Projet Predit Mobilletic : Analyse de la mobilité par les données billettiques

Livrable 1 : Etat de l'art
Partie A : Données billettiques et analyse de la mobilité

Table des matières

Introduction

Depuis quelques années, de nombreuses villes en Europe et dans le monde disposent d'une nouvelle source de données, les données billettiques. Celles-ci sont collectées via des cartes à puce (smartcard en anglais) qui sont identiques en taille à une carte bancaire et utilisées en remplacement (ou en complément) des titres papier. Elles peuvent permettre de stocker des titres de transports (tickets) et/ou de l'argent (débité en cas de validation). Chaque activité étant enregistrée, cela permet d'avoir accès à un plus gros volume de données, de pouvoir lier chaque activité à un individu, d'avoir des informations continues dans le temps et dans l'espace et de connaître une plus grande proportion des clients du réseau.

La plupart des cartes utilisées par les réseaux de transport dans le monde sont des cartes sans contact. Pour capturer les données, l'usager doit donc placer sa carte près du lecteur pour valider la transaction. Les informations de son trajet ou bien les informations de son achat, sont alors stockées dans la machine. Les informations collectées en cas de validation sont l'heure, la date, le lieu et parfois les informations de l'utilisateur. Cela fournit de gros volumes de données en raison du grand nombre de passagers utilisant le réseau. En revanche, de manière à respecter la vie privée des individus, aucune donnée socio-économique relative à l'usager n'est disponible. De la même manière, ces données ne fournissent pas les motifs des déplacements effectués, ni les raisons des choix modaux ou d'itinéraires. De plus dans la plupart des réseaux l'information sur le lieu de descente n'est pas disponible.

L'utilisation de ce type de données présente des avantages intrinsèques intéressants comme l'exhaustivité, une finesse spatiale et temporelle, l'absence de biais de réponse que l'on peut rencontrer dans les données d'enquête. Ces données soulèvent néanmoins quelques challenges et questions :

- Les données manquantes : pour des systèmes comme le métro ou le bus, seules les données de l'origine du déplacement sont disponibles (l'usager valide son ticket uniquement à la montée).
- Le volume important de ces données (environ xx de validations par jour ouvrable à Rennes), qui fait leur richesse, mais qui pose le problème de leur stockage et de leur traitement.
- Les données billettiques permettent d'avoir un flot de données longitudinales. En revanche, aucune donnée socio-économique relative à l'usager n'est disponible. De la même manière, les motifs des déplacements et les raisons des choix modaux et d'itinéraires ne sont pas connus.

Cette première partie du livrable dresse un état de l'art des travaux menés pour l'analyse des données billettiques en commençant par une description des différents jeux de données utilisés dans la littérature et par rappeler les potentiels de ces données. Plusieurs volets de cette problématiques sont ensuite détaillés, à savoir l'estimation des destinations et des matrices Origines-Destinations (OD), l'étude des distances et fréquence d'utilisation, les méthodes de fouille de données qui ont été mise en oeuvre pour l'analyse de ces données, l'enrichissement des données billettiques et leur croisement avec des données socio-économiques ou encore l'étude des pratiques et des motifs de déplacement.

1 Présentation des jeux de données utilisés dans la littérature

Pour des raisons de clarté, nous commençons par présenter un grand nombre de jeux de données utilisés dans la littérature de manière à ne pas avoir à le refaire ultérieurement.

1.1 Jeux de Brandford et Southport

Dans les articles [Bagchi and White, 2004, 2005], les auteurs disposent de deux jeux de données pour leurs études. Le premier provient de Brandford qui est un opérateur majeur dans l'ouest de Yorkshire (Angleterre) dans et autour de Brandford. Il possède 250 bus et une carte qui stocke les titres de transport et sert en même temps de carte fidélité. Le deuxième jeu de données provient de Southport. Ce dernier contient uniquement des cartes qui ont été données à des personnes âgées (gratuitement), afin de voir si les données ainsi obtenues pouvait remplacer celles fournies par les enquêtes et augmenter le nombre total de données. Ces données ont été recueillies sur une durée de 4 ans. Pour traiter ces données, la première étape a été d'extraire les voyages. Un échantillonnage de 10% de toutes les cartes des 2 jeux de données a ensuite été effectué, avec extraction des données de voyage. Au total 480 cartes, soit 90 062 voyages ont été conservés pour Southport et 3028 cartes soit 396331 voyages pour Bradford. De plus, le je jeu de données de Brandford contient également les échanges monétaires.

1.2 Jeu de Seoul

Les données utilisées dans l'article [Park et al., 2008] proviennent de la carte de transport en commun de Seoul en Corée. 90% des utilisateurs de bus et 72% des utilisateurs de métro disposent de cette carte pour un total d'environ 20 millions de transactions. Depuis 2004, un nouveau système de paiement a été développé, celui-ci consiste en un payement à distance : on paye un tarif de base de 800 won pour les 10 premiers km puis 100 won supplémentaires par tranche de 5 km. Le pass de base comprend 3 transferts dans les 30 minutes suivant la première validation. Contrairement à de nombreuses villes européennes, le système de transport de la ville de Seoul nécessite un contrôle en entrée et en sortie (nécessaire à l'application d'un tarif kilométrique). Les distances parcourues sont connues grâce à des GPS présents sur tous les bus et métros.

