Contributions of passive data to the understanding of travel behaviours? : Implications for the planning and organisation of public transit system.

Oscar Egu-Festas

N° d'ordre NNT : 2020LYSE2055

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 486 Sciences Économique et de Gestion

Discipline : Sciences économiques

Soutenue publiquement le 28 octobre 2020, par :

Oscar EGU

Apports des données passives à la compréhension des comportements de mobilité ?

Enjeux pour la planification et l'organisation des transports en commun.

Devant le jury composé de :

Sophie MASSON, Professeure des universités, Université de Perpignan, Présidente

Latifa OUKHELLOU, Directrice de Recherche, IFSTTAR, Rapporteure

Thierry BLAYAC, Professeur des universités, Université Montpellier, Rapporteur

Catherine MORENCY, Professeure, École Polytechnique de Montréal, Examinatrice

Michel BIERLAIRE, Professeur, École Polytechnique Fédérale de Lausanne, Examinateur

Patrick BONNEL, Ingénieur de Recherche, École Nationale des Travaux Public de l'État, Directeur de thèse

Contrat de diffusion

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de :

L'Université Lumière Lyon 2

École Doctorale : ED486 - Economie et gestion

Discipline : Sciences économiques

Laboratoire : LAET - Laboratoire d'aménagement et d'économie des transports

Entreprise : Keolis Lyon - CIFRE 2017/1082

Présentée et soutenue publiquement le 28/10/2020, par :

Oscar EGU

Apports des données passives à la compréhension des comportements de mobilité ?

Enjeux pour la planification et l'organisation des transports en commun

---

Devant le jury composé de :

Sophie MASSON Professeure des universités, Université de Perpignan	**Présidente**
Latifa OUKHELLOU Directrice de Recherche, Université Gustave Eiffel	**Rapportrice**
Thierry BLAYAC Professeur des universités, Université de Montpellier	**Rapporteur**
Catherine MORENCY Professeure titulaire, Ecole Polytechnique Montréal	**Examinatrice**
Michel BIERLAIRE Professeur titulaire, Ecole Polytechnique Fédérale de Lausanne	**Examinateur**
Patrick BONNEL Docteur Ingénieur, École Nationale des Travaux Publics de l'État	**Directeur de thèse**

« Just as freemarket capitalists believe in the invisible hand of the market, so Dataists believe in the invisible hand of the dataflow. »

Yuval Noah Harari

Remerciements

Au terme de ce travail, je souhaite exprimer ma gratitude envers tous ceux qui ont contribué, de près ou de loin, à l'élaboration de cette thèse et à rendre ces trois années enrichissantes tant humainement qu'intellectuellement.

Je tiens donc à remercier dans un premier temps, l'ensemble des acteurs du monde des logiciels libres qui œuvre pour que nous puissions avoir des outils performants et gratuits.

Un grand merci à toute l'équipe du Laboratoire d'Aménagement et d'Économie des Transports de l'ENTPE pour l'accueil et la bonne ambiance.

Bien évidemment, il me faut remercier l'entreprise Keolis Lyon d'avoir accepté de financer cette thèse, de m'avoir fourni les données indispensables à sa réalisation et d'avoir répondu à mes interrogations. Sans eux, ce travail n'aurait jamais vu le jour et puis quelle chance d'avoir un environnement de travail aussi stimulant et aussi agréable.

Pour terminer, je tiens à remercier tout particulièrement Patrick Bonnel pour avoir dirigé de bout en bout cette thèse.

Selon la formule consacrée, je reste bien sur seul responsable des erreurs et des oublis qui pourraient suivre, mais aussi des propos émis dans cette thèse.

Avant-propos

Cette thèse s'articule autour de 4 articles rédigés en anglais. L'introduction et la conclusion sont rédigées en français.

Je suis l'auteur principal des 4 articles. Patrick Bonnel a accompagné la conception et la rédaction de chacun de ces articles. Chaque article est inséré tel qu'il a été soumis, avec des objectifs qui lui sont propres, mais aussi certaines redondances notamment dans la présentation des données.

Sauf mention contraire, les tableaux et figures s'appuient sur les données du réseau de transport en commun de Lyon. Lorsque les sources des figures et tableaux ne sont pas précisées, cela signifie que je les ai réalisés dans le cadre de la thèse. Il est préférable de visualiser les figures en couleurs.

Pour naviguer de manière rapide dans le manuscrit au format pdf le lecteur peut s'appuyer sur les hyperliens. Le texte permettant le renvoi est en bleu.

Table des matières

1

Introduction générale

« You can't manage what you don't measure. »
W. Edwards Deming

La thèse qui débute ici est centrée sur le système de transport en commun urbain. Ce système a pour but de permettre aux citoyens de se déplacer au sein d'un périmètre géographique donné sans avoir recours à un mode de déplacement privé. Il a un rôle central dans l'organisation des villes, dans le quotidien des urbains, et façonne les comportements de mobilité. Il est au cœur d'enjeux politiques, environnementaux et économiques, et produit des bénéfices certains pour les individus et la communauté. Ce système est par nature complexe et nécessite la mise en place de larges organisations, ainsi que l'utilisation de processus de planification. Ces processus ont pour but d'améliorer en continu l'efficience du système de transport en commun et sont alimentés par un dispositif de collecte et d'analyse de données. Ce dispositif permet de faire le lien entre le monde de la planification et le monde réel entendu comme la production de l'offre de transport. La généralisation progressive de système de transport en commun dit intelligent s'accompagne d'une multiplication croissante des sources de données. Les réseaux de transport en commun collectent désormais en continu et de manière passive, c'est-à-dire sans intervention humaine, des quantités massives de données. Ces big data sont à même de modifier les pratiques traditionnelles de planification et d'organisation d'un système de transport en commun. Ils devraient aussi permettre d'améliorer la compréhension des comportements de mobilité. C'est ce que cette thèse ambitionne d'explorer à travers quatre articles scientifiques qui constituent les chapitres 2 à 5 de ce document. Avant cela, il convient néanmoins de :

— Replacer la thèse dans un contexte général de planification des systèmes de transport commun ;
— Présenter les motivations de la thèse et sa problématique ;
— Détailler les contributions scientifiques et le plan du présent manuscrit.

1.1 Le contexte : système de transport en commun, planification et collecte des données

Cette première section a pour objectif de contextualiser la thèse et se veut volontairement brève. Pour ce faire, nous commençons par présenter une vue stratégique des systèmes de transport en commun. Celle-ci montre que ces systèmes sont complexes et doivent être planifiés. Nous définissons donc ce que nous entendons par planification et nous présentons le schéma classique de planification opérationnelle d'un système de transport en commun. Enfin, nous montrons que la collecte des données est une étape indispensable pour la planification et l'exploitation des réseaux de transport en commun.

1.1.1 Les systèmes de transport en commun, des systèmes complexes

Il est possible de concevoir l'espace urbain en le décomposant en plusieurs systèmes tels que le système de localisation, le système de pratiques et relations sociales ou bien le système de transport [Bonnafous and Puel, 1983]. Cette thèse se focalise sur un sous ensemble du système de transport : le système de transport en commun. L'adjectif en commun est ici utilisé pour faire référence aux modes de transports publics et collectifs qui permettent le déplacement de personnes au sein d'un réseau urbain articulé autour de lignes de métro, bus et tramway.

Le système de transport en commun ne peut pas être dissocié du contexte sociétal dans lequel il opère et des objectifs qui lui sont assignés. La figure 1.1 est une adaptation libre en français de [Fielding, 1987]. Elle permet de replacer dans une perspective plus générale l'objet d'étude principal de cette thèse : le système de transport en commun.

Tout d'abord, il nous faut considérer les facteurs qui pourraient avoir une influence sur cet objet. Ces facteurs, souvent considérés comme externes, peuvent être économiques, démographiques, culturels, spatiaux, juridiques, etc. Ils vont influencer les comportements de déplacements et donc les caractéristiques de la demande, générer un certain nombre de contraintes et façonner les choix politiques. Des structures organisationnelles et décisionnelles auront alors pour but de mettre en relation ces éléments, afin de fixer des objectifs et prendre les décisions majeures concernant l'allocation des facteurs de production vers le système de transport en commun. C'est le sens de la partie haute de la figure 1.1.

Ensuite, le système de production d'un réseau de transport en commun peut être représenté à l'aide d'un triangle visible au centre de la figure 1.1. Le sommet supérieur de ce triangle symbolise les coûts de production. Ceux-ci dépendent inévitablement du capital alloué aux transports en commun qu'il soit physique (immobilier, infrastructure, matériel roulant et outils de production), immatériel

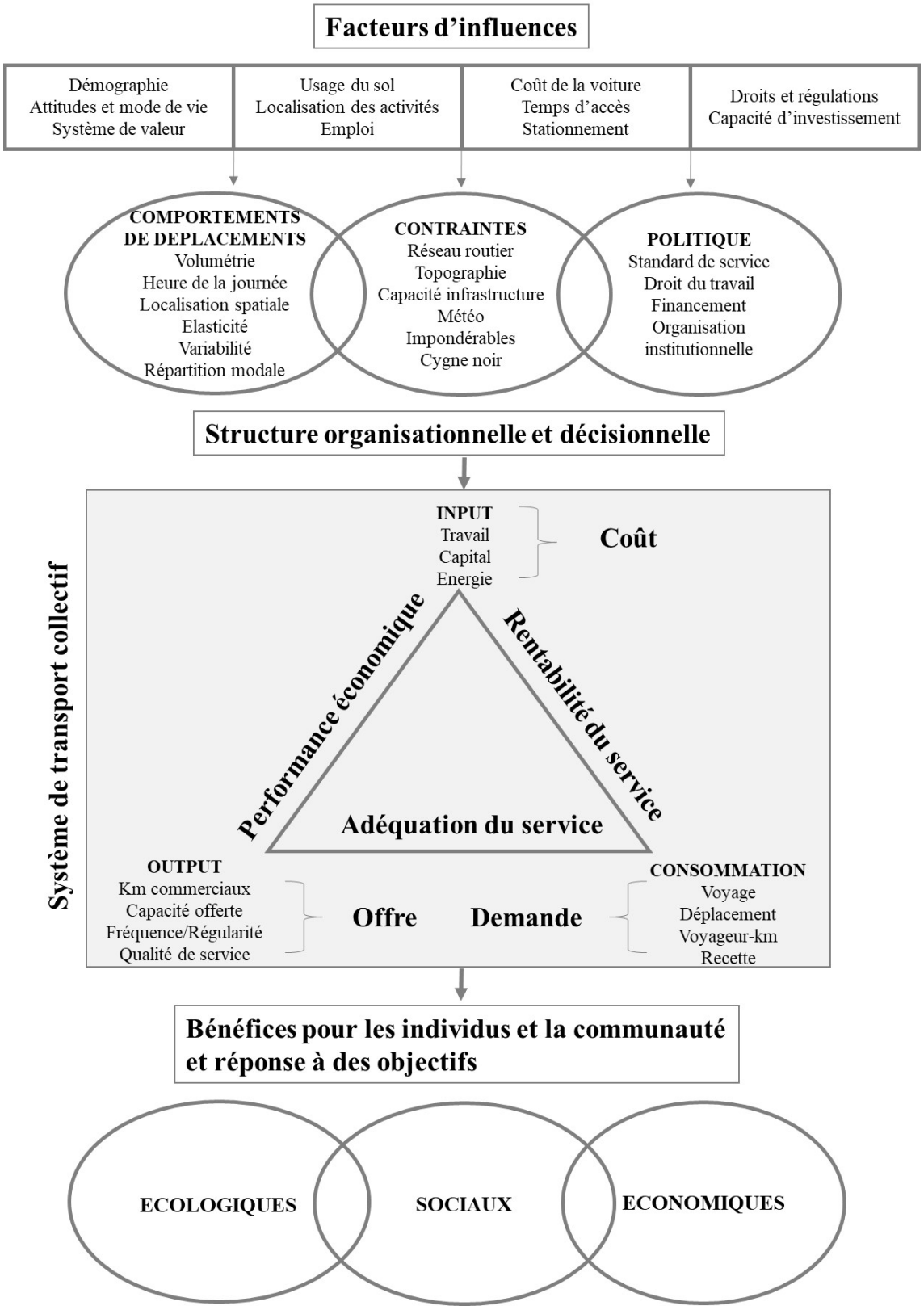


FIGURE 1.1 – Vue stratégique et contextuelle d'un système de transport en commun, Source : Adaptation libre en français de Fielding [1987]

(savoir-faire), humain (force de travail) ou bien énergétique (électricité, thermique). Ensuite, un deuxième sommet symbolise les outputs, c'est-à-dire les

services produits résultant de l'utilisation des facteurs de production. Ceux-ci sont quantifiables avec des indicateurs d'offres tels que le nombre de kilomètres produits, la capacité (places-km offertes) ou bien la vitesse commerciale. Enfin, le troisième sommet symbolise les services consommés qui peuvent se quantifier en termes de nombre de déplacements, nombre de voyages, nombre de titres vendus, etc. Les services consommés correspondent donc à la demande observée sur le réseau pour un niveau d'offre donné. La performance économique d'un système de transport en commun est alors définie comme le rapport entre les coûts et les quantités de service produit, c'est-à-dire l'offre. La rentabilité du service se définit comme le rapport entre les coûts et la consommation qui est faite du service. L'adéquation du service se mesure comme le rapport entre l'offre et la demande.

Les systèmes de transport en commun ont bien pour but d'engendrer des bénéfices individuels et collectifs qui dépendent des objectifs initiaux assignés au système. Ces objectifs sont nombreux et parfois opposés. Ils varient selon les villes et évoluent forcément avec le temps. Sans en faire la liste exhaustive, ces objectifs peuvent être de nature variée :

— **Économique :** réduction des coûts, réduction de la congestion, minimisation des temps de déplacements, augmentation de la rentabilité, réduction des externalités négatives, réduction de la dépense publique, accompagnement des développements fonciers, etc. ;
— **Environnementale :** amélioration de la qualité de l'air, réduction de la dépendance aux énergies fossiles, limitation des émissions de polluants et du bruit, limiter l'étalement urbain, etc. ;
— **Sociétale :** donner un accès à la ville ou aux aménités à ceux qui ne peuvent pas conduire, faciliter la mobilité quotidienne des plus vulnérables, améliorer la qualité de vie, répondre aux inégalités et réduire les divisions en fournissant un service universel, s'adapter aux nouveaux modes de vie, etc.

L'étude de la figure 1.1, bien que très générale et volontairement brève nous montre bien qu'un système de transport en commun ne se construit pas *ex nihilo* et est le résultat d'interactions multiples. Il fait intervenir une multitude de parties prenantes que nous pouvons, pour simplifier, grouper en trois grandes catégories. Les pouvoirs publics et les collectivités locales qui définissent le cadre institutionnel et légal, fixent les objectifs de la politique de transport et participent largement à leur financement. Les opérateurs ou exploitants (privés ou public) qui assurent la production opérationnelle du service. Les voyageurs qui utilisent le système pour satisfaire leur besoin de mobilité, et sont donc à ce titre, les consommateurs du service et principaux bénéficiaires. Plus généralement, l'ensemble des individus qui se déplacent peuvent retirer des bénéfices de la présence d'un système de transport en commun. Nous observons donc au sein de ce système l'émergence d'une organisation avec une multitude de parties prenantes et d'interrelations, un grand nombre d'interactions, des boucles de rétroaction et de causalité non triviales. Ce système peut donc être envisagé comme un système complexe [Morin, 2015].

1.1.2 Planification et organisation des systèmes de transport en commun

Pour organiser de manière optimale les systèmes de transport en commun, répondre aux objectifs fixés et donc produire les bénéfices escomptés, l'organisation des services et la planification sont des outils indispensables. L'action de planification peut se définir comme l'organisation selon un plan et des méthodes d'un service dans le temps et dans l'espace. Elle vise à réfléchir aux moyens à mettre en oeuvre pour réaliser des objectifs, et à anticiper l'action [Bonnel, 2002]. Dans le cas des systèmes de transport, le processus de planification est le résultat de la rencontre entre des méthodes de mesure et d'ingénierie, des méthodes d'évaluations socio-économiques et des logiques de décision politique [Commenges, 2013].

Puisque la planification est un processus temporel, il est d'usage de distinguer trois grands horizons de planification [van de Velde, 1999, Pelletier et al., 2011, Ortúzar and Willumsen, 2011]. La planification stratégique qui organise le système de transport en commun à des échelles de temps long (5 ans et plus). C'est à ce niveau de planification que sont formulés les grands objectifs (par exemple à travers un plan de déplacements urbains) et que les décisions sur les choix d'investissements majeurs sont prises (par exemple la construction d'une nouvelle infrastructure de transport). La planification tactique vise à assurer l'adéquation des moyens et la bonne mise en œuvre des objectifs. Elle se concentre sur des horizons temporels de 1 à 2 ans et vise à définir les caractéristiques précises du service de transport en commun comme l'ajustement des fréquences, les adaptations des tracés et les modifications tarifaires. Enfin, le niveau de planification opérationnelle qui a pour but d'assurer la production du service au jour le jour, et opère donc à des horizons temporels beaucoup plus courts (inférieur à 6 mois et parfois en temps réel).

Une autre manière de décrire le processus de planification, centrée cette fois uniquement sur le système de transport en commun, est de le décomposer en une succession d'opérations. La figure 1.2 est une adaptation en français de Ceder [2016] initialement proposée par Ceder and Wilson [1986]. Sur ce schéma, l'activité de planification d'un système de transport en commun est décomposée en quatre étapes : la définition topologique du réseau, le dimensionnement de l'offre (tables horaires), l'affectation des véhicules aux horaires (graphicage) et enfin l'affectation des ressources humaines aux véhicules (habillage). Les sorties de chaque étape viennent alimenter l'étape suivante formant un système décisionnel rétroactif qui doit être envisagé dans son ensemble pour maximiser l'efficience globale du système [Ceder, 2016]. Au plus haut niveau de cette chaîne, nous trouvons les activités de nature plus stratégique et tactique et qui visent surtout à améliorer l'adéquation du service (cf. triangle, figure 1.1). Ce sont des activités peu automatisées dans le sens où chaque scénario nécessite la prise en compte des facteurs externes, la collecte et l'analyse de données décrivant la demande et l'évaluation des impacts. Ces activités sont largement guidées par la modélisation et l'expertise métier (« jugement d'expert »). Les changements introduits dans le système de transport à ces niveaux sont donc plus conséquents pour l'organisation

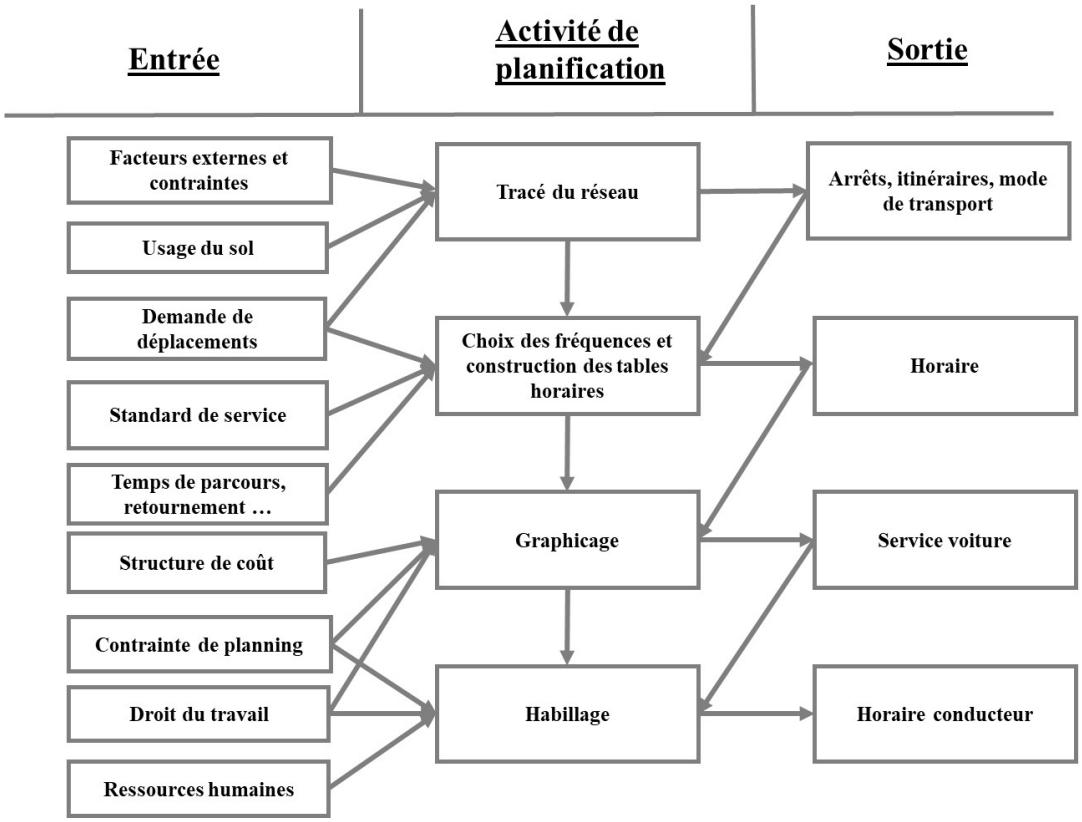


FIGURE 1.2 – Planification opérationnelle d'un système de transport en commun, Source : Adaptation en français tirée de Ceder [2016]

et de ce fait plus rare. À l'inverse, les modifications concernant les deux dernières étapes sont plus fréquentes. Ces deux étapes sont largement automatisées grâce à des logiciels commerciaux de gestion et d'optimisation (par exemple, hastus qui est utilisé à Lyon, https ://www.giro.ca). Ces deux étapes se concentrent principalement sur l'amélioration de la performance économique du système (cf. triangle, figure 1.1) et visent donc à minimiser le nombre de véhicules en ligne, les kilomètres non commerciaux (haut le pied) et les temps de battement dans le but d'optimiser les coûts de main-d'œuvre (c'est-à-dire le principal poste de dépense). Alors que les activités de design du réseau et de dimensionnement de l'offre peuvent faire intervenir plusieurs acteurs de la structure organisationnelle et décisionnelle, les activités de graphicage et d'habillage restent avant tout l'apanage de l'exploitant du réseau (ou opérateurs ce qui rappelle bien le caractère opérationnel de ces activités).

1.1.3 La collecte et l'analyse des données un maillon indispensable

Le processus de planification que nous venons de définir et décrire a pour but la mise en place d'un système de transport en commun optimal au sens où il répond aux objectifs, et il est efficient du point de vue de la performance économique, de la rentabilité et de l'adéquation du service (les trois côtés du triangle de la figure

1.1). Ces mécanismes qu'ils soient stratégiques ou opérationnels ont besoin d'être alimentés par des données. C'est ce que traduit la figure 1.3 qui vient compléter la figure 1.1 en ajoutant les deux éléments manquants de notre puzzle : le processus de planification des transports en commun et le dispositif de collecte et d'analyse des données.

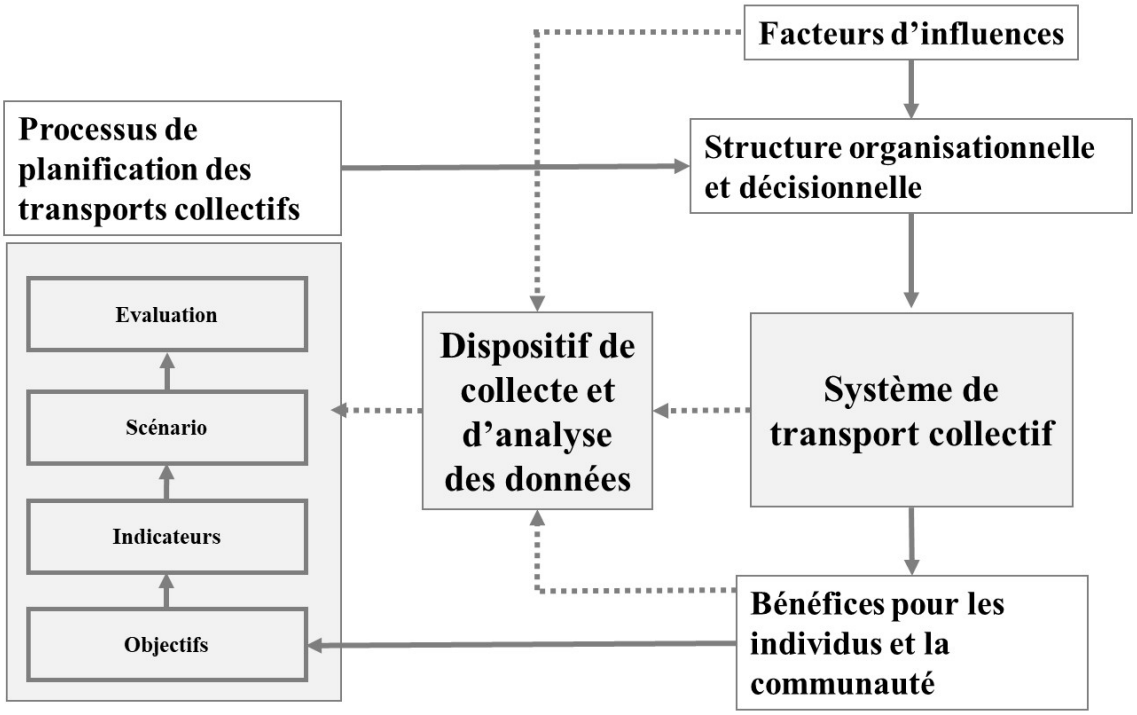


FIGURE 1.3 – Mise en relation des trois piliers de la thèse, Source : Auteur sur base de figure 1.1

Sur la figure 1.3, le processus de planification est schématisé à travers quatre étapes qui peuvent s'appliquer à toutes les échelles et tous les types de planification (stratégique, tactique ou opérationnelle). Premièrement, des objectifs sont fixés en relation avec les bénéfices escomptés. Deuxièmement, des indicateurs aussi variés soient-ils, permettent de matérialiser et d'articuler ces objectifs. Enfin, différents scénarios sont testés, puis évalués, ce qui permet ensuite d'alimenter les réflexions de la structure organisationnelle et décisionnelle qui peut ensuite sélectionner ou bien prioriser les évolutions à apporter au système de transport en commun. Même si la figure 1.3 laisse penser que les interactions entre le processus de planification des transports, le dispositif de collecte des données et le système de transport en commun sont unidirectionnelles, en pratique l'ensemble de ces briques sont en interaction bidirectionnelle et continue. Comme l'indique la figure 1.3, les données sont au cœur de la mécanique permettant de planifier un système de transport en commun plus efficient. Elles permettent de faire le lien entre la production des transports en commun et le processus de planification, et s'avèrent donc indispensables pour plusieurs raisons. Tout d'abord, elles servent à identifier et à quantifier les facteurs d'influences et plus particulièrement à améliorer la compréhension des comportements de déplacements. C'est le sens de la flèche reliant les facteurs d'influences avec le dispositif de collecte de données. Ensuite, elles permettent de formuler et quantifier des objectifs à atteindre par exemple

sous forme d'indicateurs. Ces indicateurs sont bien évidemment construits avec des données et doivent refléter l'état du système de transport en commun et les bénéfices qu'il produit. Par ailleurs, les données servent à évaluer différents scénarios de planification que ce soit en alimentant des modèles de simulation et de prévision qu'à travers des calculs socio-économiques (rentabilité, coût-bénéfice, etc..). Enfin, les données doivent être mises à profit pour mesurer les évolutions dans le temps du système, pour surveiller la performance du réseau et pour l'améliorer en continu.

La figure 1.3 permet donc de clore cette mise en contexte de la thèse en clarifiant les trois piliers de la thèse : le système de transport en commun, le processus de planification et le dispositif de collecte et d'analyse des données. Cette section volontairement conceptuelle et générale montre que le système de transport en commun est complexe et que la planification guidée par la collecte des données est indispensable pour faire évoluer de manière positive le système.

1.2 Motivations

Le contexte étant posé, nous introduisons maintenant les motivations de cette thèse. Pour ce faire, nous partons du constat que nos sociétés collectent de plus en plus de données. Ce constat est particulièrement vrai pour les systèmes de transports en commun qui deviennent de plus en plus connectés et capables de mesurer de manière fine et continue différents aspects de l'activité des réseaux de transport en commun. Ces nouvelles données viennent ainsi compléter des dispositifs de collecte anciens s'appuyant principalement sur des enquêtes. Il convient donc d'interroger l'apport de ces nouvelles sources de données. C'est la question centrale de cette thèse autour de laquelle gravitent les quatre articles qui la composent.

1.2.1 L'émergence des big data

Les quantités de données récoltées par l'humanité explosent de jour en jour à tel point qu'il est dorénavant nécessaire d'utiliser des échelles qui semblent bien éloignées de notre quotidien. En 2025, il est estimé qu'une quantité de 463 exabytes de données seront générées chaque jour [1]. Un exhabyte correspond à 10^{18} bytes, soit 1 milliard de gygabytes. À titre de référence, l'ensemble des mots prononcés par l'humanité jusqu'à ce jour équivaudrait à environ 5 exhabytes [2]. L'ensemble des validations enregistrées en 2017 par le système billettique du réseau de transport en commun de Lyon (TCL) pèse environ 100 gygabytes ($\sim$ 330 millions de transactions).

Cette explosion de données est intimement liée avec la numérisation progressive des processus industriels (industrie 3.0, usine 4.0), la multiplication des services et

1. https ://www.visualcapitalist.com/how-much-data-is-generated-each-day/, consulté le 16 mars 2020
2. https ://www.nytimes.com/2003/11/12/opinion/editorial-observer-trying-measure-amount-information-that-humans-create.html, consulté le 16 mars 2020

plateformes numériques (web social : Twitter, Facebook, WhatsApp ; fournisseurs de contenu : Netflix, YouTube ; plateforme communautaire : Airbnb, blablacar, etc.), la généralisation des objets connectés (web des objets), l'avènement du commerce électronique (e-commerce), le développement des villes et des systèmes de transports dits intelligents... Ces mutations graduelles et récentes des manières de produire, de consommer et d'interagir font que les objets et les individus sont de plus en plus connectés, à tel point que nous générons en permanence des traces numériques. Cet afflux récent de données dites massives et plus souvent dénommées via l'anglicisme big data questionne et interroge. Dans une note d'analyse consacrée au big data, le Commissariat général à la stratégie et à la prospective affirme que « *le traitement de ces masses de données, ou big data, jouera un rôle primordial dans la société de demain, car il trouve des applications dans des domaines aussi variés que les sciences, le marketing, les services clients, le développement durable, les transports, la santé ou encore l'éducation* » (Hamel and Marguerit [2013], page 1). De même, la Harvard Buisness Review dans un article désormais célèbre titrait que le big data engendrerait une révolution managériale [McAfee et al., 2012]. Cette révolution qui doit s'appuyer sur l'exploitation des données massives devrait permettre d'améliorer les prises de décisions et donc la performance des organisations. Elle ne concerne pas seulement les entreprises dont le cœur de l'activité réside dans la manipulation des données (par exemple Google ou Amazon), mais bien l'ensemble des secteurs d'activité et des chaînes de valeurs [McAfee et al., 2012]. Cette révolution est rendue possible par le fait qu'il est dorénavant facile de mesurer plus finement les activités. Comme le dit si bien Dominique Cardon, « *les capteurs numériques sont en train de jeter leur filet sur le monde pour le rendre mesurable en tout* » (Cardon [2015], page 8). Ce récent phénomène serait donc à même de bouleverser nos organisations qu'elles soient industrielles, administratives ou scientifiques. Mais de quoi parlons-nous exactement, comment définir les big data ?

Il est d'usage de caractériser les big data avec des adjectifs commençant par un V [McAfee et al., 2012, Kitchin, 2014b, Hamel and Marguerit, 2013]. Premièrement, l'adjectif **volume** qui fait référence à la quantité de données, et ce même si l'appréciation du volume de données est subjective et fonction du contexte. Néanmoins, il est courant que les volumes soient si importants qu'il devient nécessaire d'utiliser des technologies spécifiques comme le « *cloud-computing* » ou bien des outils tels que les architectures distribuées de calcul. À minima, les infrastructures matérielles et les logiciels nécessaires pour traiter des big data sont plus complexes qu'un ordinateur de bureautique grand public et un tableur. Deuxièmement, l'adjectif **vélocité** qui fait référence à la vitesse de création et d'actualisation de ces données. Les données sont collectées en temps réel ou quasi réel et les flux parfois traités en temps réel (« *streaming* ») par exemple dans les véhicules autonomes. Troisièmement, l'adjectif **variété** qui fait référence à la diversité de la nature des données. Certaines données sont souvent structurées à travers des modèles conceptuels relationnels de type base de données. D'autres sont non structurées comme les enregistrements audio, des images, des vidéos ou bien des textes. Nous avons donc une diversité sémantique importante et des formes variées de données. Quatrièmement, l'adjectif **véracité** qui fait référence à la qualité des données. C'est un aspect important, parfois négligé et qui pose des nouveaux défis [Cai and Zhu, 2015]. En effet, de par le volume, la variété et la

vélocité des données, il est souvent difficile de juger de la qualité des données. Cela est d'autant plus vrai quand elles ne suivent aucun standard et quand le processus de collecte est mal contrôlé et administré (champs manquants, capteurs défectueux, imprécision, données erronées, définition floue des populations cibles...). La rapidité d'évolution des dispositifs de collecte ne permet pas non plus de procéder à des travaux de validation définitifs. De ce fait, l'authenticité, la précision et l'exactitude avec laquelle les big data représentent ce que nous souhaitons analyser sont des aspects à ne pas négliger. À ces 4 précédents V, nous trouvons aussi dans la littérature d'autres caractéristiques associées au big data. Kitchin and McArdle [2016] s'intéressent par exemple à l'ontologie de 26 jeux de données et utilisent une grille d'analyse comprenant en plus des 4 précédents V les caractéristiques suivantes : l'exhaustivités c'est-à-dire le fait que des données sont capturées pour l'ensemble de la population, la finesse de la résolution, l'aspect relationnel des données qui permet des jointures et le caractère évolutif et extensible des données (« *scalability* »). Cet article montre que toutes les sources ne possèdent pas toutes les caractéristiques : l'identification et la définition *stricto sensu* des big data n'est donc pas possible. Dès lors, il convient plutôt de s'interroger sur ce que ces nouvelles sources de données changent pour la production de connaissances scientifiques ?

En 2008, Anderson écrivait que « *The data deluges makes the scientific method obsolete* » [Anderson, 2008]. En effet, d'après lui « *With enough data, the numbers speak for themselves* » [Anderson, 2008]. Cet article a évidemment fait polémique. Il a néanmoins le mérite de nous interroger sur la manière dont les données massives pourraient améliorer la production de connaissances. Nous pensons que les données ne parlent pas seules. Tout d'abord, les données doivent être collectées, identifiées, puis le chercheur doit se les approprier (ce qui est d'autant plus vrai quand il ne participe pas à la définition du processus de collecte). Ensuite, elles doivent être nettoyées et mises en forme. Enfin l'analyse des données à proprement dit peut commencer en s'appuyant sur des modèles, des théories ou bien des méthodes qui permettent d'obtenir des résultats et d'aboutir à des conclusions. Ce n'est donc pas de la « magie » et il y a beaucoup d'étapes à franchir pour valoriser les données massives [Ollion and Boelaert, 2015, Labrinidis and Jagadish, 2012, Kitchin, 2014b]. L'analyse des big data nécessite donc bien *ex ante* des questions à résoudre, des modèles, des hypothèses, des théories et l'usage des connaissances antérieures [Frické, 2015, Kitchin, 2014a]. Par ailleurs, ce n'est pas parce que le nombre d'observations est élevé que les problèmes de représentativité disparaissent automatiquement, car il n'y a le plus souvent pas formellement d'échantillonnage [Boyd and Crawford, 2012]. Même si ces données offrent la promesse d'une mesure objective, elles sont comme toutes les données dépendantes du contexte, des technologies et de l'environnement réglementaire [Kitchin, 2014b, Kitchin and Lauriault, 2015] et de ce fait elles sont aussi sujettes à interprétation. Bref, nous collectons des données massives, mais accumuler plus de données ne sert pas à grand-chose si nous ne savons pas comment les traiter, les interroger et si elles ne permettent pas de répondre à des questions pertinentes. Cela est d'autant plus vrai lorsque la collecte de données est une pratique antérieure à l'avènement du big data comme dans le cas du système de transport en commun.

1.2.2 Quelles nouvelles opportunités pour les systèmes de transport en commun ?

La collecte de données est une pratique ancienne dans le domaine de l'analyse de la mobilité urbaine. Traditionnellement, les données sont collectées grâce à des enquêtes réalisées sur un seul jour. C'est le cas des enquêtes ménages déplacements (EMD) réalisées à des intervalles plus ou moins réguliers (de l'ordre de 10 ans) dans la plupart des agglomérations françaises selon le standard CERTU [CERTU, 2008]. Ces enquêtes constituent souvent le principal matériel d'investigation de nombreuses recherches sur les comportements et les flux de déplacements [Bonnel, 2002, Commenges, 2013, Stopher and Greaves, 2007]. Elles servent à caler la plupart des modèles de demande développés en France et à ce titre sont une source clef du processus de planification des transports. C'est le cas aussi des enquêtes Origine-Destination (OD) qui ont pour but de connaître les arrêts de montée, de descente, les correspondances, les lignes empruntées ainsi que les modes de rabattement à l'origine et à la destination. Les enquêtes OD constituent souvent la principale donnée utilisée lors de restructuration de réseau ou de lignes de transport en commun. Elles fournissent des éléments clefs pour la connaissance de la demande de déplacement en transport en commun et pour adapter l'offre en particulier les fréquences. Les enquêtes OD portent également sur un seul jour de semaine souvent le mardi ou le jeudi et sont elles aussi réalisées de manière épisodique (tous les 4/5 ans à Lyon). En plus de ces deux sources clefs d'analyse des systèmes de transport, d'autres données sont depuis longtemps collectées soit grâce à des enquêtes soit via des processus administratifs. Sans en faire une liste exhaustive citons néanmoins, les enquêtes de satisfaction, les enquêtes fraudes, les enquêtes de préférences déclarées, les comptages manuels et autres mesure de charge (à un arrêt, pour une course), le recensement de la population et des emplois, etc. Pour une vision plus détaillée des sources de données classiques utiles à la planification des transports, le lecteur est invité à se référer à Bonnel [2002] (chapitre 4) ou bien à Ceder [2016] (chapitre 2) pour une vision centrée sur le système de transport en commun et complémentaire à la figure 1.2. Une caractéristique importante de ces données traditionnelles est que l'objectif de la collecte de données est fixé en avance pour répondre à un besoin précis défini *ex ante*. Elles sont aussi obtenues grâce à des méthodes d'enquêtes éprouvées et maîtrisées. Les outils traditionnels de la statistique peuvent être utilisés pour contrôler et quantifier les niveaux d'erreurs par exemple à travers des méthodes d'échantillonnage et de calcul de marge d'erreur [Kitchin and Lauriault, 2015, Callegaro and Yang, 2018, Stopher and Greaves, 2007, Bonnel, 2002] rendant ainsi possible l'inférence statistique sur les populations cibles. Pour collecter ces données, il est souvent nécessaire d'établir une relation d'engagement entre un enquêteur et un enquêté soit en face à face, soit par téléphone. Dans ce cas, la collecte est qualifiée d'active et la précision des données dépend de l'exactitude avec laquelle les personnes reportent ou déclarent leurs comportements de déplacements [Stopher and Greaves, 2007]. Par contre, cette relation d'engagement permet en principe la collecte de données sociodémographiques détaillées, souvent bien utiles pour expliquer et interpréter les comportements de déplacements [Pas and Koppelman, 1987, Bonnel, 2002]. Les enquêtes permettent aussi de mesurer des opinions et sont donc souvent bien utiles pour répondre à la question

« pourquoi » [Callegaro and Yang, 2018]. Pour des raisons logiques de coût, ces processus traditionnels de collecte de données peuvent difficilement être exhaustifs et sont souvent limités dans le temps et dans l'espace [Chen et al., 2016, Gärling and Axhausen, 2003]. Ces données ne permettent donc pas de saisir de manière très détaillée les dynamiques temporelles et spatiales qui sont pourtant fondamentales pour planifier de manière efficiente un système de transport en commun [Chen et al., 2016, Pelletier et al., 2011, Bagchi and White, 2005]. Il convient donc de s'intéresser au potentiel d'autres sources de données et notamment les sources présentant des caractéristiques big data.

En effet, le développement récent des systèmes de transports intelligents, la démocratisation des technologies de l'information et de la communication font que les systèmes de transports en commun collectent eux aussi des big data [Koutsopoulos et al., 2019, Welch and Widita, 2019]. Ils produisent dorénavant de manière continue (vitesse) une grande quantité de données (volume), qui peuvent prendre des formes diverses (variétés) et méritent que nous nous interrogeons sur leurs qualités (véracité). C'est à ces nouvelles sources de données que nous nous intéressons maintenant.

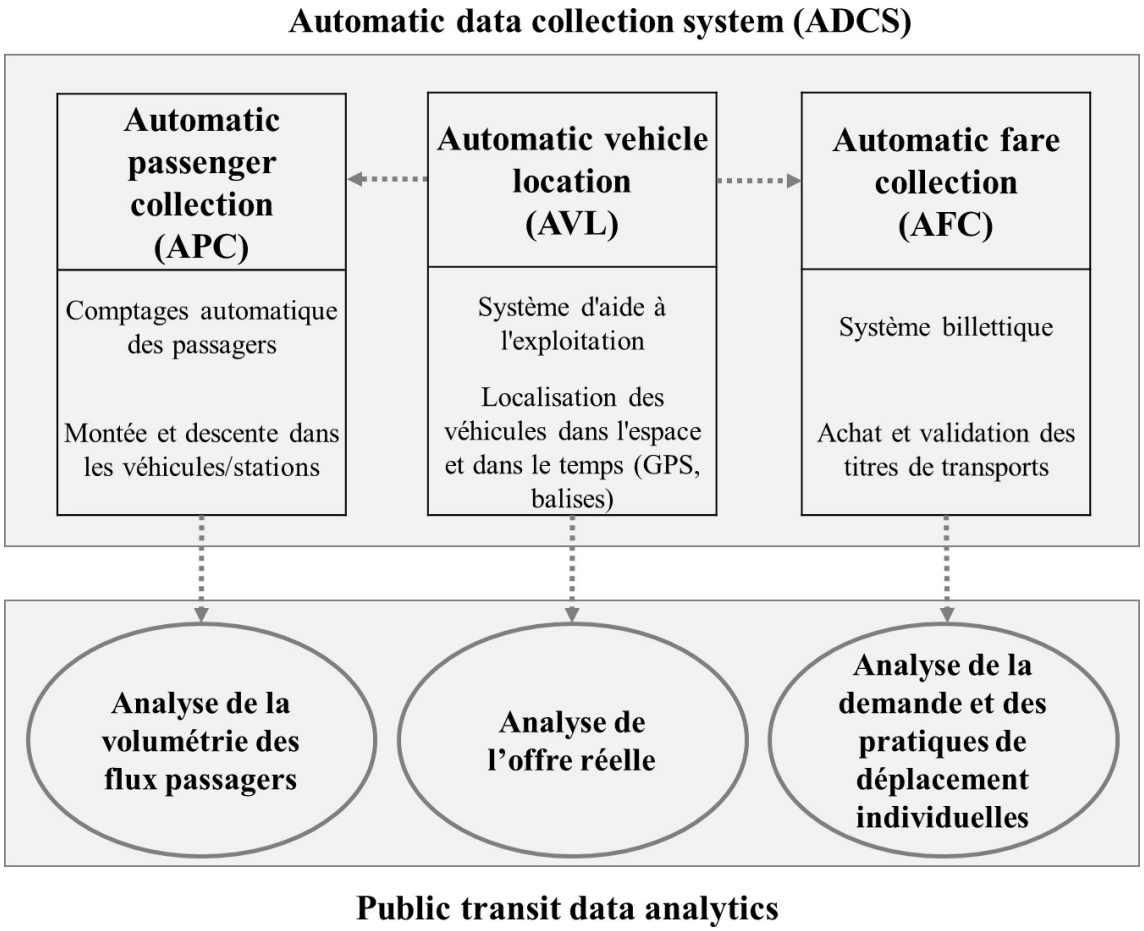


FIGURE 1.4 – Sources de données passives collectées automatiquement par les systèmes intelligents de transports en commun, Source : Auteur

Nous dénombrons trois sources principales (figure 1.4) qui forment conjointement

un dispositif de collecte automatique de données (ADCS). Le système de localisation des véhicules (AVL) fonctionne selon des technologies de type GPS ou bien pour les modes guidés grâce à l'échange d'informations avec la voie (balises). Ces données permettent une localisation des véhicules dans le temps et dans l'espace. Le système billettique permet la collecte automatique des titres de transport via des valideurs qui peuvent être embarqués dans les véhicules - dans ce cas, ils sont dits mobiles - ou bien fixes par exemple dans les stations et gares. Les systèmes billettiques modernes s'appuient principalement sur des supports de type cartes à puce sans contact avec identification unique (numéro de carte) [Pelletier et al., 2011]. Cette technologie fonctionne par excitation de signal radio courte distance (RFID : radio frequency identification). D'autres modes de perception des titres peuvent aussi coexister comme par exemple la technologie magnétique (ticket papier), la technologie des codes barres et la téléphonie mobile. Le système billettique est aussi en charge de la vente et du chargement des titres via différents canaux (vente en ligne, distributeurs, agences, paiements automatiques...). Les systèmes de comptage automatique des passagers se basent eux sur des capteurs avec des technologies plus ou moins modernes (marche compteuse, infrarouge, caméra stéréoscopique, pesée dynamique à l'essieu, etc.). Selon les installations, ces trois systèmes peuvent être connectés et échanger des données entre eux. Par exemple, dans un véhicule intelligent, le système de localisation est le maître de l'unité centrale et il vient alimenter la billettique, ou les instruments de comptage pour transmettre la localisation du véhicule. De même, le système de localisation alimente les systèmes d'informations voyageurs pour par exemple indiquer les heures de passage aux arrêts à l'aide de bornes dynamiques ou bien pour afficher sur des écrans embarqués les arrêts desservis par le véhicule. Dans la plupart des cas, seules les données de localisation sont disponibles en temps réel [Koutsopoulos et al., 2019]. Les données billettiques et de comptages sont encore, dans la majorité des cas, enregistrées localement puis stockées de manière centralisée à la fin du service ou selon des logiques d'échanges fractionnées par lot plus ou moins fréquentes (« *batch* ») qui peuvent prendre parfois jusqu'à plusieurs jours. Néanmoins, la tendance est à la généralisation du temps réel [Koutsopoulos et al., 2019] qui est facilitée par l'émergence de nouveaux protocoles de télécommunication (Lora, Sigfox, etc.) et l'accroissement des débits de la téléphonie mobile (4g et bientôt 5g). Parmi ces trois sources de données, la billettique s'avère pour le moment être la source la plus prometteuse pour la planification des transports en commun [Pelletier et al., 2011]. Lorsque le taux de pénétration des cartes est suffisamment élevé, ces systèmes permettent de collecter des données individuelles et désagrégées dans le temps et l'espace. Ce sont donc des données à caractère personnel et même si ce n'est pas l'objet de cette thèse cela pose évidemment des questions éthiques quant au respect de la vie privée et de la confidentialité [Pelletier et al., 2011, Sánchez-Martínez and Munizaga, 2016]. Les données collectées par le système billettique proviennent des interactions entre les usagers et les valideurs et se matérialisent sous la forme de transaction. Ces transactions sont directement liées aux comportements de déplacements des usagers et c'est pour cela que nous pouvons les qualifier de données intrinsèques à la mobilité [Zhao et al., 2018]. Elles peuvent donc permettre de mesurer la demande sur un réseau de transport en commun. Plus précisément, une fois enrichies à l'aide d'algorithmes, fusionnées avec les données de l'offre (AVL) et

redressées sur la base de comptages automatiques supposés exhaustifs (APC), il est théoriquement possible de mesurer de manière très précise le fonctionnement d'un système de transport en commun et de développer des approches analytiques devant guider les processus de décision et de planification [Koutsopoulos et al., 2019].

