machine learning* dans le cadre de données massives. Parmi les nouvelles méthodes, Vayatis (2017) distingue celles qui permettent une grande transparence (boîte blanche) dans l'interprétation des résultats (arbres de décisions, modèles linéaires parcimonieux) de celles qui ne permettent qu'une transparence limitée (boîte grise) (forêts aléatoires, méthodes d'ensemble).

Autre méthode apparentée au *machine learning*, le *deep learning* met en œuvre des algorithmes inspirés du fonctionnement des réseaux neuronaux dans le cerveau humain, ce qui leur vaut l'appellation d'intelligence artificielle. Déjà théorisés (et appliqués) dans les années 80, ces outils connaissent aujourd'hui un regain de popularité dans le contexte d'émergence des données massives et d'une amélioration notable des puissances de calcul. On évoque aujourd'hui des algorithmes profonds (*deep learning*) en raison de l'usage de plusieurs couches de réseaux neuronaux qui effectuent des traitements en parallèle. S'ils sont très performants, leur interprétation est opaque et ne permet généralement pas de mettre en évidence les ressorts fonctionnels qui motivent la décision prise par l'algorithme et on parle donc plus volontiers dans ce cas de boîte noire (Vayatis, 2017).

Tous ces outils peuvent être rangés dans les méthodes de classification et d'association évoquées dans le corps du texte. On remarquera qu'ils invitent à une reconfiguration du rapport entre chercheur (ou expert plus généralement) et outil dans le cadre d'un processus d'aide à la décision (Vayatis, 2017). Si les méthodes boîtes blanche ou grise autorisent un certain regard du premier sur le second (experts augmentés), l'émergence de l'intelligence artificielle tend à reléguer l'expert au rang de facilitateur technique.

1.3. Une position particulière dans la production de savoir

Globalement, ces techniques procèdent davantage d'une production inductive de connaissance, où l'information est dans un premier temps reconfigurée pour en révéler les structures, les singularités, en agissant comme une loupe ou un microscope sur le corps de données (Miller, 2010 – Allemand, 2013). La formulation théorique du phénomène observé ne viendrait ainsi que dans un second temps, en accord avec les structures mises en lumière, qui font émerger l'interprétation. Ce processus se démarque de standards plus "classiques" de production de savoir, qui relèvent d'une démarche généralement plus déductive, où le savoir se construit sur une proposition théorique préalable à la collecte et l'analyse de données, qui s'inscrivent dans un dispositif permettant de tester la proposition initiale. On y retrouve souvent la volonté de mettre en évidence les liens de causalité et les facteurs qui influencent les phénomènes mesurés.

Cette distinction est cependant davantage complémentaire qu'excluante. La remontée d'information et de savoir par les *big data* est intéressante en tant que telle et peut nourrir un cadre théorique qui servira de point de départ à une démarche plus déductive. De l'autre côté, une approche plus inductive est nécessairement limitée par les données à disposition et si le cadre *big data* permet de zoomer heureusement sur certaines problématiques, le risque est important d'en négliger d'autres en évacuant toute réflexion théorique préalable.

Cette position de révélateur de connaissance des *big data* ne manque cependant pas de réchauffer une certaine fibre positiviste incarnée à l'extrême par la position d'Anderson (2008) qui en appelle à la "[fin de la théorie]"[6] (cité notamment par Graham et Shelton, 2013 – Boullier, 2015 – Batty, 2013, Boyd et Crawford, 2012) et accueille avec enthousiasme l'avènement d'une nouvelle manière de produire de la connaissance, plus objective car se contentant de laisser parler les données. Ce discours fait face à de nombreuses objections. Premièrement, on peut arguer qu'il n'existe pas de dispositif épistémologique purement inductif. Le chercheur (au sens large) agit en effet toujours dans un cadre de pensée particulier, préexistant, et est guidé dans ses actions par la poursuite d'objectifs spécifiques et par les hypothèses et postulats qui les sous-tendent. Ce facteur humain est, comme nous l'avons vu, bien présent à chacune des étapes de production de savoir dans le cadre *big data* et celui-ci ne peut se prévaloir d'une objectivité supérieure. Deuxièmement, selon plusieurs auteurs, ce type de discours ne se situerait pas à la marge du *big data* et occuperait au contraire une place centrale au sein de ce nouveau paradigme. Pour Boyd et Crawford (2012 : 663), il agit ainsi à la manière d'un mythe entretenant l'idée que "[… les bases de données massives apportent une forme supérieure de compréhension et de savoir, générant une connaissance jusque-là impossible, avec une aura de vérité, d'objectivité et de précision]." Pour Graham et Shelton (2013), ce rapport de force épistémologique s'exprime dans le "[même] *big data*", soit la reproduction automatique, virale, d'une croyance en la supériorité des nouvelles sources de données et des vérités qu'elles découvrent. Ces auteurs insistent au passage sur le danger de minorisation qu'une telle position représente pour les autres manières de produire du savoir, et singulièrement les approches critiques. Nous abordons plus loin la manière dont cette autorité symbolique attachée au cadre *big data* est émulée par le développement du référentiel urbain de la *smart city*.

***

En synthèse des propos développés jusqu'ici, nous reportons sur le tableau suivant (Tableau 1) les principales différences qui nous semblent distinguer données d'enquête, données administratives et *big data*.

[6] Chris Anderson écrit "The end of theory" en 2008, article paru dans le magazine Wired, dont il est alors rédacteur en chef. Cet article, amplement repris et commenté par la littérature scientifique semble être devenu un point de fixation de la controverse sur l'usage des big data dans les sciences sociales et le rapport que cette nouvelle manière de produire du savoir entretient avec les modes de production traditionnels.

Tableau 1. Comparaison des données d'enquête, des données administratives et des *big data*

	Données d'enquête	Données administratives	*Big data*
Distance données – signal	Faible	Moyenne	Lointaine
Collectées dans un but d'étude	Oui	Non	Non
Participation/recrutement des sujets (Chen *et al.*, 2016)	Actif	Passif	Passif
Démarche épistémologique	Davantage déductive	Déductive / inductive	Davantage inductive
Finalité principale de leur usage	Causale	Causale / caractérisation (et classification)	Caractérisation (et classification) et prédiction
Aura d'objectivité dans le cadre *big data*	Non	Non	Oui
Producteurs de savoir – dans le cadre des politiques publiques*	Acteurs publics principalement	Acteurs publics principalement	Acteurs privés, Acteurs publics
Temporalité de l'action publique**	Lente / traitement statique	Lente / traitement statique	Courte ou lente / traitement statique ou dynamique

* Voir **Section 2. *Big data* et action publique : entre opportunités et promotion dans le cadre de la *smart city*.**
** Voir **Section 3. Défis posés par les *big data*.**

2. *Big data* et action publique : entre opportunités et promotion dans le cadre de la *smart city*

2.1. Opportunités offertes par les *big data* : exhaustivité et granularité spatio-temporelles des données et gestion dynamique

Les *big data* se prêtent bien à l'estimation de données de flux, qu'il s'agisse de volumes de personnes qui traversent un lieu ou d'individus suivis au cours du temps (données de téléphonie mobile et données des opérateurs de transport). De ce point de vue, leur grand avantage consiste en une échelle de mesure spatiale et temporelle particulièrement fine (Chandesris *et al.*, 2017 – Ermans *et al.*, 2017 – Debusschere *et al.*, 2017).

Pour les données *GPS*, produites par les acteurs du marché de l'aide à la navigation embarquée, la finesse temporelle est évidemment également bien présente mais elle n'apporte pas nécessairement de plus-value par rapport à des capteurs fixes qui mesurent aussi les flux de manière continue. Par rapport à ces derniers, c'est l'exhaustivité et la granularité spatiales de la donnée *GPS* (les flux automobiles permettent de récolter une information sur l'ensemble du réseau routier) qui présentent le plus de potentiel, en élargissant le domaine d'analyse des congestions et des temps de parcours à des portions du réseau routier pour lesquelles on ne dispose pratiquement d'aucune information.

La dimension temporelle ouvre, en particulier, une fenêtre d'opportunités nouvelles pour l'étude et la gestion de la mobilité, et ce à au moins deux titres. Premièrement, elle permet des analyses jusque-là en dehors du domaine des grandes enquêtes. En effet, d'une part, les flux repérés finement dans le temps rendent davantage possible l'étude des mobilités en heures creuses (le soir, le week-end, en dehors des heures de pointe) qui, par définition, sont des pratiques moins susceptibles d'être enregistrées dans des enquêtes qui se veulent représentatives et dont la profondeur d'échantillonnage est généralement limitée. D'autre part, ce type de données apporte des possibilités de caractériser les territoires et les lieux en fonction des temporalités de leur occupation. D'une certaine manière, ceci ouvre des perspectives de recherche sur la mobilité envisagée d'un nouveau point de vue, soit non plus comme l'étude des individus en mouvement (ou l'étude des déplacements) au sein de territoires mais comme la variabilité temporelle de l'occupation des territoires par des individus (Commenges, 2014).

Deuxièmement, cette grande précision temporelle, associée à une capacité de traitement des données en temps réel autorise de penser à un retour dynamique et une gestion des réseaux en des temps très courts (interventions lors de pics et pointes de congestion, etc.). Cette vision d'une action publique dynamique et en temps réel n'est pas nouvelle (voir Brandeleer et Ermans (2016) pour un exemple en Région de Bruxelles-Capitale) mais elle se retrouve aujourd'hui promue dans le cadre de l'émergence des *big data* et de la *smart city*.

2.2. *Smart city* et *big data* : un nouvel imaginaire urbain

La définition d'une *smart city* n'est pas clairement figée et d'ailleurs, la notion de *smart city* se retrouve elle-même dans des expressions diverses : *ville intelligente*, *ville digitale*, *ville virtuelle*, etc. Toutes cependant évoquent un imaginaire urbain où l'environnement et les individus qui y vivent et s'y déplacent, sont en interconnexion permanente en vertu de capteurs et d'instruments relevant des technologies de l'information et la communication (TIC), qui détectent et mettent en réseau l'ensemble des actions qui s'y produisent. On parle aussi de l'internet des choses (*the internet of things*), accumulant au passage une quantité de données gigantesque. Pour être *intelligente*, il faut également que cette mise en connexion donne lieu, au travers d'algorithmes et de fonctions, à une forme de rationalisation des actions, des déplacements, des flux qui permet "[une amélioration de l'efficacité, de l'égalité, de la durabilité et de la qualité de vie des citoyens en temps réel]" (Batty *et al.*, 2012 : 482).

Le cadre *big data*, avec tout ce qu'il charrie comme défis en termes de justice sociale et de protection de la vie privée, intervient dans cette vision au niveau des modalités de ce passage entre *ville connectée* et *ville intelligente* pour faire remonter un signal qui sera aussitôt interprété et "opérationnalisable" en une action exprimant une forme de gouvernance urbaine sur une échelle temporelle très courte. En termes de mobilité, il s'agirait typiquement d'optimiser les flux en temps réel, en redistribuant le trafic, les voyageurs, en limitant ou autorisant certains accès, en modulant l'offre de transport, la capacité des voiries, en organisant dynamiquement le remplissage des parkings, etc. L'objectivité apparente des *big data* (puisque c'est l'algorithme qui distingue le signal du bruit) leur confère une validité symbolique importante auprès des décideurs. Celle-ci s'étend à la *smart city* pour apporter la promesse d'une connaissance objective et en temps réel au profit d'une gestion optimisée de la vie urbaine.