Mise à part le système de comptage automatique dont la finalité première est bien la collecte de données, les systèmes billettique et de localisation des véhicules ont été mis en place pour des besoins avant tout opérationnels. La billettique permet de faciliter la collecte des titres de transport (et donc des recettes), de fluidifier les échanges passagers et d'implémenter des structures tarifaires plus complexes et plus flexibles [Pelletier et al., 2011, Trépanier et al., 2009]. Les systèmes de localisation des véhicules ont été mis en place pour permettre d'améliorer la performance et la sécurité de l'exploitation, car ils rendent possible la régulation en temps réel des circulations par les postes de commande centralisés. Les données collectées par ces systèmes sont donc des « *by product* » d'autres activités comme c'est souvent le cas avec les sources de données émergentes [Kitchin, 2014a]. À l'inverse des données d'enquêtes, ces données émergentes ne sont donc pas systématiquement collectées à des fins *ex post* d'analyse de la mobilité et de planification des transports en commun. Ce sont généralement des données de capteurs et le dispositif de production et de collecte des données n'est pas conçu formellement. Ces données ne mesurent donc qu'une partie de ce qui intéresse le chercheur et peuvent s'avérer plus pauvres qu'espérer [Ollion and Boelaert, 2015, Callegaro and Yang, 2018]. Par exemple, ces données manquent quasi systématiquement d'informations contextuelles comme la socio-démographie ou bien le motif des déplacements. Plus généralement, elles manquent de sémantique propre à la socio-économie des transports [Chen et al., 2016]. De ce fait, il est souvent nécessaire de recourir à des méthodes d'inférences et d'enrichissements qui peuvent aussi introduire des biais [Chen et al., 2016, Pelletier et al., 2011]. De plus, il n'existe pour le moment pas de standard international pour ces trois sources de données [Sánchez-Martínez and Munizaga, 2016]. Ainsi, en pratique l'architecture et les données collectées par ces trois systèmes peuvent différer d'une ville à une autre. Par exemple, dans certaines villes, il est nécessaire de valider seulement en entrée (tap-in) alors que dans d'autres villes, il est nécessaire de valider en entrée et en sortie (tap-in tap-out). Cela complique donc la réutilisation des algorithmes développés dans d'autres réseaux de transport en commun [Sánchez-Martínez and Munizaga, 2016]. Qui plus est, il arrive fréquemment que les différents systèmes de récoltes automatiques aient été conçus par des vendeurs différents et que les structures des bases de données soient incompatibles [Zhao et al., 2007] ou comme c'est souvent le cas peu documentées [Sánchez-Martínez and Munizaga, 2016]. Par contre, à l'inverse des données traditionnelles dont la collecte est limitée par le coût de la main-d'oeuvre, ces systèmes automatiques une fois installés procurent des données en continu et à un coût très faible [Zhao et al., 2007] qui diminue de jour en jour [Frické, 2015]. Il est donc maintenant possible de concevoir des analyses plus détaillées dans le temps et dans l'espace. De plus, le caractère continu des données rend possible la détection des tendances et l'identification des « *patterns* » [Kitchin and Lauriault, 2015]. Il est ainsi possible d'observer des changements plus régulièrement et limiter le risque de péremption que connaissent

les enquêtes par questionnaire [Ollion and Boelaert, 2015]. La conservation d'historiques profonds de données permet d'alimenter des outils de simulation et d'entraîner des modèles prédictifs qui n'auraient jamais pu voir le jour avec des données traditionnelles [Einav and Levin, 2014, Kitchin and Lauriault, 2015]. La continuité de la mesure est donc un des principaux avantages de ces nouvelles sources. Dans le jargon des transports, nous disons aussi que ces données sont collectées de manière passive dans le sens où le rôle de l'usager et plus généralement de l'humain dans la collecte des données est restreint. Cela permet donc de réduire les sources humaines d'erreurs [Bagchi and White, 2005], ce qui ne signifie pas pour autant que ces données ne sont pas entachées d'erreurs. Elles sont aussi sujettes à des problèmes de qualité et de représentativité [Sánchez-Martínez and Munizaga, 2016, Trépanier et al., 2007, Welch and Widita, 2019]. Ces systèmes collectent une quantité élevée de données, mais le processus d'échantillonnage n'est pas contrôlé ce qui génère souvent des « *convenience samples* » pas toujours représentatifs [Einav and Levin, 2014]. Par exemple, il est possible que seule une partie de la flotte soit équipée du système de localisation (AVL) ou que seulement une partie des usagers valide leur titre de transport réduisant ainsi le potentiel de ces données.

1.3 Problématique, contributions et organisation du manuscrit

Cette section vient clôturer l'introduction en trois temps. Tout d'abord, nous rappelons la problématique centrale de la thèse que nous déclinons en plusieurs questions générales qui guident les travaux de cette thèse. Ensuite, nous détaillons comment les contributions de cette thèse permettent de répondre à cette problématique. Pour ce faire, pour chacun des articles constituant cette thèse, nous introduisons la question de recherche, nous justifions son intérêt et synthétisons les résultats. Enfin, nous précisons l'organisation du présent manuscrit.

1.3.1 Problématique

Nous avons fait le constat que les réseaux de transport en commun collectent automatiquement et passivement de plus en plus de données. Ces nouvelles sources de données viennent s'ajouter à des sources de données dites traditionnelles basées principalement sur des enquêtes. Pris dans leur ensemble, ces sources hétérogènes forment le dispositif de collecte des données. Ce dispositif est central, car l'analyse de ces données doit permettre de mieux comprendre les comportements de mobilité et de planifier en conséquence les réseaux de transports en commun. Il est donc indispensable de s'interroger sur la pertinence de ce dispositif et plus particulièrement sur les opportunités offertes par les nouvelles sources de données passives. C'est la problématique centrale de la thèse. Elle peut se décliner en plusieurs questions de recherche générale (GQ) qui guident et motivent les 4 articles constituant le présent manuscrit :

— **GQ1 :** Comment mieux exploiter ces nouvelles sources de données ?
— **GQ2 :** Les différentes sources de données qu'elles soient passives ou actives sont-elles commensurables ? Quels sont les avantages et les inconvénients de ces nouvelles sources de données ? Le dispositif de collecte est-il redondant ?
— **GQ3 :** Les nouvelles sources de données passives permettent-elles d'enrichir le dispositif traditionnel pour mesurer la demande de déplacements en transports en commun ?
— **GQ4 :** Ces nouvelles données permettent-elles de mieux comprendre les comportements de mobilité ? Permettent-elles d'observer des aspects du comportement humain jusqu'ici très difficiles à capturer ?
— **GQ5 :** Ces nouvelles données permettent-elles enrichir la connaissance sur la fréquentation d'un réseau ? Permettent-elles d'appuyer le processus de planification de l'offre ? De développer des modèles prédictifs ?

L'ambition de cette thèse n'est pas de répondre de manière exhaustive à ces questions qui sont nombreuses et vastes. Toutefois, ce sont là les questions de fond qui sous-tendent les travaux de cette thèse. C'est en partant de ces questions générales et de l'état de l'art, que nous avons pu définir des thématiques et des questions de recherche précises et propres à chaque article. Elles sont détaillées dans la section suivante (§ 1.3.2) et s'articulent avec les questions générales comme décrit dans la figure 1.5.

		GQ1	GQ2	GQ3	GQ4	GQ5
Contribution 1	**Mesure de la fraude**	✓	✓			
Contribution 2	**Production de matrice OD**	✓	✓	✓		
Contribution 3	**Variabilité des comportements**	✓			✓	
Contribution 4	**Prévision du trafic**	✓				✓

FIGURE 1.5 – Articulation entre les chapitres du manuscrit et les questions de recherche sous-jacentes qui guident et motivent les travaux de la thèse, Source : Auteur

1.3.2 Questions de recherche et contributions scientifiques

Les contributions scientifiques de cette thèse prennent la forme de quatre articles acceptés ou en cours de révision dans des revues internationales à comité de lecture. Selon le standard scientifique, dans chaque article les lacunes de la littérature existante, la question de recherche et les objectifs sont identifiés de manière explicite. Le but de cette section n'est donc pas de répéter ces éléments, mais plutôt d'en proposer une synthèse concise. Celle-ci doit permettre au lecteur de comprendre comment les contributions s'articulent avec la problématique que nous venons d'exprimer.

- **Contribution 1**

Chapitre 2 : Can we estimate accurately fare evasion without a survey ? Results from a data comparison approach in Lyon using fare collection data, fare inspection data and counting data.

Pouvoir mesurer la fraude de manière précise est une tâche importante pour optimiser et planifier les actions de lutte contre la fraude, mais aussi pour évaluer les impacts recettes. Ce phénomène est difficile à quantifier précisément [Reddy et al., 2011, Dauby and Kovacs, 2007, Lee, 2011, Larwin, 2012] et les enquêtes terrain restent souvent les seules sources de données disponibles. Ces enquêtes sont par nature limitées dans le temps et dans l'espace alors qu'une mesure plus désagrégée serait bénéfique pour les opérateurs de transports en commun [Lee, 2011, Reddy et al., 2011, Multisystems et al., 2002]. Dès lors, il convient de s'interroger sur l'apport des nouvelles sources de données collectées en continu pour quantifier la fraude. Malheureusement, la littérature existante ne permet pas de cerner clairement le potentiel et les limites de ces nouvelles sources de données ni d'ailleurs leurs validités. L'objectif de cet article est donc de répondre à la question de recherche suivante :

— **RQ1 :** Pouvons-nous estimer la fraude de manière précise sans enquêtes ? Quels sont les compromis entre les différentes sources de données et méthodes d'estimation ?

Pour ce faire, nous proposons de mettre en relation un ensemble hétérogène de données collectées sur le réseau de Lyon avec une typologie de la fraude orientée opérateur. Plusieurs indicateurs de fraude issus de la fusion des données sont construits et comparés sur différents aspects. Ce travail permet d'examiner les différentes pistes pour la mesure de la fraude. Les avantages et les inconvénients de chaque source de données et de chaque indicateur sont ensuite discutés. Les résultats empiriques suggèrent que les données récoltées par les contrôleurs présentent des limites importantes pour mesurer avec précision le niveau de fraude. Ils suggèrent également que la fusion des données billettiques et de comptage est une direction de recherche plus prometteuse. Néanmoins, cela ne suffira probablement pas à remplacer complètement les enquêtes de terrain. Les résultats de cette recherche sont utiles aux opérateurs de transport en commun qui cherchent de nouvelles façons de mesurer en continu la fraude. Ils permettent de mieux cerner le potentiel des données passives pour la mesure de la fraude.

- **Contribution 2**

Chapitre 3 : How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination survey and household travel survey ? An empirical investigation in Lyon.

Les matrices origine-destination (OD) permettent de quantifier la distribution spatiale de la demande et sont l'un des éléments clefs de la planification des transports. Elles servent d'intrant à la plupart des modèles de prévision de la demande et sont indispensables pour évaluer les politiques de transport et les investissements en infrastructures [Ortúzar and Willumsen, 2011, Bonnel, 2002]. Il est donc impératif de les estimer correctement. Traditionnellement, ces matrices

sont obtenues avec des enquêtes qui peuvent prendre la forme d'enquêtes ménages déplacements ou d'enquêtes origine-destination. Récemment, des auteurs ont montré qu'il était possible d'obtenir ces mêmes matrices en reconstituant les déplacements individuels sur un réseau de transport en commun à partir des données passives [Munizaga and Palma, 2012, Trépanier et al., 2007, Gordon et al., 2018]. Les planificateurs ont donc maintenant accès à un corpus hétérogène de données pour estimer la demande de déplacement [Bagchi and White, 2005, Chen et al., 2016]. Nous pensons qu'à l'avenir, la mise en commun de ces sources de données pourrait permettre d'estimer plus précisément la demande de déplacements. Cela nécessite au préalable une bonne compréhension des avantages et des inconvénients de chaque source qui passe forcément par la réalisation d'études empiriques comparatives. L'objectif de cette recherche est donc d'évaluer les différences entre trois sources de données indépendantes pour estimer des matrices de déplacements sur le réseau de transport en commun de Lyon. La question de recherche peut s'exprimer de la manière suivante :

— **RQ2 :** Les différentes sources de données permettant d'estimer la demande de déplacement en transport en commun sont-elles commensurables ?

Pour ce faire, nous mobilisons des données passives de comptage et de billettique, les résultats de 5 années d'enquêtes origine-destination et les résultats les plus récents de l'enquête ménages déplacements de l'aire métropolitaine lyonnaise. Chaque source de données est traitée indépendamment avec une méthodologie spécifique afin d'obtenir une matrice OD représentative d'un jour moyen. Différents éléments des matrices résultantes sont ensuite comparés. Bien que toutes les matrices partagent certaines caractéristiques, il existe également des différences substantielles qui doivent être prises en compte si ces données sont utilisées pour la planification des transports. Les données billettiques ne sont pas exemptes d'erreurs et doivent être complétées par des données provenant d'autres sources afin de construire des matrices OD représentatives. En effet, toutes les destinations ne peuvent pas être inférées, le taux de pénétration des cartes à puce est inférieur à 100% et la fraude ne peut être ignorée. Nos résultats empiriques suggèrent que la mise à l'échelle de ces matrices à l'aide des données de comptage est une solution viable. Les résultats indiquent également que l'enquête ménages déplacements sous-estime considérablement le volume de déplacements en transports en commun par rapport aux autres sources. Cette étude valide de manière externe la pertinence des données passives pour l'estimation de matrice OD tout en questionnant la pertinence des méthodes traditionnelles. Les résultats de cette recherche contribuent ainsi à une meilleure compréhension des sources de données disponibles pour estimer la demande en transport en commun. Ils peuvent aider les praticiens à améliorer la qualité et la précision des matrices de déplacements.

- **Contribution 3**

Chapitre 4 : Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon.

L'analyse de la variabilité des comportements de déplacement est un domaine de recherche important, car il a des applications pratiques telles que l'évaluation de

l'impact des politiques de transport [Jones and Clarke, 1988], la modélisation du comportement de déplacements [Pas, 1986, 1987], la mise en oeuvre de stratégies de marketing individualisées [Gärling and Axhausen, 2003] ou encore la segmentation clients [Hanson and Huff, 1986]. Traditionnellement, les données d'enquêtes utilisées pour l'analyse des comportements de mobilité portent sur un seul jour qualifié de « jour moyen ». Elles ne permettent donc pas de mesurer la variabilité individuelle et font implicitement l'hypothèse forte que les individus ont des comportements stables au fil des jours. Cependant, les activités et les désirs qui génèrent les déplacements varient à la fois selon les individus, mais aussi selon les jours [Pas, 1987]. Les classifications basées sur le comportement d'une seule journée sont donc susceptibles d'être instables [Hanson and Huff, 1986]. Il est alors nécessaire de s'appuyer sur des données pluri journalières pour affiner la compréhension du comportement de déplacement. Grâce aux systèmes billettiques, nous collectons dorénavant en continu des traces de mobilité qui peuvent permettre des études longitudinales désagrégées. Ces sources de données peuvent être utilisées pour mesurer la variabilité [Morency et al., 2007] et ont entraîné une multiplication des recherches sur les habitudes d'utilisation des transports en commun. Toutefois, comme le soulignent Chen et al. [2016] les recherches basées sur les sources de données passives manquent souvent du cadre conceptuel classique de l'analyse des comportements de déplacements. Plus particulièrement, la revue de la littérature montre que la variabilité n'est jamais mesurée à un niveau journalier alors que les auteurs fondateurs de la discipline considèrent qu'elle devrait être mesurée à ce niveau [Hanson and Huff, 1988, Pas and Koppelman, 1987, Schlich and Axhausen, 2003]. L'objectif de cette recherche est donc de proposer et d'appliquer des méthodes permettant de mesurer la variabilité intra et inter-individuelle du comportement d'usage journalier d'un réseau de transport en commun. Notre question de recherche est la suivante :

— **RQ3 :** Comment les données billettiques permettent-elles de mieux comprendre la variabilité d'usage journalière d'un réseau de transport en commun sur des échelles de plusieurs mois ?

Pour répondre à cette question, nous proposons de combiner deux méthodes flexibles permettant de mesurer la variabilité d'usage. La première méthode s'appuie sur une technique de clustering que nous avons spécialement adaptée à notre problème. Elle permet de visualiser et d'identifier les patterns d'utilisation quotidiens les plus courants et ainsi d'explorer la variabilité inter-individuelle de manière simple. La seconde méthode est un indice de similarité conçu pour mesurer la variabilité intra-individuelle d'un jour à un autre en tenant compte de trois caractéristiques fondamentales du patron de déplacement quotidien : l'espace, le temps et le nombre de déplacements. Ces deux méthodes sont appliquées aux transactions billettiques de 40 000 cartes sur une période de 6 mois. Les résultats sont ensuite croisés avec le profil tarifaire de chaque carte pour comprendre les déterminants potentiels de la variabilité du comportement de déplacement. Les résultats peuvent aider à mieux comprendre la dynamique individuelle de l'utilisation des transports en commun. Ils peuvent également aider les opérateurs de transport en commun et les spécialistes du marketing à définir des segmentations clients basées sur le comportement d'usages.

- **Contribution 4**

Chapitre 5 : Medium-term public transit route ridership forecasting : what, how and why ? A case study in Lyon

La prévision de la demande est une tâche essentielle dans de nombreux secteurs économiques et le secteur des transports en commun ne fait pas exception. Prévoir de manière précise la demande future est une composante essentielle des systèmes de transport dits intelligents [Vlahogianni et al., 2014, Koutsopoulos et al., 2019] et un aspect fondamental de tout processus de planification rationnelle [Bonnel, 2002, Ortúzar and Willumsen, 2011]. Comme décrit en section 1.1.2, il est d'usage de distinguer trois grands horizons de planification qui ont chacun des objectifs et des méthodes de prévisions spécifiques. Le niveau stratégique traite des décisions et des objectifs à long terme. Dans ce cadre, les méthodes les plus courantes de prévision reposent sur les modèles à quatre étapes ou l'usage de coefficient d'élasticité [Boyle, 2006, Ortúzar and Willumsen, 2011, Bonnel, 2002]. Ces approches sont souvent calibrées pour un ensemble limité de situations, comme un jour de semaine moyen, une heure de pointe typique, etc. Le niveau opérationnel requiert lui des prévisions à court terme c'est-à-dire de quelques minutes à quelques heures dans le futur [Vlahogianni et al., 2014]. Ce type de prévision peut être utilisé pour aider la prise de décision opérationnelle et l'information clients en temps quasi réel et continu [Vlahogianni et al., 2014, Noursalehi et al., 2018]. Malheureusement, les réseaux de transport en commun ne peuvent pas toujours réagir efficacement sur des temps courts pour des raisons de flexibilité limitée des ressources (par exemple, véhicules, personnels) et de contrainte d'infrastructures. Étonnamment, peu de recherches portent sur le développement de prévisions tactiques à moyen terme, c'est-à-dire sur les 12 prochains mois. Nous pensons cependant que ce type de prévisions est nécessaire pour une planification efficace des systèmes de transport public. En effet, la génération de prévisions à des horizons de 1 an peut être d'une grande utilité pour les opérateurs qui veulent non seulement pouvoir adapter le fonctionnement en temps réel (prévisions à court terme) et évaluer des plans stratégiques (prévisions à long terme), mais aussi surveiller tactiquement et en continu l'évolution de la fréquentation. Les systèmes de comptage et de billettique permettent aux opérateurs de transport en commun d'accumuler en continu des historiques détaillés de fréquentation. Ces données sont essentielles pour suivre l'activité et aider à la prise de décision. Ces historiques peuvent aussi venir alimenter des algorithmes de prévision moyen terme. L'objectif de cet article est donc de proposer une méthode qui permet, à partir d'un historique de fréquentation détaillé, d'apprendre le comportement historique et de prévoir la fréquentation sur les prochains mois. La question de recherche sous-jacente peut se résumer de la manière suivante :

— **RQ4 :** Pouvons-nous développer une approche générique et multi niveau de modélisation permettant de prévoir la fréquentation à 12 mois d'un réseau de transport en commun ? Comment les prévisions ainsi générées peuvent-elles permettre d'améliorer la planification tactique de l'offre et le monitoring de la fréquentation ?

Pour répondre à cette question, les données pertinentes sont identifiées, préparées

et décrites. Puis, nous formulons une approche de modélisation qui combine un ensemble d'arbres de régression avec une projection de tendance. Nous effectuons ensuite une prévision à 12 mois de la fréquentation par jour et par heure pour les 36 lignes les plus importantes du réseau de Lyon. La qualité des prévisions est évaluée en comparant les résultats de différents modèles sur des données de test. Enfin, nous explorons à travers quelques cas d'usages l'intérêt de ces prévisions dans une perspective de planification tactique. Les résultats montrent que l'approche de modélisation proposée est valable et permet de réduire les erreurs. Les opérateurs de transport en commun peuvent exploiter ces prévisions pour mieux surveiller et anticiper la fréquentation de leur réseau. Les résultats confirment que des méthodes statistiques plus sophistiquées peuvent faciliter un certain nombre d'analyses qui reposaient auparavant principalement sur le jugement d'expert.

1.3.3 Organisation du manuscrit

Les contributions énoncées ci-dessus forment les chapitres de 2 à 5 du présent manuscrit. Chaque chapitre prend la forme d'un article scientifique en Anglais d'une longueur allant de 7 000 à 11 000 mots. Le choix retenu est de présenter une version éditée par nos soins de chaque article soit sous sa forme finale lorsque celui-ci a été accepté pour publication soit comme il a été soumis. Comme nous l'avons vu, chaque article possède des objectifs qui lui sont propres et à ce titre peut être consulté indépendamment du reste du manuscrit. Comme l'ensemble des articles s'appuient en grande partie sur les données de l'opérateur du réseau de transport en commun de Lyon, il existe parfois des redondances dans la présentation des données ou du cas d'étude. Certaines références bibliographiques sont communes à plusieurs chapitres, le choix est donc d'avoir une seule bibliographie pour l'ensemble du manuscrit.

Chaque chapitre commence par deux pages introductives contenant la référence de l'article, les principales conclusions (highlights), les mots clefs (keywords) ainsi qu'un court résumé en Anglais (abstract). Le corps du texte proprement dit est ensuite inséré accompagné des tables et des figures dans un ordre permettant une consultation quasiment sans renvoi. Chaque article possède sa revue de la littérature et sa section méthodologique. Les symboles mathématiques et les abréviations sont propres à chaque article.

Le chapitre 6 vient clore ce manuscrit en rappelant les principaux résultats de la thèse, en décrivant quelques applications opérationnelles et en discutant les limites et perspectives de ces recherches. Une table des figures, une liste des tableaux et la bibliographie sont insérées après la conclusion. Les résumés en français et en anglais sont en quatrième de couverture.

2

Can we estimate accurately fare evasion without a survey ? Results from a data comparison approach in Lyon using fare collection data, fare inspection data and counting data.

This chapter is an edited version of the following article :

O.Egu and P.Bonnel (2020). Can we estimate accurately fare evasion without a survey ? Results from a data comparison approach in Lyon using fare collection data, fare inspection data and counting data. *Public Transport*, 12(1), 1-26.

DOI : https ://doi.org/10.1007/s12469-019-00224-x

Highlights

- Fare evasion could be considered as a threat that needs to be quantified accurately
- Little research examines how new data sources could help in estimating the fare irregularity rate
- This research followed the operator's viewpoint and used a comparative approach to analyse the potential of those new data sources
- Results suggest that the fare inspection logs might have significant limitations and that the merging of automated count and farebox transactions is a more promising direction of research
- Findings can help operators in identifying the pros and cons of each data sources and implement new measurement methods

Chapter 2 : Can we estimate accurately fare evasion without a survey ? Results from a data comparison approach in Lyon using fare collection data, fare inspection data and counting data.

Abstract

In a context of worldwide urbanization and increasing awareness for environmental issues, it is undeniable that public transport will play an important role in the cities of the future. This will require increased attractiveness of public transit and adequate funding. In this regard, fare evasion could be considered as a threat that needs to be quantified accurately. To do this, transit operators often rely on on-site surveys that are limited in terms of spatiotemporal coverage. Yet, new data sources such as farebox transactions, fare inspection logs and automated passenger counter are now available and little research examines how they could help in estimating the fare irregularity rate. In this paper, we initiate research in this direction. To do this, we followed the operator's viewpoint and used a comparative approach to analyse the potential of those new data sources. We introduced a classification of fare irregularities and then applied data fusion methods to derive two fare irregularity rates. Results are then compared to a survey and the area of relevance of each data source is discussed. The research is done with data from the public transport network of Lyon which is an interesting case study because different access control types coexist (open and closed environment). The research results suggest that the fare inspection logs might have significant limitations to measure accurately the level of fare evasion. They also suggest that the merging of automated count and farebox transactions is a more promising direction of research. Still, it will probably not be enough to completely replace on-site manual survey. These findings can help operators in identifying the pros and cons of each data sources and implement new measurement methods.

Keywords— Public transport, Smart card, Fare evasion, Survey, Data collection, Lyon

2.1 Introduction

The growing rise in the amount of automatically collected data makes public transit operators (PTO) eager to leverage data to increase operational efficiency. It has been proved that passive data such as automatic fare collection data could help operators improve the public transportation system [Pelletier et al., 2011]. In other areas such as law enforcement data have also been proved to be very useful. For instance, it is used in various cities to help identify and predict the potential location of crime activity [Pearsall, 2010].

Fare evasion is one area of serious concern for PTO since it could reduce the financial viability of the system [Barabino et al., 2015, Delbosc and Currie, 2018, Reddy et al., 2011, Troncoso and de Grange, 2017], have a negative impact on attractiveness and security [Barabino et al., 2015, Reddy et al., 2011] and greatly affect the statistical representativeness of smart card data [Munizaga and Palma, 2012, Sánchez-Martínez, 2017]. Unfortunately, this phenomenon is difficult to quantify accurately [Reddy et al., 2011, Dauby and Kovacs, 2007, Lee, 2011, Larwin, 2012] and transit operators often rely on surveys to measure fare evasion. With an appropriate methodology, they should provide a global and relatively accurate measurement of the phenomenon but they remain costly and often have limited spatiotemporal coverage. However, PTO would benefit from continuous and disaggregated measurement of fare evasion [Lee, 2011, Reddy et al., 2011, Multisystems et al., 2002]. This is needed to ensure more precise monitoring of this key indicator, to define targeted fare inspection operations, to quantify the effectiveness of countermeasure but also to improve the understanding of the nature of fare evasion and its evolution. To reach this goal, big data sources and data fusion techniques may be an appropriate solution [Delbosc and Currie, 2018, Pourmonet et al., 2015]. Yet, their potential and limits remain unclear and their validity must be examined through comparison with other sources.

The objective of this paper is to develop and test a methodology based on automatically collected data sources that enable the continuous measurement of the level of fare evasion and answer the following questions :

— Can we estimate accurately fare evasion without a survey ? What are the trade-offs between different data sources and estimation methods ?

More precisely in this study, automated fare collection data, automated passenger counting data and fare inspectors' handheld unit logs are processed and merged to estimate different levels of fare evasion. The methodology is applied to data from Lyon and results are then compared to an on-site survey specifically designed to measure fare evasion. Empirical results suggest that data sources will not produce similar measures of fare evasion and that the physical infrastructure has a high impact on the level of fare evasion. Indicators based on fare inspection logs tend to produce lower irregularity rate. On the contrary, the merging of automated count and farebox transactions may indicate a higher level of irregularity. Those differences are carefully analysed and could be explained by potential errors in data collection but more importantly by the fact that fare evasion phenomenon can not be measured with the same definition when using different data sources. The potential and limits of each data sources are then discussed. To the best of our knowledge, this is the first

attempt in providing such a detailed comparative study. The outcome of this research may be useful for public transport managers looking for new ways to continuously measure fare evasion. It will also help them to identify the pros and cons of each data sources.

The paper is organized as follows. Section 2.2 provides a literature review of related work. Section 2.3 details the case study, the data sources and the methodology. Section 2.4 examines the results by comparing different data sources. Section 2.5 discusses and synthesizes the findings of this study. Finally, section 2.6 draws the conclusions and perspectives of this work.

2.2 Literature review

Fare evasion studies are not widely known among transit planners [Reddy et al., 2011]. A recent paper by Delbosc and Currie [2018] presents a literature review on fare evasion research focusing on three perspectives : the conventional perspective, the customer profile perspective and the customer motivation perspective. This review shows that fare evasion is a complex phenomenon. From the customer's motivation perspective, the definition of what is considered as fare evasion may vary between passengers and the spectrum of fare evasion is largely determined by the degree of intent [Delbosc and Currie, 2016]. Suquet [2010] in an ethnographic study of inspection workers also found that the distinction between deviant and conform behaviour is not always clear even for transport operator workers. From the criminological perspectives fare evasion is seen as a type of crime that may be facilitated by lack of supervision [Smith and Clarke, 2000] and influenced by situational factors [Hauber, 1993].

To measure fare evasion different methodologies were used by researchers. Lee [2011] have used field survey both on board and at stop requiring customers to show proof-of-payment. They found that a minimum of 9.5% of surveyed riders lacked valid proof-of-payment but this percentage varied greatly by time and location. In New York subway, fare evasion is measured using field count [Reddy et al., 2011]. Surveyors are asked to observe discretely subway entry gates and to report the count of most common methods of illegal entry. Counts are done for half-hour periods and for a random sample of 300 location-time pairs. As noted by the authors this observation methodology cannot monitor all revenue losses as surveyors do not have authority to check passengers proof-of-payment. With this method, a rate of 1.9% of illegal and questionable entries are observed [Reddy et al., 2011]. In Victoria (Australia), the fare evasion is measured with an onboard stratified survey that covers tramways, trains and buses [Public Transport Victoria, 2018]. Survey team consist of a mixed of authorised officers and survey staff who move through selected vehicles. Officers are in charge of checking proof-of-payment while surveyors record passengers counts and type of fare evasion encountered. In October 2018, the rate of fare evasion was found to be higher in the metropolitan buses (8%) than in tramways (3.2%), metropolitan trains (2.5%) or regional trains (4.9 %) [Public Transport Victoria, 2018]. A report on fare

evasion on King County Metro in Seattle notes that measuring fare evasion is challenging [King County Department of Transportation, 2010]. They decided that the best option to measure fare evasion would be to rely on operators (drivers) that were instructed to use keys on the farebox to count three types of fare evasion : adults who paid no fare, youth, seniors and disabled who paid no fare and partial payment. They found a total of 4.8% of boarding without valid payment which represents a 2.5% loss of total fare revenue. In Santiago buses, fare evasion measurements are carried out using field count done by plain clothes surveyors that sample regular daytime services [Troncoso and de Grange, 2017, Guarda et al., 2016b,a]. Their mission is to report the count of people boarding and alighting, the count of fare evasion and other variables such as bus occupancy. Those surveyors are stationed at each door of the bus and do not interact with passengers making it impossible to know exactly the type of fare evasion. Using aggregated results from those surveys Troncoso and de Grange [2017] have explored the long-term economic relationships in Santiago between fare evasion level, fare price, fare inspection level, and unemployment rate. They conclude that the relation between evasion and fare price is positive, that fare inspection is insufficient for reducing fare evasion and that there is a negative correlation between unemployment and evasion. With disaggregate data from this same survey enriched with variables related to the level of income, Guarda et al. [2016b] proposed an econometric approach to study the level of fare evasion. They used a negative binomial regression to quantify the joint influence of various variables such as headway regularity, crowding level or boarding by rear doors on the level of fare evasion.

Researchers have also used fare inspection data to measure fare evasion. To study the relationship between fare prices, penalty fees, frequency of inspection and fare evasion level in different European cities Hauber [1993] used two collection methodologies. The first method was to send surveyors that observed inspectors checking tickets and estimate the number of escaping passengers and the number of detected evaders. The second method was to rely on the declared transport company information. They state that the first method was more realistic because escape mechanisms often take place during inspections. Their results suggest that to reduce fare evasion, it is more effective to increase the frequency of ticket checks than to increase penalty fees or decrease fare prices. In Switzerland, Killias et al. [2009] in a natural experiment study how a change in the number of working hours devoted to ticket checks affects the percentage of passengers on trains found without a valid ticket. They question the fact that using this ratio as a dependent variable is valid as an indicator of fare dodging just like "offences known to the police never perfectly measure the actual numbers of crimes committed" [Killias et al., 2009]. Nevertheless, their data indicate that the concentration of checks on certain critical hours might be an efficient option to tackle fare evasion. Clarke et al. [2010] have also investigated the change of fare evasion rate in response to change in enforcement level. Their research examined how a reduction in the number of checks and an increased in the risk of being fined affect the evasion rate of Edmonton Light Rail Transit. Their study used weekly counts of the number of passengers checked, the number of passengers checked without a valid ticket and the number of penalties issued. The evasion rate is defined as the percentage of passengers checked without a valid ticket. Their results suggest that the evasion

rate remains at the same level despite the change in enforcement but the authors acknowledge that more data may be needed to better understand the dynamic of fare evasion. In a European study of the relation between design solution and enforcement strategy in light rail system, Dauby and Kovacs [2007] report that there are often discrepancies between the measured and the estimated fare evasion level and call for an improvement of the measurement methodology. Some technical reports have also discussed the subject of fare evasion measurement in relation to inspection. Multisystems et al. [2002] consider three main approaches : use of the inspection results, use of the results of 100% inspection "sweeps" and/or conduct special field survey periodically. In terms of industry practice, they found that about 40% of transit agencies use special field survey and that those surveys generally measure higher fare evasion rate. They also state that those surveys are "presumably more accurate" [Multisystems et al., 2002]. Thus, they recommend the use of surveys to produce official evasion rate especially when there is significant variation in the operation procedure employed by inspectors. In another technical report, Larwin [2012] identify a number of issues related to measuring fare evasion such as definitional controversy (what is included as fare evasion) or issue with regard to sampling techniques. They survey 22 operators using proof-of-payment and found no direct correlation between evasion rate and inspection rates. Recently some researchers in the operational field have proposed to used Stackelberg game [Correa et al., 2017, Delle Fave et al., 2014] to optimize fare inspection strategies on transit network. Delle Fave et al. [2014] have used data collected by transit inspectors to compare different allocation models in terms of number of passenger checks but also number of captures (sum of the number of warnings, citations, and arrests). To assign job and duties of security guards operating on trains and stations, Snijders and Saldanha [2017] formulate an optimisation model that use as input the reports of aggression and penalties written by train inspectors. Other authors have tried to establish the optimum level of inspection that maximizes profit with theoretical modelling [Boyd et al., 1989] and various levels of refinement and empirical support [Kooreman, 1993, Barabino et al., 2013, 2014, Barabino and Salis, 2019]. For instance, Barabino and Salis [2019] have proposed a formal economic framework to estimate the optimum level of inspection. This framework includes a segmentation of the passenger demand, a probability of being caught and fined, an estimation of the percentage of passengers who decide to fare evade and constraints in the profit function. Their model is then applied using data gathered from an Italian public transport company. Their results indicate that to maximize profit the inspection rate should be in the range of 3.4%-4.0% [Barabino and Salis, 2019]. Pourmonet et al. [2015] are to the best of our knowledge the first authors to have used automatic passenger counting and automatic fare collection to characterize and estimate fare evasion in Montreal. They show that the ratio between the number of validations and the number of counted passengers may vary spatially and temporally but did not compare those results to survey or inspection data. Another quite new approach to estimate fare evasion is the one from Sánchez-Martínez [2017]. They proposed a stochastic model that quantifies for each card the willingness of a passenger to make a trip without tapping-in his or her card based on the card transaction history. This approach is promising but limited because it is inadequate to detect non validation of occasional users or paper ticket.

In view of this broad body of literature, it is clear that fare evasion should be addressed from different perspectives and also measured with various data sources and methods. In this regard, the emergence of automatically collected data offers new opportunities. However, in the context of fare evasion studies, those new data sources have scarcely been explored by researchers and to our knowledge have never been compared to more traditional data sources such as a survey. For this reason, the main contribution of this paper is to provide a comparative study between data sources to assess if fare evasion level could be measured with automatically collected data and to identify the potential trade-offs. To achieve this, we adopted the operator's viewpoint and proposed a typology of fare irregularity and some methods to merge the different data sources. The pros and cons of each data source are then discussed.

2.3 Materials and methods

2.3.1 Case study

TCL ("Transport en Commun Lyonnais") is the commercial name of the public transport network of Lyon. This network is currently run by a private operator under the supervision of the public transport authority of the Lyon metropolitan area (Sytral). The network consists of 4 lines of metro, 2 lines of funicular, 5 lines of tramway and more than 100 regular lines of bus.

The current fare transaction system of TCL was implemented in 2002. Smart card and magnetic paper ticket can be used. Cardholders can access to a broad range of fares available from annual pass to standard single tickets. Cardholders can also benefit from reduced fare pricing (student, children, low income etc.). On the system, all fareboxes acknowledge with a sound when a correct fare is tapped-in.

In the TCL network, different access control types coexist. On the one hand, the metro and the funicular can be viewed as a closed system. All entry and exit sections have a physical barrier formed by multiple gates. Each gate is equipped with an infra-red counting system technically able to detect intrusion. Entry gates can only be opened by the presentation of a valid transport title and are designed with evasion preventing doors. In this system, travellers are not asked to validate their fare when making connections between lines and when exiting the station (no tap-out). To illustrate the system a picture of a metro entry section is depicted in figure 2.1b. As highlighted in red in the top right corner, in this specific station, stickers are in place to stimulate the validation process. The slogan on the stickers "Je valide, je suis serein" could be translated into "I tap-in, I'm serene".

On the other hand, tramways and buses are operated as an open proof-of-payment system and thus fareboxes are onboard equipment. In tramways, fareboxes are located at each door all along the vehicle. On the bus, front gate boarding is the current policy and a minimum of two fareboxes are available at the front door. An example of such a farebox is given in figure 2.1a. In practice, front gate boarding is not fully respected and rear gate boarding is still tolerated for some lines with high

(a) A bus farebox

(b) A metro gate view from the entry side

FIGURE 2.1 – Illustrations from Lyon automated fare collection system, Source : Author's pictures

load factors. In this case, fareboxes are located in front of each bus door. Note that in the bus, passengers can purchase tickets directly from the drivers (at an increased price) which is not possible in tramways where the only onboard staff is the driver isolated in his or her cabin.

With the current fare regulation in the TCL network [1], whatever the fare support, passengers are required to validate every time they board a vehicle except in the metro and funicular network where the validation is only needed when entering the system (not tap-in is needed for connections between underground lines). Single tickets are valid up to one hour from the first stamping. The operator in charge in

1. Available on-line at http ://www.tcl.fr/Media/Librairie/Encarts-contextuels/Reglement-du-reseau

Lyon has classified fare irregularity behaviours in two groups (figure 2.2) :

— Fare evasion with loss of revenue. This includes not having any transport title, not having a validated ticket, travelling with an expired fare support and using a reduced fare without the required justification ;

— Irregularity without loss of revenue. This is when customers do not respect systematic validation rules but are in possession of a valid transport title.

	Fare evasion with loss of revenue	**Fare irregularity without loss of revenue**
	-Not having any fare support	
Pay-as-you-go (paper tickets or charge into smart card)	-First boarding without validation -Expired duration of tickets -Reduce fare without required justification	-No validation when making connection but support still valid
Pass (only smart card)	-Pass out of date -Using pass of others (pass sharing)	-No validation for first boarding or connection but contract is valid

FIGURE 2.2 – Classification of fare irregularity, Source : Authors

Some may argue that this typology is simple-minded, especially from the passenger's viewpoint. It is nonetheless necessary for the PTO to be able to measure distinctly the two types of irregularity. The first one is a threat to the financial viability of the public transit networks, the second one could greatly impact the representativeness of smart card data. This distinction is also important for fare control operations. Penalties prices are set to 60€ for fare evasion with loss of revenue and to 5€ for irregularity without loss of revenue.

In Lyon, during the last few years a lot of effort has been put into the fight against fare evasion. Fare inspection is considered as a major operational component for the operator in charge. In 2017, a total of approximatively 316,000 hours of repressive fare inspections by approximatively 200 full-time equivalent staff were carried out. Inspection rate is around 1.3% and 305,000 penalties were issued which also contribute to revenue if paid. Fare inspections are done by certified staff members, either in uniform or in plain clothes and once in a while, in joint operations with the police. Inspectors are assigned to patrol of four including the leader. For a given hour and a given type of day (weekday, Saturday, Sunday) the number of controlling patrol is defined according to a fixed schedule. To conduct fare inspections a variety of approaches and strategies are used. In the closed environment, inspectors can move on randomized paths and inspect passengers inside the vehicle. They can also inspect passengers directly at stations or platforms. When they inspect people entering the system, they may decide to remain visible behind the gates to make sure that all passengers validate their title. They can also try to be as discrete as possible and target passengers trying to sneak through without paying (especially when working in plain clothes). In the

open environment, each patrol normally received at the beginning of their shift a list of lines that must be inspected. Inspections are then performed on board or at bus stops. In buses and tramways, plain clothes patrol tend to work on board and can try to target their inspections to specific passengers while inspectors in uniform do not make distinctions between passengers. Multiple patrols may also have to work together for specific operations that require a higher volume of control. A typical example is when they want to control simultaneously all passenger inside a vehicle at a given stop (100% sweep) or control all the entry and exit gates of a given metro station. In both environments, inspection patrols keep a certain level of flexibility to select particular approaches or specific locations to control. They must also adapt dynamically to operating constraints without disrupting service. In general, inspectors try to work in a peaceful manner, to avoid aggressive situations and areas deemed to be risky. If a passenger in irregularity is found the choice to impose a penalty always rests with the inspector. When a penalty is written, passengers are offered to pay directly without having to show a proof of identity. Otherwise, the inspectors try to get the passenger's identity which cannot be done systematically as some passengers refuse or use a false name. Apart from their enforcement role, fare inspectors also play an important role to promote correct usage of the system, to educate the passengers and to ensure public safety on the network. In addition to their actions, large preventive operations are often organized, communication campaigns are common and dedicated signage extensively covers the network to remind passengers of the appropriate behaviour. The current legal framework in France has recently evolved with the approbation of the law n° 2016-339 [2]. Among other things, this law allows PTO to increase the price of penalties in a certain range, creates a recidivist evader status ("délit de fraude d'habitude") and gives the possibility for PTO to communicate with administrative databases to verify the identity and residence of fare evaders. Despite this combination of measures fare evasion continues to cause concern in Lyon.

2.3.2 Data description

In the following section, we describe the data that we use to estimate fare evasion. We make the distinction between the data that are automatically collected by the PTO during the day-to-day operational activity and the survey data whose specific purpose is to estimate fare evasion.

Automatically collected data

The first provider of data is the fare collection system. Every fare transaction is recorded and stored in a table. The transactions are time stamped and normally associated with location information. In our case : entry station for metro, line plus direction and stop for buses and tramways. In the literature, automatic fare collection data are also known as smart card data and have been used by

2. Available on-line at https ://www.legifrance.gouv.fr

transportation researchers for more than a decade. A thorough description of potential and type of data can be found in Pelletier et al. [2011].

The second providers of data are the different counting systems. In the metro, this task is done by infra-red hardware integrated on each entry and exit gate that deliver directional count measure for time interval of one minute. Each tramway door is also equipped with a passenger counting system that is integrated with the automatic vehicle location system. In each tramway door, there are between two and three infra-red cells that count both boarding and alighting. Thus, counting measure for tramways are available by vehicle and by stop. Finally, buses are also instrumented with counting cells. The counting system right now is not integrated with automatic vehicle location and different generations of technologies coexist. Thereby, for the buses, only the number of boardings per line and per days is currently available from the system. The reliability and accuracy of data generated by such an automatic counting system can vary greatly from one technology to another and is highly dependent on the effort put into the supervision and the maintenance of the system. It is thus very hard to estimate. In Lyon, the operators in place consider that the margin of errors is around 2% for the Metro and between 5% and 10% for buses and tramways but no statistical estimations are available.

The third provider of data is the fare inspection system. Here the equipment that collects data is the fare inspector handheld unit. This equipment is a mobile terminal with magnetic and RFID readers that can verify electronically smart card and paper ticket validity and also issue penalties. Every time fare inspectors start to check passengers they open a new session on the device and have to input their current position. A session is a set of contiguous control done by one inspector. Each day, one inspector may open as many session as he/she needs. The concept of session is used to distinguish between time spent checking passengers and time spent on other tasks. During a session, every check is saved and when the inspectors return the device to the warehouse, all the data is transmitted to a database using wireless connections. For each check, context data are available. The internal handheld unit software implements current TCL fare regulation rules and therefore automatically assigns the checks in different categories. We can then track the results of each check and if an irregularity is found, the nature of this irregularity is also recorded. Checks are time stamped and geolocated on the network based on the information manually input by inspectors. Thus, the quality of the geolocation data is highly dependent on the fare inspection operations and operators. More precisely, inspection environment (open or closed) and line information are always available because the inspectors must verify that the line of the last validation is equal to the line where the inspection takes place (except for metro where validation between lines is not mandatory). Transit stop code also needs to be manually inputted by the inspectors but because it is not mandatory for inspection operations, this information is not always updated when inspectors are moving from one location to another (especially when checks are done within a moving vehicle or within the underground system). Penalty data are collected in the same way. Every time an inspector issues a penalty, a record is saved in another table with detail about the fine such as the reason for the fine, the price of the fine, immediate payment or not and session location at the time of the fine. To illustrate our words, an extract of the inspection log data is given in table 2.1. The column CTRL_RESULT is a boolean with value one when an irregularity is detected

by the handled unit's software and in this case, the CTRL_INFO column indicates the nature of this irregularity based on the fare regulation rules implemented in the software. ID_CARD is an encrypted and unique number for smart card and is empty for a paper ticket. Other pieces of information available are context data regarding checks such as line number (LINE) or stop (STOP_CODE). In table 2.1, the second observation is a smart card that was inspected on line C24. As indicated by the CTRL_INFO column, the last tap-in of this card was more than one hour before being inspected i.e before 07 :34 a.m still there is no loss of revenue otherwise the description would have been "contract not valid". This observation could, therefore, be classified as an irregularity without loss of revenue. The last observation in this table is a paper ticket that was not stamped by a farebox meaning that this passenger was travelling with a ticket which was still blank. For this reason, this passenger would be classified by the operator's rules as fare evasion with loss of revenue.