De ce point de vue, le rapport entre la construction empirique de la représentation de la ville et l'action publique urbaine prend une tournure nouvelle. Classiquement, cette construction empirique prenait la forme de rapports, d'études, visant la production d'indicateurs qui expriment, décrivent et construisent une représentation de la ville et de son évolution. La démarche relève de la description mais aussi de l'explication, de la recherche de causalités, qui représentent autant de leviers pour l'action publique (voir par exemple van der Loop *et al.* (2017: 10-11). La temporalité du processus est nécessairement longue, à tout le moins non immédiate, en vertu des temps de la recherche, de la décision politique et des effets de celle-ci sur le réel. Au contraire, dans la vision présentée au paragraphe précédent, les temporalités sont très brèves et les réponses quasi immédiates. La connaissance produite relève moins de l'explication que d'une structuration de l'information, qui appelle à l'optimisation. Le signal, le savoir remonté, apparaît ainsi performatif en ce qu'il semble dicter directement, de lui-même, l'action publique.

En réalité, l'outil n'est jamais neutre et la décision politique, l'action publique, se niche dans la préprogrammation des réponses à apporter aux différents stimuli de la ville intelligente (quel trafic est détourné ? quel usager reçoit la priorité ?, etc.) (Brandeleer et Ermans, 2016), voire dans le choix des quantités à optimiser par les algorithmes apprenants. Par ailleurs, on peut arguer que le choix même de privilégier cette forme d'action publique plus directement opérationnelle, en marginalisant les approches longues qui visent à identifier les mécanismes causaux, réduit les possibilités de problématisation de la mobilité (par l'aménagement spatial des fonctions ou par l'investissement dans l'offre de transport public par exemple) et constitue à ce titre une orientation politique en soi. Enfin, cette approche de l'action publique porte en elle le risque de ne s'adresser qu'aux personnes connectées.

Concluons ici en signalant que si la numérisation du monde entraîne de nouvelles manières de mesurer les mobilités, elle constitue également le support de nouvelles manières de se déplacer (Chandesris *et al.*, 2017 : 143). Pensons par, exemple, aux informations sur les véhicules en temps réels de la STIB ou aux calculateurs d'itinéraires. Avec l'essor des smartphones et des systèmes de navigation, on voit ainsi se développer des services de mobilité pour les personnes connectées. Ces évolutions soutiennent le développement de nouveaux services aux voyageurs qui tendent à se regrouper sous l'appellation de *Mobility as a Service* (*MaaS*). Outre la fourniture d'un service de calculateur d'itinéraire inter-réseaux, la *MaaS* se conçoit idéalement comme une porte d'entrée pour la comparaison et le paiement des services de mobilités disponibles (il s'agit pour le client d'optimiser le prix aussi bien que le temps de trajet).

Implicitement, les possibilités de résolution des problèmes de mobilité urbaine que comporte la *MaaS* reposent sur l'éventuelle harmonie globale qui se dégagerait des comportements des individus, qui sont autant de réponses aux informations proposées par les services de calculs d'itinéraires intermodaux et inter-opérateurs. La maîtrise de l'information fournie aux particuliers constitue ainsi un enjeu de gestion de la mobilité urbaine non négligeable.

2.3. À l'origine de la ville intelligente : le secteur des TIC

Si ce concept de *ville en réseau*, de *ville intelligente* a émergé au cours des quinze dernières années, l'unanimité contemporaine autour de l'anglicisme *smart city* n'est sans doute pas étrangère au processus récent d'incorporation de celui-ci dans les politiques urbaines. Ces dernières ont ainsi vu l'implication progressive d'entreprises opérant à l'échelle globale pour le développement et le placement de solutions TIC (infrastructures, softwares) telles qu'IBM, CISCO, Microsoft, Oracle, SAP, participant ainsi à la création de marchés de la ville numérique (Batty *et al.*, 2012 – Douay et Henriot, 2016). Batty *et al.* (2012) insistent sur l'exemple d'IBM, qui a fait le choix stratégique d'investir dans le *smart* avec sa campagne *pour une planète plus Smart* lancée à partir de 2008[7], en repositionnant ses produits et services pour offrir des solutions TIC permettant de rendre les villes plus intelligentes mais également en développant un volet de services en matière de conseil et d'aide aux autorités locales sur diverses problématiques, dont les solutions s'appuient, sans grande surprise, sur les TIC. C'est dans ce cadre qu'un grand nombre de villes ont participé au *Smarter Cities Challenge*. Dans ce cadre, IBM dépêche dans chaque ville un groupe d'experts chargés de dresser un constat sur une thématique urbaine choisie au préalable (mobilité mais aussi environnement, sécurité, administration, services sociaux, développement économique, etc.). Ce rapport est assorti de diverses recommandations pour les pouvoirs publics[8].

Ce rôle des entreprises privées qui, par leurs collaborations avec les pouvoirs locaux, tendent à infléchir les politiques urbaines au bénéfice du référentiel de la *ville intelligente* est également mis en avant par Douay et Henriot (2016), qui analysent l'expression de cette tendance globale au sein des villes chinoises. On notera qu'en conclusion, les auteurs questionnent la substance même des effets de la mise en intelligence de la ville, dont l'opérationnalité laisse souvent à désirer, et suggèrent très nettement l'hypothèse que la *smart city* participe surtout d'un processus de *storytelling* (la ville se raconte, se met en récit), dans le cadre d'une logique de marketing urbain où l'argument *smart* supplanterait celui de *durable*. Les auteurs font remarquer que ce passage de témoin s'opèrerait à la faveur d'une logique qui confère aux villes *smart*, par leur action de rationalisation sur l'ensemble de la société, la propriété d'être nécessairement des villes de basse émission carbone, durables donc.

À Bruxelles, outre l'épisode du *Smarter Cities Challenge*, le tournant de la *smart city* est perceptible dans l'accord de gouvernement 2014-2019, qui ambitionne de faire de Bruxelles une "capitale du numérique" (Gouvernement bruxellois, 2014 : 25) et qui a été suivi, entre autres mesures, de la nomination d'une secrétaire d'État de l'informatique et de la transition numérique en janvier 2015 et de la mise sur pied d'un site expressément dédié à renforcer le caractère smart de Bruxelles[9].

L'importance du CIRB (Centre d'Informatique pour la Région Bruxelloise) vis-à-vis de l'émergence d'une vision et d'une programmation de la *smart city* à Bruxelles doit également être soulignée. Cet organisme d'intérêt public œuvre en effet à la promotion de ville intelligente selon diverses modalités d'action et notamment par son Livre Blanc 2014-2019 (CIRB, 2014), l'organisation d'un *smart city summit*, d'un *smart city event*, de *smart breakfast*, etc. Ensuite, dans sa qualité de centre de compétences et de gestionnaire de réseau de télécommunications, il participe activement à la mise en œuvre de la *smart city* par le soutien qu'il apporte aux pouvoirs publics et aux citoyens en termes de solutions TIC et notamment, en ce qui concerne la collecte, la gestion et la mutualisation des données.

On notera que, plus globalement, le développement d'une vision *smart city* constitue l'un des volets seulement de la politique pour le numérique à Bruxelles, aux côtés des efforts de formation dans le domaine du travail et de l'éducation (Impulse) et de l'innovation (Innoviris). Ces trois piliers de cette politique sont repris sous la "marque ombrelle" Digital Brussels (CIRB, s.d. : 1), qui vise à garantir la cohérence et la bonne coordination de l'ensemble.

[7] Voir le site internet : http://www.ibm.com/smarterplanet/us/en/, qui ouvre sur le titre "IBM construit une planète plus smart" (IBM builds a smarter planet).

[8] Bruxelles a par exemple fait appel à IBM en 2014 pour un diagnostic relatif à la mobilité dans la ville. Les résultats peuvent être retrouvés sur https://smartercitieschallenge.org/cities/brussels-capital-region-belgium. Il s'agit d'une synthèse des principaux constats effectués par les acteurs locaux (administrations, universités, bureaux d'études) suivie de recommandations qui se bornent le plus souvent à un développement plus approfondi des infrastructures TIC, supposé engendré naturellement des gains en termes de fluidité du trafic. Le "*Smarter Cities Challenge*" n'est par ailleurs pas restreint à la mobilité et touche à de multiples problématiques urbaines : environnement, sécurité, administration, services sociaux, développement économique, etc.

[9] http://smartcity.brussels/

Sans prétendre à l'exhaustivité, on peut citer quelques initiatives bruxelloises qui s'alignent sur ce référentiel urbain montant. Ainsi, depuis 2017, digitYser (*www.digitYser.org*) s'impose comme un acteur de premier plan et souhaite positionner Bruxelles en tant que "capitale digitale de l'Europe" (DigitYser, 2017). Lancée officiellement en décembre 2017, l'organisation est financée par la Région de Bruxelles-Capitale ainsi que par divers acteurs privés (au premier rang desquels la société d'investissement Sofina) et a pour objectif de faciliter le développement de l'*IoT*, des *big data* ou encore de la réalité virtuelle en offrant un espace commun pour l'hébergement de cours, d'événements et pour "l'incubation" de jeunes entreprises. Elle organise également des *hackatons*, événements au cours desquels les participants sont mis en compétition (souvent en équipe) et tentent d'apporter une solution par l'usage des données à un problème qui leur est soumis, en un minimum de temps (généralement un week-end)[10]. On peut également citer dans cette veine l'initiative de la plateforme pour les entreprises bruxelloises *BECI* (*Brussels Enterprises Commerce and Industry*) qui a, depuis avril 2018, mis sur pied un *concept store* dans un espace *pop-up*[11] qui vise à rassembler les entreprises bruxelloises et à accélérer les *start-up* autour de solutions de mobilité innovantes : "Les visiteurs [peuvent] y découvrir et y tester des produits et des services de mobilité efficaces et innovants, dans un cadre éphémère et événementiel, propice à l'expérimentation"[12].

[10] Hackathon sur la génération de contenu musical à partir de formats MIDI par exemple (mars 2018).

[11] Espace (généralement) commercial éphémère.

[12] http://www.beci.be/centre_de_connaissance/mobilite/urban_mobility_pop_up/, page consultée le 2 mai 2018.

3. Défis posés par les *big data*

3.1. La quantité au détriment de la qualité ?

On l'a vu, une des caractéristiques des *big data* est de faire émerger de l'information signifiante à partir d'un ensemble volumineux de données a priori peu signifiantes. Pour autant, volume n'est pas forcément synonyme de fiabilité. La récolte automatique des données n'est pas garante de leur exactitude et leur fiabilité dépend grandement de la qualité et de la disposition des capteurs des données et du type d'information qu'il est possible de collecter (Rouvroy, 2016). À l'image des données administratives, les données massives sont sujettes à des biais qui ne sont pas maîtrisés au moment de la collecte et qu'il faut interpréter et corriger a posteriori (au contraire des enquêtes et comptages).

3.1.1. Représentativité ou répétitivité ?

Ces deux aspects ne constituent pas que des aléas qu'une solution technique viendrait à terme résoudre. Ils posent la question de la représentativité des informations produites par les *big data*. Les données sont ainsi plus faciles à collecter auprès de certains publics, ce qui constitue un risque dans l'extrapolation des résultats à l'ensemble de la population. Par exemple, si l'on ne mesure que les flux de personnes qui disposent d'un *GPS* ou d'un smartphone, on n'envisagera la ville ou la mobilité qu'en fonction du comportement de ces personnes (Miller, 2010).