SESSION_NUMBER	DATE	TIME	LINE	STOP_CODE	ID_CARD	CTRL_RESULT	CTRL_INFO
3641479	2017-03-29	19 :26 :40	33	1588	328681	0	
3831316	2017-07-11	08 :34 :31	C24	41074	324E8C	1	Travel time exceeded
4098957	2017-12-13	12 :42 :30	68	41994		0	
4120338	2017-12-28	17 :01 :43	C1	276		1	No validation

TABLE 2.1 – Data sample of inspection log

Survey data

In Lyon, as part of the contract between the transport authority and the operator, the fare evasion rate is measured by an independent on-site survey. This survey is done by a private company specialized in surveys and commissioned by the transit authority. Three points of measures are carried out each year. Annual results are part of the contractual indicators and associated with financial incentives or penalties which stress out the need for a reliable methodology. The measurement takes the form of a stratified random survey where a sample of passengers are asked to show their proof-of-payment. This sample aims to be representative of the network usage and it is stratified between the transportation mode (bus, tramway and metro), three types of day (weekdays, Saturday and Sunday) and the different lines of the network. Surveyors work alone in plain clothes and are requested to behave as a normal passenger and be as discreet as possible. All interviews are done on board. In the subway network surveyors follow random path and select boarding passengers at different sections of the vehicle. For buses and tramways, surveyors work on selected vehicle run. They board at the terminus and select boarding passengers randomly along the route. Once the passenger is selected the interview is carried out in three steps. First, to ease the process routine questions are asked regarding the current trip : boarding stop, travel purposes and frequency of usage. Those informations are recorded manually by the surveyor in a tablet computer. Then, the surveyor requests the passengers to show his/her proof-of-payment and verifies it electronically using the same handheld units as fare inspectors. When doing so, the non-punitive aspect of the survey is reminded to the passenger. In this step, the surveyor is responsible to ensure the accuracy of the check. In the case of irregularity, the surveyor has to identify clearly the nature of the infraction and report the

result on the tablet computer. Finally, complementary questions related to socio-demographic characteristics of the passengers are asked and also input in the tablet computer. If passengers refuse to answer the questions, the surveyor asked only to verify the proof-of-payment emphasizing the non-punitive aspect of the investigation. To ensure that the chosen passengers are not aware of the finality of the survey only passengers that board the vehicle after the end of the previous interview must be selected. Surveyors also try to hide the fare inspection handheld unit behind the tablet computer before interacting with passengers. Once the field collection is complete the logs from the handheld units and the information reported on the tablet computer are merged to ensure the validity of the collected data. This is done using the chronological sequence of face-to-face interviews. Each record is then classified according to the rules in figure 2.2. Denial of response or denial to show proof-of-payment are recorded and are equal to 2.5%. They are reintroduced in the final dataset as passengers without any fare support in the proportion of two times the average percentage by transportation mode of passengers surveyed without any fare support.

For this research, the final dataset resulting from this methodology was made available to us with the following variable month, transportation mode, type of day, line and if any the type of irregularity classified according to the figure 2.2. In 2017, the three waves of the survey took place in January, March, and November. In table 2.2, the number of surveys available by wave and by type of day is given. For 2017, a total of 56 746 passengers were interviewed. The estimated total traffic by type of day and month is also given. The resulting sampling rate of the survey varies between 0.2% and 0.9% of the estimated number of trips per type of day.

		Number of surveyed passengers	**Estimated traffic**	**Sampling rate**
January				
	weekdays	15 044	1 655 631	0,9%
	Saturday	2 495	1 056 578	0,2%
	Sunday	2 462	533 978	0,5%
March				
	weekdays	14 143	1 747 759	0,8%
	Saturday	2 388	1 111 961	0,2%
	Sunday	1 602	589 001	0,3%
November				
	weekdays	14 856	1 786 742	0,8%
	Saturday	2 474	1 126 490	0,2%
	Sunday	1 282	580 219	0,2%

TABLE 2.2 – Descriptive statistics of the survey sample, Source : Author's calculations

2.3.3 Measurement method

In this research, our goal is to measure the prevalence of fare evasion within the population of trips undertaken [Delbosc and Currie, 2018]. Two distinct measures that encompass different behaviours are estimated, i.e fare evasion with loss of revenue and a more global rate of fare irregularity. Based on the data previously

described, we define in table 2.3 space and time sets of boardings and link them to the available data sources.

Symbol	**Description**	**Source of data**
Ω	{Set of all boardings}	APC : Automatic passenger counting
V	{Set of all boardings with a fare transaction}	AFC : Automatic fare collection
Ct	{Set of all boardings with a fare inspection}	FIS : Fare inspection system
Cp	{Set of all boardings with a fare inspection resulting in an irregularity}	FIS : Fare inspection system
Cpp	{Set of all boardings with a fare inspection resulting in an irregularity with loss of revenue}	FIS : Fare inspection system

TABLE 2.3 – Sets of boarding and linked data sources

We should have the following relations between sets $V \subset \Omega$, $Ct \subset \Omega$, $Cp \subset Ct$ and $Cpp \subset Cp$. Then, if we consider only the set of boardings with a fare inspection we can compute the following rates :

$$C_i = \frac{|Cp|}{|Ct|} \tag{2.1}$$

$$C_r = \frac{|Cpp|}{|Ct|} \tag{2.2}$$

Those two indicators used as a denominator Ct a subset of Ω whose size depends on the intensity of inspection. C_i can be interpreted as the total irregularity rate derived from the fare inspection activity. C_r is created using Cpp as the numerator to take into account only fare evasion with loss of revenue. C_r should be interpreted as the rate of fare evasion with loss of revenue derived from the fare inspection activity.

Then, if we consider the complete universe of boarding, using fare collection data and counting data we can define the following ratio :

$$V_i = 100 - \frac{|V|}{|\Omega|} \tag{2.3}$$

V_i should simply be interpreted as a rate of fare non validation i.e a percentage of boardings without a corresponding fare transaction.

All the above measurements are based on automatically collected data and due to their nature, no confidence interval can be provided here. We will confront them with the survey described in section 2.3.2. In this dataset, each record is classified according to figure 2.2. It is therefore straightforward to calculate with the survey

two measures of fare irregularity. In this paper, S_r will refer to the fare evasion rate with loss of revenue estimated with the survey. S_i will refer to the fare irregularity rate estimated with the survey. As in all stratified random surveys, the proportion should be recalculated using the weight of each stratum. In our case, the weights are the estimated total number of trips in each stratum (month, type of day, transportation mode and lines) and they are based on the counting data. The weighting process allows to increase the survey representativeness of the true population. To determine the confidence interval of a proportion estimated from stratified random sampling we need to compute the standard error. Given a stratified sample, each stratum sample proportion $\hat{p}_h$ is an unbiased estimator of the stratum proportion p_h. An estimate of the variance of the proportion within a stratum h can be estimated from the sample,

$$\hat{s}_h^2 = \frac{n_h}{n_h - 1}\hat{p}_h(1 - \hat{p}_h) \tag{2.4}$$

In order to compute the margin error at the level of the population, we estimate the variance of the proportion at the level of the population which can be expressed as follows [Ardilly, 2006],

$$\hat{V} = \sum_{h=1}^{H} W_h^2 \frac{(N_h - n_h)}{N_h} \frac{\hat{s}_h^2}{n_h} \tag{2.5}$$

where H is the total number of stratum, n_h is the number of surveyed people in stratum h, N_h is the estimated total number of trips in stratum h, N is the estimated total number of trips and $W_h = \frac{N_h}{N}$ is the weight of stratum h. The confidence interval for the population is then defined using the standard normal distribution and a level of confidence set at 95%,

$$\hat{p} \pm 1.96\sqrt{\hat{V}} \tag{2.6}$$

where $\hat{p}$ is calculated as follows,

$$\hat{p} = \sum_{h=1}^{H} W_h \hat{p}_h \tag{2.7}$$

Assuming a random sampling in each stratum, the computation of the confidence interval can be done both for the proportion of surveyed passengers who fare evade with loss of revenue S_r and for the proportion of passengers surveyed with irregularity S_i. From the previous definitions, we can expect that $S_r \approx C_r$ and $S_i \approx C_i \approx V_i$.

2.3.4 Data pre-processing

In order to calculate the previous measures, it is important to ensure that the data is correctly prepared. To derive measurements of fare evasion with the fare inspection logs we need to analyse each inspection and to classify them according to the nature of the irregularity. Observations have shown that many duplicates are present in the database. Fare inspectors often swipe multiple times the same fare support

inadvertently or out of necessity (for instance to double-check). A de-duplication procedure is therefore needed. As described in section 2.3.2, in the dataset, a smart card can be identified with an anonymized unique number whereas tickets are not unitary identifiable. Fare inspectors shifts are also anonymized and unique per day and per fare officer. Having this in mind, the following two hypotheses are suggested to perform this data cleaning task :

— The same smart card cannot be inspected more than one time within 10 minutes. Only the first inspection for this smart card is retained ;
— A fare inspector cannot inspect two different tickets within an interval of 3 sec. Only the first inspection for this fare inspector is retained.

Those two rules are applied to our dataset and 2.4% of the inspections are discarded.

The second step is to merge the control log and the penalty logs because those two data sources are stored in two distinct tables. In the penalty logs, we have the reason for the penalty. To make sure that we do not reintroduce a duplicate we have to keep only the record where the reason for the penalty was "lack of travel ticket". In fact, when an officer is inspecting a passenger that has neither a card nor ticket to show to the handled unit the only record we could expect from the fare inspection activity is a penalty. Otherwise, the penalty should be theoretically preceded by the inspection of the fare support and therefore already in the control logs. In 2017, 79% of the penalties were for "lack of travel ticket", an additional 18% for irregularity but with a transport title that was theoretically previously shown to the fare officer (such as no validation or ticket overtime). The rest was not to be considered as pure fare infraction but rather behaviour infractions ("Class 4") such as quest for money, smoking inside vehicle or station, violent behaviour etc.

The last step is to affect each record from the resulting table between valid behaviour, fare irregularity without loss of revenue and fare irregularity with loss of revenue. As explained earlier, irregularities are already flagged in the control table with a description that makes reference to the current regulation rules (see figure 2.2). By combining the fare product (pay-as-you-go or pass) and the description of the irregularity we can classify each record. Penalties for "lack of travel ticket" are of course classified as irregularity with loss of revenue and are also taken into account when computing the total number of inspected boardings $|Ct|$.

In this case study, we also rely on automatic fare collection data and automatic passenger counting data. Disaggregated data have been processed and validated by local operators analysts. For this research,we have had access only to aggregated data by line (or station) and by day for 2017. Those data were then used to compute the V_i ratio at different levels of granularity such as type of day, transportation mode or line.

2.4 Results

We begin this section by a presentation in table 2.4 of some important annual figures. In buses and tramways, the total number of trips is considered equal to the number

of boardings as given by the counting system. The yearly number of boardings is equal to 166 million for buses and to 95 million for tramways. In the metro and funicular network, around 160 million entries are numbered and the connection rate between lines is estimated around 1.3 which leads to a total of 208 million trips. The percentage of paper ticket greatly varies from 29% in the closed environment to less than 20% in the bus system. This percentage also varies with location and time. The annual number of inspections and the annual number of penalties for lack of travel ticket are also given by transportation mode. The absolute number of inspections is maximum for the closed environment with 2.4 million inspections. However, if we relate those numbers to the number of trips by computing inspections rates, tramways are proportionally more inspected with an inspection rate around 1.7% compared to around 1.2% in the bus and metro. On the contrary, the ratio between the number of penalties versus the number of inspections is maximum in the bus which could indicate a higher propensity to fare evade compared to metro or tramway.

	Metro and funicular	**Tramway**	**Bus**
Number of boardings (APC)	159.6 M	95.2 M	166.1 M
Number of validations (AFC)	148.1 M	63.9 M	111.6 M
Non validation rate (Vi)	7%	33%	33%
Percentage of paper ticket (AFC)	29.1%	23.8%	19.6%
Number of inspections (FIS)	2.4 M	1.6 M	1.9 M
Number of fines issued for lack of travel ticket (FIS)	68.1 K	68.8 K	102.6 K

TABLE 2.4 – TCL 2017 annual figures, Source : Author's calculations

In the following sections, we compare the different estimations we obtain using the measurement method presented in section 2.3.3. For compatibility reason with the survey, all calculations are done only for non-holiday days covering the 2017 period of the survey (January, March, and November). For brevity, only the most important results will be described.

2.4.1 Comparison by transportation mode

In table 2.5, the value of each indicator is given by mode of transportation. The table shows that in the closed environment (metro and funicular) the phenomenon of fare irregularity is less pronounced than in open proof-of-payment system. Measures range from 4.2% to 9.6% whereas in the open environment (tramway and bus) measures range from 7.6% to 33.5%. This result is in agreement with other studies that demonstrate that a closed environment can reduce fare evasion [Dauby and Kovacs, 2007, Smith and Clarke, 2000, Clarke, 1993]. Nonetheless, even with closed gates, fare evasion could not be considered as marginal and the estimated rates are much higher than what is reported in other closed environments such as New York city subway [Reddy et al., 2011] or London underground [Clarke, 1993]. Table 2.5 also shows that dispersion between indicators is bigger for buses and tramways. For instance, for the tramway between the maximum estimation V_i and the minimum estimation C_r, the ratio is bigger than 4. This could indicate that the range of misbehaviours is wider in an open environment but could also be due to the fact

that the different estimations do not necessarily fit in with each other.

		Metro and funicular	**Tramway**	**Bus**
Fare inspection (FIS)	Cr (with loss of revenue)	4,2%	7,6%	9,7%
	Ci (irregularity)	4,9%	14,7%	22,2%
Fare collection and passenger counting (AFC + APC)	Vi (non validation rate)	7,5%	33,6%	32,9%
Survey	Sr (with loss of revenue)	7,5% ± 0,55%	17,9% ± 0,62%	19,9% ± 0,54%
	Sr (irregularity)	9,6% ± 0,61%	25,4% ± 0,7%	28,4% ± 0,61%

TABLE 2.5 – Estimation by mode, Source : Author's calculations

More precisely, in the metro environment, C_r and C_i are relatively close in value (4.2% vs 4.8%) meaning that most of the passengers found with an irregularity are also passengers generating a loss of revenue. Using data from counting and fare collection system we have obtained a ratio V_i of 7.5% equal to S_r the percentage of fare evasion with loss of revenue obtained with survey data (7.5% ± 0.55%). Finally, the survey estimation of fare irregularity is the upper-bound estimate with a value equal to 9.6%.

In tramways and buses, more discrepancy is observed between measures. Estimation based on fare inspection data indicates levels of irregularity quite low compared to the survey. For the tramway, C_r is equal to 7.6 % when S_r almost reach 18 %. For the bus, C_r is equal to 9.7 % when S_r is equal to 19.7 %. The fare non validation rate, V_i is measured at 33.5% for tramways and 33% for buses which is sensibly higher than what is obtained with the survey (25.4% and 28.4% respectively). In open proof-of-payment system, the ratio between C_r and C_i are around 2 meaning that when inspecting passengers one out of two found with an irregularity generates a loss of revenue. When calculating this same ratio with survey data results are closer to 1.4 meaning that in the survey the proportion of fare evasion with loss of revenue within the passengers in irregularity is much higher than in fare inspection data.

This first comparison shows that the different indicators can be different in absolute value. Divergences are especially large in the open environment which stresses out the need for more disaggregate comparison. Even with different absolute levels, the measurement could be related in their variation. That is why in the following sections, we focus on two important dimensions for transport operators by line and by day of the week. We will analyse variation rather than absolute value, therefore, confidence intervals will not be provided for the survey data.

2.4.2 Comparison by line

To deepen our comparative study, we now examine the relationship between indicators at the level of transportation lines. We compute for the study period, the different indicators for each line of the open system only (bus and tramway) because the non validation rate cannot be computed by line for the metro and funicular (validation and counting take place at stations). For reason of

representativeness, we eliminate lines with less than a total of 30 passengers interviewed by surveyors in 2017. We obtain 81 lines that represent 96% of the open environment counted trips for the study period. We draw on separate scatter plots in figure 2.3 the relationship between indicators that should approximately measure the same phenomenon. Regression line and R-squared are also indicated in the graphs.

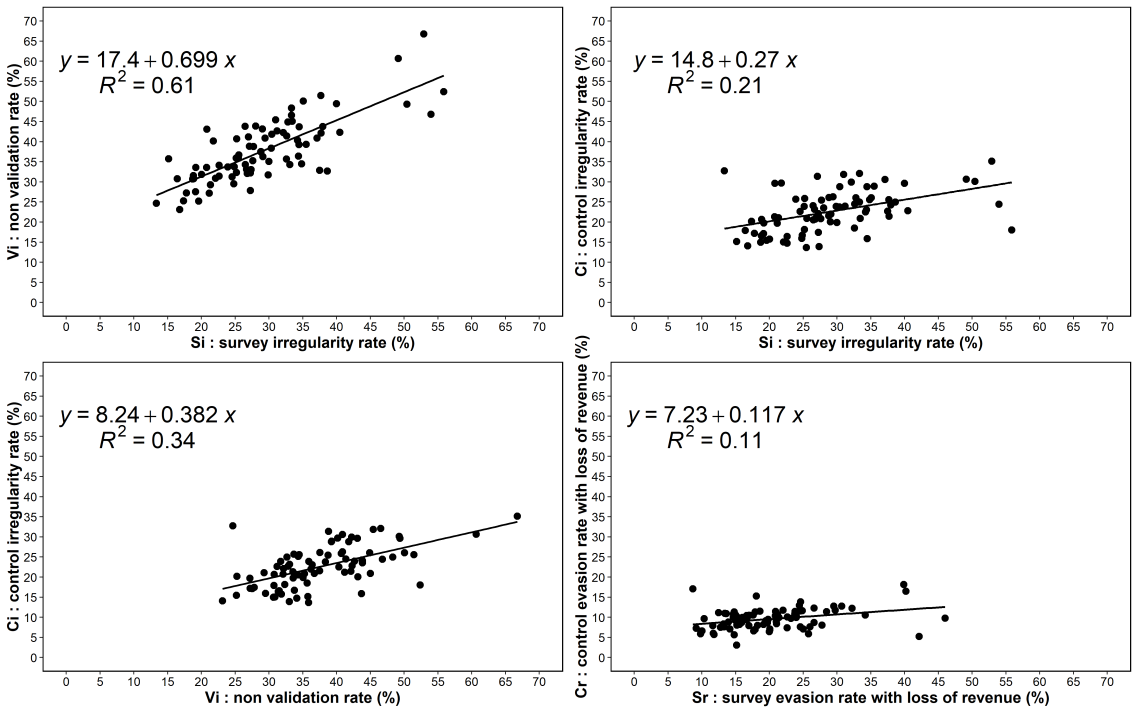


FIGURE 2.3 – Comparative scatterplot of indicators by line, Source : Author's calculations

In figure 2.3, it can be seen that the surveyed rate greatly varies from one line to another, for instance S_r varies from 10% to almost 55%, which means that fare evasion is not a homogeneous phenomenon. In the top left plot the regression line indicates that 61% of the variance in the survey irregularity rate S_i can be explained with the non validation rate V_i. The two ratios are correlated in their variations within lines despite their difference in magnitude (slope of 0.69). In the top right plot, the regression between the control irregularity rate C_i and the survey irregularity rate S_i suggest a weaker correlation with an R-squared equal to 0.21. The relation between the control irregularity rate C_i and the non validation rate V_i yield an R-squared a bit higher (0.34). Thus, both plot involving C_i suggest that the proportion of inspected passengers found with irregularity may fluctuate among lines but not in proportion as significant as in the survey or in the ratio between validations and counted boardings. In the bottom right plot, we compare the fare evasion rate with loss of revenue compute with survey S_r and fare inspection logs C_r. The linear relation between those two measures is very weak with an R-squared equal to 0.11. C_r is almost constant within the different lines meaning that the proportion of people not paying the legal fee found by inspectors is also quasi-constant. If we recall that the number of penalties issued is a major component of C_r then we may think that there is a saturation effect in the number of penalties issued by inspectors which results in C_r being almost constant. Taken together the plots in figure 2.3 indicate

that they may not always be strong correlations between the different measurement methods. Especially, there is some evidence that the fare inspection log may not produce evasion rates that significantly vary between lines.

2.4.3 Comparison by day

It is well known that days of the week have a great impact on public transportation usage and therefore it is likely that fare evasion is also variable according to the day of the week. All measurements presented in section 2.3.3 are computed for three types of day in the study period : Monday to Friday (weekdays), Saturday and Sunday. Results are given in figure 2.4 and show that fare misbehaviour rates are not constant throughout the week and that the prevalence of fare irregularity is higher during the weekend. Especially, S_i, S_r, and V_i are minimum during weekdays and maximum on Sunday. Those three measures vary in the same direction along the types of days. On the contrary, C_i and C_r are maximum on Saturday and decrease on Sunday. This may corroborate the previous observations that the estimation based on the processing of fare inspection is not well correlated in variation with other data sources. Especially, it is likely that the fare inspectors schedule and control plan changes according to the day of the week. This may impact the value of C_i and C_r without revealing a real change in the intensity of fare evasion.

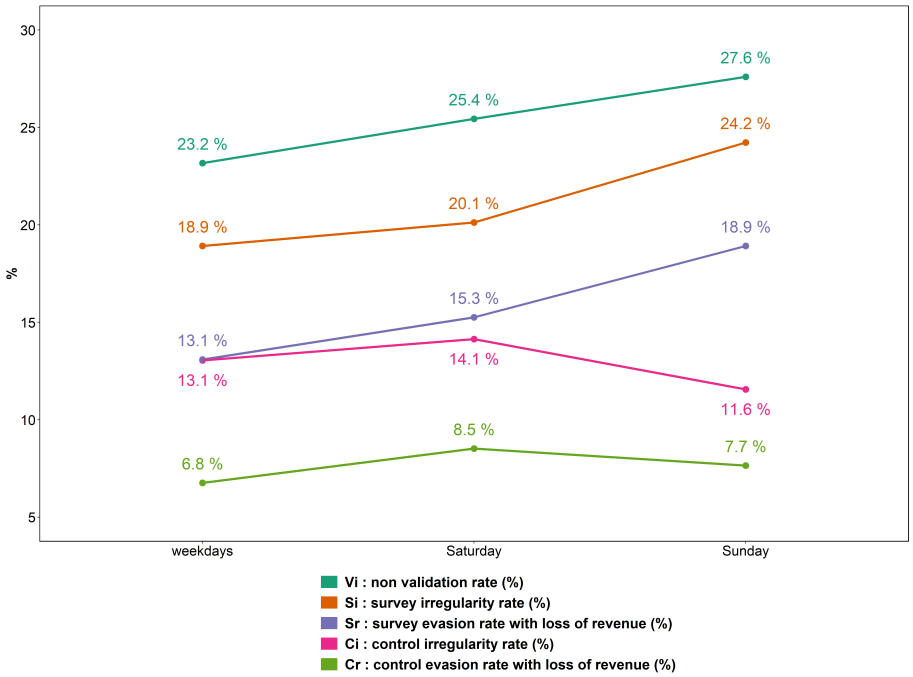


FIGURE 2.4 – Indicators by type of day, Source : Author's calculations

2.5 Discussion

Compared to the studies mentioned in the literature review, the results presented in the previous section indicate levels of fare evasion somewhat high. In the list of cities reported by Troncoso and de Grange [2017], only South American cities present levels of fare evasion bigger than 10% with Santiago clearly showing the highest level of fare evasion (27.6%). In the 22 American cities survey by Larwin [2012] the average evasion rate was measured at 2.7%. In San Francisco, Lee [2011] found a 9.5 % minimum systemwide fare evasion rate. In Montréal, Pourmonet et al. [2015] found an average non validation rate of 3.3% on the 16% of bus equipped with automatic counting. Dauby and Kovacs [2007] also report interval of fare evasion rate for European cities which ranges from 1% to 25%. However, comparisons between cities are often complex because there is a number of elements that come into play such as measurement method, cultural disparities, physical environment, fare equipment, pricing policy, legal framework, enforcement strategies etc. One important point is also that the inspection rate in Lyon is around 1.3% but some cities have much higher inspection rate as reported by Dauby and Kovacs [2007], Larwin [2012], Hauber [1993]. The inspection rate in Lyon is also much lower than what could be the optimum inspection level that maximises profit i.e 3.4%-4.0% according to Barabino et al. [2014], Barabino and Salis [2019].

In the previous section, we have also highlighted discrepancy and lack of coherence between data sources. To explain these differences and clarify the potential trade-offs of each estimation method we have summarise in table 2.6 the main pros and cons of each data sources. This table serves as a guide for the rest of the discussion.

	Pros	**Cons**
Survey	-Specifically designed to measure fare evasion -Randomization in passenger selection and stratification should ensure representativeness -Can provide precise information on irregular behaviour and socio-demographic of passengers	-Depend on the ability of surveyors to correctly interview passengers -Costly and limited in terms of sample size but also spatiotemporal coverage -Difficult to use for day-to-day operational planning and monitoring
Fare inspection logs	-Continuous data collection -Measure the inspectors production -Possibility to identify different types of fare evasion	-Presence of inspectors could change the behaviour of passengers -Inspection strategies cannot be identified clearly and not all operational aspects can be taken into account -Not necessarily random and representative of network usage
Passenger counting and fare collection	-Can cover the entire network continuously with no human interventions -Easy to calculate once the data has been consolidated -Could be used to measure variations and feed models	-Does not identify types of fare evasion and not all aspects of ticketing can be controlled -May have different meanings in closed and open environment -Quality of data and accuracy of counting instruments may be difficult to evaluate

TABLE 2.6 – Pros and cons of each data sources, Source : Authors

In Lyon, a specifically designed survey is used to measure fare evasion and produce official figures. This survey is based on a stratified random sample in order to be representative of the network usage. The stability of the data collection methodology

over time ensures the consistency of the measurement at the macroscopic level. This survey is based on on-board face-to-face interviews with passengers which makes it possible to collect precise information concerning the nature of irregularities and passenger profiles. These elements are useful to have a realistic view of the typology of fare misbehaviours. Nonetheless, this data source also has several disadvantages. As in all surveys, the data is subject to bias. The accuracy of data relies on the ability of surveyors to interview passengers while they are travelling, and on their capacity to build a relationship of trust. This task can be complex in a crowded vehicle, when passengers are in a hurry or when dealing with passengers that speak other languages. Moreover, it is expensive to send surveyors to the field. Thus, the amount of data gathered during each survey wave is limited. It is therefore difficult to consider to use this type of data collection methodology to monitor continuously fare evasion and to support operational planning.

Fare inspection logs are appealing data sources to investigate and measure fare evasion. In Lyon, this data is continuously generated by inspectors during their day-to-day activities (average of 16,000 records per day). The collected data are essential to evaluate the productivity of inspectors in terms of number of checks, number of penalties etc. They are also needed to measure the inspection rate. This variable is critical as some authors found that increasing the inspection intensity is one of the most effective ways to prevent and reduce fare evasion [Hauber, 1993, Dauby and Kovacs, 2007, Killias et al., 2009, Delbosc and Currie, 2018]. In this case study, we proposed a methodology to use those logs as a way to survey the network in a disaggregated manner. Because each unitary check is linked to a specific fare support (smart card or ticket), and associated with contextual information, those logs can be used to calculate two levels of fare irregularity. Our comparative results suggest that those indicators may underestimate the prevalence of fare irregularity and fare evasion with loss of revenue. They also indicate that variations of measures based on inspection data are not tightly correlated by line and by day with other data sources. Inspection logs are closely related to inspection strategies and inspectors do not systematically follow randomization in selecting passengers or areas to inspect. This intrinsic link with operational processes may cause desynchronization between real and inspected traffic and may impact the phenomenon we want to measure. For instance, inspectors in plain clothes tend to target their control whereas inspectors in uniform are likely to miss potential fare evaders that escape or validate when seeing inspectors [Delbosc and Currie, 2018, Hauber, 1993]. Some areas of the networks with high levels of violence are hardly inspected because the safety of inspectors is at risk, but those areas may experience higher levels of fare evasion. On board inspection and at stop inspection may not results in a similar volume of control and similar reactions of passengers. Furthermore, vehicle size and crowding level also have an impact on the collected data. For example, longer vehicles such as tramways cannot be inspected completely between two stops. Apart from the repressive aspect of their activity inspectors play a key role in preventing fare evasion. Their presence can deter some evaders or force them to move from a non-paying condition to a paying one. When evaders are found, writing a penalty always remains at the discretion of the inspector and they do not systematically do it. They may be lenient and judge that this will be counterproductive. They may negotiate one penalty to be paid immediately for multiple fare evaders or ask passengers to purchase a ticket. Those

events are not recorded in the logs even though they may have an impact on the measured evasion rate. This is also the case if inspectors do not use the handled units (eyesight verification) or do not systematically ask for justification when checking holder of fare with discounted price. In theory, all those elements should be taken into account to measure accurately fare evasion. Unfortunately, this cannot always be done in a satisfactory manner with the raw inspection data. For those reasons, we believe that in Lyon fare inspection logs have significant limitations to measure accurately the level of fare evasion. One way to overcome some of those limitations would be to implement beforehand a clear codification of inspections strategies (on board/at stop, plain-clothes/uniform, 100% sweeps etc.).

In Lyon, all vehicles and stations are equipped with an automatic passenger counting system and an automatic fare collection system. Those systems generate data automatically and continuously at a very low cost with fine spatiotemporal resolution. In contrast to survey or inspection, no human intervention is needed to collect the data, and once the data is consolidated, it is quite easy to calculate the ratio between both quantities. This ratio can be an interesting indicator of the level of irregularity. It can be used to monitor continuously the network, analyse variation, and potentially feed predictive models. However, this ratio has also some limitations. By definition, all passengers that do not validate are considered as evaders and no precise information on the type of evasion can be derived from this indicator. In a closed environment with physical barriers such as station with gates, passengers who enter without valid fare need to enforce the system showing a certain level of intention to evade. In this context, the amount of non-intentional evasion and opportunistic evasion should be greatly reduced and we could expect that the ratio of evasion with loss of revenue and the ratio of non validation will be very close. In practice, this may not exactly be the case. First, because some passengers may enter by ways that are not equipped with the counting system (such as exit section, elevator with direct access to the platform or by climbing above the gates). Second, because the fare system cannot detect people entering with a transport title that will expire once inside the system or without the proper justification. Third, because in a gesture of solidarity some passengers who leave the metro donate their paper tickets (when still valid) to entering passengers. In open environments, the temptation to fare evade may be higher because there is no physical constraint. This can increase the proportion of opportunistic fare evaders and non-intentional evasion. For instance, in a very crowded vehicle, passengers cannot always access the farebox, especially when boarding by rear doors not equipped with fare equipment. When making connections, some passengers do not feel the need to validate or may forget to do so. Thus, in open environments, the non validation rate aggregates in one indicator a diversity of misbehaviours and situations that could have very different meanings and implications. The relevance of this indicator also depends greatly on the quality and the accuracy of counting instruments. Errors can result from several factors. There can be mechanical problems such as defective counting cells or cells that are not connected to the door opening system and count continuously. There can be systematic errors when the system constantly counts incorrectly for instance with old equipment that need to be recalibrated. Passengers behaviours can also have an impact on the counting data. For instance, when several people are entering

successively forcing a metro gate the number of entry can be underestimated. Passengers sometimes need to go on and off of vehicle at certain stops to let other passengers alight when it is too crowded which can generate double count. Passengers may carry bulky things like bags that are also counted as passengers. Finally, post-processing of data can generate additional errors such as wrong allocation of counts or even data loss. All those factors have an impact on the non validation rate but they are hard to evaluate and correct especially when the number of equipment is large like in Lyon. However, as noted by Furth et al. [2005], the correction of raw APC data stream to ensure accurate counts is key to the usefulness of APC data. Thus, particular attention should be given to the quality of APC data if used for fare evasion measurement purpose.

All these elements lead us to believe that no single data source meets all needs. Each source has advantages and disadvantages and does not measure the phenomenon with the same definition and in the same way. This can create significant discrepancies between indicators and sometimes contradictory results. This also confirms that fare evasion is a complex phenomenon that can be hard to measure precisely and continuously. Just as it seems necessary to combine a package of measures to increase fare compliance [Tirachini and Quiroz, 2016], it may be necessary to combine a variety of data sources to improve the quality and reliability of fare evasion measurement methods.

2.6 Conclusions

Fare evasion is a common phenomenon in transit system. Information to study it can be hard to obtain [Reddy et al., 2011] and PTO are not always willing to give publicity to fare evasion [Hauber, 1993]. Fare misbehaviours are also complex and there is a great deal of variation within the spectrum of fare evasion [Delbosc and Currie, 2016]. Yet, fare evasion needs to be measured accurately to assess revenue at risk, to measure the efficiency of enforcement strategies, to limit potential crime, to design delegation contracts with incentive mechanism and to get a clear understanding of representativeness issue in fare collection data. In many cases, a specifically designed survey may be the best way to achieve this but this approach has also some limitations. In this research, we have examined the viability of alternative data sources to measure fare evasion. Our results suggest that using data from fare inspection activities could lead to important underestimation of the fare evasion phenomenon. Using fare non validation rate derived from automatic counting and fare collection system may be a more promising direction. It could help to explore the structure and variability of fare evasion. Nonetheless, this indicator may not be fully satisfactory because not all aspects can be controlled and not all passengers who do not validate their transport ticket or smart card should be considered as wrongdoers [Delbosc and Currie, 2018]. For those reasons, we believe that in Lyon it will be hard to completely replace surveys with only automatic data collection.

In the near future, the multiplication of connected devices will continue to increase

the quantity and the variety of data collected by PTO. Those new data sources are appealing to measure fare evasion but they require great attention from operators to ensure quality and completeness. They also require careful preprocessing and data cleaning to transform raw data into relevant information. Once pool together those data sources can help to improve the quality and reliability of fare evasion measurement methods. They can also help to improve the understanding of the nature of fare evasion and its evolution if the potential bias are identified. This requires that we compare them and systematically question their validity. In this domain, we believe that there is still room for improvement especially when studying phenomena that are very sensitive to the local context such as fare evasion.

3

How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination survey and household travel survey ? An empirical investigation in Lyon.

This chapter is an edited version of the following article :

O.Egu and P.Bonnel (2020). How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination surveys and household travel survey ? An empirical investigation in Lyon. *Transportation Research Part A : Policy and Practice*, 138, 267-282.

DOI : https ://doi.org/10.1016/j.tra.2020.05.021

Highlights

- Researchers and practicians now have access to a more heterogeneous corpus of data sources to estimate travel demand
- We perform an empirical comparison of three independent data sources to estimate public transit origin-destination trip matrices
- Automatic fare collection data is not error-free and needs to be supplemented with automatic passenger counting to estimate the full demand
- Household travel survey may significantly underestimate the volume of public transit trips
- The results provide a better understanding of the available data sources for public transit demand estimation and the potential trade-offs between them

Chapter 3 : How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination survey and household travel survey ? An empirical investigation in Lyon.

Abstract

Origin-destination (OD) matrices are one of the key elements in travel behaviour analysis. For decades, transportation researchers have mostly used data obtained by active solicitation such as surveys to construct these matrices but new data sources like automatic fare collection (AFC) are now available and can be used to measure OD flows. As a result, a more heterogeneous corpus of data sources is now available to estimate travel demand. However, little research examines how comparable the estimated demands may be. In this paper, three data sources namely a household travel survey, a large scale origin-destination survey and entry only automated fare collection are processed to derive typical weekday public transit OD trip matrices. Various elements of the resulting matrices are then compared. While all the matrices share some common characteristics, there are also substantial differences that must be addressed. AFC data is not error-free and needs to be supplemented with data from other sources to construct a representative OD trip matrix. This is because not all destinations can be inferred, the smart card penetration rate is far less than 100% and fare evasion cannot be ignored. Our empirical results suggest that scaling an AFC matrix with automated passenger counts may be a viable solution. The results also indicate that the household travel survey significantly underestimates the volume of public transit trips compared to the other sources. The findings of this research contribute to a better understanding of the available data sources for public transit demand estimation. They can help practitioners to improve the quality and accuracy of OD matrices.

Keywords— Public transportation, big data, smart card data, travel survey, OD matrices

3.1 Introduction

Origin-Destination (OD) matrices play a key role in travel behaviour analysis. They are essential for building travel demand models which are often used to evaluate public planning policy and costly infrastructure investments [Ortúzar and Willumsen, 2011, Bonnel, 2002]. For decades, transportation researchers have mostly used data obtained by active solicitation such as surveys to construct these matrices but the rapid rise of intelligent transport systems, the widespread implementation of passive sensing and recent advances in data collection and data processing are changing this portray [Munizaga and Palma, 2012, Pelletier et al., 2011, Chen et al., 2016, Bonnel and Munizaga, 2018].

Both types of data have their pros and cons [Bagchi and White, 2005, Chen et al., 2016]. Active solicitation data have the advantage of being specifically designed to answer a research question and meet given objectives. However, they rely on people's ability to report precisely on their travel behaviour [Stopher and Greaves, 2007]. They are also costly and often based on small samples both in terms of observations and spatiotemporal coverage [Bagchi and White, 2005, Chen et al., 2016]. On the contrary, passive data may have larger sample sizes but are rarely collected for travel behaviour analysis which may result in poor semantics [Chen et al., 2016, Pelletier et al., 2011, Bagchi and White, 2005]. To compensate, inference methods are needed but these rely on hypotheses and can introduce bias. In addition, passive data often suffer from the quasi-absence of sociodemographic information.

Thus, researchers and practicioners now have access to a more heterogeneous corpus of data sources with a variety of collection methodologies and semantics. We believe that to estimate travel demand more accurately it may be beneficial to pool these data sources. However, such pooling can only be of practical use if the area of relevance of each source is identified and the differences between them are known. This requires comparative studies and we argue that this is an important task, although so far there have been very few empirical investigations.

In this context, the aim of this paper is to compare and evaluate the differences between three independent data sources for transit demand estimation. The research focuses on estimating typical weekday demand. Three data sources are used : a large scale origin-destination survey (ODS), a household travel survey (HTS) and automatic fare collection (AFC). They are first processed independently to estimate travel demand which is synthesized in the form of a public transit OD trip matrix. Various aspects of the resulting matrices are then compared to explore and explain possible differences. The results empirically validate the usefulness of AFC data for estimating public transit travel demand at the network level. They also help us to achieve better classification and understanding of the available data sources and their potential bias.

The remaining part of the paper is structured as follows. A literature review is presented and the research needs are identified in Section 2. Materials and methods are described in Section 3. The results are presented in Section 4. Implications of the findings, conclusions and areas for further research are set out in Section 5.

3.2 Previous work and research needs

3.2.1 Literature review

AFC systems were originally designed to facilitate fare collection and speed up the boarding process [Pelletier et al., 2011, Trépanier et al., 2007]. In most flat fare networks, AFC systems only record passenger boarding information but not alighting information (referred to as an entry only system). Considerable research efforts have therefore been directed at estimating the alighting stop of these boardings and identifying trips as these two elements are required to build OD matrices. The most common technique relies on the trip chaining method where it is assumed that the alighting stop of a given transaction is close to the origin of the next transaction stop and that no trip leg involved other transportation modes (walking, car, bicycle etc.). Barry et al. [2002] were among the first researchers to develop such an algorithm for the subway network of New York City considering only successive subway validations. Their algorithm was then further extended by Zhao et al. [2007] to build an OD matrix for the subway network of Chicago considering also bus transactions and using iterative proportional fitting to estimate the full matrix. A mathematical formalization of the trip chaining algorithm that considers multiday data was proposed by Trépanier et al. [2007]. Their algorithm was applied to the bus network of Gatineau (Canada) ; a tolerance distance of 2km was used and 66% of the alighting stops were inferred. Munizaga and Palma [2012] merged AFC data with real-time bus GPS data and proposed a modified version of the trip chaining algorithm. Instead of minimizing the distance between successive transactions, they chose to minimize the generalized travel time to take account of possible bus circuity. Their model was then applied to the large scale network of Santiago considering both bus and subway transactions. In London, using Oyster card data, Gordon et al. [2013] proposed a series of rules to identify transfer transactions more accurately and a methodology to scale passenger trip flow using additional information such as count data [Gordon et al., 2013, 2018]. They showed that on the subway network this method produces results similar to iterative proportional fitting but were not able to validate the method exogenously at the network level due to a lack of data. Alsger et al. [2016] used data from the South-East Queensland network (Australia) that included both boarding and alighting times to validate the trip chaining method assumptions at an individual level. They concluded that increasing the allowable walking distance beyond 800 meters did not significantly improve the matching rate and that the allowable transfer time may have little impact. Nunes et al. [2016] applied the trip chaining method to the case of the main operator of Porto buses. They proposed additional endogenous validation rules for distance-based fare structure but failed to provide any sort of exogenous validation. Nassir et al. [2015] built upon the work of Gordon et al. [2013] to suggest a more complex set of rules to identify transfer. These rules are based on the concept of off-optimality whose goal is to capture the deviation from the optimal path.

Household travel surveys have existed for a considerable time and remain an essential data source for transport planning. Recently, a number of issues were

raised regarding such surveys. These include their problematic cost, rising non-response rates, the omission of a significant number of trips and the accuracy of the collected data [Stopher and Greaves, 2007, Stopher et al., 2007, Jones and Stopher, 2003]. In addition, for public transit planning applications, they also suffer from small sample sizes, inadequate spatial resolution and temporal incompatibility as they are collected only once every few years [Chapleau et al., 2008]. To compensate for these shortcomings they are often supplemented with manual surveys such as OD surveys. In some studies, these additional data sources have been used in conjunction with AFC data. Using a travel diary survey, Barry et al. [2002] have for instance confirmed that the alighting assumptions for the subway network are correct for 90% of surveyed riders. Munizaga et al. [2014a] have validated part of their methodologies using metro OD surveys and data from volunteers where the respondent accepted to divulge their smart card ID. They found that the alighting stop was correctly estimated in 79% of the cases. OD surveys were also used at an aggregated level to validate results from AFC for a specific line [Wang et al., 2011] or for the subway network of Santiago [Pineda et al., 2016]. In London, Seaborn et al. [2009] have explored various time thresholds to identify transfer with AFC and compare the results to the London Travel Demand Survey highlighting some inconsistencies. Riegel [2013] also in London, have used a subset of respondents from the London Travel Demand Survey who agreed to disclose their card number to investigate the potential of AFC data to enhance travel surveys. They found that only 44% of the sample had perfectly matching survey and smart card trip legs and recommended taking account of the AFC record earlier in the survey interview process using prompted recall methods. In Montreal, Spurr et al. [2015] have proposed a method to match household travel survey respondents with smart card data and have provided three examples that demonstrate the possibility of discrepancies at the individual level. Chapleau et al. [2018] have also compared the Montreal household travel survey with AFC to identify, at the aggregated level, potential bias in the HTS data. Their analysis shows that the household travel survey may capture simplified travel patterns such as work- and home- based trips. With the same data, Spurr et al. [2018] have explored the effects of using either data source for metropolitan transit financing. They concluded that neither HTS nor AFC data may be sufficiently representative of actual transit use for this task. Tamblay et al. [2016] have proposed a methodology that can be applied to stop-to-stop OD matrices obtained from AFC to infer a zonal OD matrix that captures full trips (with their access and egress link). The model requires land use information and a zoning system. It was calibrated with a passenger access survey conducted at public transit stops and the total generated and attracted trip share per commune was compared with a large scale OD survey to provide external validation of the results. Finally, some authors have tried to enrich the AFC dataset with these additional data sources. For instance, Kusakabe and Asakura [2014] used trip survey data to calibrate a naïve Bayes probabilistic model that is then used to infer the trip purpose of AFC trips. In a similar vein, Alsger et al. [2018] used HTS survey, land use and OD surveys to calibrate a rule-based model that estimates the purpose of smart card trips in Brisbane, Queensland (Australia).

3.2.2 Research needs and objectives

This broad literature survey shows that although much research already exists on the use of AFC for travel demand estimation, there is also a clear need to compare and integrate the resulting demand with more traditional data sources. Yet, it is not clear how comparable the resulting demand estimates may be and we believe that more research needs to be done in this direction.

In fact, it is commonly accepted that once an entry-only AFC dataset has been processed to infer alighting points and identify transfer it is possible to derive OD matrices at the desired disaggregation level [Munizaga and Palma, 2012]. However, many factors may affect the quality and usefulness of the resulting matrices, such as the violation of the trip chaining assumptions, fare non-interaction, the proportion of passengers using smart cards and self-selection bias among smart card users [Zhao et al., 2007, Munizaga and Palma, 2012, Kurauchi and Schmöcker, 2017, chap. 2]. Such issues need to be considered and empirical validation of the matrices derived from AFC must be undertaken. To the best of our knowledge, no previous exogenous validations have examined these aspects in full or validated the results at the macroscopic level (i.e. for the whole public transit network). To do this, traditional data sources are required. However, these data sources may not be comparable and will probably also have biases. It is therefore important that the validation process helps us in turn to examine their accuracy. This means we should introduce as many data sources as possible in the comparative study to improve the value of the experiment.

In Lyon, it is possible to undertake this task because three data sources that provide a picture of public transit usage are collected independently : a large scale OD survey, a household travel survey and automatic fare collection. These three data sources differ in terms of the information they contain, their collection methodologies and their level of spatial and temporal resolution. However, we expect them all to describe in a more or less similar fashion the way people use the transit network and move within the city on an average weekday. From the travel demand modelling perspective, this usage can be summarized with OD trip matrices. Thus, the goal of this paper is to provide an empirical comparison of the public transit OD trip matrices that can be derived from three data sources. As far as we know no previous work has advanced so far in this comparative task. In particular, no study compares OD matrices derived from HTS surveys with what can be derived from AFC or other actively collected data such as OD surveys. This research will examine the validity and accuracy of each data source in order to improve the quality of public transit demand estimation. The next section begins with a presentation of the data sources.

3.3 Materials and methods

3.3.1 Context, study area and terminology

TCL ("Transport en Commun Lyonnais") is the commercial name of the urban public transit network of Lyon. The network consists of 4 metro lines, 2 funicular lines, 5 tramway lines and more than 100 regular bus lines. This network is currently run by a private operator under the supervision of the public transport authority of the Lyon metropolitan area (SYTRAL).