On fera remarquer que si les biais de représentativité ne sont pas circonscrits au domaine des *big data*, dans le cas de données d'enquête, ils sont généralement connus et maîtrisés dès la conception (Chandesris *et al.*, 2017).

Vayatis (2017) insiste également sur la notion de répétitivité, soit le fait que toutes les catégories de valeurs apparaissent suffisamment fréquemment dans les flux captés. Cette caractéristique prend toute son importance dans le cadre de l'usage d'algorithmes apprenants, qui ne peuvent classer et prédire avec précision que les situations qu'ils ont entraînées auparavant.

Représentativité et répétitivité sont contradictoires dans une certaine mesure (un flux de données représentatif fera apparaître plus rarement les modalités les moins fréquentes). Dans une finalité descriptive, on privilégiera la représentativité tandis que les exercices de classification et de prédiction pour l'appui à la décision préféreront une répétition fréquente de toutes les instances afin d'entraîner les algorithmes à un ensemble le plus divers possible de situations.

3.1.2. Faire face au déficit contextuel

Les *big data* se caractérisent également par un déficit de données contextuelles. Ceci s'explique premièrement par l'absence de question de recherche en amont du processus de collecte mais aussi par la sélection de données compatibles avec une modélisation mathématique. Cette approche mécanique de la collecte des données implique une perte de certains facteurs explicatifs (Boyd et Crawford, 2012). Typiquement, dans le domaine de la mobilité, il s'agit d'information sur les individus (âge, genre, catégories socio-économiques, etc.) ou sur les déplacements mêmes (modes de déplacement, motifs et vécus de déplacement). Cette limite pousse les *big data* à se cantonner à leur caractère opérationnel : le succès d'un algorithme se mesure à la rapidité d'obtention d'une information rationnelle pour un coût minimum. La logique est celle du rendement, de l'optimisation, pas de la validité (Rouvroy, 2016).

Par exemple, l'optimisation de la fréquence, des horaires, des trajectoires des véhicules de transport en commun se ferait en fonction d'intérêts collectifs déduit de la géolocalisation des personnes (Rouvroy, 2016). Les *big data* seraient alors au service d'une ville efficiente, où les problèmes mesurés trouveraient une solution technique rapide et optimale (un itinéraire alternatif en cas de congestion, par exemple), sans forcément chercher la cause des problèmes, ni agir sur celle-ci (la cause de la congestion, par exemple). Le risque de cette recherche d'efficacité est donc également de minorer certaines problématiques sociales (inégalités socio-économiques, problématiques environnementales, etc.) (Miller, 2010).

Les déficits de contexte (absence de dimensions importantes dans les données) ne sont cependant pas sans poser problème aussi pour les exercices à visée opérationnelle et peuvent mener à des classifications et prédictions notablement faussées. Pour remédier à ces éventuels manquements, il existe essentiellement deux démarches possibles : soit l'inférence des données manquantes (par exemple le genre de la personne sur base de son adresse mail) ou alors le croisement avec d'autres sources de données (éventuellement massives) (Vayatis, 2017). Ces solutions posent évidemment la question de la préservation de l'anonymat des personnes et de la protection de la vie privée (voir plus loin).

3.2. Un outil difficile à maîtriser

L'émergence des *big data* représente un enjeu de gouvernance important pour les pouvoirs publics dans la mesure où la maîtrise même de l'outil est délicate.

Premièrement, le savoir technique qu'elles nécessitent est extrêmement pointu et en réserve l'accès à un nombre limité de personnes. On voit ainsi se développer en parallèle à l'essor des *big data* l'émergence d'un nouveau type d'expert : le *data scientist* [13]. Sa légitimité repose davantage sur la maîtrise d'une palette large d'outils techniques (programmation, mathématiques, statistiques avancées, *data mining*, *machine learning*, etc.) que sur le savoir propre à certaines disciplines (économie, sociologie, sciences politiques, etc.). Ce type de profil est par ailleurs très demandé sur le marché du travail et représente de ce fait un investissement à la fois important et pas nécessairement évident pour les pouvoirs publics.

Deuxièmement, de nombreuses sources de données *big data* proviennent d'acteurs privés, ce qui suscite en soit plusieurs enjeux.

Pour commencer, on le verra par la suite, les données issues du secteur privé sont en général déjà retravaillées pour être prêtes à l'emploi sous la forme d'indicateurs classiques. Le processus d'extraction de signal, d'information a déjà été réalisé par l'acteur privé. Ceci implique l'absence de droit de regard sur les opérations réalisées, les choix posés et les limites éventuelles qui entachent le produit finalement délivré.

Ensuite, l'opérateur public n'est souvent pas propriétaire des données. En acceptant de baser l'action publique sur une source qui ne lui appartient pas, il se met en dépendance de logiques de production de données, en ce compris les traitements qui précèdent la livraison d'indicateurs clé-sur-porte, qui lui échappent et peuvent à tout moment lui imposer une refonte de ses outils de travail.

Enfin, il ne fait aucun doute qu'aussi bien les technologies qui permettent la captation de traces (réseaux télécoms, GSM, smartphones, *GPS*, etc.) que les populations connectées grâces à celles-ci vont évoluer et que les traitements visant à faire émerger le signal des données récoltées devront être adaptés également. L'utilisation de ces indicateurs pour la constitution de séries temporelles est dès lors fortement compromise (van der Loop *et al.*, 2017).

[13] On notera que les universités belges francophones semblent avoir pris acte de l'émergence d'un tel profil sur le marché du travail et la plupart d'entre elles (Université Catholique de Louvain, Université de Liège, Université de Namur) proposent des masters en *science des données* ou des finalités de Master en *data science* depuis la rentrée 2017.

3.3. Le respect de la vie privée : un défi gigantesque

L'enregistrement des données personnelles semble être la contrepartie inévitable de l'usage d'une multitude d'applications et appareils numériques. Par les nouvelles possibilités de ré-identification qu'elles offrent, les *big data* floutent la distinction entre données anonymes et données personnelles. Dans ce contexte, l'anonymisation des données n'est plus une condition suffisante au respect de la vie privée (Rouvroy, 2016).

Des chercheurs (De Montjoye *et al.*, 2013) ont montré, par une étude de 15 mois sur des données de mobilité de 1,5 million de personnes, que les traces de mobilité sont particulièrement uniques. Dans une base de données de Proximus, l'opérateur historique de téléphonie en Belgique, où les localisations individuelles sont enregistrées toutes les heures, avec la granularité permise par le réseau d'antennes actuelles, seuls quatre points spatio-temporels sont nécessaires pour individualiser 95% des traces. Deux suffisent déjà pour identifier plus de la moitié des utilisateurs. Les bases de données étaient pourtant anonymisées, ne contenant ni nom, ni adresse, ni numéro de téléphone. Particulièrement en matière de mobilité, les individus tendent à adopter des schémas uniques qui rendent, d'une certaine manière, leurs données de déplacements et de géolocalisation plus personnelles que leurs empreintes digitales (qui nécessitent douze points de référence pour identifier un individu) (Grosjean, 2015).

De plus, l'anonymisation ne protège pas contre les possibilités de profilage des individus. On peut définir le profilage comme une méthode de *data mining* permettant de classer, avec une certaine probabilité, un individu dans une catégorie particulière afin de prendre des décisions individuelles à son égard (Grosjean, 2015). L'exemple le plus connu de profilage est celui de la publicité ciblée sur internet en fonction des habitudes de navigation et des sites consultés par l'utilisateur, permettant de déduire ses caractéristiques individuelles (genre, catégorie d'âge, localisation, etc.) mais également ses goûts ou envies. Par exemple, grâces aux données collectées, un site pourra envoyer de la publicité pour des produits liés à une grossesse à une jeune femme que l'algorithme aura catégorisée comme enceinte (Floridi, 2012). Ce processus pose évidemment des questions éthiques importantes en matière de catégorisation des individus et de respect de la vie privée. A partir du moment où le profilage permet, dans une certaine mesure, de prédire des comportements individuels, cette méthode peut être également utilisée pour la surveillance des masses (profilage criminel, mais aussi religieux, social…) (Grosjean, 2015).

Pour l'instant la collecte des données par les entreprises privées est très peu encadrée. La législation européenne actuelle soumet la qualité des données à un certain nombre de principes : un traitement loyal et licite, un principe de finalité, un principe de proportionnalité, de pertinence et d'exactitude, une durée n'excédant pas le temps nécessaire à la réalisation des finalités. Une difficulté majeure, dans le cadre *big data*, est que la finalité d'un traitement n'est pas nécessairement connue a priori. La législation en matière de protection de la vie privée considère la collecte, le transfert, la modification de données personnelles, etc. mais pas la signification nouvelle des données grâce à leur agrégation (Bensamoun et Zolynsky, 2015). Force est de constater que les évolutions technologiques dépassent largement les évolutions réglementaires.

Le *General Data Protection Regulation – GDPR* (ou Règlement général en matière de protection des données – RGPD), adopté en mai 2016, vise à devenir le cadre réglementaire européen en la matière. Cette directive a été transposée au niveau des États membres en mai 2018 et devrait permettre la création d'un guichet unique au niveau européen, chargé de l'application effective des règles adoptées, tant pour les entreprises et organisations basées dans l'UE que pour leurs sous-traitants responsables des traitements.

Les *big data*, en raison notamment des perspectives qu'elles ouvrent grâce à la segmentation très fine de la clientèle ou du ciblage publicitaire, ont une valeur économique gigantesque (Rouvroy, 2016). Selon certaines estimations, la valeur des données à caractère personnel des citoyens européens est susceptible d'augmenter pour atteindre près de 1.000 milliards d'euros par an d'ici à 2020 (Jourova, 2016). L'Europe considère dès lors le renforcement des normes en matière de protection des données comme une opportunité à la création de débouchés commerciaux, et non comme un obstacle à l'innovation.

Ce texte de compromis tente donc d'améliorer la protection des personnes en tenant compte de l'évolution technologique, sans pour autant entraver l'innovation (et la valeur économique de celle-ci) en la matière (Leonard, 2016).

Les principaux changements de ce nouveau règlement sont (voir Jourova, 2016 et Leonard, 2016) :

- Le *droit à l'oubli* : toute personne pourra obtenir la suppression de ses données afin que celles-ci ne soient plus traitées.

- Le *droit à la portabilité des données* : le fait de pouvoir accéder facilement à ses propres données et de les transférer gratuitement d'un prestataire de service à un autre.

- Le *renforcement de la transparence* qui oblige le responsable du traitement à prévoir des mécanismes clairs permettant à la personne concernée d'exercer ses droits. Lorsque le consentement est requis, il doit être demandé par un acte clair. La même transparence est demandée pour la façon dont les données sont traitées.

- Le *droit de ne pas être soumis à une décision automatisée*, ce qui inclut le profilage, sauf si la personne a donné son consentement explicite ou pour des motifs d'intérêt public.

- Les principes de *data protection by design* et de *data protection by default* : ces deux principes visent le renforcement de la responsabilité et de l'obligation de rendre compte pour les entreprises et organisations qui traitent des données à caractère personnel. Celles-ci devront prendre des mesures appropriées, tant dans la conception du traitement que dans sa mise en œuvre, pour le rendre conforme aux règles de protection de la vie privée (*by design*). En ce qui concerne le principe *by default*, les responsables des traitements s'engagent à limiter le traitement des données personnelles à ce qui est strictement nécessaire.

On le voit, le *GDPR* constitue une avancée importante en matière réglementaire pour la protection des données personnelles. Les personnes concernées auront droit à un recours juridictionnel effectif contre le responsable du traitement ou un sous-traitant en cas d'atteinte à leurs droits. En cas de non-respect de ses provisions, le *GDPR* prévoit une série de sanctions. L'autorité de contrôle pourra prononcer des amendes administratives allant jusqu'à 20 millions d'euros ou 4% du chiffre d'affaires annuel pour les violations les plus graves (Leonard, 2016).

Cependant, plusieurs dispositions restent floues dans leur application. Par exemple tout l'intérêt des *big data* est justement de réutiliser les données à des fins qui n'avaient pas été prévues initialement. Cette aspiration se heurte aux principes de détermination des finalités de la collecte et de proportionnalité des données collectées à ces finalités ainsi que leur durée de conservation (Cytermann, 2015). L'un des défis du développement des *big data* sera donc de déterminer la frontière entre utilisation statistique et profilage des individus. Par ailleurs, bien que le consentement puisse se faire de façon plus transparente et explicite, il conditionne souvent l'accès à de nombreux services. Il est probable que beaucoup d'individus ne mettront que peu en balance la valeur de leurs données personnelles par rapport à l'usage d'un service.

4. Regard sur des exemples d'exploitation de *big data*

Dans cette section, nous nous penchons de manière plus concrète sur les obstacles et les opportunités qui surgissent à l'usage des *big data* pour l'étude de la mobilité. Nous passons en revue ici trois types de données fréquemment cités : les données de téléphonie mobile ou *floating mobile data* (*FMD*), les données *GPS* collectées par les opérateurs de services de navigation embarqués dans les véhicules motorisés, également appelées *floating car data* (*FCD*) et les données de validation automatique des opérateurs de transport en commun.

Le domaine des traces numériques propres à l'établissement d'indicateurs pour l'étude de la mobilité des personnes est évidemment plus large. On peut notamment penser aux données Google Map qui apportent des informations sur les temps de parcours selon le mode de déplacement, aux traces enregistrées par des capteurs Bluetooth ou le suivi des personnes connectées à un réseau wi-fi public, qui permettent de représenter les déplacements des individus dans l'espace public, ou encore aux données Viapass qui enregistrent les déplacements des poids lourds sur le territoire belge. Cette liste est loin d'être exhaustive.

Dans les développements qui suivent, nous restreignons le champ des applications couvertes à un usage statique des *big data*, soit un usage pensé dans le cadre d'une action publique sur le temps long.

4.1. Données des opérateurs de téléphonie mobile – *Floating Mobile Data* (*FMD*)

Les *FMD* sont produites et détenues par les opérateurs de téléphonie mobile et il s'agit évidemment d'un sous-produit de leur activité principale. Leur intérêt réside en ce qu'elles permettent de construire une image très fine de la distribution spatiale des usagers des réseaux de téléphonie, aussi bien en termes de répartition spatiale que de répartition temporelle pour un coût et des délais très concurrentiels vis-à-vis, par exemple, de recensements de la population ou d'enquêtes sur les déplacements.

En Belgique, les principaux fournisseurs de données sont Proximus, Orange et Base. Bruxelles Mobilité s'est lancée dans une première exploitation des données Proximus pour l'analyse des déplacements en lien avec la Région à une échelle spatio-temporelle très fine (Determe, 2018).

4.1.1. De la donnée brute aux déplacements

a) Données de signalement vs données de facturation

On compte deux types de données de téléphonie mobile exploitables pour représenter les déplacements : les données de transactions et les données *passives*. Le produit final sera très dépendant de la sélection d'un de ces types ou, éventuellement, des modalités de leur association.

Les données de transactions recouvrent les transactions qui nécessitent une facturation, tels que les appels ou les sms, soit l'enregistrement détaillé des appels ou *call detail records* (*CDR*) en anglais. Les échanges de données via internet peuvent également entrer dans cette catégorie. Ceux-ci sont situés précisément dans le temps (au moment de l'envoi, au début de l'appel) et peuvent être repérés dans l'espace à l'aide de la zone, ou cellule, couverte par le réseau qui permet la communication[14]. Dans ce cas de figure, l'utilisateur est identifié au centre de la cellule.

Des données sont également récoltées indépendamment des communications de chaque utilisateur. Celles-ci sont liées à la communication qu'entretient le réseau avec les téléphones mobiles indépendamment des actions de l'utilisateur. On parle alors plus volontiers de données passives (*sighting data* (Chen *et al.*, 2016) ou *signalling data* (Bonnel *et al.*, 2015)). Il s'agit de communications qui ont pour but de savoir à tout moment où le téléphone mobile se trouve sur le réseau afin de pouvoir le localiser au plus vite dans le cas d'un appel, de l'envoi d'un sms ou autre communication par internet. Le niveau spatial requis pour la localisation est cependant beaucoup plus grossier que dans le cas de transactions donnant lieu à facturation. On parle ici d'aires de localisation (*location area*) qui agrègent plusieurs cellules[15]. Ces communications passives mobile-réseau interviennent à l'allumage et à l'extinction de l'appareil, lorsque le téléphone change de *location area* et, lorsque le téléphone reste fixe, sa position est actualisée à intervalles de temps réguliers.

[14] Chaque zone est liée à une station de base fixée à une antenne. Ces stations ne sont cependant pas nécessairement placées au centre de chaque cellule et peuvent même être situées à l'extérieur de celles-ci et couvrir plusieurs cellules simultanément.

[15] Pour donner un ordre de grandeur, Bonnel *et al.* (2015) recensent 32 *location area* pour 10.000 stations de base en région Île-de-France.

Figure 1. Exemples de données de téléphonie : *call detail records* (en haut) et transactions de "signalement" (en bas)
Source : Chen *et al.* 2016

Table 1
Sample records in CDR data. [a]

X	Y	ID	Time	Duration (sec)
195925	32464	J000001	82141	81
195925	32464	J000001	82456	75
195018	31555	J000002	82100	140

[a] XY coordinates are transferred from geographical coordinate system. A conversion can be made to convert them into the absolute latitude and longitude coordinates.

Table 2
An example of the sightings data.

ID	Time [a]	Location [b]
3X35E90	1319242582	34.044162|–112.454400
3X35E90	1319242583	34.044059|–112.455550
3X35E90	1319301785	34.044392|–112.453519

[a] Time is Unix timestamp-defined as the number of seconds that have elapsed since 00:00:00 Coordinated Universal Time, Thursday, 1 January 1970.
[b] Location is the longitude and latitude coordinates of mobile phones.

b) Effet ping-pong

Deuxièmement, les utilisateurs ne sont pas toujours effectivement localisés dans la cellule qui assure la communication. Plusieurs effets perturbateurs peuvent ici jouer un rôle. En particulier, une station peut renvoyer une partie de sa charge sur les stations avoisinantes lorsque la charge du trafic l'impose. Les conditions météorologiques ou la topographie locale peuvent également réorganiser la prise en charge des flux. Du point de vue de l'interprétation des données, ces perturbations se traduisent par une imprécision sur la localisation réelle de l'utilisateur mais peut aussi donner lieu à un effet d'oscillation (Chen *et al.*, 2016) ou effet ping-pong (Bonnel *et al.*, 2015), de l'utilisateur qui est perçu comme parcourant des va-et-vient à très grande vitesse entre plusieurs cellules. Dans ce dernier cas, il existe différentes options pour corriger les mesures et notamment pour éviter la surestimation des déplacements[16].

c) Localiser les appareils mobiles dans l'espace

La représentation du réseau de téléphonie mobile en une structure hiérarchique constituée de cellules, au niveau de désagrégation le plus poussé, enchâssées dans des aires de localisation (location areas) qui les surplombent, est une construction simplifiée de la réalité, adaptée aux besoins de l'exercice. Le réseau est en effet plus complexe, composé à la fois d'une architecture 2G (qui permet le transit des seuls échanges GSM et SMS), d'une infrastructure 3G (qui permet également les échanges de données et supporte les communications internet) et depuis peu d'une infrastructure 4G (qui autorise de nouveaux objets à communiquer avec le réseau). Il s'agit donc dans un premier temps de simplifier le référentiel spatial[17], qui prend alors généralement la forme d'un diagramme de Voronoï (Figure 2) qui partitionne l'espace de telle manière à ce que chaque point à l'intérieur de chaque polygone soit plus proche de la station de base à laquelle il se réfère que de toute autre station.

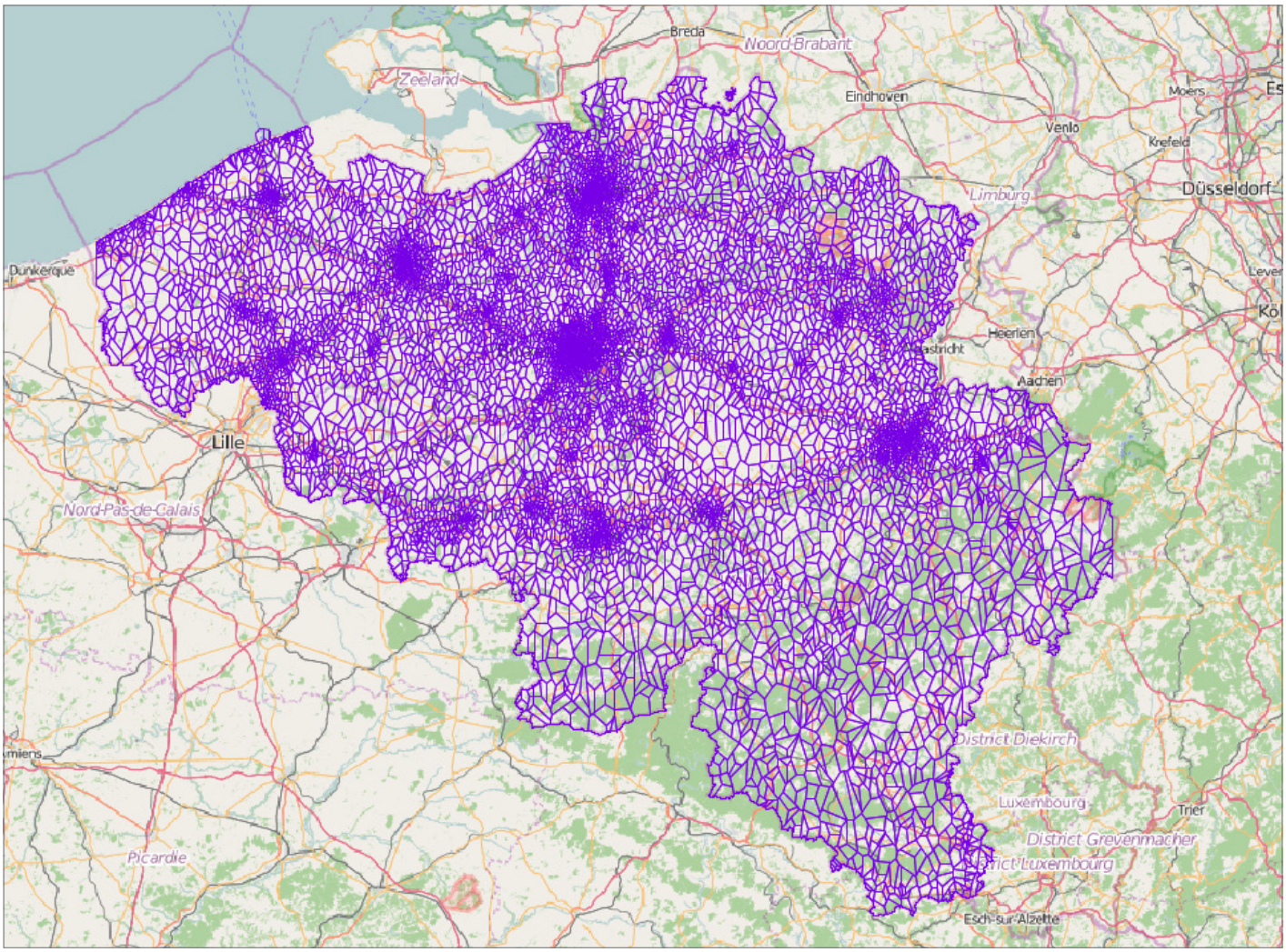
On constate directement que les mailles de ce découpage sont fortement variables. Le niveau de couverture du réseau est en effet très sensible à la densité de population et les mailles se relâchent à mesure que celle-ci baisse. Plus généralement, c'est l'intensité de transactions qui détermine le dimensionnement du réseau. Celui-ci, adaptable dans le temps, n'aura donc pas la même précision en heures creuses qu'en heures de grande activité (Ricciato *et al.*, 2015).