The chosen spatial level of aggregation to build the different OD matrices is depicted in Figure 3.1. The entire zone covers an area of 746 square kilometres and has a population of approximately 1.3 million inhabitants. In the central area, the public transit network is denser with the presence of mass transit infrastructures. This area has been divided into districts whose borders are shown in black. This results in 10 central area zones : the districts of Lyon numbered from 1 to 9 plus Villeurbanne. In the peripheral ring, the municipalities were aggregated to form 8 zones corresponding to known catchment areas whose names are given in Figure 3.1. This leads to a total of 18 zones and therefore 324 OD pairs for the matrices. This number of zones was defined to be compatible with the HTS sample size. In fact, since the sample of the HTS is limited, it is preferable to build larger zones to observe a sufficient number of surveyed transit trips in most OD pairs.

Before going into detail on how each OD trip matrix is built we shall clarify the terminology used in this paper. A trip will refer to the movement of a passenger from an origin to a destination. On buses and tramways, a trip leg is the movement of a passenger on a single vehicle between a boarding point and an alighting point. In the subway networks (metro and funicular) no validation is needed for transfers. To ensure comparability between the datasets, in this environment a trip leg is defined as the movement of a passenger between an entry station and an exit station. A trip is made up of at least one trip leg. Transfer between trip legs can include waiting time and walking time but not any activities. An itinerary is a set of lines or entry stations (in the case of the subway) used for a trip. A day (or service day) is defined as running from 4.30 a.m. to 4.30 a.m. on the next day when there is no activity on the transit network.

3.3.2 Large scale origin-destination surveys (ODS)

The first available datasets were manually collected face-to-face origin-destination surveys (ODS). In Lyon, all the lines are manually surveyed every five years on a continuous basis. During the survey, each passenger is asked to describe their current trip. The following details are recorded : (1) origin-destination of the trip leg on the surveyed line ; (2) the connecting line before the surveyed one, if any (up to 3) ; (3) the connecting line after the surveyed one, if any (up to 3) ; (4) origin-destination of the full trip coded at the stop level ; (5) socio-demographic of the

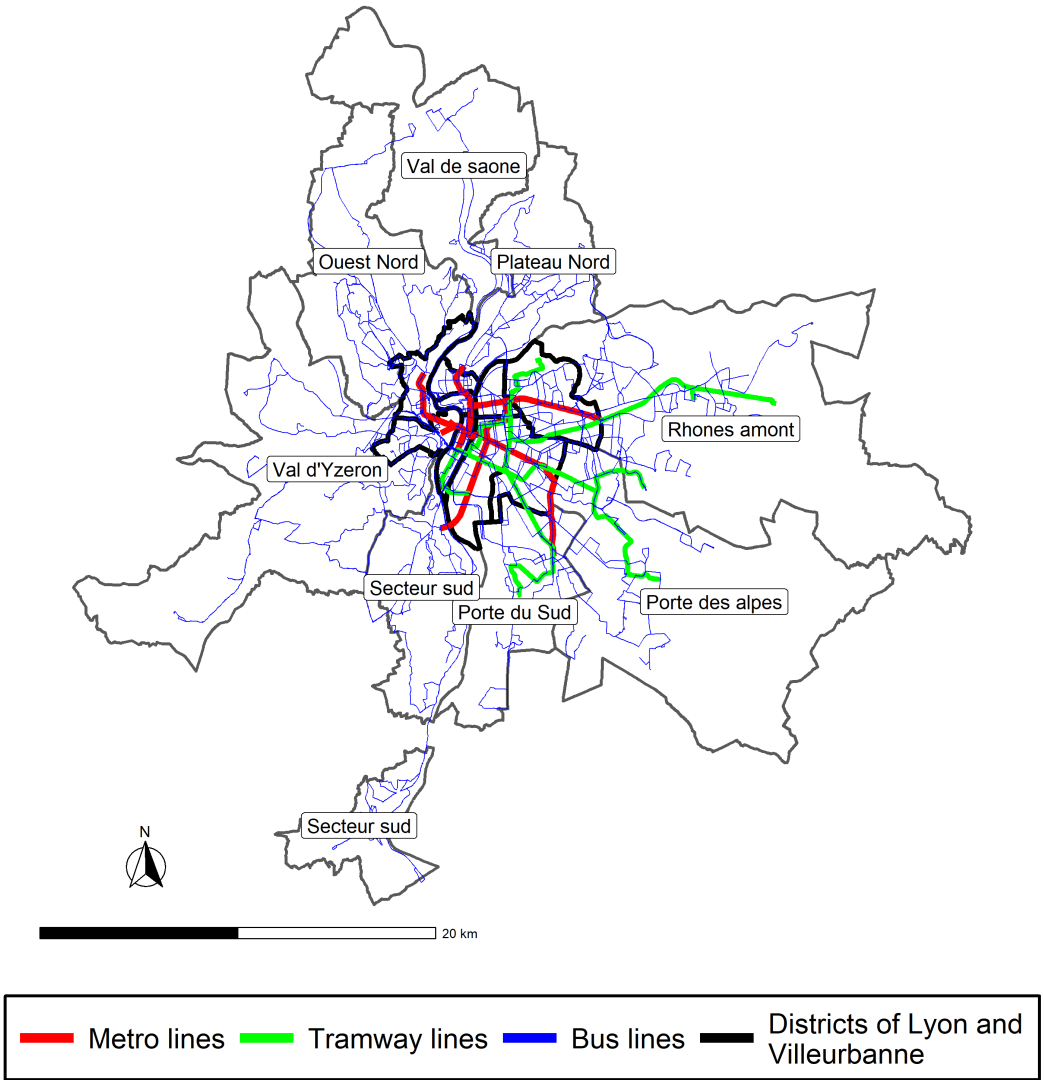


FIGURE 3.1 – Study area and zoning system, Source : Authors

passenger, trip purpose and public transit access modality. Each line is surveyed for three types of days : weekdays, Saturdays and Sundays. On bus lines, this survey is conducted on board and assumed to be exhaustive (all the daily runs on each type of day are surveyed). For tramway and subway lines, this survey is conducted at stations. A random sample of approximately 35% of passengers for tramway lines and 25% of passengers for subway lines are surveyed. The results are then scaled according to half-hour counts at each station derived from the automatic passenger counting system. The goal of this survey is to obtain accurate information about transit usage on each line. It is used for tactical and operational level planning (bus route itineraries modification, network redesign, route level transfer analysis etc.). It is also used to calibrate local transit assignment models.

To obtain data for the complete networks, all the OD surveys collected between 2013 and 2017 on weekdays were compiled into a single database. In this database, each observation corresponds to a trip leg, and in theory each trip should generate a number of records equal to the number of trip legs it contains. For instance, a

passenger making a trip that involved a transfer between line A and line B will be surveyed both on line A and line B. To obtain a database that reflects an average network weekday and not an average weekday of each line we must deduplicate. To do so, all records containing a transfer line before the surveyed trip leg were discarded. Therefore, the remaining records in the database only contain transfer lines after the surveyed trip leg (if any and up to 3). Each record should then correspond to a unique trip on an average weekday over these five years. In this database, the origin and destination of all trips are encoded by surveyor team at the stop level so that it is possible to build the ODS trip matrix by intersecting with the zoning presented in Figure 3.1. It should also be noted that between 2013 and 2017 there have been no major changes on the network.

3.3.3 Household travel survey (HTS)

In the Lyon metropolitan region, the last household travel survey was performed between October 2014 and April 2015. This survey was done according to the French standard methodology fixed by the CEREMA [1] [CERTU, 2008]. For a large zone like the Lyon metropolitan area, this survey consists of a combination of face-to-face and telephone interviews. Telephone interviews are reserved for areas where resident densities are low while face-to-face interviews are conducted in denser areas. As indicated by Figure 3.2, the survey perimeter extends considerably beyond the TCL network (i.e. the study area) and all the households living in the area served by TCL were interviewed face-to-face. During the last survey, a total of almost 16,400 households were surveyed (approximatively 10,000 of them face-to-face). The survey focused on weekdays and was performed from Tuesday to Saturday.

This survey takes the form of a geographically stratified random survey of households [CERTU, 2008]. First, a zoning system that is compatible with administrative data is defined. It consists of sampling sectors that are further divided into finer zones to allow more detailed encoding of the geographic attributes of trips. In each sampling sector, administrative and fiscal data are used to build a sampling frame of primary residences where a sample of households is drawn without replacement. In each household, all individuals over five years old are then requested to describe all the trips they made the day before the interview. A large number of socio-demographic and socio-economic variables regarding households and individuals are also collected. To ensure statistical accuracy at the spatial level two conditions need to be met for each sampling zone : at least 70 households and at least 160 persons have to be surveyed [CERTU, 2008]. Once all the field data are collected, a quality control check is performed to ensure that the sample data is coherent and exhaustive. Finally, a data expansion process is implemented to correct for potential sample biases (as the respondent sample may not be perfectly representative of the actual population) and to derive weights that can be used to obtain statistics for the entire population of residents. This requires the use of external data such as census data or administrative data (provided by the French National Institute of Statistics known as INSEE). This expansion procedure is conducted in two stages. First, the distribution of household

1. Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement.

sizes (in 5 categories) by sampling sector is used to derive a household sample weight. Then, the distribution of the population by age segment (9 categories) and by sampling sector is used to determinate a balancing weight for each surveyed individual. Once combined, the two weights can be used to expand the data collected from individuals to the known characteristics of the population of residents in terms of household size and age segment. For this research, the final dataset resulting from this process and duly validated by the CEREMA was made available to us.

To build a public transit OD trip matrix with this data source, all trips that were reported as made on the TCL network were extracted (without considering the household's residential zone). The unweighted sample consisted of 10,570 trips made by 4708 distinct individuals. The origin and destination of trips were known at the level of the most fine-grained zoning. As shown in Figure 3.2 the size of these zones ranges from the municipality to small zones of less than 100m^2 in areas with a high population density. They are therefore always much smaller than the zoning system shown in Figure 3.1. The barycentre of each of the finest zones was used to encode the trip origins and destinations. By combining the resulting trip table with the zoning presented in Figure 3.1 an OD trip matrix can be derived from the HTS survey. The expansion of the matrix to the full population of residents is done using the individual calibration weight estimated from the methodology described above.

3.3.4 Automatic fare collection (AFC)

Destination inference and trip identification

In Lyon, the current fare transaction system was introduced in 2002. It is an entry only fare transaction system. Passengers can either use smart cards or magnetic paper tickets which must normally be validated every time they board a vehicle (tramway or bus) or enter the subway network (metro and funicular station). Smart card or ticket also needs to be validated to open some park-and-ride barriers. For this study, all fare transactions from the 13th to the 17th of March 2017 were extracted. This period corresponds to 5 standard weekdays in a month deemed to be representative. They should represent an average synthetic weekday and capture the structural elements of public transit demand. In the fare transaction database only the smart cards have a unique ID (all paper tickets share the same ID).

To enrich this data source the following steps were implemented :

— **Data correction**. In moving vehicles the automatic fare transaction system is normally integrated with automatic vehicle location (AVL) making it possible to identify the boarding stop, line, direction and route. However, we found that these details were sometimes missing. In this case, a search within the AVL dataset was conducted using the vehicle number and the timestamp. If the time difference between the closest AVL record and the AFC record was under 3 minutes the missing data were replaced based on AVL data.

— **Data deduplication**. To ensure integrity we made the hypothesis that two cards should not be validated within 10 minutes of each other at the same

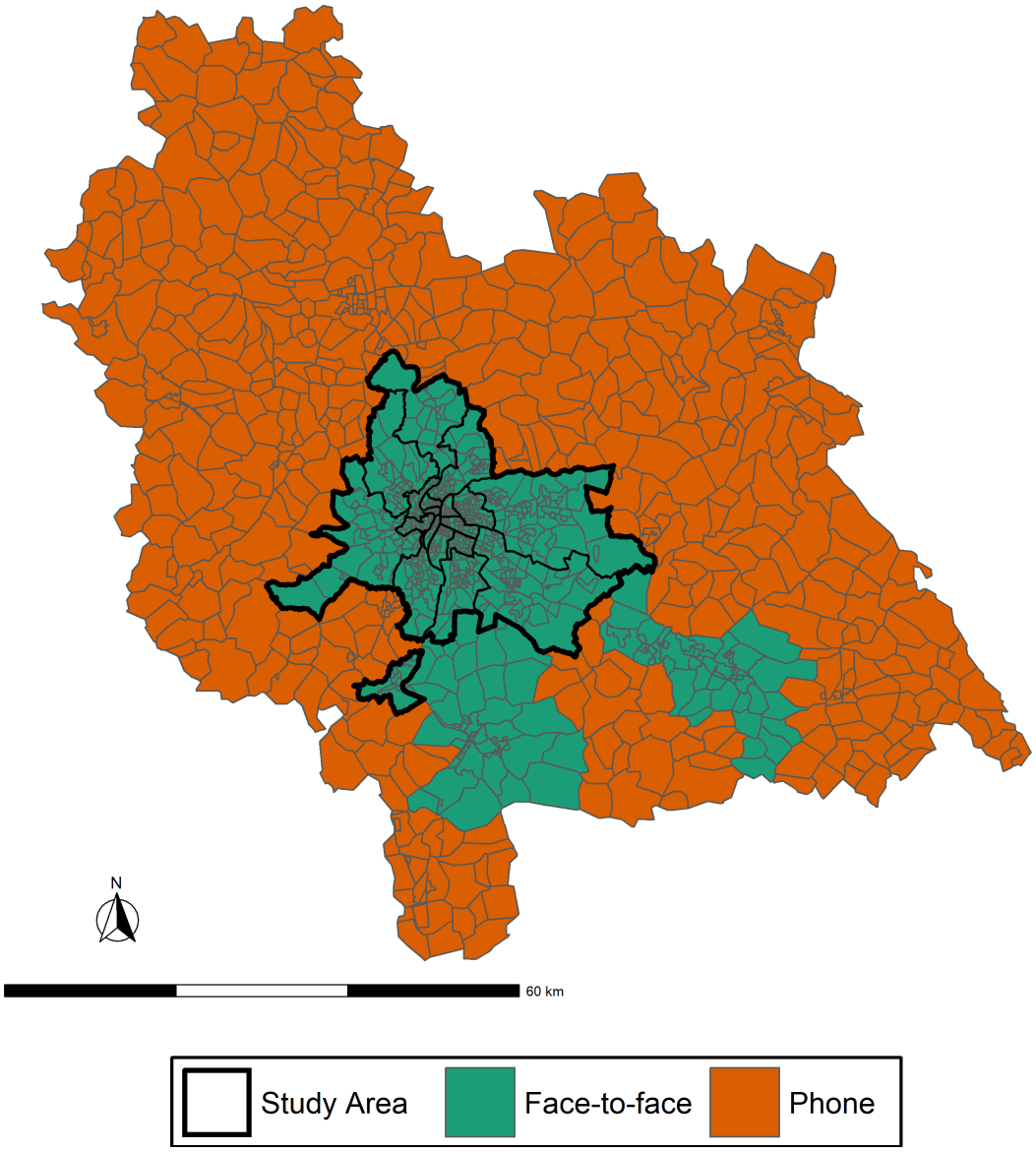


FIGURE 3.2 – Map of the area showing the finest zoning and the collection method for the Household Travel Survey, Source : Authors

subway station or in the same vehicle (tramway and bus). This procedure identified 0.56% of transactions as duplications.

— **Transfer identification**. To identify the beginning of a new trip and isolate transfer we implemented the following rules : (1) The first transaction of a day is always the beginning of a new trip, (2) two boarding transactions that occur within 60 minutes and that are not made on the same line or at subway stations are considered to be part of the same trip [Munizaga and Palma, 2012, Spurr et al., 2015, Devillaine et al., 2012, Seaborn et al., 2009]. For paper tickets, the system is designed so it is possible to identify whether a ticket has already been validated with the 60 minutes rule (fare policy allows transfers for up to 60 minutes). Therefore, it is possible to isolate the ticket transactions corresponding to the beginning of a trip. However, it is not possible to apply the binary criteria on the line and the station because

paper tickets do not have a unique number for tracing. As this is the only available information in the remainder of this paper we will assume that all paper ticket transactions flagged in the database as first stamp-ins correspond to a trip departure.

— **Destination inference.** The alighting stop or exit station was inferred using the trip chaining method, minimizing the distance between successive boarding transactions [Trépanier et al., 2007, Nunes et al., 2016, Zhao et al., 2007, Li et al., 2018]. The maximum walking distance was set at 600 meters. For the last transaction of the day, a return to the closest stop of the first transaction of the day was considered.

	# Transactions	% Transactions
Data error	272,641	5,6
Single trip	241,250	5,0
Inferred but not valid	417,730	8,6
Inferred and assumed to be valid	3,920,500	80,8
Total	4,852,121	100

TABLE 3.1 – Destination inference performance and source of errors, Source : Authors

Table 3.1 gives the detailed results of the destination inference process. This procedure yields a percentage of alighting points estimated of 80.8 % which is quite comparable with a number of studies of entry only system [Munizaga and Palma, 2012, Trépanier et al., 2007, Li et al., 2018, Nunes et al., 2016]. 5.6 % of alighting points cannot be inferred due to data errors such as stop missing for the next transaction, inability to identify the route number, bus not equipped with AVL and/or unknown boarding stop location (around 2.3% of the transactions). 5% of the transactions were either a single daily trip or the last trip leg of a single daily trip. In both cases, no estimation was performed [Nunes et al., 2016]. Lastly, 8.6 % of the estimated alighting points were either too far (more than 600 m) or the origin was the same as the destination.

Matrices scaling

After the four steps described above, we obtain a trip database where not all the trip attributes are available for each record. Firstly, for paper tickets, we can identify the trip origin and starting time but we cannot have any information regarding trip destination and transfers during trips. We will refer to these trips as TAFC (Ticket Automatic Fare Collection). Second, for SCAFC trips (Smart Card Automatic Fare Collection) the available elements are the trip origin, the corresponding trip leg transactions and, when successfully inferred, the trip destination. The subset of SCAFC trips with an inferred destination will be referred to as SCAFCD (Smart Card Automatic Fare Collection with Destination). Unfortunately, the full OD trip matrix cannot be estimated based only on this subset because (1) there are trips where the destination is unknown but the origin is identified such as TAFC (2) there

are trips where neither the destination nor the origin is known but they appear in the database (3) there are trips that do not appear in the database because of fare non-interaction and fare evasion [Munizaga and Palma, 2012]. To estimate the full demand, these three issues need to be tackled and the SCAFCD matrix needs to be scaled accordingly. In this research, we have examined two methods to do so.

The first method does not require additional disaggregated data but is based on strong assumptions. The goal is to build a zonal expansion and two uniform expansion factors as proposed by Munizaga and Palma [2012]. More precisely, if we assume that the distribution of the destination of trips with no inferred destination is the same as those from the same origin with an inferred destination we can obtain a zonal expansion factor f_o. Let T_{od} denote the set of trips contained in the full AFC database (tickets and smart card) with origin o and destination d (inferred or not). Then a zonal expansion factor f_o for each trip $T_{od} \in$ SCAFCD can be calculated as follows :

$$f_o = \frac{\sum_{d} T_{od}}{\sum_{d \neq null} T_{od}} \tag{3.1}$$

Then a uniform expansion factor can be calculated based on the total of trips contained in the AFC database :

$$f_2 = \frac{\sum_{o,d} T_{od}}{\sum_{\substack{d \neq null \\ o \neq null}} T_{od} \times f_o} \tag{3.2}$$

Lastly, another uniform expansion factor f_3 must be fixed to account for trips not recorded by AFC, for instance, using the rate of fare non-interaction derived from a survey or other external data sources. If we assume that the fare non-interaction rate is constant throughout the network, each trip is then weighted by multiplying the three factors. In this research, the uniform expansion factor f_3 was set using the results of a local fare evasion survey performed by a private company commissioned by the transit authority. In March 2017, the total rate of fare non-interaction was measured at 21% [SYTRAL, 2017].

When additional disaggregate data such as automated passenger counting (APC) are available, a second method for estimating the full demand would be to compute a scaling factor for each trip based on expansion factors for linked-trip itineraries [Gordon et al., 2018]. This requires a set of control nodes[2] where the total flow is known. The sample flow on each linked-trip itinerary is then scaled to correspond to the control node total such that,

$$\sum_{i \in I} B_{ni} \alpha_i t_i = \Delta_n \quad \forall n \in N \tag{3.3}$$

where I is the set of itineraries, B_{ni} is an incidence matrix indicating whether node n is traversed by itinerary i, t_i is the sample flow on itinerary i, α_i is the scaling factor for itinerary i and Δ_n is the difference between the control total and the

2. Defined by the authors as objects of the network where AFC transactions can occur.

sample total at node n [Gordon et al., 2018]. The problem is solved iteratively until the difference between the control flow on each node and the scaled flow is deemed marginal.

In our case, itineraries are defined as the sequence of bus lines, tramway lines and subway entry stations used for each SCAFCD trip (53,490 distinct itineraries observed). This definition was derived from the available APC data in Lyon where the daily counted boardings for each bus and tramway line as well as the daily number of entries at each station are recorded. The different volumes for the period from the 13th to 17th of March 2017 were used to define control node flow. This results in 155 control nodes (43 subway station, 5 tramway lines and 107 bus lines). To initialize the algorithm the seed flow in each itinerary is determined using the SCAFCD trips and then each α_i is computed recursively until all the control totals (APC volume) are matched within $\pm 0.1\%$. At the end of this process, the itinerary scaling factor α_i is applied to each trip with an itinerary $i \in$ SCAFCD and is taken into account to construct the OD trip matrix.

In the rest of this paper, the matrix resulting from the first method will be referred to as WOAFC (Weighted on Origin Automatic Fare Collection) and the matrix resulting from the second method will be referred to as WIAFC (Weighted on Itinerary Automatic Fare Collection). The structure of each of these matrices is inspected and compared in detail in the following section.

3.4 Results

3.4.1 Structural comparisons

Descriptive statistics regarding each dataset presented in the materials and methods section are given in Table 3.2. All the results from passive data were obtained by dividing the results for all the weekdays by five to mimic an average weekday. Table 3.2 shows that in terms of volume there are substantial differences. The total number of trip legs in the AFC database (SCAFC + TAFC) is around 1,21 million compared to 1,56 million in the APC database confirming that the total rate of fare non-interaction is close to 21%. When we compare the modal share of APC with the modal share of SCAFC and TAFC there are proportionally more fare transactions in the subway network. This could indicate a lower non-interaction rate for the subway probably because this environment is closed by barriers. SCAFCD trips contain less than 50% of the likely real number of trip legs as given by APC (1,56 million) or ODS (1,51 million). The share of subway trip legs in the SCAFCD trips is also higher (46% of subway trip legs). This may be because the destination inference method performs better for subway transactions. Weighted AFC trip databases have a similar number of trip legs but their modal composition is quite different. The AFC matrix scaled on origin (WOAFC) is based only on the distribution of observed transactions without taking into account the structure of fare non-interaction and therefore contains a large proportion of subway trip legs (46%). On the contrary, the AFC matrix scaled on itinerary flow (WIAFC) matches the structure of APC and includes 37% of

		APC	SCAFC	SCAFCD	TAFC	WOAFC	WIAFC	ODS	HTS
Volume									
	Trip legs (in million)	1,56	0,98	0,74	0,24	1,55	1,56	1,51	1,11
	Trip (in million)	-	0,70	0,54	0,17	1,11	1,10	1,16	0,80
Modal share of trip legs									
	bus (%)	41	38	34	29	34	41	39	43
	tram (%)	23	20	20	19	20	23	22	21
	subway (%)	37	42	46	52	46	37	39	36
Trip transfer									
	Mean trip legs per trip	-	1,39	1,39	1,45	1,39	1,42	1,30	1,39
	1 leg (% of trips)	-	68	68	-	68	65	73	66
	2 legs (% of trips)	-	27	27	-	27	29	24	30
	3 + legs (% of trips)	-	6	5	-	5	6	3	4
		-							
Trip euclidean distance (m)									
	Mean	-	-	3,551	-	3,550	3,545	3,288	3,706
	Median	-	-	2,891	-	2,888	2,836	2,568	2,967
	Q1	-	-	1,651	-	1,646	1,590	1,426	1,704
	Q3	-	-	4,718	-	4,718	4,718	4,392	4,934

TABLE 3.2 – Descriptive statistics regarding each data sources, Source : Authors

APC : Automated passenger couting, SCAFC : smart card automatic fare collection, SCAFCD : smart card automatic fare collection with inferred destination
TAFC : ticket automatic fare collection, WOAFC : automatic fare collection matrix weigthed on origin , WIAFC : automatic fare collection matrix weighted on itinerary
ODS : origin-destination survey, HTS : household travel survey

subway trip legs. In the ODS dataset, the number of trip legs is fairly close to what can be obtained with the scaled AFC (1,5 million vs 1,55 million). The modal share of ODS trip legs is also quite similar to the modal share of WIAFC (and thus APC) with 39% of subway trip legs. Finally, there is a significant difference in the volume of trips and trip legs contained in HTS. In this data source, there are only 1,11 million trip legs which is almost 30% less than in the weighted AFC matrices or ODS matrix.

As far as transfers during trips are concerned, there are some interesting differences between the data sources. In the OD survey, there is a higher proportion of single-leg trips (73%) compared to the SCAFC (68%) or HTS data (66%). The proportion of two-leg trips is highest in the HTS data source (30%). These differences are further confirmed by the analysis of the Euclidean trip distance. The ODS survey tends to capture smaller trips with a mean distance of 3,288 meters compared to 3,551 meters in SCAFCD and 3,706 meters in the HTS. This analysis is not possible for the TAFC database. However, the mean number of legs per trip is much higher than in the rest of the data sources (1.45). This is probably because in practice tickets are used for round trips that the sole use of a time criterion to flag first boarding cannot identify. As expected, the two weighting methods result in different trip transfer structures. In particular the weighting method based on trip itinerary increases the percentage of trips with transfers. This can be explained by two factors. First, trips that involve a transfer are more likely to appear in the SCAFCD database because there is a high probability of the destination being inferred. Secondly, itineraries that include a bus or tramway transfer have greater weighting factors to compensate for the higher rate of fare non-interaction. These differences can also be analysed by simultaneously comparing the volume of trip legs and the volume of trips. For instance, in the ODS the number of trip legs is lower than the WIAFC or WOAFC but the number of trips is 5% higher.

In Figure 3.3, the distribution of the number of trips per day and per person (or card) is given for the SCAFC and HTS trip database. There are major differences between the two sources for this dimension. The share of passengers making two trips a day is much higher in the HTS database (67%) than in the SCAFC (50%). In contrast, a high percentage of cards are used for only one trip per day (18%). This confirms observations by Spurr et al. [2015], Seaborn et al. [2009]. We also compare the AFC with the HTS database in terms of the number of passengers (or distinct cards in the case of AFC). For the 5 days, in the SCAFC database an average of 290,117 distinct cards were observed on the network compared to a total of 355,146 individuals for HTS. This difference is only informative if we consider that a number of passengers actually travel with paper tickets, evade fares or are not residents of the metropolitan area (therefore not included in the HTS database). For instance, if we make the hypothesis that the mean trip rate per ticket passenger is equal to the mean trip rate in SCAFC and HTS (2.33) then approximatively 71,000 passengers may travel with a paper ticket.

To conclude this structural analysis, the temporal distribution of the trip starting time for each data source is given in Figure 4. The trip distribution obtained from the HTS survey is markedly different from the rest of the trip database. It shows a higher concentration of trips during the morning and the afternoon peak period.

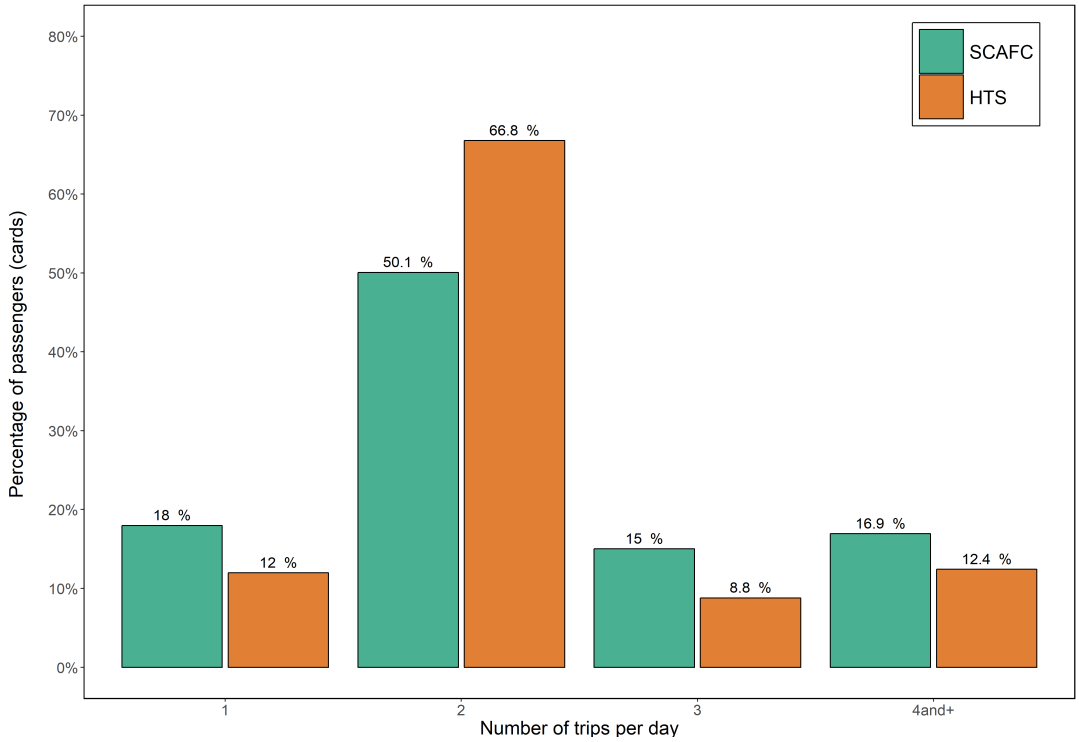


FIGURE 3.3 – Distribution of the number of trips per day and per person (or card), source : Authors

SCAFC : smart card automatic fare collection, HTS : household travel survey

The same discrepancy between AFC and HTS data has already been observed in the city of Montreal [Spurr et al., 2015, 2018, Chapleau et al., 2018]. The distributions resulting from the two weighting methods are quite close to the initial distributions of SCAFC and TAFC even if the temporal dimension was omitted from the scaling process (the focus is on the structure of the daily matrices). These distributions are also fairly similar to the trip distribution in ODS. The main difference is observed in the morning peak where there are more trips in the AFC database. Here, the difference could simply be the result of seasonal variation since ODS data are collected throughout the year.

3.4.2 Comparisons between the matrices

This section begins with an examination of the matrix margins. Because the total estimated flow in each matrix is different, the results are given in Figure 3.5 as a share of trips per zone and per matrix. This plot confirms that, overall, the margins are quite similar. The largest zone is the 3rd district of Lyon (where the central business district and the city's main railway station are located). For this zone, the number of trips is higher in the HTS matrix than in the others matrices, both in terms of origin and destination. In contrast, in the 7th district, demand seems to be underestimated in the HTS matrix. Smaller differences appear when we look at the peripheral areas with tramway lines such as Porte Des Alpes, Rhones Amont and Porte du Sud. In particular, in these areas, the number of trips in the WOAFC

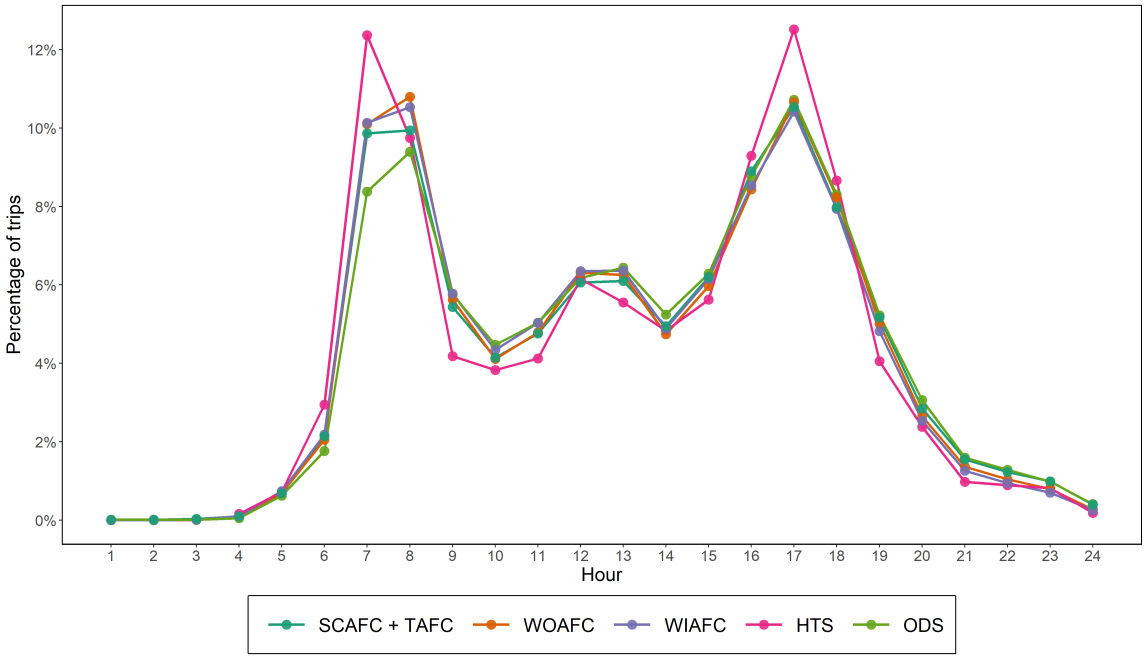


FIGURE 3.4 – Hourly trip distribution in each data source, source : Authors

SCAFC : smart card automatic fare collection, TAFC : ticket automatic fare collection,
WOAFC : automatic fare collection matrix weigthed on origin ,
WIAFC : automatic fare collection matrix weighted on itinerary,
ODS : origin-destination survey, HTS : household travel survey

database is lower than in other matrices.

To extend our analysis of the matrices we have aggregated the trip distribution to form two macro zones. The central area encompasses the 9 districts of Lyon plus Villeurbanne while other areas are considered to be part of the peripheral ring. The trip distribution between these two macro zones is given in Table 3.3. Despite the coarse level of aggregation, it can be seen that the matrices have differing trip distributions. First, the two matrices derived from AFC have a distinct macrostructure. The matrix that is weighted on origin (WOAFC) has more trips in the central area (63%) which is in line with previous observations regarding the modal composition of these trips (especially the high percentage of subway trip legs). On the contrary, the matrix that is obtained from weighting on itineraries (WIAFC) has more trips in the peripheral area and more cross-sectional trips. The structure of the HTS matrix is fairly similar to the WIAFC matrix with fewer trips in the central area. Finally, the distribution obtained from the ODS is intermediate between the other matrices with 61% of trips in the central area.

The differences were further examined using regression analysis on the 324 OD pairs. All the regression plots are shown in Figure 3.6, and to supplement them the main statistical parameters of OD flow distribution in the matrices are given in Table 3.4. The regression plots show that all the matrix structures are globally coherent but each matrix has some specific characteristics. As shown by the regression equations, the intercepts do not exceed 413 which can be considered as small compared to

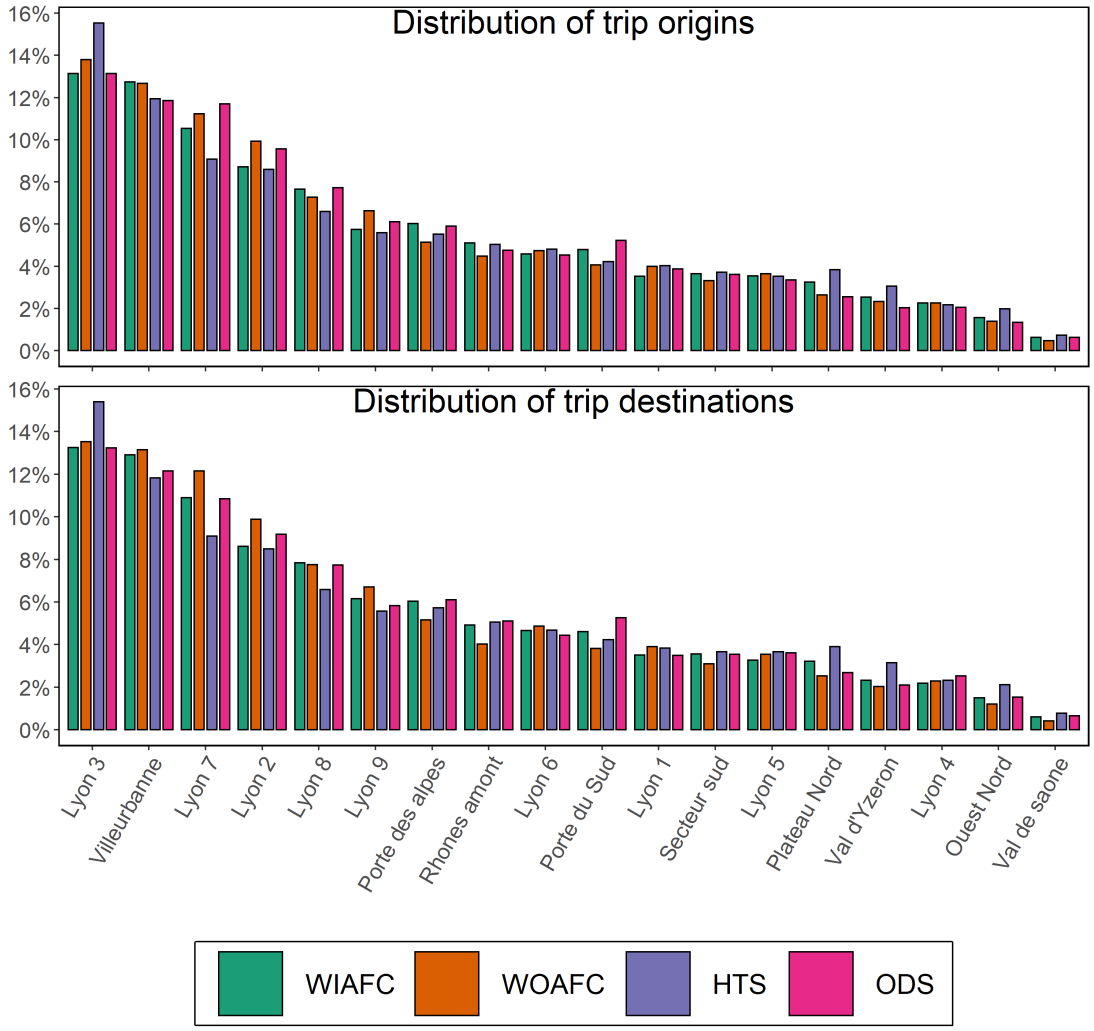


FIGURE 3.5 – Percentage of trips according to origin zone and destination zone, source : Authors

WIAFC : automatic fare collection matrix weighted on itinerary
WOAFC : automatic fare collection matrix weighted on origin
HTS : household travel survey, ODS : origin-destination survey

Origin	**Destination**	**WOAFC (%)**	**WIAFC (%)**	**ODS (%)**	**HTS (%)**
Central area	Central area	63	58	61	57
Central area	Peripheral ring	13	14	13	15
Peripheral ring	Central area	15	15	12	14
Peripheral ring	Peripheral ring	9	13	14	15

TABLE 3.3 – Macrostructure of the matrices, source : Authors

WOAFC : automatic fare collection matrix weighted on origin
WIAFC : automatic fare collection matrix weighted on itinerary
ODS : origin-destination survey, HTS : household travel survey

the mean value for OD pairs given in Table 3.4. The slopes vary from 0.6 to 1 as a result of different total trip volumes. As shown in Table 4, the HTS matrix has lower flow while the flow distributions in the other matrices are generally quite similar. In terms of R-squared, the best fit is between the ODS, and the WIAFC matrices with

an R-squared value of 0.97. In the top-left plot, there are four important OD pairs with a cross symbol where the flow in the WOAFC matrix is much lower than in the ODS pushing the regression line downwards. The points in question are closer to the diagonal line in the WIAFC matrix resulting in an R-squared value that is 0.04 points higher (top right plot). These four points correspond to intrazonal OD pairs in peripheral areas with large trip volumes (Portes des Alpes, Portes du Sud, Rhone Amont and Secteur Sud). This confirms our previous observation that some peripheral trips may be missing from the WOAFC matrix, in zones where the level of non-interaction (and fare evasion) is probably higher. The plot constructed with the HTS data have to be taken more cautiously as there are some OD pairs that have few surveyed trips. They show that the matrix built from HTS has a structure that is fairly similar to the other sources but also some OD pairs that significantly deviate from the regression line.

	Mean	**Median**	**Q1**	**Q3**
WIAFC	3394	1689	512	3780
WOAFC	3583	1847	482	4253
ODS	3584	1616	441	4094
HTS	2466	1174	336	2894

TABLE 3.4 – Main parameters of OD flow distribution by matrices, source : Authors

WIAFC : automatic fare collection matrix weighted on itinerary
WOAFC : automatic fare collection matrix weighted on origin
ODS : origin-destination survey, HTS : household travel survey

To conclude this section the differences between each matrix were evaluated in an aggregate manner using the Root Mean Square Error (RMSE) and the Symmetric Mean Absolute Percentage Error (SMAPE) :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{1,i} - y_{2,i})^2} \tag{3.4}$$

$$\text{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{\frac{1}{2}|y_{1,i} - y_{2,i}|}{(y_{1,i} + y_{2,i})} \tag{3.5}$$

where $y_{1,i}$ is the flow for pair i in dataset 1 and n is the number of OD pairs. The matrices were first normalized so they all contained the same number of trips as the ODS matrices. In Table 3.5, the error measures were first calculated using all the OD pairs (n=324). Then, in Table 3.6, the error measures were calculated only for pairs where the HTS was considered to be statistically significant i.e. with more than 30 surveyed trips (n=97) [CERTU, 2008]. Both tables are given as two matrices where the upper triangle corresponds to the RMSE and the lower triangle to the SMAPE. If we consider all the pairs the minimum RMSE value is between ODS and WIAFC with a value of 959, very close to the RMSE between WIAFC and WOAFC (1049). In terms of the percentage of errors, the two closest matrices are WOAFC and WIAFC with an SMAPE of 13%. If we consider only the most important pairs, the differences between WIAFC and WOAFC are greater and the WIAFC matrix becomes closer to the ODS matrix (SMAPE equal to 12%). This is largely due to the peripheral OD pairs with large flows where the WOAFC captures fewer trips.

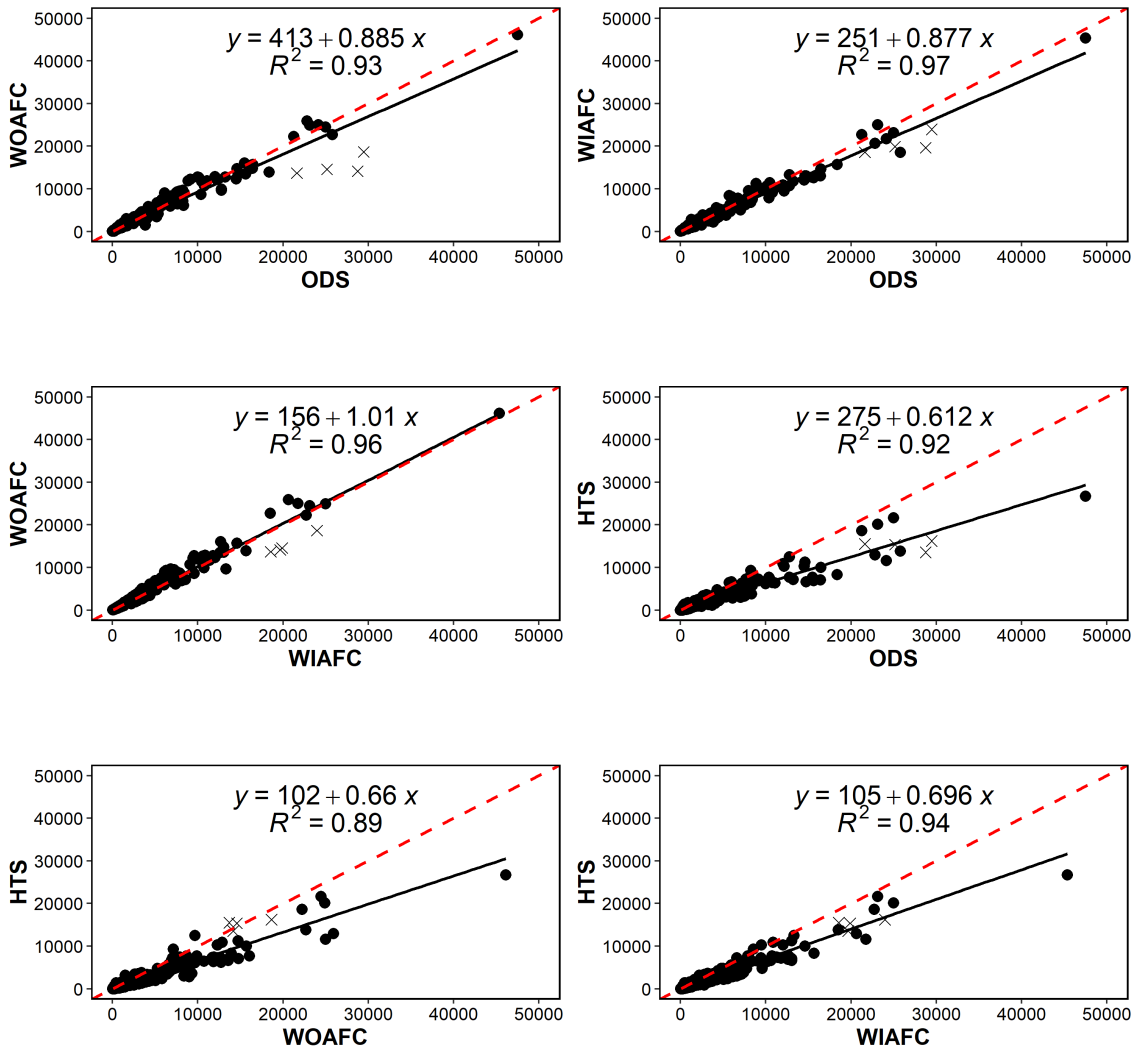


FIGURE 3.6 – Regression analysis between the matrices, source : Authors

WOAFC : automatic fare collection matrix weigthed on origin
WIAFC : automatic fare collection matrix weighted on itinerary
ODS : origin-destination survey, HTS : household travel survey

All the comparisons that involve the HTS matrix have higher measures of errors both in percentage and in absolute value. The SMAPE values are always over 20% and the RMSE values range from 2335 to 3065 compared to the lowest value of 1664 that is obtained by comparing ODS and WIAFC (table 3.6).