[16] Il s'agit généralement de supprimer les déplacements apparents en se référant à une limite de vitesse, sur la base de la perception d'allers-retours consécutifs entre paires de cellules ou sur la base des deux méthodes citées simultanément.

[17] Debusschere *et al.* (2017) recourent au concept de "technology-agnostic cell sector" ou secteur cellulaire indépendant de la technologie sous-jacente (TACS). Chaque TACS est constitué de toutes les cellules présentant le même azimut, quelle que soit la technologie qu'elle emploie.

Figure 2. Découpage de l'espace belge en diagramme de Voronoï centré autour des stations de bases du réseau Proximus

Source : Ermans *et al.*, 2017



d) De l'occupation du territoire aux déplacements

Déterminer les déplacements passe par la distinction des positions spatio-temporelles observées entre celles considérées comme étant à l'arrêt ("*staying points*") et celles considérées en mouvement ("*transit points*"). Deux ou plusieurs mesures successives sont considérées à l'arrêt, elles font partie du même staying point, si elles se trouvent dans la même maille et que la durée entre la première observation et la dernière observation est séparée par une valeur-seuil de durée minimale. Pour les données Proximus, l'opérateur se base sur un intervalle d'une heure (Ermans *et al.*, 2017). Autrement, les observations sont labellisées comme transit point.

Les déplacements sont ainsi définis entre deux *staying points* successifs et leurs trajectoires peuvent être décrites plus finement (dans le temps et dans l'espace) dans le cas où l'on dispose de transit points intermédiaires. Tout déplacement réalisé à l'intérieur d'une maille ou dans les limites de la valeur-seuil de durée minimum ne sera donc pas enregistré. Bonnel *et al.* (2015) ont notamment montré que les résultats (matrices origines-destinations) sont très dépendants de la valeur choisie.

4.1.2. Représentativité et opacité : principales contraintes à l'usage des *FMD*

Le problème principal des *FMD*, quand elles sont mobilisées pour des objectifs de planification territoriales, s'articule autour de la question de leur représentativité. Divers facteurs peuvent être signalés à ce propos (Chen *et al.*, 2016) :

- Taux de pénétration de l'opérateur qui varie fortement selon les catégories de population.
- Toute la population ne dispose pas de téléphones portables.
- L'intensité d'usage des téléphones qui varie également fortement en fonction des usagers mais aussi du type de téléphone (les smartphones connectés à internet peuvent être détectés à tout moment).
- Certains usagers disposent de plusieurs téléphones portables.

Bonnel *et al.* (2015) ont, par exemple, tenté de valider des matrices origines-destinations pour des données passives sur le territoire de la Région Île-de-France. La comparaison avec les données sur les navettes (de travail et scolaires) issues du recensement débouche sur un résultat mitigé avec des décalages importants dans la structures des matrices. La comparaison avec l'Enquête Globale Transport (2010), qui englobe tous les motifs de déplacements, donne de meilleurs résultats en termes de structure des matrices. De manière générale, l'utilisation des *CDR* apporte de meilleurs résultats.

Afin de corriger les matrices obtenues, il peut être fait usage de données auxiliaires, telles que les comptes de population des recensements pour caler les matrices avec un facteur d'échelle permettant d'approcher au mieux les volumes réels de déplacement (Chen *et al.*, 2016). Dans les produits proposés par Proximus, l'opérateur réalise un redressement de l'information (l'utilisateur final peut se baser sur un facteur de redressement associé à chaque flux dans le cadre d'une matrice O/D) mais ne diffuse pas d'information quant à la nature des traitements sur lesquels ce redressement se construit.

Les produits fournis clé sur porte revêtent une double opacité qui touche, d'une part, au processus de collecte de l'information et, d'autre part, aux retraitements des traces avant fourniture des fichiers, qui ne sont pas sans poser problème pour la constitution de séries temporelles. En effet, la téléphonie mobile est une technologie en mouvement et aussi bien les appareils mobiles que le réseau évoluent chaque année. De plus, l'opérateur a des objectifs d'optimisation des réseaux qui peuvent influencer la couverture réseau indépendamment de l'infrastructure. Sans métadonnées claires et processus de validation (avec éventuelles corrections) par des données externes, l'usage des *FMD* pour le monitoring des tendances constitue un exercice périlleux.

En matière de respect de la vie privée, on a vu que l'accès aux traces individuelles ouvrait des possibilités de ré-identification aisée des individus. Pour les produits délivrés par Proximus sans passage de la demande auprès de la Commission de protection de la vie privée, les données sont agrégées par paire origine-destination. Celles qui concernent moins de 30 traces ne sont pas communiquées.

4.1.3. Quels usages concrets pour les *FMD* ?

On peut classer l'exploitation des *FMD* en trois catégories principales.

On peut classer les produits *FMD* en essentiellement trois catégories propres à l'étude des mobilités. Premièrement, le débouché principal consiste en matrices origines-destinations des déplacements (voir Chen *et al.* (2016) et Bonnel *et al.* (2015) par exemple). Ces matrices origines-destinations peuvent être utilisées pour caractériser des lieux (par exemple le profil de fréquentation horaire autour d'un nœud de transport) ou des espaces (Debusschere *et al.*, 2017).

Deuxièmement, les *FMD* se prêtent bien au suivi longitudinal des individus dans l'espace et dans le temps. De cette manière, il est possible de dégager des profils horaires de déplacement ou des séquences d'occupation spatio-temporelles (Bayir *et al.*, 2010).

Enfin, sur base d'un suivi suffisamment long des individus au cours du temps, il est possible d'inférer l'activité pratiquées dans les espaces qu'ils fréquentent le plus souvent. Typiquement, la localisation du lieu de domicile et du lieu de l'activité principale diurne (travail ou éducation) sont relativement aisés à identifier (Chen *et al.*, 2016, Debusschere *et al.*, 2017).

4.2. Données des opérateurs de services *GPS* embarqués – *Floating car data* (*FCD*)

Les données que nous évoquons ici sont des sous-produits des outils de navigation embarqués sur les véhicules. Leurs propriétaires les collectent et les agrègent afin de fournir à leurs clients des informations sur les conditions de circulation ainsi que, généralement, un service de calcul d'itinéraire permettant, le cas échéant, de minimiser les temps de parcours. En Belgique, les principaux fournisseurs de ces données (et éventuellement de leurs services) sont TomTom, Inrix, Be-Mobile, Waze ou encore Coyote.

Contrairement aux données de téléphonie mobile, les *FCD* mesurent principalement les conditions de trafic. En néerlandais, on parle de "mobiele data"[18] pour souligner que les capteurs sont embarqués dans les véhicules en déplacement, en opposition aux capteurs fixes situés sur, en dessous, à côté ou au-dessus de la route. Ces données peuvent potentiellement servir pour mesurer les charges de trafic mais les déficits de représentativité rencontrés apparaissent aujourd'hui encore trop complexes pour pouvoir être utilisés de manière fiable et robuste.

4.2.1. Des traces mobiles aux temps de parcours

Les données *GPS* provenant de systèmes embarqués sont multiples. On peut lister les capteurs (on parle souvent de "*probe network*", soit un réseau de sondes) les plus fréquemment cités :

- Systèmes de navigation intégrés aux véhicules personnels.
- Applications de trafic sur smartphones.
- Systèmes de navigation intégrés dans une gestion de flotte (d'entreprises, notamment pour le transport de bien et de personnes, etc.).
- Certains opérateurs feraient usage de données de capteurs fixes disposés le long des voiries (Ermans *et al.*, 2017).

Les données de positionnement *GPS* sont transmises à l'opérateur via le réseau de téléphonie mobile et sont ensuite assemblées par des algorithmes tenus secrets, pour des raisons commerciales évidentes, et propres à chaque opérateur. L'objectif pour celui-ci est alors de renvoyer vers les véhicules connectés une information en temps réel pour l'aide à la navigation. En temps réel, de nombreux segments ne sont pas parcourus par un flux suffisant de véhicules permettant de produire des estimations jugées fiables et ce sont alors les données historiques qui prennent le relais. La fiabilité des données en temps réel est donc fortement variable d'un segment ou d'une période à l'autre. TomTom propose ainsi des valeurs de confiance associées à chaque donnée mais le processus de construction de ces valeurs reste relativement opaque (Ermans *et al.*, 2017).

[18] Ce terme désigne également les données issues de téléphonie mobile.

Pour le chercheur ou l'expert, cette production prend la forme d'indicateurs de temps de parcours (ou de vitesse de déplacement) par segment de route couvert par l'opérateur et généralement agrégés par intervalles de durées relativement courts (5, 15 ou 30 minutes par exemple). Il peut s'agir soit de données correspondant à un jour et à un moment particuliers de la journée, soit de données dites historiques, qui sont agrégées dans le temps.

4.2.2. Opacité et représentativité : principales contraintes à l'usage des *FCD*

L'absence de transparence, à la fois sur la composition des capteurs et sur les algorithmes utilisés pour produire l'information constitue un obstacle de taille pour l'utilisateur final des données qui ne dispose de la sorte pas d'une vision claire de la population concernée ou des choix méthodologiques effectués. Ceci génère un déficit de connaissance en ce qui concerne la maîtrise des inévitables biais qui touchent tout indicateur (Van Der Loop *et al.*, 2018).

Premièrement, il est impossible d'évaluer le niveau de représentativité des capteurs connectés (le trafic connecté constituerait 5% des véhicules aux Pays-Bas (Van Der Loop *et al.*, 2018), 8 à 10% en Île-de-France (Remesy et Belloche, 2018)) ce qui rend toute mesure du volume du trafic périlleuse.

Deuxièmement, la comparabilité dans le temps des données est incertaine. Les opérateurs tendent, en effet, à modifier régulièrement les modes d'extraction de l'information (de même que la modélisation du réseau routier) afin d'optimiser leurs services en temps réels. Ceci a pour conséquence de fausser les comparaisons dans le temps : les évolutions observées résultent-elles de modifications des conditions de trafic, des prétraitements de l'opérateur ou de la composition des capteurs ? Une solution, pas toujours applicable et forcement limitée, peut être de corriger les données par le couplage de celles-ci avec des données de capteurs fixes (Van Der Loop *et al.*, 2018).