These results show that while all the matrices share a globally similar structure there are also significant differences. These differences are more easily understood if we bear in mind that the trip structure in each of the datasets may vary in terms of transfer rate, geographical distribution and modal share. In this section, all these aspects were examined when we moved from the aggregated structure of the matrices to the detailed analysis of trip distribution. The next section discusses the implications of these results and the lessons to be learned from this investigation.

	WOAFC	**WIAFC**	**HTS**	**ODS**
WOAFC	-	1049	1741	1484
WIAFC	**13%**	-	1350	**959**
HTS	40%	39%	-	1656
ODS	20%	21%	40%	-

TABLE 3.5 – Error measure between matrices for all OD pairs (n=324), source : Authors

WOAFC : automatic fare collection matrix weigthed on origin
WIAFC : automatic fare collection matrix weighted on itinerary
ODS : origin-destination survey, HTS : household travel survey

	WOAFC	**WIAFC**	**HTS**	**ODS**
WOAFC	-	1874	3065	2663
WIAFC	15%	-	2335	**1664**
HTS	24%	22%	-	2903
ODS	16%	**12%**	21%	-

TABLE 3.6 – Error measure between matrices for pairs with more than 30 surveyed trips in HTS (n=97), source : Authors

WOAFC : automatic fare collection matrix weigthed on origin
WIAFC : automatic fare collection matrix weighted on itinerary
ODS : origin-destination survey, HTS : household travel survey

3.5 Discussion and conclusions

We are at a time where more and more data are collected on a day-to-day basis leading to interesting new opportunities for transport planning and modelling. These new data sources add up to an already rich tradition of data collection for travel demand estimation. In this research, we have studied empirically the difference between three independent data sources that provide information on public transit usage in Lyon. The results show that, overall, the structure of the demand described with OD matrices is similar in the three datasets. This is reassuring in the sense that different collection methodologies and processing methods generate a comparable measure of the same phenomenon. Nonetheless, when looking in detail in the context of Lyon there are discrepancies that may be of concern when trying to estimate the full public transit travel demand. These elements must be known and taken into account if these data are used to inform transportation policy.

Our results confirm that it is possible to construct an OD trip matrix with entry only automatic fare collection data. However, the trip database that results from the destination inference cannot always be used as it is to estimate travel demand. For instance, in the specific conditions that apply in Lyon, the processing of smart card transactions results in a trip database that may contain less than 50% of the true number of trips because of fare non-interaction, a limited smart card penetration rate and errors in the destination inference procedure. Thus, proper scaling needs to

be implemented to estimate the full demand, and this task is crucial [Gordon et al., 2018]. This process may require additional data sources and major assumptions that could produce distorted matrices. In this research, two methods were investigated and the empirical results show that they create distinct matrices. The first method requires no additional data sources and is based on the spatial distribution of the recorded transaction and on the assumption that the percentage of unrecorded trips is uniformly distributed. Compared to other data sources, this leads to a matrix that underestimates peripheral flow and exaggerates the importance of the subway network in shaping transit demand. The second method proposed by Gordon et al. [2018] is based on disaggregate automatic passenger counting. This method matches the flow inferred from smart card trips to the counted flow. Our empirical comparison suggests that the resulting matrix contains more trips in peripheral areas and more bus and tram trip legs where the fare non-interaction rate is higher. This matrix is also very similar to the large scale origin-destination survey. Thus, we believe that this method will produce a matrix that is more representative of the real flow on the network. In all cases, the estimation of the full matrices from automatic fare collection relies heavily on the inferred characteristic of smart card holders' trips. Despite all the methodological improvements, these inferred characteristics can still contain errors because in some cities passengers do not validate their ticket every time they board a vehicle, because some passengers may make only one trip a day, or because their trip chain may include other transport modes especially in dense areas where several forms of mobility are available. Furthermore, even if they account for an important proportion of the passengers, smart card holders may not be representative of the entire population. Therefore, an estimation of demand based only on their transit patterns could miss other practices. From a practical point of view, this means that smart card data is not error-free and representativeness issues should not be put aside because the initial sample is large. Whenever possible, it is recommended to combine these data with automatic passenger counting to make the best of those two passive and continuous data sources.

With regard to the household travel survey, our results show that in Lyon, compared to the other matrices this data source underestimates the number of public transit trips on an average weekday by about 30%. Various hypotheses could explain this difference. The first relates to the fact that only residents are surveyed, so trips made by non-residents are absent. However, the area covered by the household travel survey is much larger than the public transit network perimeter (see Figure 3.2). It is therefore hard to believe that people living outside this large area such as tourists or daily visitors arriving by train generate 30% of the daily trips made on the network. A more likely hypothesis is that people under-report their number of public transit trips. This tends to confirm previous studies that used GPS to assess the accuracy of household travel surveys at the individual level [Wolf et al., 2003, Wolf, 2006, Zmud and Wolf, 2003, Stopher et al., 2007]. This could also help explain the differences between AFC and HTS in terms of the number of trips per person and per day (see Figure 3.3). The last hypothesis is that there could be other issues in the sample due to selection bias or inaccurate weighting coefficients which mean that some of the population that uses the public transit system are not captured or are underestimated. In addition to this underestimation of trip volume, our results also show that trips reported in the household travel survey have certain distinctive

characteristics. First, they are temporally more concentrated in the morning and evening peak period. Secondly, they tend to be longer in terms of distance travelled and have a higher share of two-leg trips. Thirdly, the geographic distribution of trips has specific features. On the one hand, there is a higher share of trips in peripheral areas which also coincides with a higher percentage of bus trip legs. On the other hand, analysis of the matrix margins reveals that there are also more trips in the district of Lyon 3. This district is central, where the main railway station is located (Part-Dieu), and also has the highest job density. In the light of these considerations, it is possible that the HTS matrix may capture simple travel patterns such as work and home-based trips as suggested by Chapleau et al. [2018]. Taken together, these factors may lead to important biases in the resulting public transit OD trip matrix, even at a coarse level of aggregation. Therefore, several possibilities need to be explored to improve the ability of HTS to represent transit trips. First, an extension of the sampling frame to non-residents should be considered and a specific survey methodology needs to be developed to target sub-populations that might be missed when residents only are selected. Second, more effort should be made to reduce respondent burden and to ensure the completeness of trip reporting at the individual level, for instance, using a prompted recall method based on smart card transactions or smartphone GPS. Third, the expansion of travel survey should include more variables of interest (such as car availability, socio-occupational category or gender) to ensure a better synthesis of the known resident population. Finally, while this research shows that in practice in Lyon it would be more prudent to use other sources to create a comprehensive public transit trip matrix this does not mean that HTS will not still be needed in the foreseeable future. Household travel surveys will probably remain crucially important in the modelling process because they collect a large number of socio-economic and socio-demographic variables needed to calibrate behavioural choice models. Household travel surveys are also essential to study long term changes, to analyse travel behaviour with reference to all transportation modes and in relation to activities, household structure or car ownership.

Finally, this research also highlights some interesting opportunities in terms of data fusion and data collection cost optimization. OD surveys are part of the traditional toolkit of public transit operators. In Lyon, they have proved to be the most reliable source for all operations related to public transit planning. They have the advantage of covering all transit users (including fare evaders and non-residents). Unfortunately, they are costly, do not age well when there are substantial changes in supply and, because not all the interviews are done on the same day, the resulting demand refers to a composite time frame that does not capture the temporal variability of demand. Once processed and correctly scaled, AFC data might be a valuable complement or replacement for OD surveys. One promising possibility is that instead of surveying the different lines of the network continuously and regularly (every 5 years in Lyon) the agency would only conduct an OD survey when a specific need appears and the data inferred from AFC is deemed insufficient. These data collection strategies would use AFC data as the core component because it produces a continuous estimate of demand. OD survey costs would thus be reduced and OD surveys conducted in a "responsive manner" as opposed to the current "preventive approach". Household travel surveys could also be integrated within this approach, to provide further information on public

transit travel behaviour. To do this, it is necessary to identify or derive common features. If these features provide an unambiguous link such as the card number, an exact matching procedure can be implemented. Otherwise, features such as spatial or temporal attributes should be made compatible so that data fusion techniques could be implemented. One possibility would be to train statistical models using HTS that would then be applied to AFC to infer new information (e.g. as in Kusakabe and Asakura [2014]). Another possibility would be to apply statistical matching procedures and calibration methods to generate a more comprehensive and detailed synthetic population of public transit trips. Survey weights would then be continuously reviewed and readjusted based on AFC. We believe that ultimately the integration of multiple data sources will improve the quantity and quality of the available data and permit more accurate assessment of transport policies. However, achieving this will be a laborious process that will require a clear understanding of the limitations of each data source as well as a comprehensive analysis of comparability issues such as those presented in this paper.

There is no doubt that in the future practitioners and researchers will have access to a more heterogeneous corpus of data sources to estimate travel demand. To ensure the consistency of the knowledge extracted from these data sources, comparative studies will become increasingly important and challenging because of inconsistencies over various dimensions in the different data sources. This paper represents an initial contribution in this area, but we believe that more work is required. For instance, it would be interesting to test the influence of the data processing methodologies and parameters on the comparability of the matrices inferred from smart card data. More research could also be done to improve the scaling method, for instance, considering simultaneously the spatial distribution of paper ticket transactions and the counted flow. Finally, it would be interesting to enrich the pool of data by adding other passive sources such as mobile phone data, or do the same sort of analysis in other cities.

4

Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon.

This chapter is an edited version of the following article :

O.Egu and P.Bonnel (2020). Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon. *Travel Behaviour and Society*, 19, 112-123.

DOI : https ://doi.org/10.1016/j.tbs.2019.12.003

Highlights

— Travel behaviour variability has important implications in terms of modelling, policy evaluation or marketing.
— Smart card data allows us to undertake longitudinal analysis of day-to-day variability of transit usage.
— Combining clustering algorithm and day-to-day similarity metric is a valuable approach to efficiently mine smart card data.
— Very distinct level of intrapersonal variability can be found within each cluster and each fare profile.
— Findings can help in identifying new passenger segmentation and in tailoring information and services.

Chapter 4 : Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon.

Abstract

To examine the variability of travel behaviour over time, transportation researchers need to collect longitudinal data. The first studies around day-to-day variability of travel behaviour were based on surveys. Those studies have shown that there is considerable variation in individual travel behaviour. They have also discussed the implications of this variability in terms of modelling, policy evaluation or marketing. Recently, the multiplication of big data has led to an explosion in the number of studies about travel behaviour. This is because those new data sources collect lots of data, about lots of people over long periods. In the field of public transit, smart card data is one of those big data sources. They have been used by various authors to conduct longitudinal analyses of transit usage behaviour. However, researchers working with smart card data mostly rely on clustering techniques to measure variability, and they often use conceptual framework different from those of transportation researchers familiar with traditional data sources. In particular, there is no study based on smart card data that explicitly measure day-to-day intrapersonal variability of transit usage. Therefore, the purpose of this investigation is to address this gap. To do this, a clustering method and a similarity metric are combined to explore simultaneously interpersonal and intrapersonal variability of transit usage. The application is done with a rich dataset covering a 6 months period (181 days) and it contributes to the growing literature on smart card data. Results of this research confirm previous works based on survey data and show that there is no one size fits all approach to the problem of day-to-day variability of transit usage. They also prove that combining clustering algorithm with day-to-day intrapersonal similarity metric is a valuable tool to mine smart card data. The findings of this study can help in identifying new passenger segmentation and in tailoring information and services.

Keywords— Public transit, travel behaviour, smart card data, passenger clustering, day-to-day variability, user segmentation

4.1 Introduction

Travel behaviour research has been prominently based on cross-sectional data where individuals are asked to report their travel behaviour on a single day [Gärling and Axhausen, 2003, Pas, 1986, Hanson and Huff, 1986]. However, using one-day observation is in general insufficient because individual needs and desires that generate travel, vary from day-to-day [Pas, 1987] and because classifications based on a single day are prone to be unstable [Hanson and Huff, 1986]. Multiday data are therefore needed to refine the understanding of travel behaviour and measure how it may vary from day-to-day [Hanson and Huff, 1981, Schlich and Axhausen, 2003]. This is an important area of research that has several practical applications such as assessment of policy impact [Jones and Clarke, 1988], implementation of travel demand management strategies and individualised marketing [Gärling and Axhausen, 2003], travel behaviour modelling [Pas, 1986, 1987] or even market segmentation [Hanson and Huff, 1986].

To increase patronage, public transit operators also need to develop and evaluate new strategies. This requires a comprehensive understanding of transit usage behaviours, but also the ability to measure the multiday dynamics of individual demand [Morency et al., 2007, Briand et al., 2017, Ma et al., 2013, Bhaskar et al., 2015, Zhao et al., 2018]. In large urban areas, travel patterns are heterogeneous [Goulet-Langlois et al., 2016] and for decades transportation researchers have been exploring this heterogeneity with data of active solicitation [Schlich and Axhausen, 2003, Chen et al., 2016]. Unfortunately, active multiday data are costly, difficult to collect and often limited in terms of sample size [Goulet-Langlois et al., 2016, Briand et al., 2017, Gärling and Axhausen, 2003, Chen et al., 2016, Schlich and Axhausen, 2003]. Thanks to recent advances in technologies, it is now possible to collect continuously and passively massive data about mobility [Chen et al., 2016] with less or no burden for respondent [Bagchi and White, 2005]. In the field of public transit, smart card data is considered to be one of the most promising passive data sources [Pelletier et al., 2011]. As opposed to extrinsic mobility data such as mobile phone data, it is an intrinsic mobility data that is generated by travel events and therefore it provides direct information about transit usage [Zhao et al., 2018]. This data sources can be used to measure variability [Morency et al., 2007] and has resulted in a multiplication of studies on transit usage pattern. However, as noted by Chen et al. [2016] studies based on passive data sources often lack the long used conceptual framework of transportation researchers familiar with active data sources.

In this context, the purpose of this study is to investigate day-to-day transit usage variability using the conventional concept of daily trip pattern. The application is done with 6 months of smart card data. Two dimensions of variability are measured in parallel and interpreted with available socio-demographic profile derived from the type of fare used by each card. More precisely, day-to-day interpersonal variability is examined with a clustering method designed for this specific analysis. Day-to-day intrapersonal variability is measured with a trip based similarity metric [Huff and Hanson, 1986] taking into account the daily trip rate and the spatiotemporal characteristic of trip pattern. This paper shows that

combining interpersonal clustering with traditional intrapersonal similarity metric is a valuable approach to mine smart card data. It can extend our knowledge of day-to-day variability of transit usage. Results of this research can assist transit marketers in defining more meaningful passenger segmentation. They can also help in tailoring information and services.

The remainder of the paper comprises four sections organised as follows. The first part will review related works and clarify the research needs. The second part will describe the data and methods. The third part will present the main results of this research. Finally, in the fourth section, the empirical findings will be synthesised and future directions of research will be given.

4.2 Literature review

4.2.1 Studies based on active data collection

The first serious discussions about travel behaviour variability emerged during the 1980s. At that time, multiday data were difficult to collect due to high response burden [Schlich and Axhausen, 2003] and risk of deterioration of data quality over the survey period [Hanson and Huff, 1981]. In a series of work based on a 35 consecutive day travel survey known as the Uppsala survey, Huff and Hanson [1986], Hanson and Huff [1981, 1988] have found significant systematic intrapersonal variability and showed that daily travel behaviour is neither totally repetitious neither totally variable. With the same dataset, they have also identified five clusters of individuals based on multiday travel characteristics [Hanson and Huff, 1986]. Even if the five groups share distinctive travel behaviour and socio-demographic attributes, the authors noted that there is still substantial intragroup variance. They suggest that future classifications should recognize multiday patterns and recognize that individuals have more than one habitual daily travel pattern. Pas and Koppelman [1987] with a 5-day travel diary survey have examined the determinant of day-to-day trip rate variability and conclude that there are large intrapersonal variability and significant differences across socio-demographic groups. In a somehow related paper, Pas [1987] has investigated the effect of day-to-day variability on model goodness-of-fit. They divide the total sum of square of a standard least square trip generation models between interpersonal variability and intrapersonal variability and report that a substantial proportion of the total variability is due to intrapersonal variability. Jones and Clarke [1988] have discussed the importance of taking into account day-to-day variability from a policy perspective notably to assess the impact of measure that affects multiday behaviour. They proposed an activity-based measure of variability and noted that different measure can lead to different conclusions. Using the data from the mobidrive six-week travel diary, Schlich and Axhausen [2003] have compared various measures of day-to-day similarity. They noted that different measures may have different interpretations in terms of variability and conclude that day-to-day behaviour is more variable if measured with trip-based methods

than with activity-based methods. They also found that travel behaviour is more stable on weekdays and recommend that surveys about travel behaviour cover period of at least two weeks. More recently, Susilo and Axhausen [2014] have used Herfindahl-Hirschman index to examine the degree of repetition of daily-activity-travel-location. Their results indicate that constraint activities such as work or school have more repetitive combination than leisure and private business trips. With a seven-day travel diary collected in Belgium, Raux et al. [2016] have proposed different measurement methods to analyse interpersonal and intrapersonal variability within different periods of the week. Their results confirmed the overall picture that emerged from those studies : there is an important intrinsic variability in daily travel behaviour.

4.2.2 Studies based on smart card data

The first research on transit usage variability with smart card data was initiated in Canada. Agard et al. [2006] have used smart card data from the Société de Transport de L'Outaouais (STO) to cluster card users over 12 weeks based on the weekly temporal characteristic of trips. Change in cluster composition was then measured to explore intrapersonal variability. Morency et al. [2007] with 10 months of data from the STO have investigated separately the spatial and temporal variability of transit users. Their sample included only cards that were observed travelling at least once in the first and last month of their study period. The spatial variability is measure through the frequency of usage of bus stops and the temporal variability is evaluated with a clustering method based on boardings times. Ma et al. [2013] have investigated the regularity of trips pattern using 5 days of data from Beijing. For each card, the data is aggregated into four scalar features : number of travel days, number of similar first boarding times, number of similar route sequence and number of similar stop ID sequence. The K-means ++ algorithm is then used to identify cluster with different level of regularity. A rough set approach is also proposed to enhance the performance of the algorithm for large dataset. Bhaskar et al. [2015] have used four months of working days data from the transit authority of SEQ (Australia) to segment passenger based on the spatiotemporal variability of their trips. The proposed methodology starts with the application of DBSCAN to identify independently regular origin-destination (OD) trips and habitual trips starting time. Then, each passenger is described with the percentage of regular OD trips and the percentage of regular trips starting time. Using a priori rules, users are then segmented into four categories. For example, transit commuters are defined as those who make more than 50% of trips within habitual time and between regular OD. Briand et al. [2017] have proposed a Gaussian mixture model to cluster typical trips temporal pattern and measure the evolution of cluster composition over multiple years to assess change in passenger behaviour. They noticed some changes in the cluster composition but conclude that the majority of cards move to cluster with similar characteristics. Manley et al. [2018] have processed three months of data from London Oyster smart card to identified clusters of travel event for each individual with DBSCAN. A bottom-up approach is then used to derive a system-wide spatiotemporal understanding of regularity and irregularity. The analysis reveals that there are

more regularities in the origin of trips in the suburbs than in central London. Goulet-Langlois et al. [2016] have proposed a longitudinal representation of passenger activity based on the sequence of location (user area) infer for each card. Principal component analysis is then used to reduce the dimension of each sequence and serve as an input to cluster analysis. They apply this method on frequent users over 29 days of data from London and found 11 clusters with distinct sequence structure. The variability of each sequence is then measured with entropy rate to take into account the order of travel events and to detect individual with more variability [Goulet-Langlois et al., 2018]. Deschaintres et al. [2019] have focused on weekly variability of daily trip rate using smart card data from Montréal. Their sample includes cards observed with an amplitude of 12 months. Using the K-means clustering algorithm a week typology is created and each card is then represented as a sequence of week cluster. Those sequences are then used to cluster interpersonal variability and measure intrapersonal variability.

4.2.3 Research gap

The literature review shows that in studies based on smart card data researchers rely on clustering techniques to investigate and measure variability. To construct clustering variables, researchers often used scalar or vector aggregation of passenger's trips attributes [Goulet-Langlois et al., 2016] or generative model (e.g [Briand et al., 2017]). Clusters are then used : (1) to group passengers with similar travel behaviour i.e to study interpersonal variability, (2) to assess intrapersonal variability by studying cluster membership through time and (3) to identify events and sequence that are not regular. Those approaches are interesting but do not provide a metric of variability such as those proposed by authors using active data. Moreover, the clustering variables are often built aggregating more than one day of data. However, the conventional paradigm of travel research is based on the concept of daily trip pattern and many previous authors have argued that variability should be measured between days [Hanson and Huff, 1988, Pas and Koppelman, 1987, Schlich and Axhausen, 2003]. Finally, authors using smart card data often limit the scope of their analysis to a certain type of users (e.g frequent users), to a reduced temporal period (e.g a month), or to only one dimension of travel behaviour (e.g temporal pattern or spatial pattern).

For those reasons, we believe that the problem of measuring day-to-day public transit usage variability with smart card data has not been addressed completely. Therefore, there is a gap that needs to be filled. To achieve this objective, two flexible methods that can be applied to any type of users, any type of days, and any temporal period, are implemented. The first one is a clustering algorithm that allows to visualize and identify the most common day-to-day usage pattern and explore intrapersonal variability in a straightforward manner. The second one is a similarity index [Huff and Hanson, 1986] designed to measure day-to-day intrapersonal variability taking into account three fundamental features of daily trip pattern : space, time and trip rate. Our approach is based on the assumption that the day is the fundamental period for travel behaviour analysis, thus the intrapersonal day-to-day variability of transit usage needs to be explicitly measured. The application is done with a rich

dataset from the public transit networks of Lyon covering a 6 months period. Results are then cross-checked with the available fare profile to understand the potential determinant of day-to-day variability. To the best of our knowledge, no previous study based on smart card data has gone so far into the analysis of day-to-day variability. Furthermore, there is no study that combines clustering methods with day-to-day similarity measurement. This work contributes to the reconciliation of traditional methods with novel datasets and clustering techniques. Findings can help to better understand the dynamics of individual transit usage. They can also assist transit marketers and operators in defining meaningful passenger segmentation.

4.3 Materials and methods

This section starts by providing a brief description of the most important aspect of this case study. Then, the methods to measure variability are specified in detail.

4.3.1 Materials

Case study

TCL ("Transport en Commun Lyonnais") is the commercial name of the public transit network of Lyon. The network consists of 4 metro lines, 2 funicular lines, 5 tramway lines and more than 100 regular bus lines. On a working day, approximatively 1.2 million trips are done on the network. The fare transaction system of TCL is an entry only system. All transactions are anonymized and stored with boarding time and location. Smart card and magnetic paper tickets can be used by passengers. Cards cost 5€, they are strictly personal and require an identity photo of the owner. Thereafter, we will assume that there is an unambiguous relation between users, cards and individuals. The three terms will refer to a single person. We will also assume that there is no lost or stolen card. Finally, because paper tickets cannot be traced through time, the analysis is strictly restricted to cards (which represent 75% of the total number of fare transactions).

Fare profile determination

Smart card data often lack socio-demographic informations [Pelletier et al., 2011, Bagchi and White, 2005]. However, when there is a large spread of fare product, it is possible to define categories that make sense from a socio-demographic point of view. In Lyon, cardholders can access to a broad range of fares like annual pass, monthly pass, weekly pass but also access to standard single trip fare. They can also benefit from reduced prices if they can show proper justifications. To assign a socio-demographic profile to each card, we have aggregated fare product according to the table 4.1. During the lifetime of a card (5 years), users can purchase different

types of fare products. In this research, the profile of each card is defined as the one that accounts for the biggest proportion of fare transactions. With this ad-hoc method we were able to add one socio-demographic dimension to the data, however, it should be acknowledged that this information may not be available in other city or at least not as specific as in this case study.

Fare profile	**Fare type**	**Required justification**	**Pricing (€)**
Student	Monthly and annual pass	Proof of university enrolment and under 28 years old.	31.5
Young	Monthly and annual pass	Student up to high school (18 years old) that reside within the TCL network perimeter.	9-31.5
Elderly	Monthly and annual pass	More than 65 years old or more than 60 years old and retired.	9-31.5
Social	Monthly and annual pass	Dedicated to individuals that can justify low revenue such as unemployed people or people that benefit from state allowance.	9-31.5
General public	Monthly and annual pass	No justification needed can be half reimbursed by the employer of cardholder.	44.1-63.2
Short duration	Multi-day pass up to one week and single trip fare	May be available at a reduced price with justification (e.g. for young people, student or large family) .	1.7-19.3
Intermodality	Monthly and annual pass	Combine with other transportation mode such as rail, regional bus or public transit network from other city. May be available at a reduced price with justification (e.g. for young people or students).	47.5-205.8
Other	Monthly and annual pass	Mainly free pass for the public transportation operators agents and family or very specific passenger (blind people, policeman etc.)	0-6

TABLE 4.1 – Fare product classification into fare profile and corresponding prices for 2017 fiscal year, source : Authors

Trip identification

Trips are the building blocks of human travel behaviour. They are defined as a movement through time and space between two locations where activities are carried out [Bonnel, 2002]. In the smart card transaction database, records include boardings that are the beginning of a trip but also transfers. To identify the beginning of trips, we implement the following rules : (1) the first transaction of a day is always the beginning of a new trip, (2) two boardings transactions that occur within 60 minutes and that are not made on the same line or on metro station, are considered as part of the same trip [Munizaga et al., 2014b, Devillaine et al., 2012, Deschaintres et al., 2019]. The 60 minutes rule was defined according to the current fare policy that stipulates that a single ticket is valid up to 60 minutes from the previous validation. Sensibility test have shown that increasing the time threshold up to 120 minutes has no impact on the results.

Study period

In this research, data from January 1st 2017 to June 30th 2017 were extracted from the fare collection database. This study period was chosen for two reasons. First, because it consists of 181 days which we believe is enough to investigate day-to-day variability as there are at least 25 days of observations for each day of the week ; second, because this period includes two school holiday periods : winter break (from 2017-02-18 to 2017-03-05) and spring holidays (from 2017-04-15 to 2017-05-01) but also six bank holidays (2017-01-01, 2017-04-17, 2017-05-01, 2017-05-08, 2017-05-25, 2017-06-05). Those events can affect individual usage pattern and may be of interest

in terms of variability. Throughout this paper, a day will be referred to as a holiday day if it is a weekday that is within the two periods of school holidays or if it is a bank holiday. The rest of the weekdays will be considered as working days. Saturday and Sunday are considered separately. Note also that a day (or service day) is defined from 4.30 a.m to 4.30 a.m of the next day when the activity of the network is null but also when the majority of people are asleep.

4.3.2 Methods

Clustering interpersonal variability

Clustering analysis is one of the most common techniques of data mining [Friedman et al., 2001]. It aims to group objects that are similar in the same cluster which makes it a valuable strategy to study interpersonal variability and give more semantic to raw data. The three main steps in clustering are the definition of a vector space, the definition of a metric distance and the grouping of objects based on their similarity in the vector space.

At the most basic level, the day-to-day transit usage pattern of a card k could be described using a boolean vector $X_k = [x_1, ..., x_i, ..., x_n]$ where x_i takes value one when there is at least one trip on the day i otherwise, it takes value zero. This simplistic representation has three important advantages. First, it is very straightforward to compute the vector X_k for each card without any enrichment of the data. Second, to be meaningful this representation doesn't require a minimum number of transactions per card. Third, the binary vector allows us to focus on the revealed choice to use public transit on a given day without taking into account the characteristics of this usage that are voluntarily incorporated later on in the investigation.

Having defined a vector space, we need a measure of dissimilarity. When studying public transit usage, it is as important to know on which day passengers do use the system than on which day passengers do not use the system. Thus, in each vector, X_k zero and one carry equivalent information. Two vectors are to be considered close in the vector space when there is mutual presence or mutual absence. The simple matching distance (SMD) is a measure of dissimilarity that has this property, and it can be expressed as follow for two users k and l,

$$D(X_k, X_l) = 1 - \frac{f_{00} + f_{11}}{f_{00} + f_{11} + f_{01} + f_{10}} \tag{4.1}$$

where :

f_{00} = number of days where X_k is 0 and X_l is 0
f_{11} = number of days where X_k is 1 and X_l is 1
f_{01} = number of days where X_k is 0 and X_l is 1
f_{10} = number of days where X_k is 1 and X_l is 0

With the above dissimilarity measure, we can calculate a dissimilarity matrix M.

In this matrix, each element M_{kl} corresponds to $D(X_k, X_l)$. This matrice is then used as an input for the clustering algorithm. Hierarchical clustering is a common approach that does not require that we commit to a particular number of clusters [Friedman et al., 2001]. It is very popular because it produces a dendrogram that illustrates how the objects are joined together. After some test, we decided to use an agglomerative approach with the Ward method [Ward Jr, 1963]. This method uses a criterion for choosing the pair of clusters to merge at each step by minimizing the change in the total sum of squares. It can be implemented recursively by a Lance Williams algorithms. As opposed to K-means, this method can be applied to dissimilarities measures that are not strictly Euclidean such as the SMD. This method also tends to produce compact clusters of approximate size.

To quantifies statistically the strength of the association between a given cluster and a given fare profile, we can use the odd ratio (OR) [Goulet-Langlois et al., 2018]. An OR bigger than 1 indicates a positive association and vice versa. For a sample of the population, OR can be estimated as follows :

$$\widehat{OR_{a,b}} = \frac{N_{a,b}.N_{a',b'}}{N_{a',b}.N_{a,b'}} \tag{4.2}$$

In the above formula, a refer to a single fare profile, b refer to a single cluster, a' refer to the aggregation of all clusters except a, b' refer to the aggregation of all fare profile except b, $N_{a,b}$ denotes the number of individuals with characteristic a and b. The log of the OR is normally distributed and can be used to statistically test whether an OR is significantly different from 1 at a given confidence level [Goulet-Langlois et al., 2018, Morris and Gardner, 1988].

Measuring intrapersonal variability

The previous method does not incorporate any information regarding how each passenger uses the network. Therefore, it does not address the question of how similar are each day for a given individual. From the perspective of transit usage, each day can be described in terms of trip rate but also considering the spatiotemporal characteristic of trips. More precisely, two days can be considered similar if they have the same number of trips and if trips share time and space attributes. Huff and Hanson [1986] have proposed a trip based similarity measure between two days i and j that can measure conjointly those aspects and can be expressed as follows,

$$S_{ij} = [1 - \frac{1}{2}\sum_k |P_{ic} - P_{jc}|]\frac{n_i}{n_j}, \quad nj \geq ni \tag{4.3}$$

where P_{ic} is the proportion of trips in days i that have the characteristic of the equivalence class c and n_i is the number of trips on day i. This measure of similarity ranges from 0 to 1. Two days having the same number of trips and identical trip pattern regarding the equivalent class c will result in a similarity of 1.

To define an equivalent class, there are several options because trips can be described with many attributes such as purpose, distance, mode, time of departure etc. With smart card data, the number of combinations is often reduced because

not all attributes are available directly from the transaction database. In Lyon, two attributes are naturally available : transaction time and boarding stop. Since we now have identified the transactions that correspond to the beginning of trips, we can use the spatial and temporal features of those transactions to define the equivalent class. Both features have high cardinality as there are many stops in the network (more than 4000) and the timestamp is known with second precision. To reduce the dimension of those features, we have decided to use two grids that make sense both from a practical and behavioural point of view :

— **Temporal grid**. Trip starting time are grouped into the following time slot : before 7 a.m, 7 a.m to 10 a.m, 10 a.m to 12 p.m, 12 a.m to 2 p.m, 2 p.m to 4 p.m, 4 p.m to 8 p.m and after 8 p.m.

— **Spatial grid**. Trip origin stops are aggregated at the district level in the city of Lyon where the network is denser and at the communal level in the peripheral areas of the urban transit perimeters. This spatial aggregation is made up of 82 zones depicted in figure 4.1.

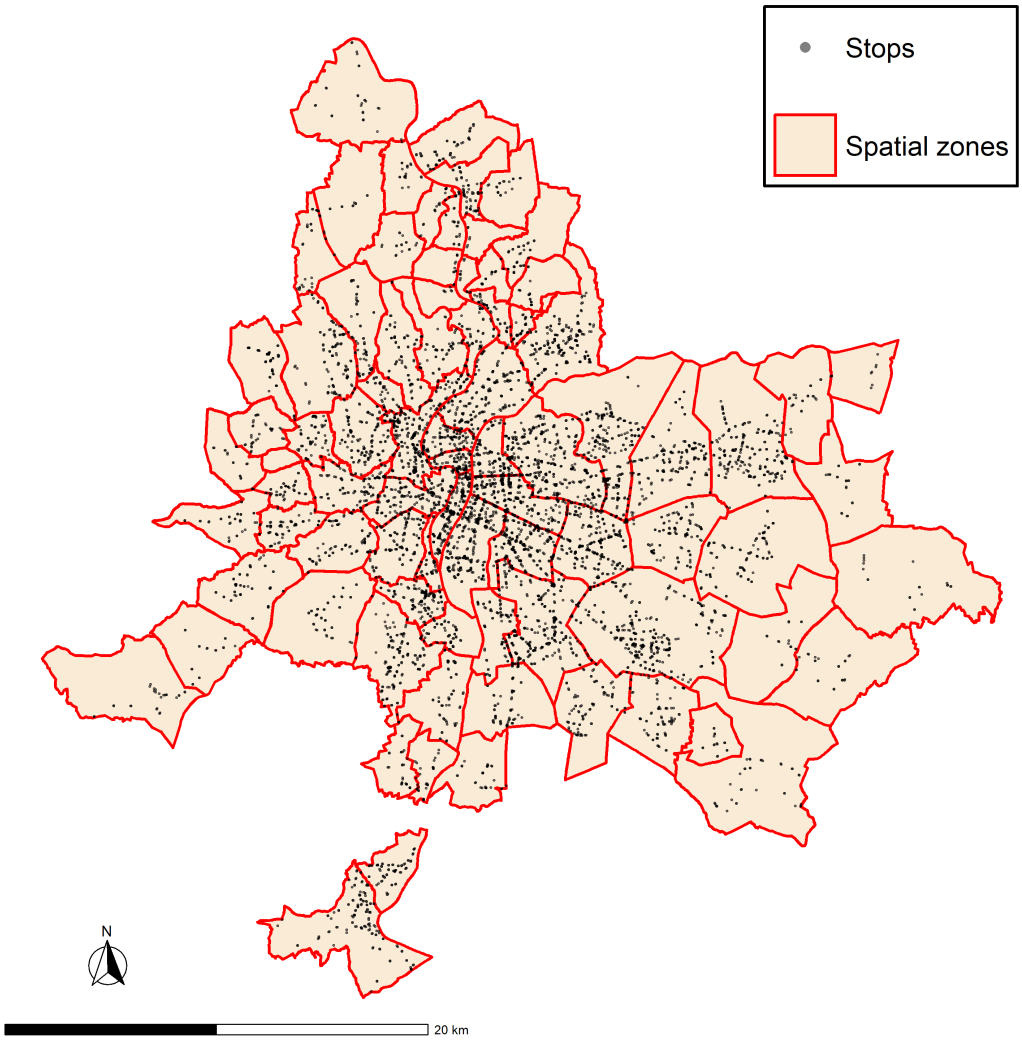


FIGURE 4.1 – Study area and aggregation of stop into a spatial grid, source : Authors

The equivalence class c are then built using a contingency table between the spatial zone and the time slot. Two trips are in the same equivalence class if they share both attributes. Each day i is then synthesised in the vector P_{ic} that transcribes

the spatiotemporal distribution of trips of day i. Therefore, S_{ij} will be equal to one if, on two distinct days, a card makes exactly the same number of trips from the same spatial zone and in the same time slot. This way, three dimensions of daily trip pattern are considered : space, time and trip rate. Moreover, to infer trip destination in entry only smart card system, it is often assumed that the destination is close to the origin of the next trip (trip chaining model). Hence, considering only the origin of trips may provide a sufficient representation of the daily trip pattern and it does meet the goal of this study.

4.4 Results

4.4.1 Data driven sample selection

To start this study with a holistic approach, a random sample of 40,000 cards among the 591,124 cards observed travelling at least once between January and June 2017, was drawn. The clustering method was then applied to a random subset of 10,000 cards to visualize day-to-day usage pattern and select a sample in a data-driven way. The results are represented on a heatmap in figure 4.2. Each row corresponds to a card, each column to a day and each cell can be either black (at least one trip) or white (no trip). Week numbers are indicated in the x-axis. The resulting dendrogram is shown on the left of the heatmap. Figure 4.2 demonstrates that even at the most basic level of days of usage, the interpersonal variability is considerable with a large diversity of pattern. Some rows are entirely white which indicate that some cardholders rarely use the transit system. Weekends generate a strong and repetitive vertical white pattern that affects a large proportion of users. Nonetheless, some rows are almost entirely black i.e some individuals use the transit system almost every day. The two holiday period in weeks 8-9 and weeks 16-17 are also visible and can lead to episodic break of usage. Lastly, figure 4.2 reveals that some users exhibit clear changes in usage intensity over the six months.

A simple interpretation of this dendogram could be to classify users in three main groups :

1. **The low frequency users (LF)**, mainly located in the middle of the dendrogram. They almost never use the system on a multimonth scale. Our hypothesis is that public transit is something that is not part of their daily routine. Those people may actually use other transportation modes, or be present in the city only during a short period of time such as tourists visiting the city.
2. **The consistent transit users (CT)**, mainly located at the top of the dendrogram. Those are individuals that used the transit system consistently over the 6 months period. They may be subject to ruptures such as holiday or weekend and may not use the transit system every day but they will not stop using the system over a long period of time. For those users, we can assume that transit usage was part of their daily routine from the beginning to the end of the study period.

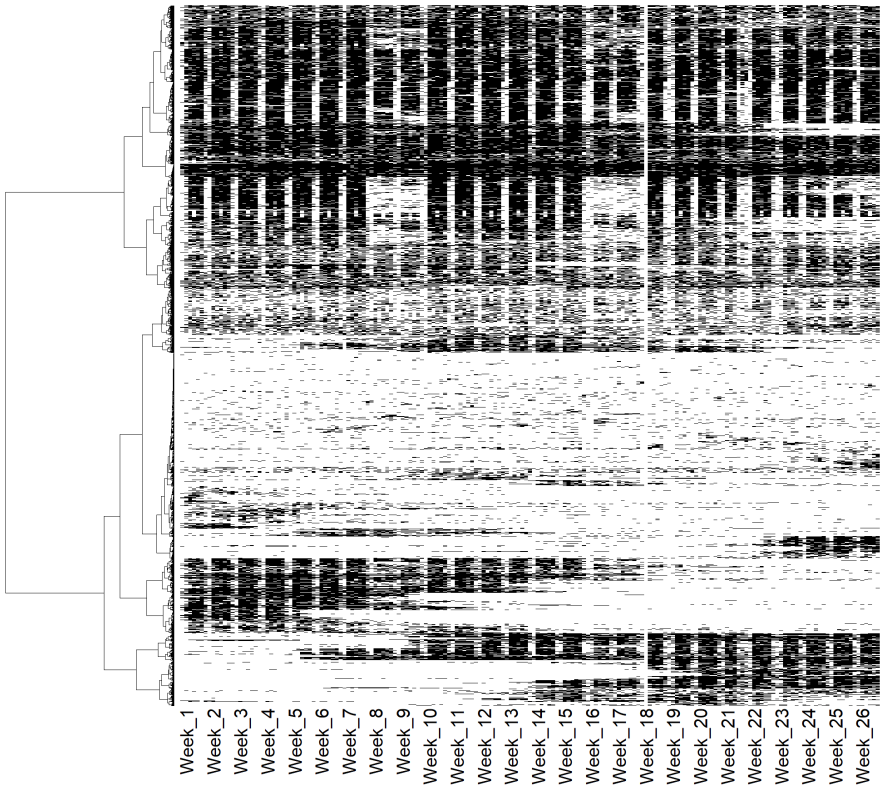


FIGURE 4.2 – Dendogram resulting from the application of the clustering method to 10,000 randomly selected cards, source : Authors

3. **The intermittent transit users (IT)**, mainly located at the bottom of the dendrogram. Those are individuals that present characteristics of low frequency users but also characteristics of consistent transit users. For those users, transit usage was part of their routine but at one point of the study period, they exhibited a marked change in day-to-day usage intensity.

To discriminate between those three groups, two attributes were computed for each card k. N_k being the number of travel days and M_k the maximum number of consecutive days without transit usage. The distribution of both variables is given in figure 4.3. N_k is spread almost uniformly between 10 and 130 meaning that few cards use the system more than 140 days out of the 181 days of the study period. There is also a concentration of cards around small values of N_k i.e cards that are observed travelling only a few days. The distribution of M_k is characterized by peaks at each multiple of 30 correspondings to usage during only a subset of months. The distribution of M_k also presents a concentration of cards between 0 and 30 meaning that a high proportion of cards will not stop using the transit system for more than 30 consecutive days. To classify users into the three proposed groups, the following rules are implemented : (1) a user is considered as LF if the number of travel days is less or equal to 10 days i.e $N_k \leq 10$, (2) a user is considered as IT if the number of travel days is bigger than 10 but there is a usage interruption of more than 30 continuous days i.e $N_k > 10$ and $M_k \geq 30$, (3)

otherwise users are classified as CT.

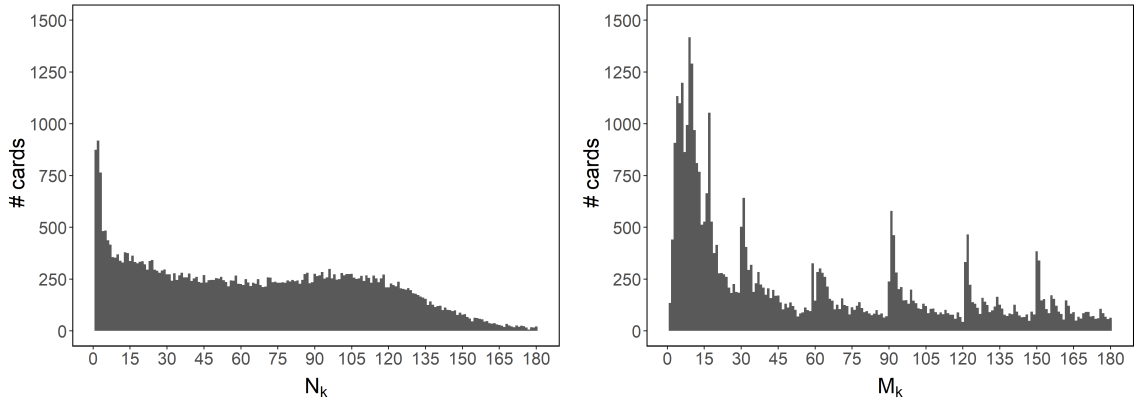


FIGURE 4.3 – Distribution of N_k and M_k, source : Authors

We have applied those rules to the 40,000 cards of our initial sample. The number of cards and the number of trips by groups are given in table 4.2. The smallest group corresponds to the LF users with a total of 5456 cards (14% of the cards), but less than 1% of the total number of trips. 16,358 users are classified as IT (41% of the cards) and account for 30% of the trips. The CT users form the biggest group with 18,186 cards which correspond to approximatively 45% of the users and account for almost 70% of the trips.

	# users	**% users**	**# trips**	**% trips**
Consistent users (CT)	18,186	45%	4,220,965	69%
Intermittent users (IT)	16,358	41%	1,840,412	30%
Low frequency users (LF)	5,456	14%	50,316	1%

TABLE 4.2 – Distribution of users and trips by group, source : Authors

To start this investigation, we have analysed the daily usage pattern of 40,000 cards selected randomly. In doing so, we were able to show the diversity of usage pattern on a multimonth scale. Based on the proposed clustering method, we have defined three groups of users with distinct multimonth frequencies of usage. The rest of this study will focus on the behaviour of the 18,186 consistent transit users. This sample selection is justified by the fact that they account for the biggest proportion of trips done on the network. It also allows us to maximise the observation period and focus on a group that share common characteristics in terms of multimonth usage routine. While this can be seen as a limit for the rest of this study, the two methods could be applied in the same way to the rest of users (as long as there is more than one day with travel events). The next section of this paper examines day-to-day regularity and intrapersonal variability at an aggregate level.

4.4.2 Aggregated analysis of intrapersonal variability

As a first step, we may be interested in evaluating the regularity in an aggregate way for each chosen dimension of variability. For example, considering the spatial dimension, for each user, we can rank the spatial zone based on the number of trips they generate and calculate for all users the average share by rank. The same can be done with the time slot, combining both dimensions (i.e using the concept of equivalent class) and also for the daily trip number. The results of those calculations are given in figure 4.4.

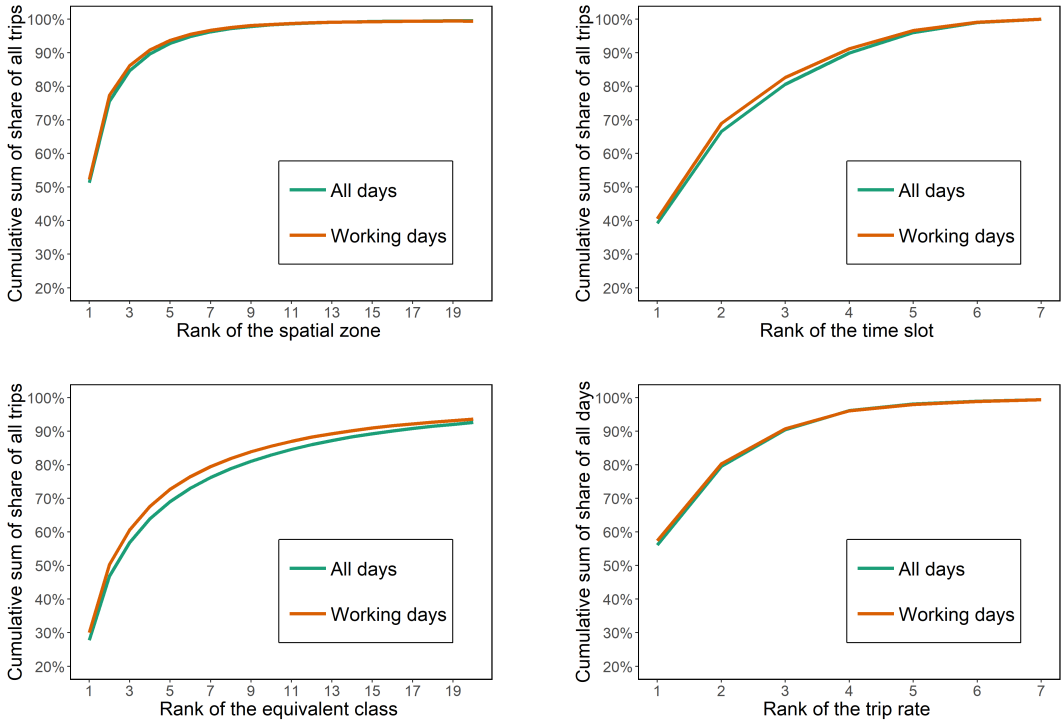


FIGURE 4.4 – Regularity of transit usage on each chosen dimension of variability, source : Authors

In average, the two most important spatial zones generate 76% of the trips and the two most important time slots generate 67% of the trips. This indicates that overall there is a high degree of spatial and temporal regularity which is in line with previous work [Schlich and Axhausen, 2003, Hanson and Huff, 1988, Morency et al., 2007]. When combining spatial and temporal dimension, the most important equivalent class generates on average 27% of the trips which is close to the 30% obtained by Huff and Hanson [1986] for an equivalent class defined with time slots and city quadrant. The five most important equivalent class generate approximatively 65% of the trips. As indicated in figure 4.4, the concentration of trips in a few equivalent class is more important during workings days. However, the difference between each curve remains thin. For instance, on working days the two most important equivalent class will generate 49% of the trips but this number only decreases to 46% when considering all days. The bottom left plot indicates that on average, 55% of the days the daily number of trips will correspond to the most frequent trip rate. In other words, the most recurrent trip rate cover on average a bit more than half of the

travel days. Again, there is no important difference when we focus on working days. This first analysis demonstrates that on average trips are repetitive in the sense that they share spatial and temporal characteristics, and that days are repetitive in the sense that on average users will make the same number of trips. However, it does not mean that all days are similar for each user, and that all users have the same level of variability. To measure those two aspects, S was calculated in a totally desegregate way i.e for each individual and each pair of travel days.

A first aggregation of S_{ij} could be to calculate for all users the mean similarity between any two days of the week. The results are given in table 4.3 and give rise to the following comments. First, we confirm the finding of Schlich and Axhausen [2003] that weekend days are less similar than other days of the week. The mean similarity within Saturday and within Sunday is equal to 0.18. The similarity of Saturday with other weekdays decrease to 0.11 and the similarity of Sunday with other weekdays decrease to 0.08. Second, even if weekdays are more similar within each other than with weekend days, there are more similarity within the same weekdays than between distinct weekdays. For instance, the similarity within Monday is equal to 0.33 but decrease to 0.27 when we compare Monday with Friday. Third, as found by Schlich and Axhausen [2003], Friday is the weekday that exhibits less similarity with the rest of the weekdays. Fourth, Tuesday is the weekday where the within-day similarity is the highest with a value of 0.36. This indicates that there is less intrapersonal variability and thus users have a higher tendency to repeat the same trip pattern every Tuesday than during other days of the week.

	Mon.	**Tue.**	**Wed.**	**Thu.**	**Fri.**	**Sat.**	**Sun.**
Mon.	0.33	0.32	0.28	0.29	0.27	0.10	0.08
Tue.		0.36	0.29	0.31	0.28	0.11	0.08
Wed.			0.33	0.27	0.25	0.11	0.08
Thu.				0.34	0.28	0.11	0.08
Fri.					0.31	0.11	0.08
Sat.						0.18	0.10
Sun.							0.18

TABLE 4.3 – Mean similarity between days of the week, source : Authors

A second analysis will be to plot the distribution of similarity among users. For each user, we compute the mean similarity for all pair of days $\overline{S}$ and the mean similarity focusing only on pair of workings days $\overline{S_{wd}}$. The distribution among users of both variables is given in figure 4.5. The median of $\overline{S}$ is equal to 0.18 and only increase to 0.22 for $\overline{S_{wd}}$ meaning that most users will have a high degree of day-to-day variability even when we focus only on working days. Both distributions are also very skewed toward higher values of similarity. In other words, there are large differences between users and some users exhibit lower levels of day-to-day variability than others. Another way to look at this problem could be to determine for each user and each travel day, the number of other travel days with the same daily trip pattern ($S_{ij} = 1$). We found that on average 62% of the daily trip pattern will reoccur at least once in the study period. We also looked for each user at the number of days where the daily trip pattern was completely unique (for i, $S_{ij} = 0 \; \forall$

j). We found that more than 96% of the cards will not have one completely unique daily pattern. In other words, while there may be in average around 40% of user-day pattern that will not reoccur in the longitudinal records, there are very few users that have a travel day that does not share any attributes with other travel days in the longitudinal record.

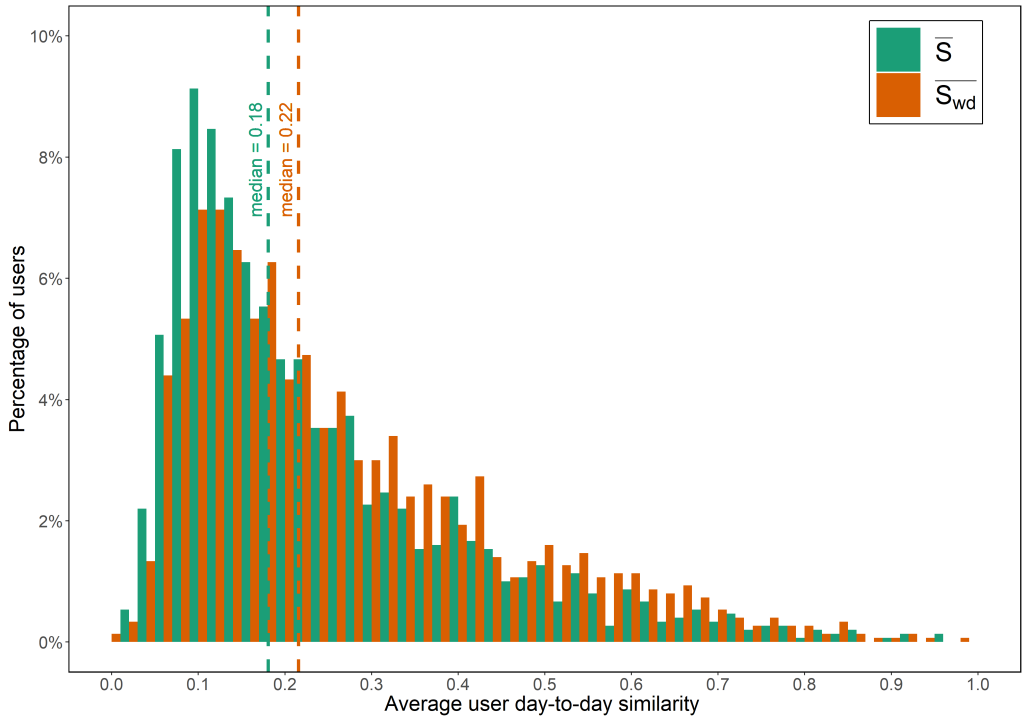


FIGURE 4.5 – Distribution of average users day-to-day similarity ($\overline{S}$: for all pair of days, $\overline{S_{wd}}$: only for working days), source : Authors

Those aggregated results support the idea of Hanson and Huff [1988] that there is a high degree of regularity in transit pattern, but there is also systematic day-to-day variability. Thus, to characterize transit usage a single day is insufficient. Moreover, if we aggregate trips characteristics over periods longer than a day, we will eclipse the daily variability of transit usage. Finally, as indicated by the large skew in the distribution of $\overline{S}$, there are reasons to believe that some users are less variable than others. The next section shows how the combination of intrapersonal variability measurement, interpersonal clustering and fare profile can offer deeper insights into day-to-day transit usage variability.

4.4.3 Combining clustering, intrapersonal variability and fare profile

The clustering method was applied to the 18,186 consistent transit users. With the help of the dendrogram, we have decided to retain 6 clusters (numbered from C1 to C6) which we believe is a good balance between the quality of the clusters and the interpretativeness of the results. To visualize the pattern of each cluster, 100 cards

were selected randomly in each cluster and plot according to the convention of figure 4.2. The graphical results are given in figure 4.6. To further help the interpretation, descriptive statistics are given for each cluster and each dimension of variability in table 4.4.

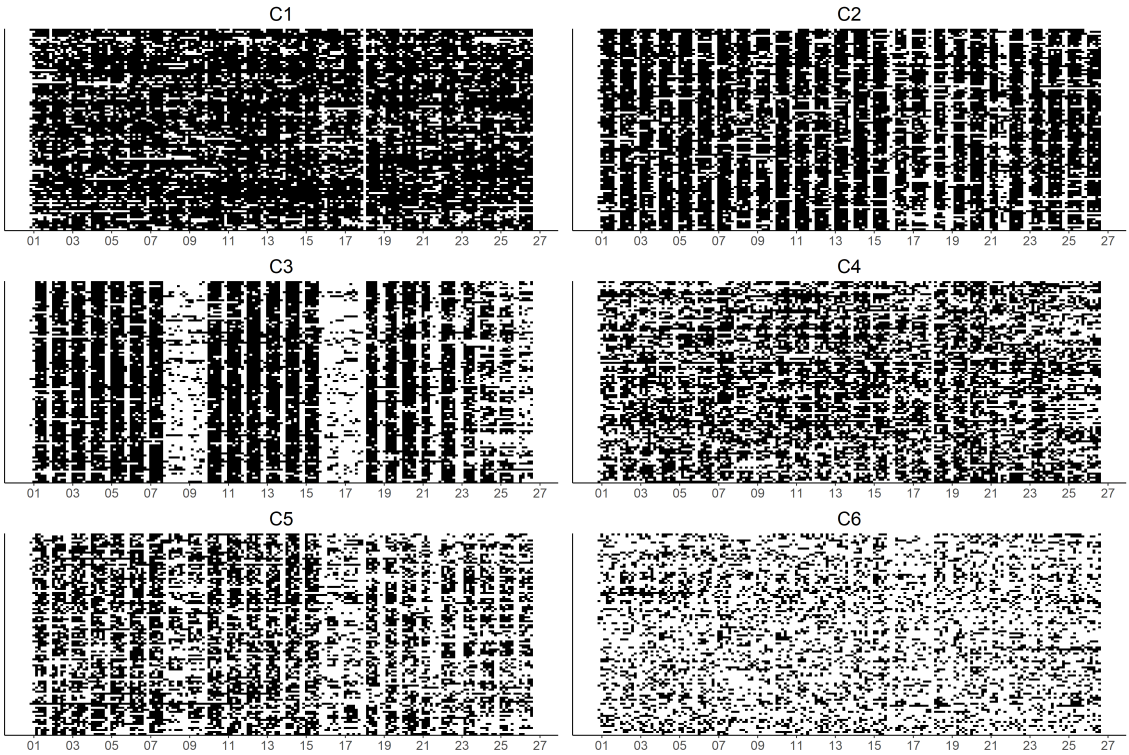


FIGURE 4.6 – Visualization of the day-to-day usage pattern of 100 random users selected from each cluster, source : Authors

In table 4.4, it can be seen that cluster 1 is characterized by the highest percentage of days of usage and almost no calendar structure as the usage rate remain quite high on holidays and weekends. Cards in cluster 1 will in average use the transit system more than one Sunday out of two. On the day they use the system, they tend to make more trips as indicated by the mean number of trips per day (2.7), but also by a high proportion of user-day with more than three trips (24%). Cluster 2 is the biggest cluster in terms of size with 4592 cards (25% of the CT users). It is characterized by a very high usage rate on workings days, few travel days during the weekend, a large proportion of travel days with two trips and a high percentage of trips in the morning and evening peak period (31% and 35%). Like cluster 2, cluster 3 exhibits a concentration of trips in the two peak period but their day-to-day usage pattern differs. In fact, in figure 4.6, it is easy to see that users in cluster 3 almost do not use the transit system during the holidays period. On this same figure, it is also possible to notice a decrease in usage at the end of June just before the summer period. Cluster 4 does not exhibit such a clear calendar structure. Cards in this cluster use the transit system on working days a bit more than 3 days out of 5 (64%) and remain largely present during the holiday (51%), Saturday (51%) and Sunday (30%). The temporal trips distribution of cluster 4 is somehow related to the one of cluster 1, with no pronounced concentration of trips in the morning peak period. Cluster 5 can be seen as an intermediate between cluster 3 and 4. As

	C1	C2	C3	C4	C5	C6
# users	3,211	4,592	2,442	2,624	2,235	3,082
% users	18	25	13	14	12	17
Day of usage						
Working days (%)	86	87	84	64	62	31
Holyday days (%)	75	63	20	51	33	23
Saturday (%)	72	31	22	51	18	23
Sunday (%)	56	13	10	28	9	12
Trip rate						
Mean trip per travel day	2.7	2.3	2.2	2.2	2.0	1.8
One trip (% user-day)	16	16	25	27	29	39
Two trips (% user-day)	41	59	52	47	52	47
Three trips (% user-day)	19	14	13	14	12	10
Four trips or more (% user-day)	24	14	14	13	10	7
Trip temporal distribution (%)						
before 7 a.m	4	4	2	3	3	2
7a.m to 10 a.m	19	31	32	19	28	17
10 a.m to 12 p.m	11	7	7	12	8	13
12 a.m to 2 p.m	13	11	15	13	12	14
2 p.m to 4 p.m	12	8	11	13	10	16
4 p.m to 8 p.m	31	35	30	32	34	32
after 8 p.m	10	5	3	9	5	7
Mean spatial indicators						
# distinct spatial zone	13.7	10.4	9.5	11.6	9.3	8.7
% of trips in the two most frequent spatial zone	71	79	81	72	78	73
Mean users similarity measurement						
All days	0.16	0.34	0.25	0.17	0.26	0.16
Working days only	0.19	0.38	0.30	0.20	0.29	0.18

TABLE 4.4 – Descriptive statistics for each cluster, source : Authors

in cluster 3, transit usage is impacted by holiday and weekend but as in cluster 4, the usage rate on working days is under 65%. The last cluster (C6), is formed by vectors that are very sparse i.e with lots of zero. In this cluster, we found users that despite the fact that they consistently use the system over the 6 months, their usage rate is less than 30% and doesn't vary much within the type of day. Cluster 6 is also characterized by a lower trip rate and a very high percentage of one trip day (39%). Finally, table 4.4 indicates that the spatial diversity of trips vary between clusters. The concentration of trips inside the two most frequent zone is higher for cluster 2, 3, and 5.

To extend this analysis, the contingency table between fare profile and clusters is computed. The results are given in table 4.5 with the corresponding odd ratio. The distribution of mean user day-to-day similarity on working days ($\overline{S_{wd}}$) for each cell of the contingency table is also given as a boxplot in figure 4.7.

		C1	**C2**	**C3**	**C4**	**C5**	**C6**	**Total**
Elderly	# users	317	135	11	375	130	549	1,517
	OR	**1.26**	0.27	0.04	**2.11**	0.65	**3.17**	
General public	# users	1,175	2,747	235	700	684	518	6,059
	OR	**1.19**	**4.62**	0.18	0.69	0.87	0.35	
Intermodality	# users	28	370	74	46	205	108	831
	OR	0.16	**2.5**	0.62	0.34	**2.47**	0.72	
Other	# users	59	77	25	90	60	199	510
	OR	0.6	0.52	0.33	**1.28**	0.95	**3.28**	
Short duration	# users	2	11	2	30	39	366	450
	OR	0.02	0.07	0.03	0.42	0.67	**24.1**	
Social	# users	815	315	63	425	149	365	2,132
	OR	**3.53**	0.48	0.18	**1.57**	0.5	1.01	
Student	# users	616	609	326	618	518	299	2,986
	OR	**1.26**	0.72	0.76	**1.72**	**1.65**	0.5	
Young	# users	199	328	1,706	340	450	678	3,701
	OR	0.22	0.23	**15.97**	0.54	0.98	1.13	
Total		3,211	4,592	2,442	2,624	2,235	3,082	18,186

TABLE 4.5 – Contingency table between clusters and fare profiles and associated Odd Ratio, bold indicate superior to 1 and statistically different from 1 at 99% confidence level, source : Authors

General public fare profile is mainly aimed at people in employment and thus work-related trip will probably shape their transit usage. Table 4.2 indicates that general public users are strongly associated with cluster 2 (OR of 4,62) and to a lesser extent with cluster 1 (OR of 1.19). Cluster 2 exhibits a usage pattern that seems to be more work-oriented than cluster 1 where users probably make more diverse use of the network. In figure 4.7 it can be seen that the median of $\overline{S_{wd}}$ for general users in cluster 2 is 0.43, but it decreases to 0.21 for general users in cluster 1 which confirms that general public users from cluster 1 tend to be more variable than those of cluster 2. Holders of intermodality pass combined public transit with other services such as trains which can constrain their usage patterns. Table 4.2 indicates

that they are mostly found in cluster 2 and 5 where they tend to show a high level of day-to-day similarity. 13% of intermodality profile are also assigned to cluster 6 where their usage of the transit system is on average more variable with a median of $\overline{S_{wd}}$ equal to 0.19 compared to 0.49 and 0.43 for intermodality users in cluster C2 and C5. As anticipated, young people dominate cluster 3 and rarely use the transit system during holidays. Young users in cluster 3 present a mean similarity between working days that is relatively high with a median equal to 0.28. They are also found in cluster 6 where their usage is less intense and with lower day-to-day intrapersonal similarity. People that are using social pass, exhibit a high level of day-to-day variability in their transit usage. Table 4.2 indicates that there is a positive and significant association between cluster 1 and social users (OR of 3.53). As pointed before, cluster 1 presents the highest intensity of usage both in terms of number of travel days and number of trips per day. Thus, users in cluster 1, may cover an important proportion of their urban travel needs with public transit. Table 4.2 also indicates that a non-marginal part of social users is assigned to cluster 4 and 6, so we can not consider all social users as very intense users. Like social users, elderly users are mostly found in clusters that do not exhibit clear calendar structure such as cluster 1, 4 and 6. In figure 4.7, it can be seen that the median of $\overline{S_{wd}}$ for elderly users in those three clusters is between 0.14 and 0.17, almost three times less than general public users of cluster 2. Student users is an interesting population because university constraints are very heterogeneous both spatially and temporally. Table 4.2 shows that this population is almost equally spread within the six clusters. The level of day-to-day similarity of students is in general low, but it can vary between clusters. For instance, students in cluster 2 and 3 have a median similarity of respectively 0.22 and 0.2 compared to median similarity below 0.15 in other clusters (see figure 4.7). Finally, as expected, short duration users are almost all assigned to cluster 6 where they exhibit a high level of variability with a median of $\overline{S_{wd}}$ equal to 0.13. Those are users who use the network from time to time with varying spatiotemporal patterns.

Our empirical results demonstrate that we can find tangible links between the intrapersonal day-to-day variability, the multimonth pattern of usage synthesise by cluster membership and sociodemographic inferred from fare profile. Those three elements are essential to carry such a fine-grained investigation of day-to-day transit usage behaviour. They can extend our knowledge of transit usage variability and are useful to define interpretable user segmentation. This is what we have done in this section. The following section takes a step back from the strict description of numerical results to discuss the lessons to be learned from this investigation.

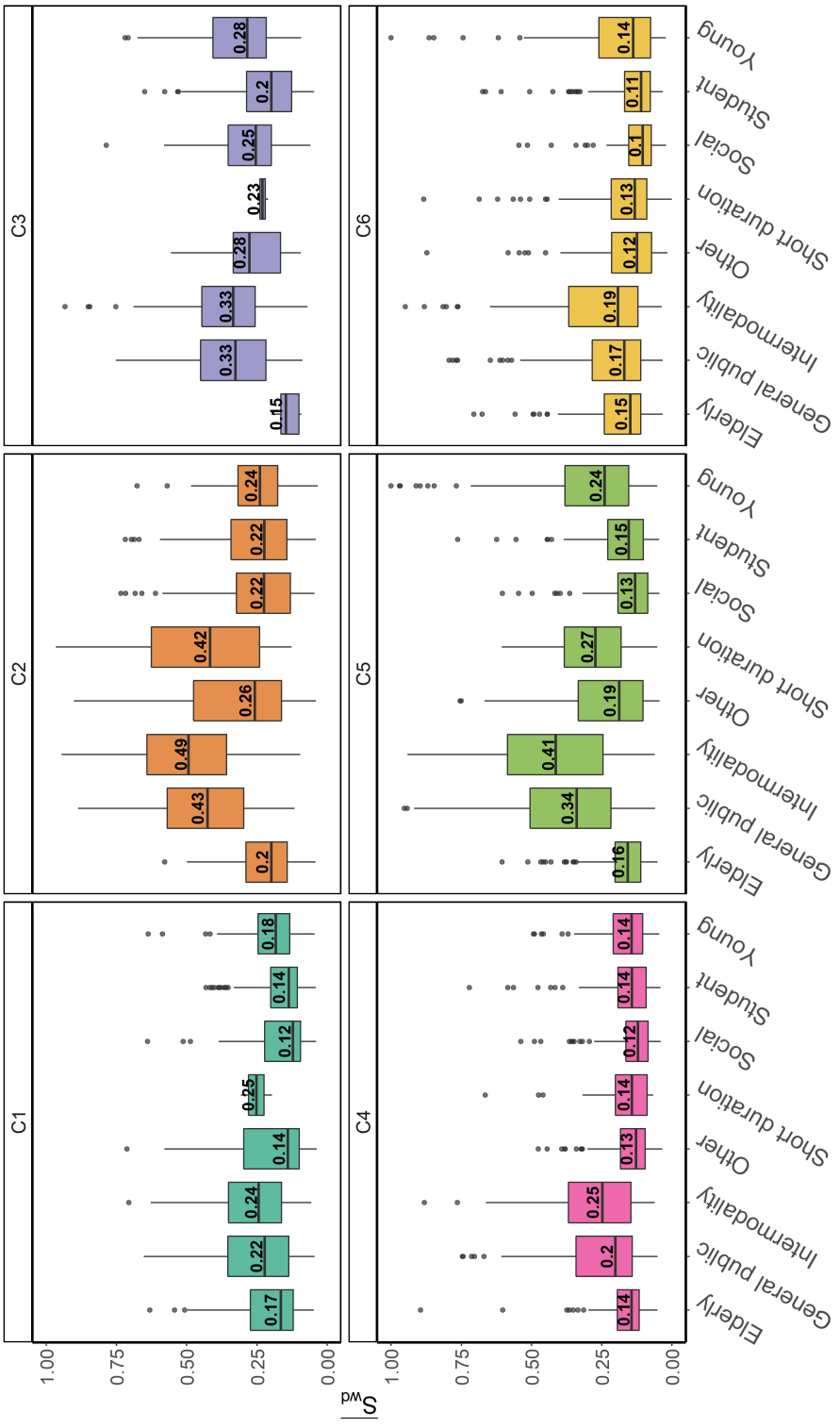


FIGURE 4.7 – Boxplot of users working days mean similarity ($\overline{S_{wd}}$) by cluster and fare profile, source : Authors

4.5 Discussion and conclusions

Researcher and transportation planners have a strong interest in understanding the day-to-day variability of transit users. Smart card data give us the opportunity to undertake this analysis over longer periods and provide a deeper understanding than the current state of the art [Manley et al., 2018]. As opposed to studies based on travel surveys, most of the research on smart card data have relied on clustering techniques to analyse variability. In this paper, we complement this approach with a traditional day-to-day similarity measure [Hanson and Huff, 1988] and prove that the combination of both technics is a valuable tool to mine smart card data. It can be used to expand traditional research on travel behaviour variability using similar paradigm and concepts.

Our empirical finding suggests that there is no "one size fits all" approach to the problem of day-to-day variability of transit usage. The simple view of transit users as solely commuting passengers from Monday to Friday is inadequate. Likewise, the classification of transit users solely based on their fare product or one-day data is incomplete. Very distinct level of intrapersonal variability can be found within each fare profile and within each usage pattern synthesised with cluster membership.

By selecting randomly our initial sample among all cards, we were able to visualize the diversity of day-to-day transit usage pattern and we have defined three main groups of users. The high proportion of intermittent transit users indicate that there is porosity between high frequency usage and low frequency usage. In other words, it is not uncommon that public transit usage habits and routines change drastically over time. This confirms the old idea that the apparent stability at the aggregate level is in reality compensated by changes at the individual level [Jones and Clarke, 1988]. Although it is not possible to understand the leading cause of ruptures using only smart card data, we believe that public transit operators should show more interest in those users and those changes if they want to influence travel behaviour and increase loyalty. This requires that they move from a purely accounting approach such as the number of pass sold per month to a customer-centric approach where each individual pattern is mined. This approach can offer a large range of new opportunities. For instance, operators could track individual pattern to identify "unsuccessful" new users, to design and evaluate new fare structure, to define more targeted marketing actions or to detect commercial opportunities.

Our results confirm that over long periods, users show spatial and temporal repeatability in their trip pattern but there is also a systematic intrapersonal variability in day-to-day transit usage [Schlich and Axhausen, 2003, Hanson and Huff, 1988]. In large urban areas such as the metropolis of Lyon, people have access to a wide variety of activities. They can move with several modes such as walking or cycling. They are not all constrained by scheduled and fixed activities such as work or school (e.g social profile or retired people). Moreover, as noted by [Schlich and Axhausen, 2003] there is a trending decline in general constraint but also greater flexibility over working hours and places of work [Manley et al., 2018, Goulet-Langlois et al., 2016]. As a result, it is not surprising to observe important

intrapersonal day-to-day variability. What is more interesting is that it is possible to correlate this intrapersonal variability to interpersonal variability using cluster analysis. In doing so, we found that there are important differences between clusters. In cluster 1, users probably cover most of their travel needs with public transit and tend to be more variable in their usage. In clusters with calendar structure such as cluster 2, 3, and 5, it can be assumed that the transit usage is driven by work or school trips which can explain the lower intrapersonal variability. Finally, in cluster 4 and 6, public transit usage is erratic and probably complementary to other modes of travel leading to higher levels of intrapersonal variability.

The analysis conducted in this paper also shows that the conventional distinction between working day, weekend and holiday may be relevant for some users but not for all. Even if on average, there are more stability in behaviour during weekdays, we have seen that the intrapersonal variability decrease only marginally when computed solely on working days. Similarly, we have observed that holiday periods have an influence only on a reduced proportion of users. This clearly shows that the traditional paradigm of transit planning focused on a set of typical days such as working days, Saturday or Sunday is not perfectly valid from the individual point of view. Thus, a more disaggregated understanding of which days individuals may or may not use the transit system is an important task. With our interpersonal clustering method, it is possible to do so. In fact, this method is designed to find homogeneous groups of passengers based only on the day they used public transit. The resulting clusters can then be used by transit planners and marketers, to inform or promote specific usage in a targeted manner. The combination of interpersonal clustering and pre/post intrapersonal similarity measurement could also assist in identifying groups of users that are more prone to change their travel behaviour after the introduction of new services or in case of specific disruption (e.g on week-end services).

This case study also demonstrates that cross-checking results with socio-demographic information derive from the fare profile, add a lot of value to the investigation. This is because socio-demographic attributes are important factors affecting intrapersonal variability [Pas and Koppelman, 1987, Susilo and Axhausen, 2014]. Unfortunately, smart card data and more generally passive data are often very incomplete on those aspects which can generate ambiguous explanation [Manley et al., 2018]. Moreover, there are increasing privacy concerns with smart card data that could jeopardise the full valuation of these data. In Lyon, the card is, for now, individual and the unique id number that identifies each card is changed every 12 months. In other cities, cards can be shared between passengers, cards id are changed more regularly and sometimes passengers have the option to pay with contactless bank cards. In those cases, it is not possible to ensure individual traceability, to perform long term analysis or to determine profile with fare product. These are additional challenges that must be addressed otherwise the potential of these data to conduct longitudinal analysis will be strongly limited.

Research around individual day-to-day transit usage variability is an important area of investigation that has several practical implications. Automatic systems such as

fare collection collect lots of data about lots of users over long periods of time and thus can be very useful to analyse the variability of transit usage. To make the best of those data, we need to develop new methodologies but also adapt the existing ones. This is what we have done in this paper, and while the results are specific to the city of Lyon, the two methods can easily be applied to other smart card datasets. It would especially be interesting to replicate this analysis in cities of distinct size, and distinct country to better understand the link between variability, city structure, and cultural context. Another direction of future research would be to further investigate the day-to-day variability of inconsistent users, and to try to understand the motivations behind the changes of transit usage intensity using targeted surveys. There are also some limits to the research presented in this paper that require additional analyses. First, representing a trip as a departure in a given time slot from a given zone is straightforward and easy to conceptualize but it is an oversimplification of the true characteristics of trips. More detail representation of trip patterns should be investigated and other measures of day-to-day similarity must be developed. Second, the calculation of a large dissimilarity matrix can be computationally expensive. More research is therefore needed to adapt this method to large datasets so it can be deployed in real-world applications. Third, our analysis is only based on smart card data but other data such as weather data or land-use data could provide further insight into the causes of variability.

5

Medium-term public transit route ridership forecasting : what, how and why ? A case study in Lyon

This chapter is an edited version of the following article under first revision in Transport Policy since 04/03/2020 :

O.Egu and P.Bonnel (2020). Medium-term public transit route ridership forecasting : what, how and why ? A case study in Lyon.

Highlights

- Examines the need for medium-term forecasting in public transit systems
- Proposes an operational modelling framework to forecast ridership one year ahead at different levels of spatiotemporal aggregation
- Outlines different applications of the forecasts for tactical planning and ridership monitoring
- Exploitation of archived passive data can enhance data-driven decision making

Chapter 5 : Medium-term public transit route ridership forecasting : what, how and why ? A case study in Lyon

Abstract

Demand forecasting is an essential task in many industries and the transportation sector is no exception. This is because accurate forecasts are a fundamental aspect of any rationale planning process and an essential component of intelligent transportation systems. In the context of public transit, forecasts are needed to support different level of planning and organisational processes. Short-term forecast, typically a few hours in the future, are developed to support real-time operations. Long-term forecast, typically 5 years or more in the future, are essential for strategic planning. Those two forecast horizons have been widely studied by the academic community but surprisingly little research deal with forecast between those two ranges. The objective of this paper is therefore twofold. First, we proposed a generic modelling approach to forecast next year ridership in a public transit network at different levels of spatiotemporal aggregation. Second, we illustrate how such models can assist public transit operators and transit agencies in monitoring ridership and supporting recurrent tactical planning tasks. The proposed formulation is based on a multiplicative decomposition that combines tree-based models with trend forecasting. The evaluation of models on unseen data proves that this approach generates coherent forecast. Different use case are then depicted. They demonstrate that the resulting forecast can support various recurrent tactical tasks such as setting future goals, monitoring ridership or supporting the definition of service provision. Overall, this study contributes to the growing literature on the use of automated data collection. It confirms that more sophisticated statistical methods can help to improve public transportation planning and enhance data-driven decision making.

Keywords— Public transit, Ridership forecasting, Machine Learning, Smart Card Data, Transport planning

5.1 Introduction

Demand forecasting is an essential task in many industries and the transportation sector is no exception. In fact, accurate prediction of future demand is an essential components of intelligent transportation systems [Vlahogianni et al., 2014, Koutsopoulos et al., 2019] and a fundamental aspect of any rationale planning process [Bonnel, 2002, Ortúzar and Willumsen, 2011]. Thus, it is not surprising that passenger demand forecasting is a widely studied subject. To summarize this area of research, we must take into account the domain of application and the type of planning issues the forecast intend to address. In many organisation including public transport, they are three commonly accepted level of planning and organisational control [van de Velde, 1999, Pelletier et al., 2011]. Strategic level deal with long-term decisions and objectives. Tactical level focus on decisions that take place in medium-term and aims to guarantee that the means to reach long-term goals are in place. Operational planning is concerned with short-term decisions that ensure the efficiency of the production. In agreement with this hierarchical order of decision-making activities, operators and transport agencies must generate different forecasts.

Short-term forecasting is a very active field [Vlahogianni et al., 2014] that deal with models that predict demand from few minutes to few hours into the future [Vlahogianni et al., 2014, Noursalehi et al., 2018]. In the context of public transit (PT), authors argue that it can enable the design of better control strategies and improve passenger experience by proactively adjust services and customer information [Koutsopoulos et al., 2019, Noursalehi et al., 2018, Ma et al., 2014, Wei and Chen, 2012]. While appealing and surely helpful, short-term forecasting can be hard to implement in real-life settings because it requires real-time data, continuous actualization of the model and trust in the model output. Short-term forecasts are also developed to support near real-time operational decision making. However, PT systems can not always react efficiently in real-time due to limited flexibility in resources (e.g vehicles, staff) and infrastructure constraints. Transport services and infrastructures are thus often designed and planned years in advance.

Long-term ridership forecasting deals with models that predict demand for time horizon ranging from 5 to 15 years ahead. In contrast to short-term forecasting, it is often a one-time exercise and forecasts are rarely generated continuously. Forecasts are produced to assess and evaluate future scenarios and support long-range strategical planning. The most common method for long-term ridership forecasting traditionally relies on four-step travel demand model [Boyle, 2006, Ortúzar and Willumsen, 2011, Bonnel, 2002]. Those models are data-intensive as they incorporate many elements such as land use, future population, choice model, network description etc. On one hand, it is a strength as this type of model can be used to forecast structural changes such as the introduction of a new route but also to evaluate different policy choice [Ortúzar and Willumsen, 2011, Bonnel, 2002]. On the other hand, it is a weakness because the resulting model can be costly, complex and hard to modify. Consequently, other forecasting approaches have been developed. For example, change in ridership in response to change in services or fares can be forecast using elasticity coefficient [Boyle, 2006, Totten and

Levinson, 2016, Paulley et al., 2006]. Other approaches include, spatial regression method [Pulugurtha and Agurla, 2012], rules of thumb method based on similar routes or professional judgement [Boyle, 2006, Diab et al., 2019]. In practice, those models and approaches are often calibrated for a limited set of situation such as an average weekday, a typical peak hour etc. They are, therefore, difficult to use in recurrent monitoring of the demand and cannot support tactical and operational planning activities.

Surprisingly, little research deal with forecast between those two horizons i.e medium-term forecasting. However, those kinds of forecasts are part of the toolbox needed for the effective planning of public transit systems. They can be of great use for PT operators and agencies that want not only to be able to adapt real-time operation (short-term forecasting) and evaluate strategic plan (long-term forecasting) but also to monitor tactically and continuously the evolution of demand. Toqué et al. [2017] have recently proposed to use machine learning models to predict PT ridership one year ahead in a disaggregate and continuous manner. However, the authors didn't introduce in their analysis a multiyear trend component neither a description of the level of supply although they may surely be needed for medium-term forecasting. Moreover, they barely discuss the practical aspect of the forecasting framework and motivate their work with use case. This paper seeks to address these research gaps and has two complementary objectives :

1. To develop a generic forecasting approach that combined trend analysis with machine learning model to predict one year in advance, at different levels of spatiotemporal aggregation, the ridership volume in a PT network.
2. To illustrate how such forecasts can support tactical planning tasks and assist PT operators and agencies in monitoring ridership.

To this end, the paper has been organized in the following way. It begins with a presentation of the modelling approach. After that, relevant data are identified, prepared and described. Then, the forecasts are validated using unseen data and bottom-up approach. Finally, use case are provided to illustrate the value of such data-driven framework. Results show that the proposed modelling approach is valid. Transit agencies can leverage the resulting forecast to better monitor and anticipate ridership in their network. Overall, this study contributes to the growing literature on the use of automated data collection to improve public transportation planning. Findings confirm that more sophisticated statistical methods can facilitate analysis that used to rely primarily on professional judgement [Hanft et al., 2016, Pelletier et al., 2011, Koutsopoulos et al., 2019].

5.2 Modelling framework

Let's denote $Y_i = (y_{i,1}, ..., y_{i,t}, ..., y_{i,n})$ as the vector of observed ridership volume on element i of a PT network for each time step $t \in 1, ., .., n$. Let's denote $X_{i,t}$ the set of features (explanatory variables) whose values are known for $t \in 1, ., .., m$ where m $(m > n)$ is the prediction horizon. Our goal is to estimate a regression model $f_i(.)$ for each element i such that $y_{i,t} = f_i(X_{i,t}) + \varepsilon$ where ε is the difference between the

observed value and the predicted value $\widehat{y_{i,t}} = f_i(X_{i,t})$. This formulation described a typical univariate regression problem where (1) each element of the network i is considered independent of the rest of the network, (2) each observation $y_{i,t}$ is considered independent of the other's observations.

As proposed by Toqué et al. [2017], to estimate $f_i(.)$ we can use machine learning models such as ensemble of decision trees. Those models have recently gained much popularities in various fields due to their ability to learn complex non-linear relation between features and provide higher accuracy than single machine learning models [Friedman et al., 2001, Opitz and Maclin, 1999]. Those models combine several decision trees into one single model that output the mean of all decision tree. A single decision tree can be seen as a stratification of the ridership volume into some simple regions using a set of splitting rules based on the available features. The underlying principle of those models is that by combining single decision trees we can significantly improve the predictive performance and reduce the variance of a single tree [Friedman et al., 2001, Breiman, 2001, Friedman, 2001].

To learn a set of decision trees effectively the two most common algorithms are random forest (RF) and gradient boosting (GB). Random forest was introduced by Breiman [2001] and is based on the concept of bagging. More precisely, each decision tree is built independently on a random subset of the training data but also considering only a random sample of features at each split. In doing so, the algorithm enforces diversity in the trees which when we average them may reduce the variance of the model [Breiman, 2001]. Gradient boosting trees were formalized by Friedman [2001] and differ from the random forest model in the sense that decision trees are built sequentially using strategically resample training data. More specifically, each new tree is built to recover the errors resulting from the model obtained with the ensemble of trees produced at the previous steps. In this approach, each of the decision trees can remain rather small with performance slightly above random guess. Fitting them sequentially can then help to improve the model in areas of the feature space that were missed by previous ensemble [Friedman et al., 2001].

One aspect of those machine learning models is that they assume independence between the $y_{i,t}$ and thus can't learn time-dependent structure and forecast potential trend. In fact, as explained above the model output are computed by averaging decision trees which are themselves based on historical ridership data. However, ridership volume may have statistical properties that evolve through time and might also be view as a non-stationary time series. To deal with this aspect, we can assume that ridership volume posses two components a yearly trend-cycle component $s_{i,T}$ and a yearly adjusted component $a_{i,t}$. Ridership level can then be expressed using multiplicative decomposition :

$$y_{i,t} = s_{i,T} \times a_{i,t} \tag{5.1}$$

were $s_{i,T}$ denotes the mean of $y_{i,t}$ in year $T, t \in T$ and $a_{i,t}$ is the yearly adjusted level ridership of element i at time step t.

The forecasting task can then be divided into two subtasks were we have to forecast $s_{i,T}$ and $a_{i,t}$. The rationale behind this process is twofold. First, $a_{i,t}$ is assumed to be stationary and free of trend and thus can be estimated with ensemble of decision

trees ($\widehat{a_{i,t}} = f_i(X_{i,t})$). Second, $s_{i,T}$ is supposed to be non-stationary and evolve slowly over time and can be forecast by taking the last value of the yearly trend-cycle component multiply by a growth factor α_i. $\widehat{y_{i,t}}$ can then be obtained in the following way :

$$\widehat{y_{i,t}} = \widehat{a_{i,t}} \times (1 + \alpha_i) s_{i,T-1} \tag{5.2}$$

In other words, the forecasted ridership for element i is equal to the mean ridership of the previous year $s_{i,T-1}$ multiply by a growth factor while the deviation from the mean yearly ridership volume $a_{i,t}$ is predicted using a regression model. α_i can then be estimated using a weighting average of the observed historical year-mean percentage evolution :

$$\alpha_i = \frac{\sum_{T=2}^{N} w_T \frac{s_{i,T}}{s_{i,T-1}}}{\sum_{T=2}^{N} w_T} \tag{5.3}$$

where N is the number of years used to train the model and $w_T = 1/(N + 1 - T)$ is the weight associated with year T defined with inverse time decay function. This weighted formulation is proposed to estimate α_i from multiyear behaviour of each element while giving higher weight to more recent years.

This forecasting formulation is well suited to our objectives for two reasons. First, it exploits the property of ensemble models to learn the influence of various features on detrended ridership data. Second, it aligned forecast with previous year observed ridership volume of element i to ensure coherency of the forecast with the most recent element state. This formulation can be seen as a conservative continuation of previous year's ridership mean volume combined with a model that learns historical behaviour of the network element under different conditions of the features space. As noted by Hyndman and Athanasopoulos [2018], traditional time series models (such as seasonal ARIMA) don't allow irregular events and have difficulty to accommodate multiple imbricated seasonality. The intuition behind the above formulation is that the decision tree model is able to capture those effects while remaining robust to potential outliers.

At this step, we have now a modelling approach that can be used to forecast the future ridership volume at time step t of each element i denoted as base forecast. To establish forecast for higher levels of the network hierarchy such as a group of elements J (e.g group of stops, group of routes) or larger temporal aggregation (e.g day, week, month, working day), we can employ the simple and straightforward bottom-up approach. In this case, the forecasted ridership volume of higher hierarchies element J for the period $M = [t1, ..., t2]$ denoted $\widehat{R_{J,M}}$ is obtained by summing the different base forecast :

$$\widehat{R_{J,M}} = \sum_{i \in J} \sum_{t=t1}^{t2} \widehat{y_{i,t}} \tag{5.4}$$

This approach is a classical way to deal with time-series collections that naturally presents hierarchical properties and different seasonal pattern [Hyndman and Athanasopoulos, 2018, Kahn, 1998]. It has the advantage to preserve in the forecasting task the dynamic and characteristics of individual elements while making it possible for the analyst to obtain forecast at the desired spatiotemporal level of the PT network.

5.3 Data preparation and exploration

Since the introduction of automated data collection systems, it is now possible to access high-resolution ridership data. Those new data sources have the potential to change the current state of practice in ridership forecasting if they are accurate and of good quality [Boyle, 2006, Diab et al., 2019]. To be used in medium-term forecasting those data need also to be correctly curated and archived so there are enough observations to train models. The proposed modelling approach requires the definition of a base forecast level. This level should not be too noisy to ensure models can properly learn patterns and at the same time sufficiently detailed for most of the required tactical analysis. This level should also be set from the perspective of the organisation managing the network which can be different from the passenger perspective. We suggest that the most appropriate spatiotemporal base level of aggregation for medium-term forecasting are :

— Temporal aggregation (t) : hourly data because it is needed to account for intra-day variation and define supply level accordingly. However, it should be noted that the same methodology could be applied to daily data if this resolution level is deemed sufficient.

— Spatial aggregation (i) : route level is most suitable than stop level because supply level is more than anything else link to route elements.

For this research, the fare transactions from 2014 to 2018 for the subway network (4 metro lines and 2 funicular lines), the tramway network (5 lines) and high-frequency buses known as line C (25 lines) were extracted from the Lyon PT operator data warehouse. Those routes were selected because they represent more than 80% of the total ridership of the network. This sample was also deemed to be enough to evaluate correctly our approach. Fare transactions were then summed by route and by hour [1]. To ensure, the completeness of the data, the following procedure was implemented recursively for each route :

— Generate the hourly time bin from the 1st of January 2014 00 :00 to the 31st December 2018 23 :00.

— Lookup for the number of fare transactions summed by hour. If no transaction is found for a given hour it was assumed that no passenger board the route and the ridership value was set to zero. This procedure is put in place to ensure the consistency of the training data which must contain periods with and without ridership to correctly learn the route historical behaviour [2].

For the rest of this paper, the dataset resulting from the above procedure will be referred to as the route ridership volume and will form the base level of our modelling framework.

Our modelling framework also required a set of features that may explain the route ridership volume variations while being available one year in advance. It is well recognized that PT ridership presents multiple temporal seasonalities and regular cycles that can be encoded with ordinary calendar attributes. Consequently, a

1. Noted that for the subway fare transactions are not linked to stations and were thus redistributed by route according to operator reporting procedure.

2. Note that in the case of potential data loss this procedure will also result in ridership equal to zero.

calendar features table common for all routes was manually crafted. Features of this spreadsheet are described in table 5.1. Ridership is also influenced by the level of supply i.e the quantity of service produced. One of the most common and simple indicator to describe this aspect of transit is the number of commercial kilometres offered. This indicator was thus extracted from the operator data warehouse and choose as an explanatory variable. Unfortunately, for this research, this feature was available with enough historical depth only at the level of day and route. Noted also that in Lyon, commercial kilometres offered by routes are set one year in advance by the operator planning team for contractual and financial reasons. Thus, like calendar features, they can be determined sufficiently in advance and be used as predictors in medium-term forecasting (which is not the case of potential other features such as weather or unplanned events).

Features	**Description**
Hour	24 categories
Month	12 categories
Week number	Numeric from 1 to 53
Day of the week	7 categories
School holiday	7 categories, one for each holiday period
Public holiday	14 categories, one for each French public holiday
Adjacent to public holiday	Binaries if weekdays and adjacent to public holiday
Light and music festivals	Binaries if major citywide events

TABLE 5.1 – Hand crafted calendar features, source : prepared by the authors

The calendar table, the supply feature and routes ridership volumes were then merged. To explore the characteristics of the resulting dataset we have aggregated the ridership volume of the 36 selected routes and perform exploratory visualizations in figure 5.1. In figure 5.1a, it can be seen that there are cyclical patterns that repeat every year such as a strong decrease of ridership volume during summer months (July and August). There are also long-term dynamics in the ridership volume with an observable increasing tendency. This is because Lyon network has been gaining popularities in the last few years. In figure 5.1c, the variation of daily ridership volume for the year 2018 are observed more finely. This figure indicated that they are multiple imbricated seasonalities and calendar effects. More precisely, it can be seen that school holiday periods (e.g in February, April, October and December) but also other events such as bank holidays or special events have an impact on the ridership volume. Samely, we observe a strong decrease in ridership during weekends and summer months. As indicated by figure 5.1a, those seasonalities seems to be recurrent every year but calendar effect are moving in time because each year events and holidays can fall in different days of the year. In figure 5.1b, the relation between the total daily number of commercial kilometres produced and the ridership volume is depicted. This figure shows that there are three main states in the network that correspond to three typical planned service output (weekday, Saturday and Sunday). Figure 5.1b also demonstrates that there are more variations in the ridership volume than in the service produced. For instance, 85,000 kilometres produced corresponds to a typical weekdays supply level but for this same amount of supply, the ridership

volume can vary between 0.8 to 1.1 million. This is because the supply is often fixed based on a set of typical days while the demand is more volatile. Finally, in figure 5.1d the hourly ridership from the 5th May 2018 to 19th May 2018 is given. This period exhibits distinct hourly patterns. During the first week, there are two public holidays on the 8th of May and 10th of May and the demand pattern on those day does not exhibit strong peak compare to the other weekday of the week. During Saturday and Sunday, the hourly ridership curve has a bell shape. Finally, from the 14th of May to the 18th of May we observe a typical demand pattern with peak periods in the morning and the evening but also higher hourly ridership volume than during the previous week.