4.2.3. Quels usages concrets pour les *FCD* ?

Les *FCD* entrent en concurrence essentiellement avec les capteurs fixes posés le long des routes pour la mesure des paramètres d'écoulement du trafic. Les *FCD* présentent, par rapport à ceux-ci, le grand avantage de couvrir une portion très complète du réseau routier là où les capteurs fixes sont généralement localisés sur les grands axes de circulation (autoroutes et voiries principales). Au contraire de ces derniers cependant, ils ne se prêtent pas bien au calcul des volumes de trafic, tout du moins dans l'état actuel des données mises à disposition. Par ailleurs, les temps de parcours et les vitesses sont assez volatiles et leur usage en valeurs absolues doit être précautionneux.

De ce fait, les usages les plus intéressants de ces données sont les suivants.

Premièrement, elles peuvent servir à évaluer la congestion sur les routes, qu'il s'agisse de la détection de points noirs en temps réels ou sur une temporalité plus longue (Van Der Loop *et al.*, 2018 – Trotta, 2016 par exemple) et surtout sur des segments du réseau routier relativement confidentiels.

Il est par exemple possible de représenter l'évolution de la congestion des voiries dans un quartier ou autour d'un carrefour important afin de mettre en évidence les dynamiques locales de propagation de la congestion (Van Der Loop *et al.*, 2018). Une autre application est de produire les temps de parcours moyens et les retards dus à la congestion sur certains trajets particuliers. On peut ensuite envisager de comparer ces trajets entre eux ou, en s'intéressant à un segment ou trajet particulier, établir des profils de vitesses et de congestion en fonction du moment de la journée (Ermans *et al.*, 2017).

La construction d'indicateurs de congestion passe généralement par la comparaison des vitesses (ou temps de parcours) observées en un segment du réseau ou pour un trajet sur celui-ci par rapport à une vitesse (ou un temps de parcours) de référence. Plusieurs approches de la congestion se distinguent autour du choix de cette valeur de référence. Celle-ci peut correspondre à une vitesse arbitrairement jugée comme acceptable (approche arbitraire), à la vitesse associée au flux maximal sur la voirie (approche technique ou "ingénieur") ou enfin, au niveau optimal d'utilisation de la route (approche économique), c'est-à-dire une situation dans laquelle le débit est maximal tout en permettant une circulation fluide, non gênée (Reymond, 2005). C'est bien souvent cette dernière approche qui est privilégiée.

On fera également remarquer qu'il faut distinguer les indicateurs de congestion diffusés par les opérateurs mêmes des indicateurs de congestion développés par des chercheurs et experts sur base des données sous-jacentes, mises à disposition par les opérateurs. Inrix et TomTom publient ainsi à intervalle régulier des données de congestion concernant les villes (Inrix index, TomTom traffic index). Au peu de confiance qu'il faut accorder aux comparaisons temporelles, s'ajoute ici la définition peu harmonisée (par rapport notamment à des critères de densité, de niveau d'urbanisation) de la notion de ville, qui impose une réserve quant aux comparaisons spatiales également (Van Der Loop *et al.*, 2018). De plus, le niveau de congestion est évalué sur base d'un trajet-type de navetteur entrant, soit les personnes qui ont à affronter le plus la congestion routière (Ermans et Brandeleer, 2016). Ceci souligne sans doute l'intérêt de communication marketing que revêt la démarche de publication de ces indicateurs (Van Der Loop *et al.*, 2018).

Deuxièmement, la connaissance des temps de parcours permet de créer des indicateurs d'accessibilité, à la manière par exemple de Lebrun (2018) dans le cadre de l'analyse des transports publics en Région de Bruxelles-Capitale. On peut ainsi évaluer l'accessibilité d'un lieu en calculant un indicateur synthétique de position (moyenne ou médiane généralement) sur l'ensemble des temps de parcours pour se rendre dans ce lieu (accessibilité à destination) ou depuis ce lieu vers tous les autres (accessibilité à l'origine), pour un territoire donné. Alternativement, il est possible de ne sélectionner qu'un certain nombre de lieux, jugés importants par rapport aux fonctions qu'ils accueillent (logement, emplois, offre scolaire, offre de culture, etc.).

Troisièmement, les temps de parcours peuvent servir dans le cadre de modèles de déplacements[19]. Ceux-ci peuvent entrer en ligne de compte dans les étapes de répartition de la demande entre lieux d'origine et de destination, à l'étape d'attribution des choix modaux ou encore à l'étape d'affectation de la demande sur le réseau. Ils interviennent alors essentiellement en tant que mesure d'éloignement, de résistance au déplacement (on parle d'impédance) qui sépare deux lieux considérés (Ermans *et al.*, 2017).

Enfin, au-delà de l'écueil encore difficilement surmontable du manque de comparabilité dans le temps (spécifiquement d'année en année), les *FCD* offrent potentiellement des possibilités d'évaluation ex-post de mesures ou de modification des infrastructures. Récemment, la destruction du viaduc Reyers, a été l'occasion de réaliser un essai de ce type (Servonnat, 2017).

4.3. Données billettiques des opérateurs de transport en commun

Nous entendons ici par données billettiques l'ensemble des données recueillies par les opérateurs de transport en commun lorsque les usagers valident leurs titres de transport sur le réseau au moyen d'un dispositif de validation automatique faisant usage de la technologie *RFID* (*Radio Frequency IDentification*). Si l'objectif premier du dispositif est de valider les titres de transport[20], il produit une manne de données en continu qu'il est possible d'exploiter pour l'étude de la mobilité des personnes sur le réseau et l'amélioration des prestations offertes par l'opérateur.

Pour l'utilisateur final des données, qui est généralement l'opérateur lui-même, celles-ci constituent une belle opportunité. En effet, celui-ci étant propriétaire des données, le coût marginal (au-delà des investissements substantiels consentis pour mettre sur pied l'infrastructure et pour la maintenir fonctionnelle) est relativement faible et il jouit de la maîtrise totale des traces recueillies et des traitements préalables à l'analyse. De ce point de vue, il s'agit d'une configuration sensiblement différente à celles qui prévalent pour les *FMD* et les *FCD*.

Les données billettiques présentent des potentialités d'usage relativement analogues aux *FMD* : elles permettent de représenter les usagers sur le réseau de manière très précise (au niveau des stations, des arrêts) avec une grande finesse temporelle en même temps qu'une fenêtre d'enregistrement très large (tant que le réseau fonctionne en fait). L'existence d'un identifiant unique attaché à chaque carte permet généralement de suivre individuellement, chaque usager sur le réseau. Elles sont en revanche évidemment restreintes aux déplacements des usagers des transports en commun concernés.

En Région de Bruxelles-Capitale, depuis 2016, la carte MOBIB équipée d'une puce *RFID* sert de support pour l'achat et la validation de tout type de billet émis par la STIB[21]. Les données récoltées par ce biais présentent un potentiel important pour l'analyse des déplacements mais leur usage se heurte encore à divers obstacles. Citons notamment la sous-validation des abonnements dans les stations et arrêts non équipés de portiques, un problème qui se pose surtout en sortie (où la validation n'est pas obligatoire et rarement contrainte par des portiques) mais aussi en entrée, où la validation n'est contrainte par des portiques que sur le réseau souterrain.

[19] Modèles d'échelles et d'applications diverses qui visent à estimer la demande de déplacement sur un territoire et, selon les cas, à l'affecter sur un réseau de transport, selon des modalités variables des paramètres de mode et de motifs de déplacements. Voir notamment Centre d'Études sur les Réseaux (2003) pour une discussion sur ce sujet.

[20] Voir Pelletier *et al.*, 2011 pour un compte-rendu des avantages et inconvénients du système évoqués dans la littérature scientifique.

[21] La STIB est le principal opérateur de transports en commun en Région de Bruxelles-Capitale mais il n'est pas le seul puisque De Lijn, le TEC et la SNCB contribuent également à l'offre globale.

4.3.1. De la trace billettique au déplacement : estimer les destinations et les correspondances

La validation en sortie de véhicule, de station, n'est généralement pas obligatoire, ce qui implique l'absence d'information directe sur le lieu et le moment de descente des usagers. Afin d'estimer les destinations sur le réseau métro rennais, Briand *et al.* (2017), les auteurs, inspirés par la solution proposée par Trépanier *et al.* (2007), ont considéré systématiquement comme station de destination d'un trajet considéré la station la plus proche, parmi les destinations possibles, utilisée lors de la validation (d'entrée en station) suivante. Dans le cas où la distance entre ces deux stations est supérieure à 500 mètres, ils considèrent que la station de destination est inconnue. Pour la destination de la dernière validation de chaque journée, la première station de validation au cours de cette même journée est prise en compte. L'hypothèse sous-jacente à cet algorithme est que les usagers se déplaçant sur le réseau n'effectuent pas à pied (ou autrement d'ailleurs) des distances qu'ils peuvent parcourir en transport en commun.

Une fois déterminée la chaîne de montées et descentes d'un véhicule se déplaçant sur le réseau, il reste à distinguer entre les descentes – montées qui s'apparentent à des correspondances de celles qui représentent des arrêts marquant la fin d'un déplacement. Dans Briand *et al.* (2017), on parle de correspondance à partir du moment où l'heure d'arrivée à la station de destination inférée et l'heure de départ à la station suivante (avec validation observée) est inférieure à 30 min. Si l'intervalle de temps est supérieur, on fait l'hypothèse qu'il s'agit de deux déplacements différents, avec des motifs propres et distincts. La même logique est utilisée dans d'autres exemples d'applications avec des intervalles de durées variables (Ma *et al.*, 2013 – Ma *et al.* 2017). Comme pour les *FMD*, le risque est ici également de négliger les éventuels petits trajets.

Notons enfin qu'il est également possible d'inférer les stations les plus proches du lieu de domicile et du lieu de travail (le cas échéant) en fonction des fréquences et moments de validation.

4.3.2. Prise en charge technique et redressement des indicateurs : principales contraintes à l'usage des données de validation

Premièrement, la gestion de volumes de données très importants n'est pas une tâche nécessairement évidente pour les opérateurs de transport en commun, *a fortiori* dans le cas de réseaux multi-opérateurs qui ajoute au volume la diversité des processus de collecte et des formats (voir Chandesris *et al.* (2017) sur le cas du réseau de transport public en Île-de-France). Les réponses techniques que ces défis appellent peuvent, dans une certaine mesure, être en décalage avec le métier, au quotidien, de l'opérateur, mais aussi avec les ressources dont il dispose. Vis-à-vis des *FMD* et *FCD*, si la maîtrise totale des données de validation permet davantage de transparence et de maîtrise des biais, elle nécessite en contrepartie un investissement en personnel qualifié important.

Deuxièmement, les données billettiques souffrent de données manquantes qui biaisent la représentativité des indicateurs. Celles-ci proviennent de différentes sources, telles que la fraude aux transports, la couverture imparfaite du réseau par des infrastructures contraignantes de validation en station ou encore la possibilité laissée à l'usager de sortir du réseau sans devoir valider. Dans certains cas, les systèmes de billettiques à validation automatiques coexistent avec des systèmes à validation manuelle. Dans ce dernier cas, une partie du trafic ne donne pas lieu à l'enregistrement de traces.