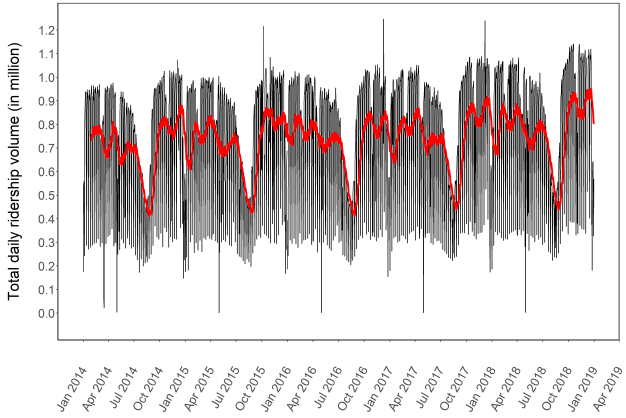
Overall, the previous plots confirm that ridership volume presents multiple moving imbricated seasonality and irregular events. Changes appear from one year to another, from one month to another, during the hour of the day, the days of the week etc. There is also a relation between the level of supply and demand. Those observations are not per se surprising but we should recall that those are the elements that each base model need to capture independently for each route. In fact, while the previous plots were obtained based on the aggregation of all route, the model will be trained for each route separately and we can expect that routes have distinct behaviour. For example, the influence of holiday may be less important for some part of the network, some routes may show higher multi-year growth than other, some routes may have specific characteristics on weekends because they serve specific areas of the city such as touristic places. Hence, it is fundamental that the model can adapt correctly and in an automatic way to varying route behaviours while capturing system-wide demand variation and structure.

5.4 Models fitting and forecast errors

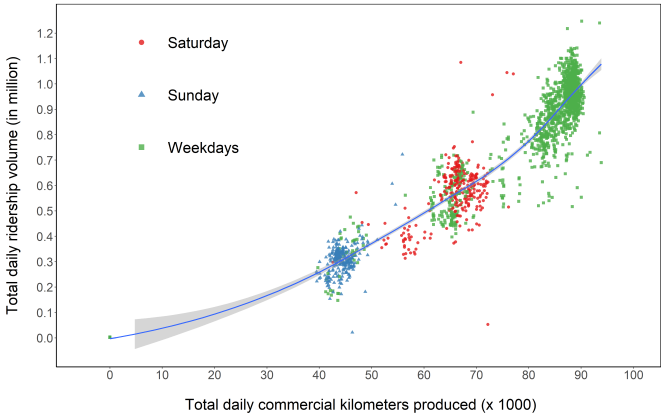
Name	**Model**	**Features**	**Decomposition**
HM	Historical median	Hour, Type of day, Month, Working day	
HMT	Historical median	Hour, Type of day, Month, Working day	x
RF	Random forest	Calendar features	
RFT	Random forest	Calendar features	x
RFTS	Random forest	Calendar features and supply feature	x
GB	Gradient boosting	Calendar features	
GBT	Gradient boosting	Calendar features	x
GBTS	Gradient boosting	Calendar features and supply feature	x

TABLE 5.2 – List of implemented models, source : prepared by the authors

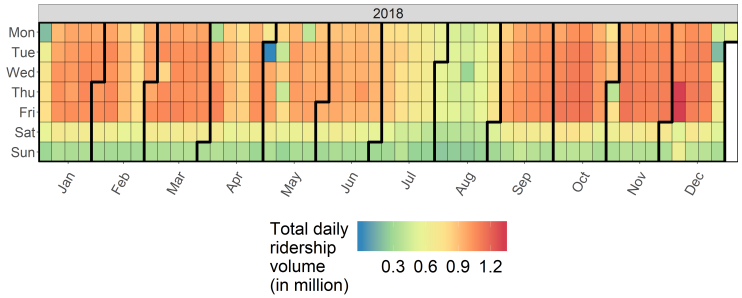
In the previous sections, we have defined the problem, proposed a modelling framework, gather the data and perform some preliminary exploratory analysis, we now present models fit. Table 5.2, depicts the list of models that we have decided to evaluate. The first model formulation is a naive historical approach. In this case, the forecast is computed using the median of all past time step over the historical observation with the same characteristic as the forecasted time step. Time step



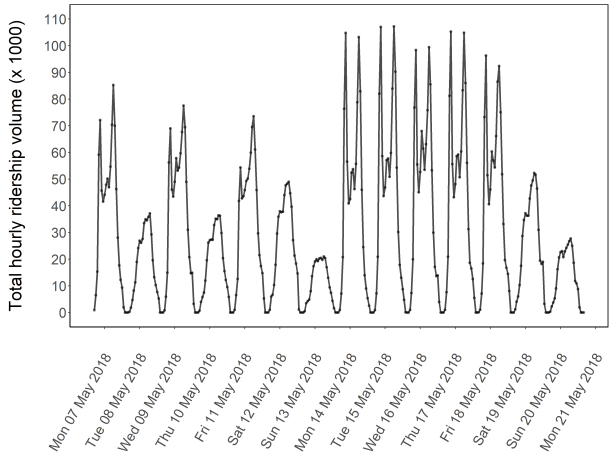
(a) Dayli ridership volume from 2014 to 2018, red curve correspond to 30 days moving average



(b) Relation between the supply feature and the daily ridership volume, points are colored by day type



(c) Calendar heatmap of daily ridership volume in 2018



(d) Hourly ridership volume from the 7th of May 2018 to the 20th of May 2018

FIGURE 5.1 – Exploration plot based on the aggregated ridership volume of the 36 selected routes, source : prepared by the authors

characteristics were in this case defined with the cartesian product between the hour, the month, the type of day (weekday, Saturday, Sunday) and a boolean indicator for working day (no public holiday, no school holiday, no special events). This formulation will serve as the baseline to evaluate the two ensemble models described in section 5.2. Table 5.2 also indicates if the models were trained with the raw ridership volume or with the decomposed ridership volume. In the second case, the equation 5.2 was used to forecast future ridership volume.

The accuracy of a forecast procedure can only be truly determined using data that were not used to fit the model [Hyndman and Athanasopoulos, 2018]. To do so, there are two common approaches : the cross-validation where several sets of tests are sequentially constructed and the hold-out approach where the data is divided into two portions. In this research, the hold-out approach was selected. The data was thus divided between a training set comprising all observations from 2014 to 2017 and a test set consisting of 2018 observations. To measure forecast errors the following standard metrics were chosen :

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_{i,t} - \widehat{y_{i,t}}| \tag{5.5}$$

$$\text{MAPEv} = \frac{1}{n}\sum_{i=1}^{n} \frac{|y_{i,t} - \widehat{y_{i,t}}|}{y_{i,t}}, \quad \forall \; y_{i,t} > v \tag{5.6}$$

$$\text{MdAPE} = \text{median}(\frac{|y_{i,t} - \widehat{y_{i,t}}|}{y_{i,t}}) \tag{5.7}$$

In table 5.3, we have reported the error measures on the train and test dataset for different levels of aggregations. First, as expected, when the aggregation level increases, errors decrease significantly. For instance, the best model at the base forecast has a MdAPE of 11.7 % almost four times bigger than the best model MdAPE once forecasts are aggregated by day (2.8 %). This indicates that the base levels are challenging to forecast but the bottom-up approach allows to generate coherent results. In other words, base forecast errors tend to compensate when forecasts are aggregated by route or by time. A second important observation is that ensemble models outperform baseline model in terms of absolute errors (MAE) but also in terms of percentage errors (MAPEv and MdAPE). Ensemble models take into account a more complex feature space but still can generalize to previously unseen data more comprehensively and without overfitting. However, there are no important differences in terms of accuracy between the gradient boosting and the random forest which is not surprising since both methods have strong similarities. What is also clear from this table is that the decomposition approach is essential when historical data with a certain depth is used to model future ridership. This is because the decomposition ensures the coherency of the forecast with the most recent observed ridership volume. Finally, the introduction of supply feature in ensemble models can reduce the errors especially large errors. If we compare the RFT (random forest with decomposition) and RFTS (random forest with decomposition and supply feature), it can be seen that median absolute percentage errors are quite similar but mean absolute percentage error can be considerably higher when the supply feature is omitted. This is because, when

there is a significant reduction in the kilometres produced the model based on calendar features will assume no change in the ridership volume.

		TRAIN SET (2014-2017)							
		HM	**HMT**	**RF**	**RFT**	**RFTS**	**GB**	**GBT**	**GBTS**
Base level (route-hour)	MAE	97	93	79	74	69	81	77	**67**
	MAPE_50 (%)	17,4	16,8	15,1	14,5	**12,9**	16,6	16,1	13,4
	MdAPE (%)	10,5	10,0	10,3	9,6	**9,0**	10,7	10,1	9,1
Aggregated (route-day)	MAE	1795	1670	1236	1063	**712**	1230	1088	859
	MAPE (%)	25,6	24,8	11,2	10,3	**5,5**	13,1	12,1	8,2
	MdAPE (%)	6,3	5,6	5,2	4,3	**3,0**	5,4	4,6	3,8
Aggregated (hour)	MAE	2682	2618	1913	1816	1741	2040	1987	**1564**
	MAPE (%)	16,0	15,7	17,1	16,9	18,0	20,2	20,1	**14,8**
	MdAPE (%)	6,4	6,0	5,5	5,0	5,0	6,4	6,2	**4,9**
Aggregated (day)	MAE	51725	49869	30793	26737	**18264**	31063	28508	21007
	MAPE (%)	67,0	67,1	14,7	14,1	**6,3**	18,5	18,4	11,5
	MdAPE (%)	4,4	4,0	3,1	2,4	**1,8**	3,3	2,8	2,2
		TEST SET (2018)							
		HM	**HMT**	**RF**	**RFT**	**RFTS**	**GB**	**GBT**	**GBTS**
Base level (route-hour)	MAE	131	115	118	99	95	123	104	**93**
	MAPE_50 (%)	21,7	21,7	19,9	19,8	**17,2**	21,1	21,4	17,4
	MdAPE (%)	14,5	12,8	13,7	11,9	**11,7**	14,4	12,6	11,7
Aggregated (route-day)	MAE	2662	2157	2339	1708	**1480**	2351	1720	1497
	MAPE (%)	36,3	36,9	26,8	26,7	**16,6**	28,1	27,8	17,7
	MdAPE (%)	10,7	7,7	9,9	6,8	**6,4**	10,0	6,9	6,5
Aggregated (hour)	MAE	3562	2878	2993	2140	2182	3226	2349	**1976**
	MAPE (%)	25,7	24,9	19,7	19,1	16,7	25,4	25,1	**16,4**
	MdAPE (%)	10,0	7,1	8,4	**5,7**	5,9	9,7	6,8	5,8
Aggregated (day)	MAE	75721	54068	63457	33658	31259	63643	35738	**30671**
	MAPE (%)	55,2	55,7	28,9	26,7	**9,0**	32,9	31,0	13,8
	MdAPE (%)	8,1	4,4	7,2	**2,8**	3,1	7,8	3,5	3,1

TABLE 5.3 – Model results on the train and test dataset for different level of aggregation (HM : historical median, RF : random forest, GB : gradient boosting, T indicates the use of the proposed decomposition approach, S indicates the incorporation of supply feature), source : prepared by the authors

5.5 Use case analysis

In the previous section, we have fitted different models and generate forecasts of ridership volume one year in advance. We have shown that it was possible to obtain rather low errors at higher aggregated levels but the forecast quality decreases at lower levels of aggregation. While forecast accuracy is in itself an important matter it is also important to recall that PT networks are complex systems that are impacted by all sort of unexpected and exogenous events. Therefore, the above models should not be seen solely for their prediction capability but also for their ability to assist PT organisations in monitoring ridership and in improving tactical decision-making activities. This is what we intend to demonstrate in this section by introducing a few use cases.

With the above modelling approach, it is possible to obtain coherent aggregated

ridership forecasts. These types of forecasts could be used for tactical decision making. For instance, they can assist senior executives in setting next year goals and monitoring their PT network with a more data-driven approach. Samely, every year, transit agencies need to prepared next year budget which obviously requires at one point some sort of ridership forecast. In practice, this process can be tedious and is often a matter of expert. In the case where different operators coexist in the same network, this task is even more complex and critical for the proper functioning and financing of the system. Agencies could use the proposed approach to generate forecasts for all route and then aggregated them to prepare budgets that match more closely the future variation of ridership and fit the financial/operational division of the network [3]. To illustrate this first use case, the weekly forecast for the two random forest models obtained through decomposition (RFT and RFTS) by type of transportation mode is plotted in figure 5.2.

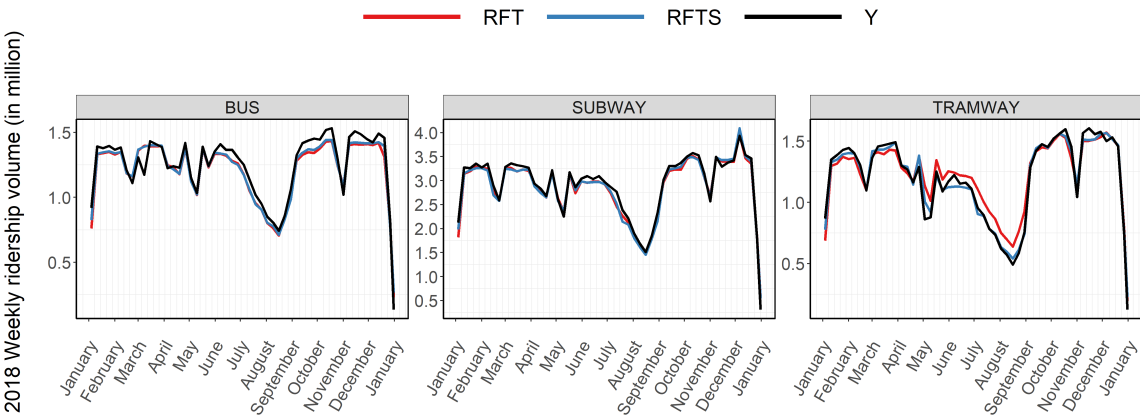


FIGURE 5.2 – Weekly forecast one year in advance for the three transportation mode (RFT : decomposed random forest, RFTS : decompose random forest with supply features, Y : real ridership volume), source : prepared by the authors

In figure 5.2, it can be seen that the different series are overall quite close and the aggregation of models captures typical calendar patterns such as the impact of school holidays, public holidays, or citywide events (i.e peak at week 50 for subway network). It can also be seen that the ridership volume of high-frequency bus network has increased more quickly than anticipated by the two models, especially after September. Finally, figure 5.2 indicates that for the tramway routes both model yield different forecast. This is because in 2018 the tramway network was impacted by important construction work of a new tramway route (T6) that leads to change in the service provision. Those type of elements can be further inspected by an analyst using models outputs in a retrospective manner. Especially, models can be used to identify routes whose behaviour deviate from historical behaviour as captured by the models. In doing so, analysts could identify and prioritize which route they need to review first in a more efficient manner than doing it based on "anecdotal knowledge" or fixed cycle [Coleman et al., 2018]. In figure 5.3, we have plotted each route as a point in a two-dimension space where the x-axis is the annual percentage error between the RFTS model and the real 2018 ridership and the y-axis is the

3. Those same forecasts could also be used to schedule workforces such as field information officers or fare officers

total real ridership volume. Using figure 5.3 it is easy to identify routes that require further analysis, balancing the importance of the route (y-axis) and the importance of the deviation from the model (x-axis). Figure 5.3 indicates that most routes errors are under 5% but some route exhibit higher discrepancies between model and real observations such as C3, C14, C15, C9 or T5. This route level analysis can be further extended to go deeper into the available data. For instance, real and model route ridership volume were aggregated by month and inspected in figure 5.4.

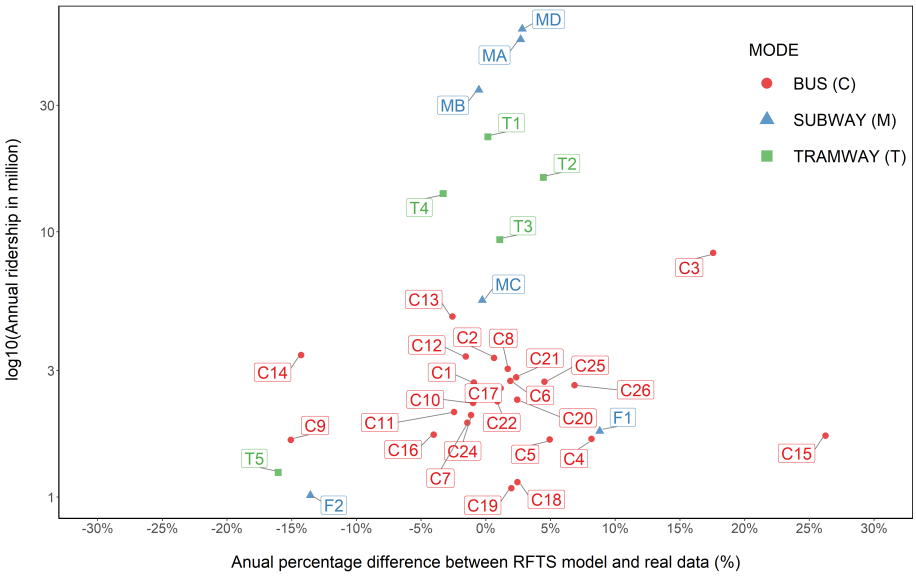


FIGURE 5.3 – Annual ridership volume forecast error for decomposed random forest with supply feature model (RFTS), source : prepared by the authors

In figure 5.4, we can observe that there is indeed a large diversity of route ridership historical pattern. For instance, route C3 is the busiest bus route in Lyon and important road work were carry on between mid-2016 to mid-2019 to provide exclusive right-of-way. The route was divided into two sections and part of the stops were relocated. As indicated by figure 5.4, it had a great impact on the 2017 ridership volume and the model forecast further decreased of the demand in 2018. However, by the end of 2018, as the route progressively returns to better operational condition, the ridership starts to increase significantly in an unforeseen manner for the model. In 2016, a strong decrease of ridership is observed on route C14[4] follow by a return at historical ridership volume during 2017. These disruptions were not well assimilated by the models, and further growth of the route was forecast. As mentioned earlier, in 2018 the tramway network was impacted by the construction of a new tramway line (T6). In figure 5.3, it can be seen that while tramway routes T1 and T3 are forecast correctly the model underestimates/overestimates the traffic on routes T2/T4. Figure 5.4 indicates that those routes exhibit a strong decrease in summer month (July/August). Models with supply feature partially anticipate this decrease (see figure 5.4 and figure 5.2) but models with only calendar features couldn't predict those changes (see figure 5.2). The impacts of those kinds of exogenous elements are in practice very difficult to understand and measure clearly without confronting the real

4. Route C14 share a portion of C3 layout and was also impacted by road work

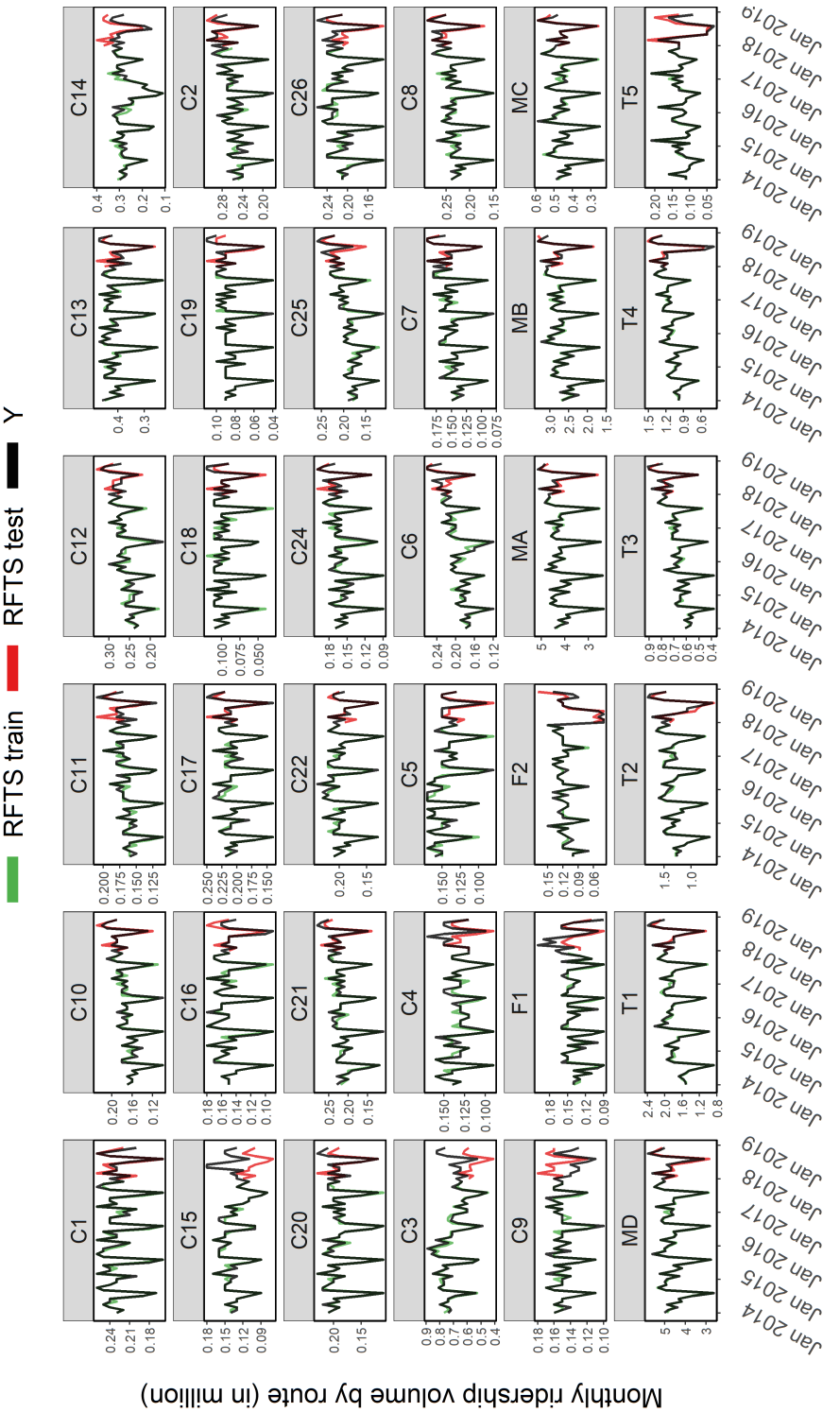


FIGURE 5.4 – Monthly time series of ridership volume by route (RFTS train : training of decomposed random forest with supply feature (2014-2017), RFTS test : testing of decomposed random forest with supply feature (2018), Y : real monthly ridership by route), source : prepared by the authors

ridership with a "business as usual" scenario. With the above modelling approach, transit agencies can now more easily do so. They can monitor route ridership with a data-driven approach using a predictive model that is easy to re-actualize but also to implement. Moreover, they can use these "business as usual" forecast to estimate in advance the number of trips impacted by a planned disruption. Once the period of disruption and the routes impacted are known, an analyst can interrogate models to retrieve the forecasted number of impacted trips. This information can then be used in the designed of replacement services or mitigation plan. In a similar vein, if different periods of disruption are being considered, models could also be used to identify the period that minimizes the impacted volume of trips. For this kind of use case, one may prefer to leverage on model using calendar features so the forecast is a pure "business as usual" prediction.

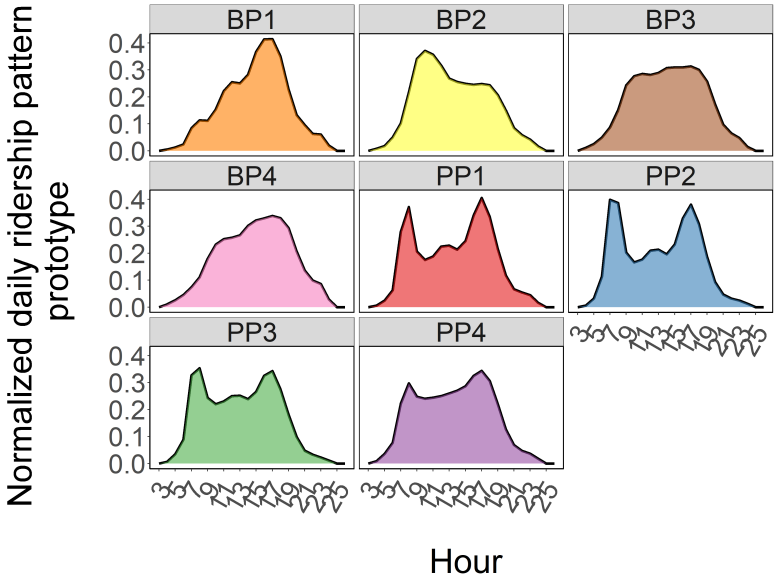
The above use cases were based on spatial and/or temporal aggregation of models outputs but the base model i.e by route and hour can also be of value to apprehend in-depth ridership pattern. To this aim, forecasts need to be synthesized in such a way that the information can directly be helpful for tactical planning purposes. One way to do so would be to cluster the model output to identify a set of typical daily ridership profiles. Spherical kmean is a popular text mining prototype-based clustering algorithm [Buchta et al., 2012]. It relies on the cosine distance to measure dissimilarity between objects which has interesting properties for our problem as the observations are cluster irrespectively of their magnitude i.e the ridership volume. It can be mathematically express as a minimization of the cosine distance of all possible allocation of objects m to clusters $c(m) \in 1, ..., k$ and over all prototypes $p_1, ..., p_k$ in the same feature space as object m denoted x_m :

$$\min \quad \sum_m 1 - \frac{x_m \cdot p_{c(m)}}{\|x_m\| \cdot \|p_{c(m)}\|} \tag{5.8}$$

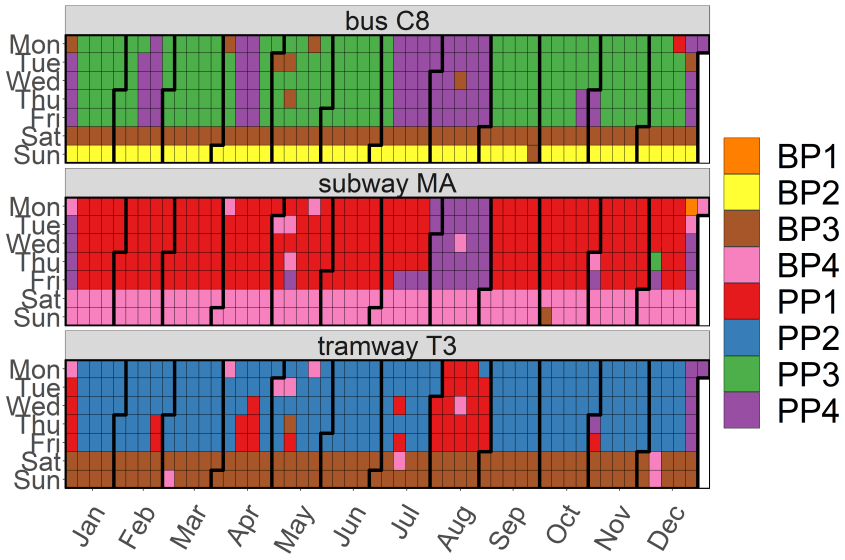
In our problem, the observation x_m is the vector of hourly forecasted route ridership volume on a given day. The features space dimension is 24 and the number of observation is equal to the number of days ahead we forecast multiply by the number of routes [5]. In prototype-based clustering, the number of clusters k need to be defined first. Then, a fixed point algorithm can iterate between optimal allocation of observations to cluster and optimal prototypes to attempt to determine a partition that minimizes equation 5.8. For the sake of illustration, we have decided to retain 8 clusters accounting for almost 80% of the dissimilarity. The fixed-point algorithm was then run 20 times and the best solution was retained. The resulting clusters prototypes are given in figure 5.5a and were named according to their shape : bell shape clusters (BP) from 1 to 4 are daily ridership profiles that don't exhibit two peak periods while peak profiles clusters (PP) numbered from 1 to 4 are daily ridership profiles that exhibit a morning and evening peak pattern.

Those 8 clusters synthesize the high-resolution data contained in the base model. They facilitate the analysis and provide more manageable information to transportation planners and analysts. They can be used to further explore the future demand of each route. A typical use case would be to produce in addition to

5. For 2018, N= 13140 = 36x365



(a) Clustering prototype derived from RFTS model (random forest decomposed with supply)



(b) Calendar of forecasted daily prototype for three selected routes : bus route C8, subway route MA and tramway route T3)

FIGURE 5.5 – Clustering analysis of model forecast, source : prepared by the authors

the previous analysis on the ridership volume a calendar of future daily ridership profile. The output of this procedure is given in figure 5.5b for three selected routes (bus route C8, subway route MA and tramway route T3). In this figure, it can be seen that the models have learned differents daily ridership profiles for each route. For instance, bus route C8 is characterized by a change from prototype PP3 to prototype PP4 during holiday weekdays. It means that the model forecast that during school holidays the peak period will be less important with regard to the daily volume. Moreover, the model has learned that a typical Saturday will exhibit a BP3 profile while Sunday will exhibit a higher concentration of trip in the

morning (prototype BP2). Subway route A daily profiles are overall more stable. Unlike route C8, Saturday and Sunday prototype are equal. During weekday the assigned prototype remains the same (PP1) except during the month of august (PP4). Finally, the tramway route T3 is characterized by weekday with a high concentration of trip during morning peaks periods (prototype PP2). Compared to the subway route the weekend prototype is BP3 meaning that in proportion less ridership is forecasted in the late evening. These observations demonstrate that the models have learned and forecast distinct daily ridership profiles. Once cluster those forecasts can be examined more easily and help analysts to adjust service plans to intra-day demand variation. Altogether, this use case section offers some insights into how the proposed modelling framework could be used in real-world applications and answer common tactical issues. One of the main advantages of the proposed solution is the flexibility to explore forecasts at different levels of aggregation.

5.6 Conclusions

This research was undertaken with two objectives. First, to design a modelling framework suitable for medium-term forecasting of route ridership volume. Second to evaluate the potential benefit of such forecasts for transit agencies and operators. The evaluation of different models has shown that the proposed decomposition formulation can learn complex patterns from historical data that can then be satisfactorily projected into the future. The resulting forecasts have proven to be quite appropriate to facilitate recurrent tactical decision making and help planning better services. The use case section outlined different applications such as setting future goals, monitoring ridership volume at different levels of aggregations, estimating the impact of future disruptions and supporting the definition of future service provision. It is therefore recommended that transit agencies complement their traditional reporting tools with these types of predictive approaches. By doing so, they can value at their full potential their historical data and enhance data-driven decision making. Moreover, the cost of implementation of such a tool is reduced and not too challenging. Models should simply be trained automatically by batch (e.g. every month) based on the above data preparation procedure. Then the provision of modern dashboarding tools with drill-down capability will make the data available through a web application for anyone interested in the organisation.

Although the study has demonstrated that the proposed modelling formulation generate coherent forecast and is operational, it also has several limitations that need to be considered and further investigated. First, the ensemble models are trained using the same weight for all historical observations. Even though this approach is simple and generic it would be interesting to elaborate methods to identify non-representative historical data that should not be part of the training set. We should also investigate mechanisms that put more weight on recent observations such as what is proposed for the estimation of trend component. Second, while we did incorporate one supply feature in the analysis, more research

is required to establish additional evidence on which variable are needed and how they should be treated in the modelling process. In this regard, one direction would be to incorporate previously calculated elasticities factors in the model. Third, for this research, we assumed that higher hierarchies forecast can be obtained simply by summing the base forecast but other approaches need to be considered such as optimal forecast reconciliation or middle-out approach [Hyndman et al., 2011].

6

Conclusions

« If you torture the data long enough, it will confess to anything. »
Ronald Coase

Les réseaux de transports en commun urbains sont des systèmes critiques pour le bon fonctionnement des villes. Ils offrent un service de mobilité indispensable à la réalisation des activités humaines et répondent à des impératifs économiques, écologiques et sociétaux. Ces systèmes sont complexes et leur exploitation est onéreuse. En matière d'investissements, les décisions stratégiques portent leurs fruits pendant des décennies et peuvent avoir un fort impact sur le développement territorial. De même, les décisions tactiques et opérationnelles ont un impact direct sur le quotidien des usagers et les comportements de mobilité. Ces systèmes doivent donc être planifiés et organisés avec rigueur à tous les niveaux. Pour ce faire, il convient de s'appuyer sur un dispositif de collecte et d'analyse des données permettant de faire le lien entre le monde de la planification et le monde réel entendu comme la production de l'offre de transport. Depuis l'émergence des big data, ce dispositif est en pleine mutation. Des sources de données collectées en continu et sans intervention humaine viennent s'ajouter à des sources de données dites traditionnelles basées principalement sur des enquêtes. L'ensemble forme un dispositif hétérogène qui est l'objet principal de cette thèse. L'ambition de cette thèse est de s'interroger sur la pertinence de ce dispositif et sur l'apport des nouvelles sources de données dans une perspective d'aide à la décision. Dans ce manuscrit, cette ambition se matérialise sous la forme de quatre articles scientifiques. Pour chaque article, une question de recherche est formulée et des réponses sont apportées. Il nous faut maintenant conclure ce manuscrit. Nous commençons donc par rappeler les principaux résultats de nos recherches. Nous décrivons ensuite quelques applications opérationnelles qui nous semblent pertinentes. Enfin, nous identifions les limites de ces travaux et proposons de futures pistes de rechserche.

6.1 Rappel des principaux résultats de la thèse

La fraude dans les transports en commun est un phénomène complexe qui possède une forte composante socioculturelle et est très sensible au contexte local. Elle est le résultat de comportements individuels variés allant de la fraude non intentionnelle à la fraude intentionnelle avec perte de recette. Du point de vue des opérateurs, la fraude est un risque sérieux, car elle engendre des pertes de recettes et peut créer un sentiment d'insécurité nuisible à l'attractivité du réseau. C'est aussi une source de biais importante pour les données billettiques. Dans le chapitre 2, nous avons cherché à déterminer quelles sont les sources de données permettant une mesure fine et précise de ce phénomène. Pour ce faire nous proposons une typologie de la fraude orientée opérateur, car elle différencie la fraude avec ou sans perte de recette. Nous proposons ensuite une méthodologie de fusion des données permettant de dériver des indicateurs de fraude. Ces travaux confirment la difficulté de mesurer précisément le phénomène et démontrent que l'infrastructure physique a un fort impact sur le taux de fraude (milieu ouvert ou fermé). L'analyse détaillée montre que les différents indicateurs ne mesurent pas le phénomène avec exactement la même définition et que cela peut créer d'importants écarts et des résultats parfois contradictoires. À Lyon, nos résultats indiquent que l'utilisation des données récoltées sur le terrain par les contrôleurs peut entraîner de forte sous-estimation du phénomène notamment dans les milieux ouverts de type bus et tramway. L'utilisation d'un taux de validation semble être une piste plus intéressante pour mesurer la fraude. Cet indicateur peut permettre d'explorer la structure et la variabilité spatio-temporelle de la fraude tarifaire. Malheureusement, c'est un indicateur qui reste partiel dans le sens où il agrège une variété de comportements et de situations. C'est aussi un indicateur qui reste très dépendant de la disponibilité et de la précision des instruments de comptages. L'utilisation d'enquête fraude par sélection aléatoire d'un échantillon stratifié de voyages apparaît donc comme la méthode la plus précise de collecte de données sur la fraude. Ces enquêtes permettent de collecter des informations détaillées sur les motifs et la typologie de fraude. Cependant, ces enquêtes sont coûteuses et la quantité de données recueillies est limitée ne permettant pas d'utiliser cette méthode de collecte pour assurer un suivi au jour le jour du niveau de fraude. Tous ces éléments nous amènent à conclure qu'en matière de mesure de la fraude aucune source de données ne répond à tous les besoins. Chaque source présente des avantages et des inconvénients qui sont appréciés dans cette thèse. Tout comme il semble nécessaire de combiner un ensemble de mesures pour prévenir la fraude, il semble nécessaire de croiser ces diverses sources de données pour améliorer la qualité et la fiabilité des méthodes de mesure de la fraude.

La bonne connaissance des origines-destinations (OD) sur un réseau de transport en commun est une information critique pour les planificateurs et un vecteur d'amélioration de l'offre. Cette information peut-être synthétisée à travers des matrices OD qui indiquent le nombre de déplacements allant de chaque origine à chaque destination. Une fois estimée, ces matrices peuvent venir alimenter les modèles de déplacements à quatre étapes, modèles qui restent aujourd'hui l'outil principal d'aide à la décision en matière de politique des transports. Ces matrices viennent aussi alimenter les études concernant les modifications d'itinéraires des

lignes et l'adaptation des fréquences. Dans le chapitre 3, nous avons comparé empiriquement trois méthodologies pour obtenir cette information pour le réseau de transport en commun de Lyon. Les résultats montrent que, globalement, la structure de la demande décrite avec les matrices OD est similaire dans les trois sources de données. Cela est rassurant dans le sens où les différentes méthodes de collecte et de traitement, génèrent une mesure comparable du même phénomène. Toutefois, si nous examinons en détail les résultats, il existe des divergences et des sources d'erreurs qui doivent être connues et prises en compte si ces matrices sont utilisées pour éclairer la politique de transport. Nos résultats laissent penser que l'enquête ménages déplacements sous-estime sérieusement le nombre de déplacements sur le réseau de transport en commun de Lyon pour un jour moyen. Cette recherche montre aussi que les données billettiques ne sont pas exemptes d'erreurs et qu'il est nécessaire de les combiner avec des données de comptage pour obtenir des matrices représentatives de l'usage réel du réseau. Les données passives ainsi traitées peuvent constituer un complément voir remplacer les enquêtes montées descentes réalisées sur le réseau. En effet, bien que ces enquêtes manuelles soient réputées précises, car elles collectent en face-à-face des données sur les déplacements des usagers, elles sont coûteuses et ne concernent en général qu'une seule journée. Elles vieillissent donc mal dans le temps quand l'offre évolue et ne permettent pas de capturer la variabilité temporelle de la demande de déplacement. Pris dans son ensemble, le chapitre 3 est une contribution importante qui permet une meilleure compréhension des sources de données disponibles pour l'estimation de la demande de transport en commun.

Les données billettiques permettent un suivi dans le temps des validations d'une même carte ce qui les rend particulièrement utiles pour analyser la variabilité des comportements d'usage. La bonne compréhension de cette variabilité est un prérequis pour développer et évaluer à un niveau individuel l'impact de nouvelles stratégies. Dans le chapitre 4, nous proposons une méthodologie qui permet d'explorer et de visualiser de manière fine la variabilité inter-individuelle, mais aussi de mesurer la variabilité intra-individuelle du comportement journalier de déplacement. Nos résultats empiriques suggèrent qu'il existe une grande variété de comportements d'usage quotidien des transports en commun. Le simple fait de considérer les individus comme des passagers machinaux qui font la navette du lundi au vendredi est insuffisant. De même, la segmentation des usagers uniquement sur la base de leur produit tarifaire ou de données relatives à une seule journée est incomplète. Des niveaux très distincts de variabilité intra-individuelle peuvent être trouvés dans chaque profil tarifaire et dans chaque cluster inter-individuel. Plus précisément, nos résultats indiquent qu'à Lyon, une forte proportion des cartes ont des usages intermittents et qu'il existe une porosité entre une utilisation fréquente et une utilisation occasionnelle. En d'autres termes, il n'est pas rare que les habitudes et les routines d'utilisation des transports publics changent radicalement au fil du temps. Nos résultats confirment aussi que sur des périodes de plusieurs mois, nous observons une répétitivité spatiale et temporelle des déplacements, mais qu'il existe également une variabilité intra-individuelle systématique dans l'utilisation quotidienne des transports en commun. L'analyse menée dans ce chapitre montre également que la distinction classique entre jour ouvrable, week-end et jour férié peut être pertinente pour certains utilisateurs,

mais pas pour tous. Même si, en moyenne, les comportements sont plus stables pendant les jours de semaine, nous avons constaté que la variabilité intra-individuelle ne diminue que marginalement lorsqu'elle est calculée uniquement sur les jours ouvrables. De même, nous avons observé que les périodes de vacances n'ont une influence que sur une proportion réduite des utilisateurs. Le paradigme traditionnel de la planification des transports en commun centré sur un ensemble de jours types tels que les jours ouvrables, le samedi ou le dimanche n'est donc pas valable du point de vue individuel. Il n'existe donc pas d'approche universelle pour comprendre la variabilité quotidienne de l'utilisation des transports en commun. Sur de longues périodes, chaque comportement individuel de mobilité est unique.

Pouvoir prévoir la demande de transport est une étape incontournable d'une démarche de planification des transports en commun. Depuis l'introduction des systèmes billettiques et de comptage automatiques, les opérateurs stockent des données de fréquentation détaillées qui peuvent fournir la matière première de l'exercice de prévision. Dans le chapitre 5, nous proposons un outil de prévision répondant à la planification tactique. Cet outil s'appuie uniquement sur les données historiques collectées automatiquement. Notre objectif est de pouvoir prévoir les fréquentations à 1 an à un niveau désagrégé tant sur le plan spatial que temporel (par jour calendaire/heure/ligne). Une fois les prévisions de bas niveau construites une agrégation ascendante permet d'analyser la fréquentation attendue à tous les niveaux du réseau de transport en commun. L'approche de modélisation que nous proposons, combine via une décomposition multiplicative, une projection de tendance et des arbres de régression qui apprennent automatiquement des relations non linéaires complexes entre la fréquentation et les variables explicatives. Les résultats empiriques montrent que cette approche permet de diminuer significativement les erreurs par rapport à des approches plus naïves. L'application de ces méthodes à l'ensemble des lignes du réseau fournit un outil puissant de planification tactique. Cet outil est utile pour prévoir la fréquentation globale du réseau et estimer les futures recettes, pour évaluer l'impact des modifications d'offre sur la fréquentation d'une ligne, ou bien pour simuler l'impact de travaux ou d'interruption de service. De manière générale, c'est un outil puissant de diagnostic et de suivi dynamique de la fréquentation du réseau. L'outil permet d'appuyer les décisions concernant les modifications de l'offre de transport. Il est donc recommandé que les opérateurs complètent leurs outils traditionnels de reporting par ce type d'outil prédictif qui valorise leur patrimoine de données. De manière générale, les résultats de ce chapitre confirment que des méthodes statistiques plus sophistiquées peuvent faciliter la planification tactique et complémenter des analyses qui reposaient auparavant principalement sur l'expérience des planificateurs et le jugement d'expert.

6.2 Quelques applications plus opérationnelles

Ces recherches montrent que les données passives peuvent permettre une mesure plus précise de l'usage du réseau et de sa performance. L'application des méthodes

présentées dans cette thèse permet de produire des informations détaillées sur l'usage passé et futur du réseau, sa performance, mais aussi sur les comportements des clients. Une fois les algorithmes déployés dans les systèmes d'information ces informations peuvent être produites de manière continue et à moindre coût. L'objectif de cette section est donc d'illustrer quelques applications pratiques de ces algorithmes une fois directement connectés sur les bases de données de l'opérateur.

- **Suivi des comportements d'usages pendant la crise du covid**

Il va sans dire que la crise de la COVID est un événement majeur pour le secteur du transport en commun. Cette crise nécessite que les réseaux s'adaptent et réagissent très rapidement à des changements brutaux de la demande de déplacement. L'usage des données passives s'est donc avéré encore plus critique pour comprendre la nouveauté de la situation et surtout son évolution. Les opérateurs ont par exemple besoin de bien comprendre les profils d'usages pendant et après la crise pour s'assurer de la pertinence des dispositifs de confinement et pour appuyer les actions de marketing et de reconquête des clients.

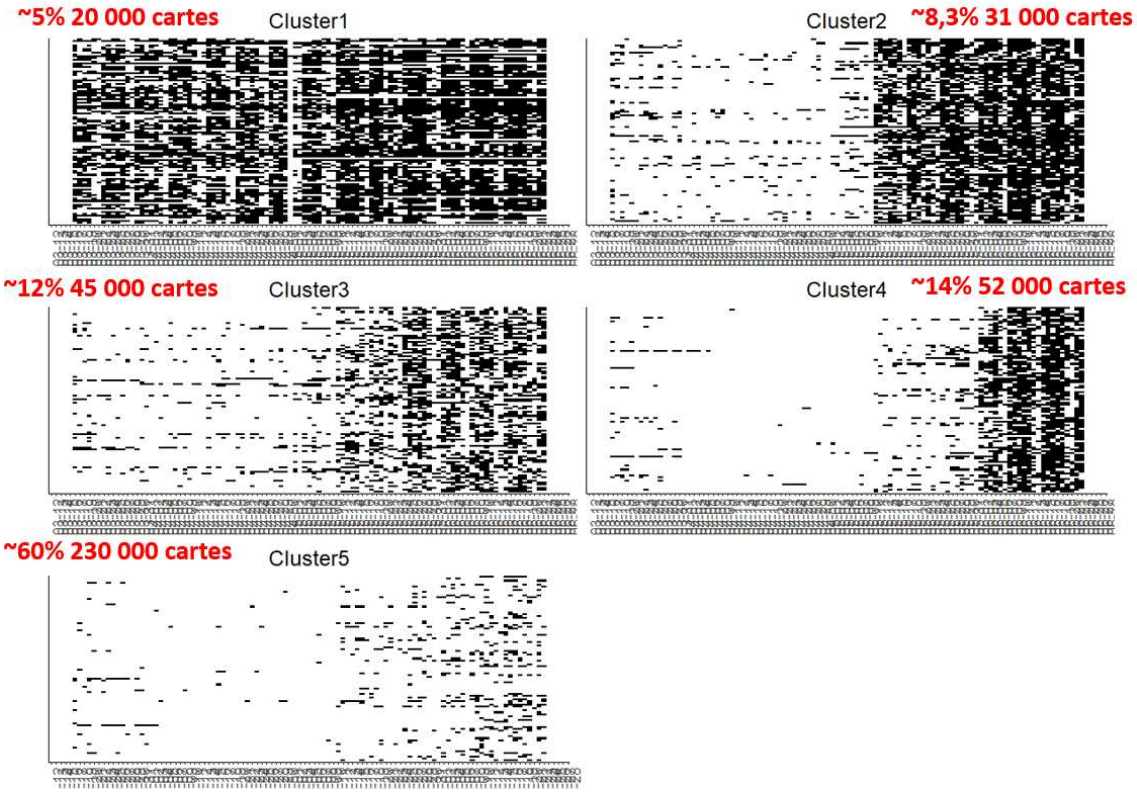


FIGURE 6.1 – Application opérationnelle de suivi des comportements d'usages pendant la crise de la COVID-19 entre le 17 mars 2020 et le 24 juin 2020, Source : Auteur

Pour répondre à ces objectifs, une application typique est d'utiliser la méthode de clustering intra-individuelle proposée au chapitre 4. Cette méthode est appliquée sur la période allant du mardi 17 mars 2020 (date de début du confinement) au mercredi 24 juin 2020 (phase 3 du déconfinement). Durant cette période, 370 007 cartes distinctes ont effectué au moins une validation sur le réseau. À titre de

comparaison, pour la période d'avril 2019, plus de 450 000 cartes différentes avaient utilisé le réseau. Une synthèse graphique des résultats après division en cinq clusters se trouve dans la figure 6.1. Ces résultats permettent d'identifier les grandes familles de comportement durant la pandémie de coronavirus à Lyon. Seulement 20 000 cartes (cluster 1) ont continué à utiliser le réseau fréquemment pendant le confinement. Les clusters 2 et 3 correspondent à une reprise de l'usage du transport en commun plus ou moins intense à partir du 11 mai 2020 (phase 1 du déconfinement). Le cluster 4 correspond lui à une reprise de l'usage à partir du mardi 2 juin 2020 c'est-à-dire dès la phase 2 du déconfinement. Enfin, le cluster 5 est le cluster le plus important en volume, car il représente 230 000 cartes qui présentent des habitudes d'usages occasionnels mêmes si nous pouvons noter une intensification des jours de présence avec la progression du déconfinement et la diminution du risque sanitaire.