À l'échelle de flux de voyageurs agrégés par paire d'origines et destinations, Chandesris *et al.* (2017) analysent le cas du réseau de transport en Île-de-France (SNCF, RATP, Optile) et se heurtent au problème de la partialité de la couverture des gares/stations équipées de capteurs à la fois en entrée et en sortie. Pour remédier à ce problème, ils opèrent un redressement des trajets enregistrés sur base d'un plan de sondage a posteriori ("Tout se passe comme si les données avaient été collectées selon ce plan") avec une stratification sur plusieurs niveaux pour tenir compte d'un découpage spatial et temporel complexe. Les données sont ensuite redressées par un calage des éléments de la matrice sur les marges, en faisant appel à des sources ancillaires permettant d'estimer la fréquentation globale du réseau (comptages, ventes de billets, etc.).

Troisièmement, les traces collectées sont dépourvues de contexte sur les caractéristiques des individus, notamment du point de vue socio-économique (âge, genre, statut socio-professionnel, etc.). Afin d'enrichir les analyses, ces données peuvent être imputées sur base du croisement des données avec des bases de données commerciales (type d'abonnement notamment), par inférence sur base des comportements sur les réseaux ou les deux en même temps (Briand *et al.*, 2017). L'inférence des données de contexte manquantes n'offre évidemment qu'une garantie limitée quant à la fiabilité des données. Dans tous les cas, ces manipulations qui augmentent le potentiel de ré-identification des individus doivent faire l'objet d'autorisations des instances compétentes en matière de protection de la vie privée.

Les données relatives aux motifs de déplacement sont évidemment également manquantes. Il est possible d'inférer la fonction de certaines destinations grâce à la fréquence et à l'horaire de déplacement (typiquement le lieu de domicile ou le lieu de travail), ce qui permet d'associer à ces destinations un motif (se rendre à la maison, se rendre au lieu de travail). Kusakabe et Asakura (2014) proposent également une méthode de qualification des motifs de déplacements issus de données billettiques à l'aide d'un couplage de données billettiques avec des données d'enquête sur les pratiques de déplacement[22].

[22] Le couplage est effectué sur les moments et stations d'entrées et de sorties du réseau et les motifs inférés des distributions observées dans l'enquête sur les pratiques de déplacements.

4.3.3. Quels usages concrets pour les données billettiques ?

Les analyses rendues possibles par ces données sont logiquement très proches de celles permises par les données issues de la téléphonie mobile. On peut les organiser selon les catégories suivantes.

Premièrement, les matrices origines/destinations selon le moment de la journée. Si l'on dispose de données contextuelles complémentaires, il est également possible d'établir ces matrices par type d'usager (étudiant, travailleur, écolier) en fonction de la disponibilité de variables complémentaires.

Deuxièmement, les profils types des usagers en fonction de l'usage qu'ils font du réseau (fréquence et horaire de déplacement) au cours d'une journée (Briand *et al.*, 2017), éventuellement selon le type de jour (ouvrable, samedi, dimanche), voire sur une période d'une semaine Ma *et al.* (2013) ou d'un mois (Ma *et al.*, 2017). Par exemple, Ma *et al.* (2013) utilisent les données billettiques du réseau pékinois afin d'identifier les routes régulièrement employées par les usagers ainsi que la régularité d'usage de ces routes (en termes d'heures de départ, d'intensité hebdomadaire, etc.).

Troisièmement, il est possible de caractériser les stations en fonction de leur profil de fréquentation dans le temps.

Pelletier *et al.* (2011) synthétisent les débouchés des études sur données de validation pour les opérateurs de transport en commun sur trois axes : (1) les études stratégiques, qui concernent la prise de décision à long terme (extension/modification du réseau, prédiction de la demande, anticipation des évolutions des comportements des usagers) – (2) les études tactiques, qui concernent l'ajustement des services à la demande (modification des horaires, adaptation des fréquences, ajustement des correspondances, etc.) – (3) les gains opérationnels (indicateurs de performance (régularité, vitesse, adhérence aux horaires), flexibilité tarifaire, amélioration du système de billettique, etc.). Ajoutons à cette liste la possibilité d'évaluer l'impact sur la demande de mesures de planification. Briand *et al.* (2017) évaluent par exemple l'effet de lissage de la pointe du matin résultant du décalage de l'heure de début des cours des étudiants de l'Université de Rennes.

Les nouvelles perspectives d'analyses ouvertes par les données de validation sont évoquées avec un enthousiasme certain dans la littérature, notamment en ce qu'elles permettent un monitoring continu, mais de nombreux auteurs soulignent la nécessaire complémentarité de ces nouvelles sources avec les enquêtes de déplacement traditionnelles, qui seules permettent des analyses multidimensionnelles riches (Briand *et al.*, 2017). Globalement, des sources de données produites en parallèle sont nécessaires, aussi bien pour l'inférence ou l'estimation des données manquantes que pour la validation de l'information remontée.

4.4. Comparaison des exemples d'exploitation présentés

En guise de conclusion de cette section, nous reprenons ici quelques caractéristiques qui nous semblent rapprocher ou distinguer les trois sources de données passées en revue. Premièrement, toutes mobilisent un processus de remontée du signal qui fait du passage de la trace à l'indicateur un processus extrêmement complexe, nécessitant des données complémentaires afin d'assurer un éventuel redressement et, surtout, une validation de l'ensemble du processus.

Deuxièmement, l'intérêt des indicateurs pour l'opérateur qui les produit et pour son *core business* est très variable et peut dès lors influer sur leur qualité. Ainsi, les opérateurs de téléphonie mobile n'ont aucun intérêt *a priori* à disposer d'une connaissance fine des déplacements des personnes par rapport à la fourniture de leurs services principaux. À l'inverse, les *FCD* sont au cœur des services de navigation embarqués, même si les objectifs des opérateurs de ceux-ci ne se confondent pas exactement avec ceux des chercheurs (recherche du trajet optimal du point de vue des temps de parcours vs représentation la plus fidèle possible des temps de parcours sur le réseau routier). Enfin, les opérateurs de transport en commun auront pour objectif principal la gestion et la planification de l'offre de transport à divers niveaux (stratégique, tactique, opérationnel) et seront donc attentifs à leur validité.

Troisièmement, le processus de production des indicateurs tendra à être moins documenté pour les opérateurs privés, notamment dans le cadre de la préservation de secrets commerciaux. Ceci engendre des difficultés supplémentaires pour la validation des données et, en particulier, pour l'utilisation de celles-ci pour la construction de tendances.

	FMD	FCD	Données de validation
Propriétaire	Privé	Privé	(para)public
Rôle des indicateurs créés pour les activités de l'opérateur	Limité	Services commerciaux	Gestion du réseau aux niveaux stratégique, tactique et opérationnel
Collecte et traitement des données	Opaque	Opaque	Transparent
Validation nécessaire	Oui	Oui	Oui
Comparabilité temporelle	–	–	+/-

Conclusion générale

Nous avons montré que l'émergence des données *big data* entraîne l'apparition d'un nouveau paradigme en termes de production de savoir. Celui-ci est lié premièrement aux caractéristiques de ces données : (1) données massives (les 3 V) qui impliquent une maîtrise technique importante et (2) distance entre la donnée brute et l'information, le signal, qu'on peut en extraire. Ces informations nécessitent un traitement préalable à l'analyse, à la fois important et spécifique, qui suit une direction davantage inductive que déductive.

Ensuite, les *big data* jouissent aussi d'un engouement parfois positiviste et d'une émulation particulière, notamment dans le contexte du développement des *smart cities*. À cet égard, les données massives se distinguent peut-être moins par leur capacité à agréger un type nouveau de données, qu'en tant que cadre émergent de production de connaissance, de nouvelle manière de *rendre le monde signifiant* (Rouvroy, 2014). L'importance dans ce cadre des méthodes de remontée du signal et d'effet loupe, peuvent alimenter un sentiment d'objectivité naturelle face auquel l'on doit rester prudent, sous peine d'assister à un appauvrissement de la diversité des données et des méthodes d'analyse.

De même, la promotion des *big data* au sein de l'imaginaire de la *smart city* promeut une action publique dynamique, où les stimuli perçus par *l'internet des choses* devraient être directement interprétés et traduits en actions, avec la visée d'optimiser la mobilité urbaine en temps réel (redistribution du trafic, modulation de l'offre de transport, de la capacité des voiries, etc.). De ce point de vue, la problématique des *big data* englobe également la question des modalités du passage du savoir produit à l'*action* sur le réel. Il nous apparaît ainsi important de continuer à valoriser, en contrepoint, une production de connaissance qui puisse servir la programmation sur des temporalités plus longues (dans le cadre d'études stratégiques notamment) et permette une problématisation de la mobilité au-delà de la rationalisation à court terme des flux de déplacement.

L'usage des *big data* peut évidemment être envisagé au même titre que les outils et indicateurs "traditionnels" (enquêtes, comptages, données administratives) dans le cadre d'une production de connaissances à visée plus stratégique. Ainsi redimensionnées au statut d'outil parmi d'autres dans l'arsenal méthodologique, elles offrent une opportunité pour les études de mobilité qui repose essentiellement sur une exhaustivité et une granularité très fine de la couverture spatio-temporelle, qu'on ne peut généralement pas espérer atteindre par des enquêtes ou capteurs fixes (notamment pour les déplacements en heures creuses ou les paramètres du trafic automobile sur l'ensemble du réseau routier), ne fût-ce que pour des raisons budgétaires.

Au-delà de ces opportunités, les *big data* ne remplacent pas pour autant les données plus classiques poste pour poste. Premièrement, elles présentent généralement un déficit d'informations contextuelles (âge, genre, caractéristiques socio-économiques en ce qui concerne les individus – modes, motifs et vécus en ce qui concerne les déplacements) qui doit être comblé par d'autres sources. De ce fait, elles se prêtent moins à l'analyse des causalités qu'à la description et à la caractérisation des flux et des déplacements.

Deuxièmement, la complémentarité avec les sources plus classiques réside également dans le nécessaire exercice de validation, et éventuellement de redressement, des données dans la mesure où celles-ci subissent de nombreux traitements en aval de la collecte (imputations probabilistes de données manquantes, de variables absentes, redressement, etc.). De ce point de vue, le recours aux *big data* semble, dans une certaine mesure, indissociable de l'existence de points de comparaison qui permettent de les valider. Ceci nous mène à relativiser le caractère bon marché (à ce stade !) des *big data* fournies clé-sur-porte par des opérateurs privés puisqu'il faut également y intégrer les coûts de validation (main-d'œuvre, traitement et production de données de comparaison, etc.).

Troisièmement, il ne fait aucun doute qu'aussi bien les technologies qui permettent la captation de traces (réseaux télécoms, *GPS*, etc.) que les populations connectées grâce à celles-ci vont évoluer. Les traitements qui visent à faire émerger le signal des données récoltées devront donc être adaptés également. Tout ceci implique un renouvellement régulier des procédures de validation. Cette instabilité du processus de production des indicateurs *big data* met également en cause l'usage de ceux-ci pour la constitution de séries temporelles de données permettant d'établir des tendances.