- **Enquête origine destination dynamique**

Les données billettiques permettent de connaître les lieux de montées. Afin d'enrichir ces données, un algorithme de reconstitution des itinéraires et des déplacements (voir chapitre 3) est nécessaire. Sans rentrer dans les détails, la figure 6.2 donne une vue simplifiée de l'architecture nécessaire pour réaliser les traitements permettant de produire des enquêtes origine destination dynamiques c'est-à-dire en continu. La chaîne de traitement comprend un certain nombre

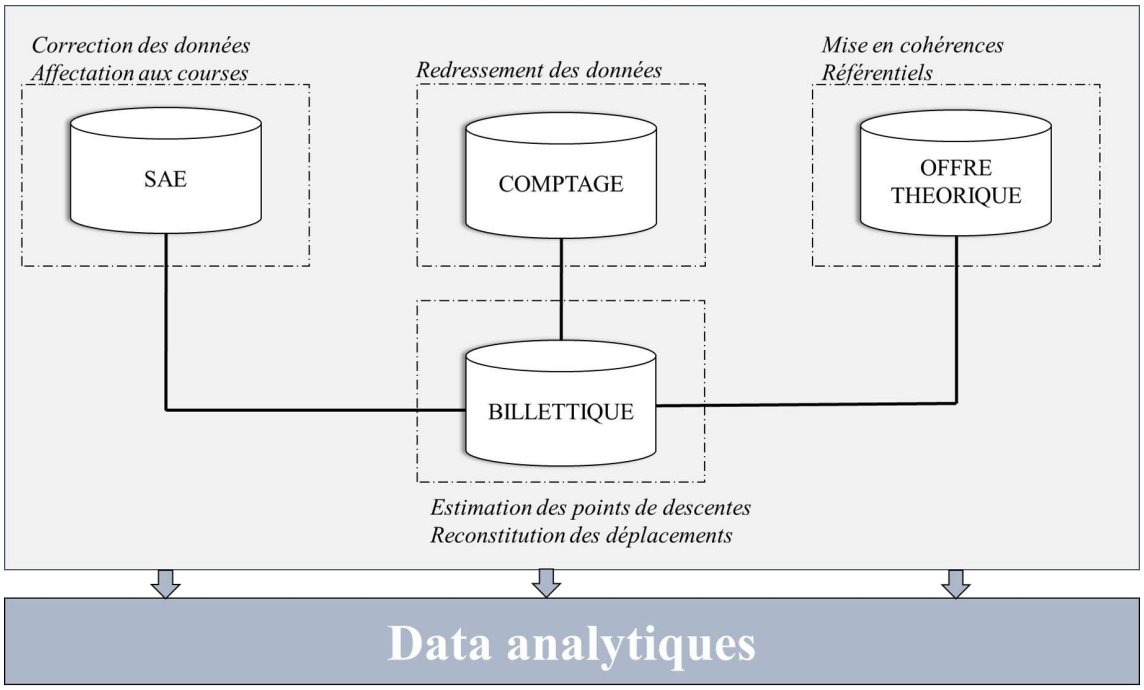


FIGURE 6.2 – Architecture de données simplifiée pour la production d'enquêtes origine-destination dynamiques, Source : Auteur

d'étapes plus ou moins complexes dont nous ne reprenons que les éléments les plus importants. Tout d'abord, la mise en commun de la billettique, du SAE et de l'offre théorique permet de corriger les données et de déterminer, pour chaque validation, la position du véhicule et la liste des arrêts desservis par ce véhicule

(potentiel arrêt de descente). Ensuite, par analyse des validations successives, il est possible d'estimer des points de descentes probables pour les validations cartes. Après application de l'algorithme, pour 80 % des validations cartes, un arrêt de descente est estimé. Suite à cela, les validations de chaque carte sont regroupées au sein de déplacements avec des règles spatio-temporelles. Enfin, les résultats sont redressés avec les comptages en utilisant des méthodes plus ou moins complexes comme le redressement par facteur ou la fusion par balancement itératif. Cette étape est cruciale, car il faut prendre en compte, les cartes dont on ne peut pas estimer la destination, les tickets papier, la non-validation et la fraude dure. Ces traitements doivent se faire de manière désagrégée. Cette procédure est donc complexe tant au niveau algorithmique qu'au niveau du volume de données à traiter. Elle permet par contre de mettre en relation la demande (billettique + comptage) avec l'offre réelle (SAE). Les résultats de cette procédure ont été validés empiriquement à l'échelle macroscopique par comparaison avec l'enquête ménages déplacements et les enquêtes origine destination de Keolis Lyon (voir chapitre 3). L'algorithme est développé en python [Van Rossum et al., 2007] et connecté directement aux bases de données de l'opérateur. Le temps de traitement pour une journée de données est d'environ 30 minutes. L'interrogation des résultats permet plusieurs applications opérationnelles et des économies importantes sur les coûts d'enquêtes.

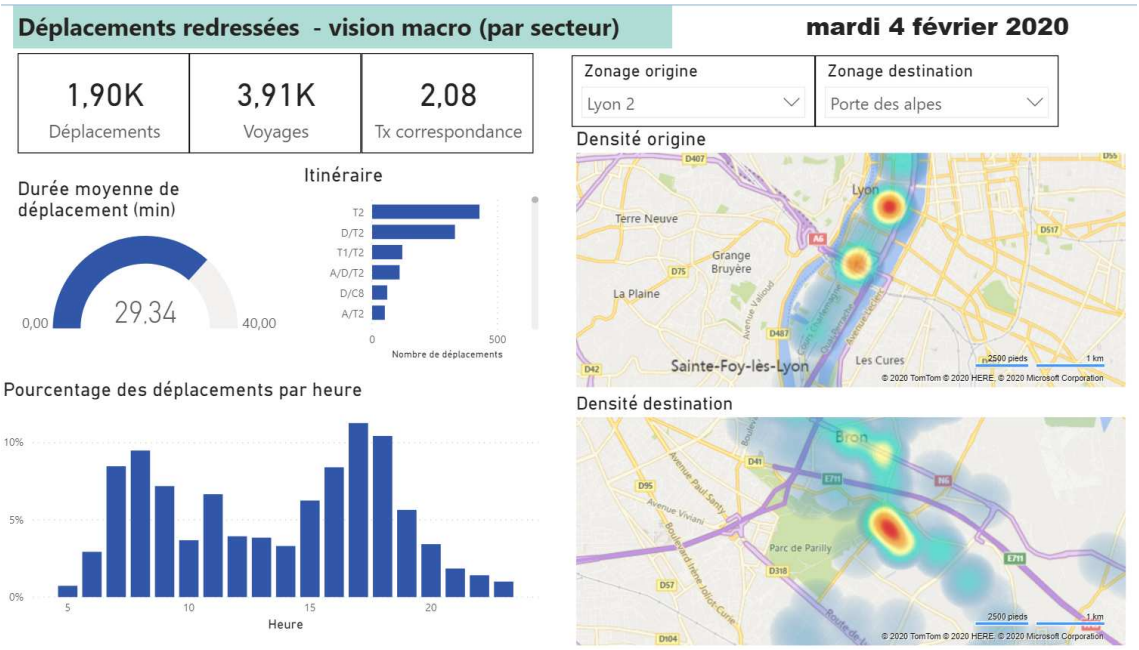


FIGURE 6.3 – Outil d'interrogation des flux de déplacements zone à zone à l'échelle du réseau de Lyon, Source : Auteur

Tout d'abord, il est possible d'obtenir les flux de déplacement horodatés d'arrêt à arrêt du réseau. L'agrégation à l'échelle spatio-temporelle souhaitée par l'analyste de ces déplacements permet d'obtenir des matrices OD. L'ensemble des indicateurs classiques peuvent ensuite être dérivés, tels que le temps de parcours moyen, la distance de déplacement, la distribution des déplacements sur les lignes (itinéraires), le taux de correspondance, etc. Ces informations doivent ensuite venir alimenter les décisions concernant les réorganisations de l'offre de transport au

niveau macroscopique. En effet, l'objectif ici est de fournir, à partir des données passives, des informations sur la demande agrégée au niveau réseau. Les données ainsi enrichies peuvent aussi venir compléter les enquêtes pour des jours/périodes atypiques voir les remplacer complètement. Un exemple de Dashboard permettant d'analyser interactivement ces résultats est donné en figure 6.3. L'analyste peut sélectionner une zone d'origine (ici Lyon 2), une zone de destination (ici Porte des Alpes) et connaître l'ensemble des éléments concernant les déplacements entre ces zones : nombre de déplacements, nombre de voyages, taux de correspondance, distribution horaire, itinéraires populaires, durée de déplacement et densité spatiale à l'origine et à la destination.

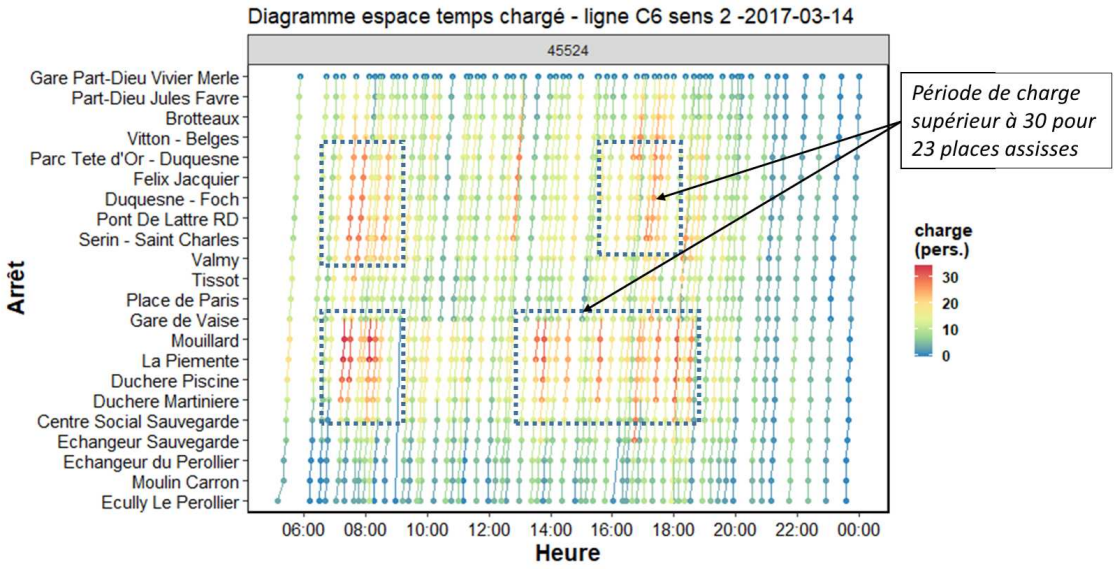


FIGURE 6.4 – Exemple d'application opérationnelle de reconstitution des charges par course pour la ligne C6 d'Écully à Part-Dieu le 14 mars 2017, Source : Auteur

Puisque les traitements présentés ci-dessus se font sur les données les plus fines, il est aussi possible d'utiliser les résultats pour effectuer des analyses par ligne. En effet, les données enrichies permettent d'analyser les comportements individuels des cartes sur le réseau et par redressement et agrégation d'estimer le comportement global. Plus précisément, pour les cartes l'arrêt de montée et de descente est connu sur chaque ligne. Il est donc possible d'estimer le temps passé par chaque individu dans le véhicule, mais aussi d'estimer pour chaque inter arrêt le nombre de personnes dans le véhicule (la charge). L'objectif ici est de confronter l'offre à la demande à des niveaux fins. La figure 6.4 est un exemple d'analyse. Cette figure est un graphique espace-temps chargé de la ligne C6 dans le sens Écully-Part Dieu. Plus précisément, chaque course enregistrée dans le SAE est représentée sous forme d'une ligne verticale. Chaque point correspond au stop horodaté d'un véhicule à un arrêt et est coloré en fonction du nombre de personnes estimé dans le véhicule à cet instant. Ce type de graphique permet de comprendre la dynamique d'utilisation de la ligne C6, mais aussi de montrer la quantité d'informations que nous pouvons extraire en continu des données passives. En effet, à partir de ces informations, un certain nombre d'indicateurs peuvent être calculés afin de mesurer l'expérience client sur le réseau. Ces indicateurs sont dits orientés client, car ils mesurent la qualité du service tel que vécue par le client (temps de parcours, durée des correspondance, taux de charge,

temps de parcours avec charge critique, etc.). Ils peuvent être calculés par ligne et mis à disposition sous forme de Dashboard synthétique permettant de suivre au cours du temps l'adéquation offre/demande d'une ligne et ainsi justifier/hiérarchiser des adaptations et des renforts d'offres. Pour ce faire, ces indicateurs doivent être couplés avec la définition de seuil d'alerte permettant l'identification automatique des situations critiques. Par exemple, un ratio passager/place assise supérieur à 160 % peut être considéré comme une situation d'inconfort. De même, un temps moyen de correspondance supérieure à 15 minutes entre deux lignes peut être considéré comme inacceptable (selon le volume de correspondance).

- **Outil opérationnel d'analyse et de prédiction du trafic journalier**

La fréquentation d'un réseau de transport est un phénomène complexe, influencé par de nombreuses variables comme la performance opérationnelle, la distribution spatiale des activités, les évolutions de l'usage du sol, les changements tarifaires, etc. C'est aussi un phénomène qui fluctue selon des dynamiques temporelles comme l'agencement des jours fériés, les vacances scolaires ou bien les heures de la journée. Cependant, les exploitants doivent être en capacité d'estimer le trafic de l'année à venir pour plusieurs raisons opérationnelles :

— Définir une offre adaptée ;
— Prévoir les recettes et construire des budgets optimisés ;
— Affecter les ressources humaines ;
— Piloter l'activité par les données.

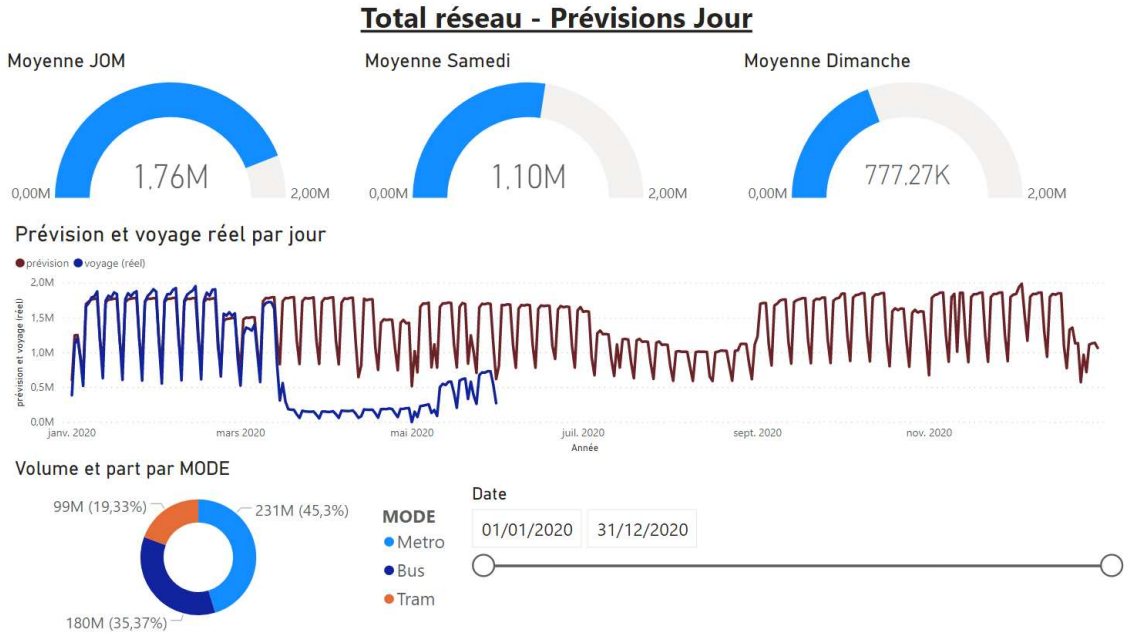


FIGURE 6.5 – Exemple d'outil opérationnel d'analyse et de prévision de la fréquentation, Source : Auteur

Pour répondre à ces besoins concrets, nous avons construit un script R [R Core Team et al., 2013] automatisant le modèle RFT (voir chapitre 5) directement depuis les bases de données de l'opérateur du réseau TCL. Le script s'appuie sur l'historique de la fréquentation journalière de chaque ligne et s'assure

que les données sont complètes et cohérentes afin d'effectuer pour chaque ligne du réseau une prévision pour les 365 jours suivants. Les résultats sont ensuite intégrés dans un modèle conceptuel de données cohérent et restitués de manière opérationnelle grâce à un Dashboard de business intelligence. Une partie de ce Dashboard est illustré en figure 6.5. Cet outil permet de connaître la fréquentation passée et future sur les périodes temporelles et les lignes choisies. L'utilisateur peut facilement procéder à des agrégations temporelles et des sélections de lignes. L'outil calcule automatiquement, les indicateurs clefs tels que la fréquentation d'un jour ouvrable moyen prévue et passée. L'outil répond à un besoin de simplicité et de fluidité dans l'analyse des données désagrégées de fréquentation. Dans l'exemple en figure 6.5, les modèles sont entraînés sur la période 2014-2019 et effectuent une prédiction pour chaque jour de l'année 2020. La courbe de prévision journalière en rouge est très proche de la fréquentation réelle jusqu'au confinement dont les effets n'ont évidemment pas été modélisés. Les prévisions ainsi construites offrent une vision de la fréquentation attendue sans la crise sanitaire du coronavirus (scénario au fil de l'eau). L'outil permet ainsi d'estimer précisément par ligne et par jour le nombre de voyages perdu durant la crise sanitaire, mais aussi de déterminer précisément quand la fréquentation aura retrouvé un niveau similaire à celui attendu sans la crise de la covid.

6.3 Limites et perspectives de recherche

Une thèse de doctorat est un projet de recherche limité dans le temps alors que l'exploitation des données passives est un domaine de recherche en pleine ébullition. Cela signifie que toutes les questions éventuelles liées au sujet n'ont pas pu être résolues. Toutefois, ces lacunes peuvent être un point de départ pour des travaux futurs. Il est donc important de les identifier et de donner des perspectives qui pourront être approfondies lors de recherches ultérieures.

L'augmentation des volumes de données, la multiplication des sources et la complexification des systèmes d'information nécessite de poursuivre les efforts méthodologiques afin de développer des algorithmes plus performants et plus flexibles. Dans le futur, une attention particulière devra être portée à la standardisation des méthodes et des structures des bases de données passives afin de faciliter la diffusion et la réutilisation des algorithmes. Les efforts concernant le développement d'outils de visualisation doivent aussi se poursuivre afin de mieux saisir les dynamiques spatio-temporelles contenues dans ces données et faciliter leur valorisation opérationnelle.

Les données passives peuvent permettre d'améliorer la mesure de la fraude et devront dans le futur être mises à profit pour mieux comprendre les facteurs influençant ce phénomène. Pour ce faire, il convient de développer des modèles explicatifs testant l'impact de différentes variables sur l'évolution du taux de fraude. Ces variables peuvent être de nature temporelle, spatiale, liées à des changements de tarification, à l'usage de méthodes spécifiques de contrôle ou bien à la réalisation de campagnes de communication. Ces modèles pourront ainsi aider à une meilleure allocation du

personnel, une meilleure compréhension de la nature de ce phénomène et la bonne évaluation des stratégies de lutte contre la fraude.

Les travaux de cette thèse montrent qu'il est impératif de procéder à des comparaisons et des croisements entre les sources de données pour améliorer le dispositif de collecte et d'analyse. Une direction de recherche importante est donc de comparer les données collectées par le système de transport en commun avec d'autres données passives émergentes par exemple, les données de requête d'itinéraire, les données GPS smartphone, les données de téléphonie mobile, les données wifi ou bien les données Bluetooth. La mise en commun de ces sources permettrait ainsi d'identifier des biais potentiels. Par ailleurs, la combinaison avec des données passives d'autres systèmes de transport comme les vélos en libre-service, les taxis, les trottinettes en flotte libre apporterait des éclairages sur les interactions entre les modes de transport. Enfin, le développement de méthodes de fusion et de croisement avec des données externes d'usage du sol, de météo ou bien d'évènements sont des pistes intéressantes pour enrichir les données et la compréhension des phénomènes.

Concernant la prédiction de la fréquentation, la méthodologie que nous proposons doit être enrichie pour mieux prendre en compte l'influence de l'offre et des évènements ponctuels sur la fréquentation. La procédure d'agrégation des modèles doit aussi être améliorée dans le but de réconcilier de manière optimale des approches de modélisation à différents niveaux. Cela permettrait d'augmenter la robustesse de l'approche de prévision. Il convient aussi de proposer des méthodes permettant de détecter de manière automatique des périodes anormales. Ces méthodes de détection de données aberrantes seraient particulièrement utiles pour identifier des évènements atypiques, mais aussi pour assurer que les modèles apprennent sur des jeux de données d'entraînement représentatifs du comportement normal de la ligne. Il va sans dire que les travaux futurs doivent aussi explorer d'autres horizons de prédiction et d'autres objectifs de prédiction. L'horizon court terme nous semble particulièrement pertinent pour améliorer l'information client par exemple en développant des méthodes de prédiction de la charge en temps réel.

Les données billettiques une fois enrichies permettent d'envisager une pléthore d'analyses concernant les comportements individuels. Parmi les analyses possibles mieux comprendre les choix d'itinéraires sur un réseau de transport en commun est une piste à explorer. Notamment, il est possible d'étudier pour une même paire origine-destination la variabilité intra-individuelle du choix d'itinéraires et la mettre en relation avec des niveaux de charge, le confort du véhicule, la variabilité observée des temps de parcours, des temps d'attentes, des temps de correspondance... Le développement d'indicateurs de performance orientés client est aussi à étudier pour justifier des améliorations d'offres plus pertinentes. Par exemple, la mesure en continu de la charge sur les lignes du réseau couplée avec des indicateurs de qualité de l'offre réelle doit permettre d'évaluer la pertinence des stratégies opérationnelles de régulation, de détecter les périodes où l'offre n'est pas optimale et donc d'agir en connaissance de cause. Plus généralement, l'usage plus fréquent d'indicateurs orientés clients doit alimenter des changements de paradigme dans les contrats de délégation de service. Jusqu'ici, les mécanismes

incitatifs se basent surtout sur la performance de l'exploitant et encore trop peu sur l'expérience vécue par les passagers. Celle-ci devrait pourtant être au coeur des préoccupations.

Table des figures

Liste des tableaux

Bibliographie

B. Agard, C. Morency, and M. Trépanier. Mining public transport user behaviour from smart card data. *IFAC Proceedings Volumes*, 39(3) :399–404, 2006.

A. Alsger, B. Assemi, M. Mesbah, and L. Ferreira. Validating and improving public transport origin–destination estimation algorithm using smart card fare data. *Transportation Research Part C : Emerging Technologies*, 68 :490–506, jul 2016. doi : 10.1016/j.trc.2016.05.004.

A. Alsger, A. Tavassoli, M. Mesbah, L. Ferreira, and M. Hickman. Public transport trip purpose inference using smart card fare data. *Transportation Research Part C : Emerging Technologies*, 87 :123–137, feb 2018. doi : 10.1016/j.trc.2017.12.016.

C. Anderson. The end of theory : The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7) :16–07, 2008.

P. Ardilly. *Les techniques de sondage*. Editions Technip, 2006.

M. Bagchi and P. R. White. The potential of public transport smart card data. *Transport Policy*, 12(5) :464–474, 2005.

B. Barabino and S. Salis. Moving towards a more accurate level of inspection against fare evasion in proof-of-payment transit systems. *Networks and Spatial Economics*, pages 1–28, 2019.

B. Barabino, S. Salis, and B. Useli. A modified model to curb fare evasion and enforce compliance : Empirical evidence and implications. *Transportation Research Part A : Policy and Practice*, 58 :29–39, 2013.

B. Barabino, S. Salis, and B. Useli. Fare evasion in proof-of-payment transit systems : Deriving the optimum inspection level. *Transportation Research Part B : Methodological*, 70 :1–17, 2014.

B. Barabino, S. Salis, and B. Useli. What are the determinants in making people free riders in proof-of-payment transit systems ? evidence from italy. *Transportation Research Part A : Policy and Practice*, 80 :184–196, 2015.

J. Barry, R. Newhouser, A. Rahbee, and S. Sayeda. Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record : Journal of the Transportation Research Board*, (1817) :183–187, 2002.

A. Bhaskar, E. Chung, et al. Passenger segmentation using smart card data. *IEEE Transactions on intelligent transportation systems*, 16(3) :1537–1548, 2015.

A. Bonnafous and H. Puel. *Physionomies de la ville*, volume 8. Editions de l'Atelier, 1983.

P. Bonnel. *Prévision de la demande de transport*. Presses de l'École Nationale des Ponts et Chaussées, Paris, 425p, 2002.

P. Bonnel and M. A. Munizaga. Transport survey methods-in the era of big data facing new and old challenges. *Transportation Research Procedia*, 32 :1–15, 2018.

C. Boyd, C. Martini, J. Rickard, and A. Russell. Fare evasion and non-compliance : A simple model. *Journal of Transport Economics and Policy*, pages 189–197, 1989.

D. Boyd and K. Crawford. Critical questions for big data : Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5) :662–679, 2012.

D. K. Boyle. *Fixed-route transit ridership forecasting and service planning methods*, volume 66. Transportation Research Board, 2006.

L. Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.

A.-S. Briand, E. Côme, M. Trépanier, and L. Oukhellou. Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transportation Research Part C : Emerging Technologies*, 79 :274–289, 2017.

C. Buchta, M. Kober, I. Feinerer, and K. Hornik. Spherical k-means clustering. *Journal of Statistical Software*, 50(10) :1–22, 2012.

L. Cai and Y. Zhu. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14, 2015.

M. Callegaro and Y. Yang. The role of surveys in the era of "big data". In *The Palgrave handbook of survey research*, pages 175–192. Springer, 2018.

D. Cardon. *A quoi rêvent les algorithmes. Nos vies à l'heure : Nos vies à l'heure des big data*. Le Seuil, 2015.

A. Ceder. *Public transit planning and operation : Modeling, practice and behavior*. CRC press, 2016.

A. Ceder and N. H. Wilson. Bus network design. *Transportation Research Part B : Methodological*, 20(4) :331–344, 1986.

CERTU. *L'enquête ménages déplacements standard CERTU*. CERTU, Lyon, 2008.

R. Chapleau, M. Trépanier, and K. K. Chu. The ultimate survey for transit planning : Complete information with smart card data and gis. In *Proceedings of the 8th International Conference on Survey Methods in Transport : Harmonisation and Data Comparability*, pages 25–31, 2008.

R. Chapleau, P. Gaudette, and T. Spurr. Strict and deep comparison of revealed transit trip structure between computer-assisted telephone interview household travel survey and smart cards. *Transportation research record*, 2672(42) :13–22, 2018.

C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C : emerging technologies*, 68 :285–299, 2016.

R. V. Clarke. Fare evasion and automatic ticket collection on the london underground. *Crime Prevention Studies*, 1 :135–146, 1993.

R. V. Clarke, S. Contre, and G. Petrossian. Deterrence and fare evasion : Results of a natural experiment. *Security Journal*, 23(1) :5–17, 2010.

M. Coleman, L. Tarte, S. Chau, B. Levine, and A. Reddy. A data-driven approach to prioritizing bus schedule revisions at new york city transit. *Transportation Research Record*, 2672(8) :86–95, 2018.

H. Commenges. *L'invention de la mobilité quotidienne. Aspects performatifs des instruments de la socio-économie des transports*. PhD thesis, Université Paris-Diderot-Paris VII, 2013.

J. Correa, T. Harks, V. J. Kreuzen, and J. Matuschke. Fare evasion in transit networks. *Operations Research*, 65(1) :165–183, 2017.

L. Dauby and Z. Kovacs. Fare evasion in light rail systems. *Transportation Research E-Circular*, (E-C112), 2007.

A. Delbosc and G. Currie. Four types of fare evasion : A qualitative study from melbourne, australia. *Transportation Research Part F : Traffic Psychology and Behaviour*, 43 :254–264, 2016.

A. Delbosc and G. Currie. Why do people fare evade ? a global shift in fare evasion research. *Transport Reviews*, pages 1–16, jun 2018.

F. M. Delle Fave, A. X. Jiang, Z. Yin, C. Zhang, M. Tambe, S. Kraus, and J. P. Sullivan. Game-theoretic patrolling with dynamic execution uncertainty and a case study on a real transit system. *Journal of Artificial Intelligence Research*, 50 :321–367, 2014.

E. Deschaintres, C. Morency, and M. Trépanier. Analyzing transit user behavior with 51 weeks of smart card data. *Transportation Research Record*, page 0361198119834917, 2019.

F. Devillaine, M. Munizaga, and M. Trépanier. Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record : Journal of the Transportation Research Board*, 2276(1) :48–55, jan 2012. doi : 10.3141/2276-06.

E. Diab, D. Kasraian, E. J. Miller, and A. Shalaby. Current state of practice in transit ridership prediction : Results from a survey of canadian transit agencies. *Transportation Research Record*, page 0361198119841858, 2019.

L. Einav and J. Levin. The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1) :1–24, 2014.

G. J. Fielding. *Managing public transit strategically. A comprehensive approach to strengthening service and monitoring performance*. 1987.

M. Frické. Big data and its epistemology. *Journal of the Association for Information Science and Technology*, 66(4) :651–661, 2015.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

J. H. Friedman. Greedy function approximation : a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

P. G. Furth, J. G. Strathman, and B. Hemily. Making automatic passenger counts mainstream : Accuracy, balancing algorithms, and data structures. *Transportation research record*, 1927(1) :206–216, 2005.

T. Gärling and K. W. Axhausen. Introduction : Habitual travel choice. *Transportation*, 30(1) :1–11, 2003.

J. B. Gordon, H. N. Koutsopoulos, N. H. M. Wilson, and J. P. Attanucci. Automated inference of linked transit journeys in london using fare-transaction and vehicle location data. *Transportation Research Record : Journal of the Transportation Research Board*, 2343(1) :17–24, jan 2013. doi : 10.3141/2343-03.

J. B. Gordon, H. N. Koutsopoulos, and N. H. Wilson. Estimation of population origin–interchange–destination flows on multimodal transit networks. *Transportation Research Part C : Emerging Technologies*, 90 :350–365, may 2018. doi : 10.1016/j.trc.2018.03.007.

G. Goulet-Langlois, H. N. Koutsopoulos, and J. Zhao. Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C : Emerging Technologies*, 64 :1–16, 2016.

G. Goulet-Langlois, H. N. Koutsopoulos, Z. Zhao, and J. Zhao. Measuring regularity of individual travel patterns. *IEEE Transactions on Intelligent Transportation Systems*, 19(5) :1583–1592, 2018.

P. Guarda, P. Galilea, S. Handy, J. C. Munoz, and J. d. D. Ortúzar. Decreasing fare evasion without fines ? a microeconomic analysis. *Research in Transportation Economics*, 59 :151–158, 2016a.

P. Guarda, P. Galilea, L. Paget-Seekins, and J. d. D. Ortúzar. What is behind fare evasion in urban bus systems ? an econometric approach. *Transportation Research Part A : Policy and Practice*, 84 :55–71, 2016b.

M.-P. Hamel and D. Marguerit. Analyse des big data. quels usages, quels défis ? *CGSP, La note d'analyse*, (8), 2013.

J. Hanft, S. Iyer, B. Levine, and A. Reddy. Transforming bus service planning using integrated electronic data sources at nyc transit. *Journal of Public Transportation*, 19(2) :6, 2016.

S. Hanson and J. Huff. Classification issues in the analysis of complex travel behavior. *Transportation*, 13(3) :271–293, 1986.

S. Hanson and J. O. Huff. Assessing day-to-day variability in complex travel patterns. *Transportation Research Record*, 891 :18–24, 1981.

S. Hanson and O. J. Huff. Systematic variability in repetitious travel. *Transportation*, 15(1-2) :111–135, 1988.

A. R. Hauber. Fare evasion in a european perspective. *Studies on Crime and Crime Prevention*, 2 :122–141, 1993.

J. O. Huff and S. Hanson. Repetition and variability in urban travel. *Geographical Analysis*, 18(2) :97–114, 1986.

R. J. Hyndman and G. Athanasopoulos. *Forecasting : principles and practice*. OTexts, 2018.

R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang. Optimal combination forecasts for hierarchical time series. *Computational statistics & data analysis*, 55(9) :2579–2589, 2011.

P. Jones and M. Clarke. The significance and measurement of variability in travel behaviour. *Transportation*, 15(1-2) :65–87, 1988.

P. Jones and P. R. Stopher. *Transport survey quality and innovation*. Emerald Group Publishing Limited, 2003.

K. B. Kahn. Revisiting top-down versus bottom-up forecasting. *The Journal of Business Forecasting*, 17(2) :14, 1998.

M. Killias, D. Scheidegger, and P. Nordenson. The effects of increasing the certainty of punishment. *European Journal of Criminology*, 6(5) :387–400, jul 2009.

King County Department of Transportation. Report on fare evasion on metro transit, april 2010. 2010.

R. Kitchin. Big data, new epistemologies and paradigm shifts. *Big data & society*, 1(1) :2053951714528481, 2014a.

R. Kitchin. The real-time city ? big data and smart urbanism. *GeoJournal*, 79(1) : 1–14, 2014b.

R. Kitchin and T. P. Lauriault. Small data in the era of big data. *GeoJournal*, 80 (4) :463–475, 2015.

R. Kitchin and G. McArdle. What makes big data, big data ? exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 3(1) : 2053951716631130, 2016.

P. Kooreman. Fare evasion as a result of expected utility maximisation : some empirical support. *Journal of Transport Economics and Policy*, pages 69–74, 1993.

H. N. Koutsopoulos, Z. Ma, P. Noursalehi, and Y. Zhu. Transit data analytics for planning, monitoring, control, and information. In *Mobility Patterns, Big Data and Transport Analytics*, pages 229–261. Elsevier, 2019.

F. Kurauchi and J.-D. Schmöcker. *Public Transport Planning with Smart Card Data*. CRC Press, 2017.

T. Kusakabe and Y. Asakura. Behavioural data mining of transit smart card data : A data fusion approach. *Transportation Research Part C : Emerging Technologies*, 46 :179–191, 2014.

A. Labrinidis and H. V. Jagadish. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12) :2032–2033, 2012.

T. F. Larwin. *Off-board fare payment using proof-of-payment verification*, volume 96. Transportation Research Board, 2012.

J. Lee. Uncovering san francisco, california, muni's proof-of-payment patterns to help reduce fare evasion. *Transportation Research Record : Journal of the Transportation Research Board*, (2216) :75–84, 2011.

T. Li, D. Sun, P. Jing, and K. Yang. Smart card data mining of public transport destination : A literature review. *Information*, 9(1) :18, jan 2018. doi : 10.3390/info9010018.

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

Z. Ma, J. Xing, M. Mesbah, and L. Ferreira. Predicting short-term bus passenger demand using a pattern hybrid approach. *Transportation Research Part C : Emerging Technologies*, 39 :148–163, 2014.

E. Manley, C. Zhong, and M. Batty. Spatiotemporal variation in travel regularity through transit user profiling. *Transportation*, 45(3) :703–732, 2018.

A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton. Big data : the management revolution. *Harvard business review*, 90(10) :60–68, 2012.

C. Morency, M. Trépanier, and B. Agard. Measuring transit use variability with smart-card data. *Transport Policy*, 14(3) :193–203, 2007.

E. Morin. *Introduction à la pensée complexe*. Le Seuil, 2015.

J. A. Morris and M. J. Gardner. Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British Medical Journal (Clinical Research Edition)*, 296(6632) :1313–1316, 1988.

I. Multisystems, I. Mundle & Associates, and I. Parsons Transportation Group. *A Toolkit for Self-service, Barrier-free Fare Collection*, volume 80. TRB, Washington D.C., 2002.

M. Munizaga, F. Devillaine, C. Navarrete, and D. Silva. Validating travel behavior estimated from smartcard data. *Transportation Research Part C : Emerging Technologies*, 44 :70–79, jul 2014a. doi : 10.1016/j.trc.2014.03.008.

M. Munizaga, F. Devillaine, C. Navarrete, and D. Silva. Validating travel behavior estimated from smartcard data. *Transportation Research Part C : Emerging Technologies*, 44 :70–79, 2014b.

M. A. Munizaga and C. Palma. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. *Transportation Research Part C : Emerging Technologies*, 24 :9–18, oct 2012. doi : 10.1016/j.trc.2012.01.007.

N. Nassir, M. Hickman, and Z.-L. Ma. Activity detection and transfer identification for public transit fare card data. *Transportation*, 42(4) :683–705, 2015.

P. Noursalehi, H. N. Koutsopoulos, and J. Zhao. Real time transit demand prediction capturing station interactions and impact of special events. *Transportation Research Part C : Emerging Technologies*, 97 :277–300, 2018.

A. A. Nunes, T. G. Dias, and J. F. e Cunha. Passenger journey destination estimation from automated fare collection system data using spatial validation. *IEEE Transactions on Intelligent Transportation Systems*, 17(1) :133–142, jan 2016. doi : 10.1109/tits.2015.2464335.

É. Ollion and J. Boelaert. Au-delà des big data. les sciences sociales et la multiplication des données numériques. *Sociologie*, (3, vol. 6), 2015.

D. Opitz and R. Maclin. Popular ensemble methods : An empirical study. *Journal of artificial intelligence research*, 11 :169–198, 1999.

J. d. D. Ortúzar and L. Willumsen. *Modelling transport*. John wiley & sons, 2011.

E. Pas. Multiday samples, parameter estimation precision, and data collection costs for least squares regression trip-generation models. *Environment and Planning A*, 18(1) :73–87, 1986.

E. I. Pas. Intrapersonal variability and model goodness-of-fit. *Transportation Research Part A : General*, 21(6) :431–438, 1987.

E. I. Pas and F. S. Koppelman. An examination of the determinants of day-to-day variability in individuals' urban travel behavior. *Transportation*, 14(1) :3–20, 1987.

N. Paulley, R. Balcombe, R. Mackett, H. Titheridge, J. Preston, M. Wardman, J. Shires, and P. White. The demand for public transport : The effects of fares, quality of service, income and car ownership. *Transport policy*, 13(4) :295–306, 2006.

B. Pearsall. Predictive policing : The future of law enforcement. *National Institute of Justice Journal*, 266(1) :16–19, 2010.

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

C. Pineda, D. Schwarz, and E. Godoy. Comparison of passengers' behavior and aggregate demand levels on a subway system using origin-destination surveys and smartcard data. *Research in Transportation Economics*, 59 :258–267, 2016.

H. Pourmonet, S. Bassetto, and M. Trépanier. Vers la maîtrise de l'évasion tarifaire dans un réseau de transport collectif. *11e Congrès International De Génie Industriel–CIGI2015*, 2015.

Public Transport Victoria. Victorian official fare compliance series. https ://www.ptv.vic.gov.au/footer/data-and-reporting/revenue-protection-and-fare-compliance/, 2018.

S. S. Pulugurtha and M. Agurla. Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods. *Journal of Public Transportation*, 15(1) :3, 2012.

R. R Core Team et al. R : A language and environment for statistical computing, 2013.

C. Raux, T.-Y. Ma, and E. Cornelis. Variability in daily activity-travel patterns : the case of a one-week travel diary. *European transport research review*, 8(4) :26, 2016.

A. Reddy, J. Kuhls, and A. Lu. Measuring and controlling subway fare evasion : improving safety and security at new york city transit authority. *Transportation Research Record : Journal of the Transportation Research Board*, (2216) :85–99, 2011.

L. K. Riegel. *Utilizing automatically collected smart card data to enhance travel demand surveys*. PhD thesis, Massachusetts Institute of Technology, 2013.

G. E. Sánchez-Martínez. Estimating fare noninteraction and evasion with disaggregate fare transaction data. *Transportation Research Record : Journal of the Transportation Research Board*, (2652) :98–105, 2017.

G. E. Sánchez-Martínez and M. Munizaga. Workshop 5 report : Harnessing big data. *Research in Transportation Economics*, 59 :236–241, 2016.

R. Schlich and K. W. Axhausen. Habitual travel behaviour : evidence from a six-week travel diary. *Transportation*, 30(1) :13–36, 2003.

C. Seaborn, J. Attanucci, and N. H. M. Wilson. Analyzing multimodal public transport journeys in london with smart card fare payment data. *Transportation Research Record : Journal of the Transportation Research Board*, 2121(1) :55–62, jan 2009. doi : 10.3141/2121-06.

M. J. Smith and R. V. Clarke. Crime and public transport. *Crime and Justice*, 27 : 169–233, 2000.

H. Snijders and R. L. Saldanha. Decision support for scheduling security crews at netherlands railways. *Public Transport*, 9(1-2) :193–215, 2017.

T. Spurr, A. Chu, R. Chapleau, and D. Piché. A smart card transaction "travel diary" to assess the accuracy of the montréal household travel survey. *Transportation Research Procedia*, 11 :350–364, 2015. doi : 10.1016/j.trpro.2015.12.030.

T. Spurr, A. Leroux, and R. Chapleau. Comparative structural evaluation of transit travel demand using travel survey and smart card data for metropolitan transit financing. *Transportation Research Record*, page 0361198118773897, 2018.

P. Stopher, C. FitzGerald, and M. Xu. Assessing the accuracy of the sydney household travel survey with gps. *Transportation*, 34(6) :723–741, 2007.

P. R. Stopher and S. P. Greaves. Household travel surveys : Where are we going ? *Transportation Research Part A : Policy and Practice*, 41(5) :367–381, 2007.

J.-B. Suquet. Drawing the line : how inspectors enact deviant behaviors. *Journal of Services Marketing*, 24(6) :468–475, 2010.

Y. O. Susilo and K. W. Axhausen. Repetitions in individual daily activity–travel–location patterns : a study using the herfindahl–hirschman index. *Transportation*, 41(5) :995–1011, 2014.

SYTRAL. *Rapport enquête fraude de mars 2017*. SYTRAL, Lyon, 2017.

S. Tamblay, P. Galilea, P. Iglesias, S. Raveau, and J. C. Muñoz. A zonal inference model based on observed smart-card transactions for santiago de chile. *Transportation Research Part A : Policy and Practice*, 84 :44–54, feb 2016. doi : 10.1016/j.tra.2015.10.007.

A. Tirachini and M. Quiroz. Evasion del pago en transporte público : evidencia internacional y lecciones para santiago. 2016.

F. Toqué, M. Khouadjia, E. Come, M. Trepanier, and L. Oukhellou. Short & long term forecasting of multimodal transport passenger flows with machine learning methods. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 560–566. IEEE, 2017.

J. C. Totten and D. M. Levinson. Cross-elasticities in frequencies and ridership for urban local routes. 2016.

M. Trépanier, N. Tranchant, and R. Chapleau. Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1) :1–14, jan 2007. doi : 10.1080/15472450601122256.

M. Trépanier, C. Morency, and B. Agard. Calculation of transit performance measures using smartcard data. *Journal of Public Transportation*, 12(1) :5, 2009.

R. Troncoso and L. de Grange. Fare evasion in public transport : A time series approach. *Transportation Research Part A : Policy and Practice*, 100 :311–318, 2017.

D. M. van de Velde. Organisational forms and entrepreneurship in public transport : classifying organisational forms. *Transport policy*, 6(3) :147–157, 1999.

G. Van Rossum et al. Python programming language. In *USENIX annual technical conference*, volume 41, page 36, 2007.

E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias. Short-term traffic forecasting : Where we are and where we're going. *Transportation Research Part C : Emerging Technologies*, 43 :3–19, 2014.

W. Wang, J. Attanucci, and N. Wilson. Bus passenger origin-destination estimation and related analyses using automated data collection systems. 2011.

J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301) :236–244, 1963.

Y. Wei and M.-C. Chen. Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C : Emerging Technologies*, 21(1) :148–162, 2012.

T. F. Welch and A. Widita. Big data in public transportation : a review of sources and methods. *Transport reviews*, 39(6) :795–818, 2019.

J. Wolf. Applications of new technologies in travel surveys. In *Travel survey methods : Quality and future directions*, pages 531–544. Emerald Group Publishing Limited, 2006.

J. Wolf, M. Loechl, M. Thompson, and C. Arce. Trip rate analysis in gps-enhanced personal travel surveys. In *Transport survey quality and innovation*, pages 483–498. Emerald Group Publishing Limited, 2003.

J. Zhao, A. Rahbee, and N. H. Wilson. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering*, 22(5) :376–387, 2007.

Z. Zhao, H. N. Koutsopoulos, and J. Zhao. Individual mobility prediction using transit smart card data. *Transportation research part C : emerging technologies*, 89 :19–34, 2018.

J. Zmud and J. Wolf. Identifying the correlates of trip misreporting-results from the california statewide household travel survey gps study. In *10th International Conference on Travel Behaviour Research*, pages 10–15, 2003.

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Oscar EGU

Apports des données passives à la compréhension des comportements de mobilité ?

Enjeux pour la planification et l'organisation des transports en commun

Résumé

Les réseaux de transport en commun sont des systèmes critiques pour le bon fonctionnement des villes. Ces systèmes doivent être planifiés et organisés avec rigueur en s'appuyant sur un dispositif de collecte et d'analyse des données. L'ambition de cette thèse est de s'interroger sur la pertinence de ce dispositif et sur l'apport des nouvelles sources de données passives. Quatre axes de recherches sont explorés : la mesure de la fraude, l'estimation de la demande sous forme de matrices origine-destination, l'étude de la variabilité des comportements de déplacements et la prédiction moyen-terme de la fréquentation. Ces travaux montrent que les données passives offrent des opportunités intéressantes pour améliorer la planification des réseaux de transport en commun.

Mots-clés— Transport en commun, données passives, big data, billettiques, enquête déplacements, comportement de mobilité, prévision, variabilité, fraude, matrices origine-destination

Abstract

Public transit networks are critical systems for the proper functioning of cities. These systems must be rigorously planned and organized based on data collection and data analysis. The ambition of this thesis is to question the relevance of this mechanism and the contribution of new passive data sources. Four research axes are explored : the measurement of fare evasion, the estimation of demand in the form of origin-destination matrices, the study of the variability of travel behaviour and the prediction of medium-term ridership. This work shows that passive data offer interesting opportunities to improve the planning of public transit networks.

Keywords— Public transportation, passive data, big data, smart card data, travel survey, travel behaviour, forecasting, variability, fare evasion, origin-destination matrices