Cette complexité du processus de production est plus problématique dans le cas de producteurs de données privés qui communiquent peu sur celui-ci, notamment dans un souci de préservation du secret commercial, ce qui amène à poser la question des rapports entre administrations publiques et opérateurs privés. De ce point de vue, il semble que les premières auraient tout intérêt à prendre langue avec les seconds afin de faire entendre leurs besoins et de coproduire le processus de production des données. Pour les pouvoirs publics, il s'agirait d'améliorer la qualité et la maîtrise des indicateurs tout en minimisant la mise en dépendance de ses méthodes de travail et de ses ressources vis-à-vis des logiques de production propres à l'acteur privé. Pour ce dernier, ce serait l'occasion d'obtenir une validation de ses produits statistiques par un organisme public, ce qui en augmenterait la visibilité et la légitimité.

Bibliographie

Allemand L. (2013), "Dossier – Les promesses du big data", *La Recherche*, décembre 2013, pp. 27-42.

Anderson C. (2008), "The end of theory?: The data deluge makes the scientific method obsolete", *Wired Magazine*, vol. 16, n°7.

Batty M. (2013), "Big data, smart cities and city planning", *Dialogues in Human Geography*, vol. 3, n°3, pp. 274-279.

Batty M., Axhausen K.W., Giannotti F., Pozdnoukhov A., Bazzani A., Wachowicz M., Ouzounis G. et Portugali Y. (2012), "Smart cities of the future", *The European Physical Journal Special Topics*, vol. 214, n°1, pp. 481-518.

Bayir M.A., Demirbas M. et Eagle N. (2010), "Mobility profiler : A framework for discovering mobility profiles of cell phone users", *Pervasive and Mobile Computing*, vol. 6, n°4, pp. 435-454.

Bensamoun A., Zolynsky C. (2015), "Cloud computing et big data. Quel encadrement pour ces nouveaux usages des données personnes ?", *Réseaux,* n°189, pp. 103-121.

Bonnel P., Hombourger E., Olteanu-Raimond A.-M. et Smoreda Z. (2015), "Passive Mobile Phone Dataset to Construct Origin-destination Matrix : Potentials and Limitations", *Transportation Research Procedia*, vol. 11, pp. 381-398.

Boullier D. (2015), "Les sciences sociales face aux traces du big data ? Société, opinion et répliques", *FMSH*, WP, n°88.

Boyd D. et Crawford K. (2012), "Critical questions for Big Data : Provocations for a cultural, technological, and scholarly phenomenon", *Information, Communication & Society*, vol. 15, n°5, pp. 662-679.

Brandeleer C. et Ermans T. (2016), "Quand gérer des feux de circulation préfigure des choix de mobilité : les enjeux stratégiques d'un outil technique", *Brussels Studies*, n° 103.

Briand A.-S., Côme E., Coulombel N., El Mahrsi M.K., Munch E., Richer C. et Oukhellou L. (2017), "Projet MOBILLETIC, Données billettiques et analyses des mobilités urbaines : le cas de Rennes", *in* André De Palma et Sophie Dantan (dir.), *Big data et politiques publiques dans les transports*, Paris, Economica, pp. 174-196.

CERTU (2003), "Modélisation des déplacements urbains de voyageurs : guide des pratiques". Centre d'études sur les réseaux, les transports, l'urbanisme et les constructions publiques, Ministère de l'équipement, des transports, du logement, du tourisme, et de la mer, Direction des transports, Lyon, 244 p.

Chandesris M., Ganansia F. et Remy A. (2017), "Les données massives au service des mobilités de demain", *in* André De Palma et Sophie Dantan (dir.), *Big data et politiques publiques dans les transports*, Paris, Economica, pp. 138-169.

Chen C., Ma J., Susilo Y., Liu Y. et Wang M. (2016), "The promises of big data and small data for travel behavior (aka human mobility) analysis", *Transportation Research Part C : Emerging Technologies*, vol. 68, pp. 285-299.

CIRB (2014), "Livre blanc 2014-2019", Centre d'Informatique pour la Région Bruxelloise, Bruxelles, 71 p.

CIRB (s.d.), "Smart city strategie" Centre d'Informatique pour la Région Bruxelloise, Bruxelles, 4 p.
https://cirb.brussels/fr/fichiers/brussels-smart-city-strategie

Commenges H. (2014), "La mobilité comme variabilité temporelle de la présence spatiale", *Flux*, 2014, 1(95), pp. 41-55.

Connelly R., Playford C. J., Gayle V., Dibben C. (2016), "The role of administrative data in the big data revolution in social science research", *Social Science Research*, n°59, pp. 1-12.

Cytermann L. (2015), "Promesses et risques de l'open et du big data : les réponses du droit", *Informations sociales*, n°191, pp. 80-90.

Debusschere M., Lusyne P., Dewitte P., Baeyens Y., De Meersman F., Seynaeve G., Wirthmann A., Demunter C., Reis F. et Reuter H.I. (2017), "Big data et statistiques. Un recensement tous les quarts d'heure...", Bruxelles, Direction générale statistique – Statistics Belgium, 22 p.

De Montjoye Y.-A., Hidalgo C., Verleysen M., Blondel V. (2013), "Unique in the crowd : the privacy bounds of human mobility", *Scientific reports,* vol. 23, n°1376, pp. 1-5.

De Palma A. (2017), "Tour d'horizon et repérages", *in* André De Palma et Sophie Dantan (dir.), *Big data et politiques publiques dans les transports*, Paris, Economica, pp. 15-35.

Determe K. (2018), "Analyse macro des flux de mobilité grâce aux données de téléphonie mobile". Présenté à la 67e Réunion de la Commission régionale de la mobilité, Bruxelles, le 23 avril.

DigitYser (2017), "DigitYser opent "digitaal clubhuis" voor Next-ondernemers in hartje Brussel", 18 décembre 2017.

Douay N. et Henriot C. (2016), "La Chine à l'heure des villes intelligentes", *L'Information géographique*, vol. 80, n°3, pp. 89-102.

Ermans T., Haynes J., Kluyskens E. et Servonnat G. (2017), "Inventaire et possibilités de croisements de sources de données statistiques sur la mobilité des personnes en Belgique", Bruxelles, SPF Mobilité et Transports, 111 p.

Fayyad U., Piatetsky-Shapiro G. et Smyth P. (1996), "From data mining to knowledge discovery in databases", *AI magazine*, 1996, pp. 37-54.

Floridi L. (2012), "Big Data and Their Epistemological Challenge", *Philosophy & Technology*, vol. 25, n°4, pp. 435-437.

Gouvernement bruxellois (2014), "Projet d'accord de majorité 2014-2019", Bruxelles, Be.Brussels.

Graham M. et Shelton T. (2013), "Geography and the future of big data, big data and the future of geography", *Dialogues in Human Geography*, vol. 3, n°3, pp. 255-261.

Grosjean A. (2015), "Le profilage : un défi pour la protection des données à caractère personnel", in Grosjean (dir.) *Enjeux européens et mondiaux de la protection des données personnes,* Bruxelles, Ed. Larcier, pp. 277-310.

IDC (2014), "The digital universe of opportunities – Rich data and the increasing value of the internet of things", EMC, 17 p.

Jourova V. (2016), "La réforme de la protection des données dans l'UE et les mégadonnées", fiche technique, Commission européenne, direction générale Justice et Consommateurs, 4 p.

Kusakabe T. et Asakura Y. (2014), "Behavioural data mining of transit smart card data : A data fusion approach", *Transportation Research Part C : Emerging Technologies*, vol. 46, pp. 179-191.

Laney D. (2001), "3D data management : controlling data volume, velocity, and variety", Meta Group, 4 p.

Lebrun K. (2018), "Temps de déplacements en transport public à Bruxelles : l'accessibilité des pôles d'activités", *Brussels Studies*, n°123.

Leonard T., Chaumont D. (2016), "Commentaire général du GDPR", Ulys, Bruxelles, 19 p.

Ma X., Liu C., Wen H., Wang Y. et Wu Y.-J. (2017), "Understanding commuting patterns using transit smart card data", *Journal of Transport Geography*, vol. 58, pp. 135-145.

Ma X., Wu Y.-J., Wang Y., Chen F. et Liu J. (2013), "Mining smart card data for transit riders' travel patterns", *Transportation Research Part C : Emerging Technologies*, vol. 36, pp. 1-12.

Malaurent J. (2017), "Big data : enjeux et applications pour appréhender la mobilité", *in* André De Palma et Sophie Dantan (dir.), *Big data et politiques publiques dans les transports*, Paris, Economica.

Miller H.J. (2010), "The data avalanche is here. Shouldn't we be digging ?", *Journal of Regional Science*, vol. 50, n°1, pp. 181-201.

Pelletier M.-P., Trépanier M. et Morency C. (2011), "Smart card data use in public transit : A literature review", *Transportation Research Part C: Emerging Technologies*, vol. 19, n°4, pp. 557-568.

Remesy R. et Belloche S. (2018), "Floating car data : quel bilan pour la gestion du trafic ? ", *TEC*, avril, n°237, pp. 38-39.

Reymond M. (2005), "La tarification de la congestion automobile : Acceptabilité sociale et redistribution des recettes du péage", Thèse de doctorat, Montpellier, Université Montpellier 1, 339 p.

Ricciato F., Widhalm P., Craglia M. et Pantisano F. (2015), *Estimating population density distribution from network-based mobile phone data*, Luxembourg, Publications Office of the European Union, 70 p.

Rouvroy A. (2016), "Des données et des hommes. Droits et libertés fondamentaux dans un monde de données massives", Rapport de recherche, Bureau du comité consultatif de la convention pour la protection des données à l'égard du traitement automatisé des données à caractère personnel, Conseil de l'Europe, 45 p.

Rouvroy A. (2014), "Des données sans personne : le fétichisme de la donnée à caractère personnel à l'épreuve de l'idéologie des Big Data", in Conseil d'Etat (dir.) *Le numérique et les droits fondamentaux,* La Documentation française, pp. 407-421.

Servonnat G. (2017), "Big data – The theory of chaos", Results Presentation BCUS Civil Society Fellowship 2017, Bruxelles, 19 décembre 2017.

Trépanier M., Tranchant N. et Chapleau R. (2007), "Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System", *Journal of Intelligent Transportation Systems*, vol. 11, n°1, pp. 1-14.

Trotta M. (2016), "Que nous apprennent les données *GPS* sur la vitesse de nos routes ? Mesure de comportement : vitesse hors agglomération 2015", Rapport de recherche, n°2016-R-03-FR, Bruxelles, Institut Belge pour la Sécurité Routière – Centre de Connaissance Sécurité Routière, 50 p.

van der Loop H., Francke J., Jorritsma P. et Moorman S. (2017), *Bruikbaarheid van floating car data voor beleidsonderzoek*, Den Haag, Kennisinstituut voor Mobiliteitsbeleid (KiM), 46 p.

Vayatis N. (2017), "La décision par algorithme", *in* André De Palma et Sophie Dantan (dir.), *Big data et politiques publiques dans les transports*, Paris, Economica, pp. 51-70.

Éditeur responsable : Camille Thiry – rue du Progrès 80 – 1035 Bruxelles

Rédaction : Thomas Ermans, Céline Brandeleer et Michel Hubert

Fonds de plan de la RBC : Brussels UrbIS® © CIRB

Photos : SPRB - p.4 : STIB-MIVB

Layout et production : Altavia ACT* - www.altavia-act.com

BRUXELLES MOBILITÉ
SERVICE PUBLIC RÉGIONAL DE BRUXELLES