Le jeu de données se compose de deux jours de données : un mercredi en octobre 2004 et un jeudi en novembre 2005. Ces choix ont pour but d'étudier les variations d'un jour sur l'autre et d'une année sur l'autre. Ils disposent du type d'utilisateur (adulte, étudiant, enfant), du mode de voyage (local, feeder, metropolitan express, circle bus, metro, main) et du temps de trajet sur 24 heures.

1.3 Jeu de Chicago, CTA

Les données utilisées dans [Zhao et al., 2007, Utsunomiya et al., 2006] proviennent de l'autorité des transports de la ville de Chicago (Chicago Transit Authority, CTA) qui est le deuxième plus grand système de transport des États-Unis et recouvre la ville de Chicago ainsi que sa banlieue et le comté de Cook avec des bus, un service ferroviaire, etc...

Deux cartes sont utilisées dans cette zone géographique, la Chicago card (depuis 2002) qui est une carte périodique pouvant être rechargée à une machine ou avec du liquide et la Chicago card plus (depuis 2004) . Celle-ci est également une carte périodique mais elle permet en plus des payements en ligne automatiques (mensualisation ou au montant libre). En tout, plus de 372 000 cartes étaient en circulation en 2004.

Plusieurs informations peuvent être tirées de ces cartes, des informations sur l'utilisateur : nom, adresse, téléphone, adresse mail de l'utilisateur ; des informations sur la transaction : numéro de carte, temps de transaction, identifiant de l'horodateur, la station dans le cas du métro et le numéro de ligne dans le cas du bus, le type de transaction (achat ou validation). En revanche on ne dispose pas de la station de descente ni pour le métro ni pour le bus.

1.4 Jeu de Gatineau, STO

Les données utilisées dans les articles [Morency et al., 2006, Agard et al., 2006, Trépanier et al., 2007, Agard et al., 2009, Devillaine et al., 2012] proviennent du STO (Système de transport de l'Outaouais) basé à Gatineau au Québec de l'autre côté de la rivière où se trouve Ottawa. Gatineau est une ville possédant 240 000 habitants, la STO possède 200 bus qui sont tous équipés de système de lecture de carte. Le STO dispose d'une carte de transport depuis 2001 et à ce jour plus de 80% des usagers en possèdent une. De plus tous les bus sont équipés de lecteurs GPS. À chaque montée, la localisation de l'arrêt, la ligne de bus ainsi que la date et l'heure sont enregistrées. La carte permet d'utiliser des titres valable au mois ou pour un unique trajet.

Les données contiennent pour chaque transaction un numéro de carte, un statut et un type d'embarquement (régulier, transfert, refusé), une date et une heure de montée, le numéro de la ligne, le numéro de conducteur, la direction et l'arrêt. On dispose également de l'heure et de la date de départ prévue.

En tout 45 millions de transactions ont été enregistrées sur 9 ans grâce à 186 milles cartes sur un réseau comportant 1700 arrêts de bus et 62 lignes.

1.5 Jeu de Santiago

Ce jeu est utilisé pour les articles [Munizaga and Palma, 2012, Devillaine et al., 2012]. Un système de cartes de transport en commun existe depuis 2007 à Santiago au Chili, ville possédant 6 millions d'habitants. Ce système nécessite de payer son titre à l'entrée et permet jusqu'à 3 transferts dans une durée de 2 heures. Pour les bus seule la carte sans contact, carte bipl, est autorisée. En revanche pour le métro, les usagers ont le choix entre la carte bipl et le ticket. Ceux-ci ne concernent que 3% des validations dans le métro. Le système ne permet pas d'avoir des traces des changements de métro et il y a de nombreux problèmes de fraude, surtout dans certaines zones. Les transactions effectuées sont enregistrées dans une base de données, laquelle contient pour chaque transaction, l'opérateur et l'instant de celle-ci. Lorsque l'usager effectue un achat, on dispose de son identifiant, le bus ou la station de métro où la transaction a eu lieu, l'heure et la date et le montant payé. 35 Millions de transactions sont ainsi enregistrées chaque semaine (environ 7 millions de transactions par jour) par 3 millions de cartes. Pour ce qui est des métros et des stations de bus, on dispose de la localisation directe. Pour les bus la localisation directe n'est pas disponible, les auteurs se servent donc d'une base de données contenant les informations géo-codées (latitude, longitude, heure, date et vitesse) pour tous les bus toutes les 30 secondes. Enfin ils utilisent le temps "théorique" pour les bus et les métros.

1.6 Jeu de Londres, TFL

Les données de [Seaborn et al., 2009, Lathia et al., 2010, 2012, Ceapa et al., 2012] proviennent des transports en commun londoniens (TFL : Transports For London), qui disposent d'une carte, la Oyster Card, fournissant des données individuelles et enregistrant l'heure et le lieu de toutes les transactions. Chaque usager possédant une Oyster Card peut soit acheter des crédits qui seront débités lors des voyages, soit acheter un pass.

Les TFL ont entrepris un plan afin de favoriser les changements, ils ont ainsi catégorisé 614 plateformes de changement en 5 groupes et les ont ensuite triées par ordre d'importance. Cela repose sur des métriques le nombre d'usagers étant indisponible. Ils ont également ajouté 700 intersections de routes de bus (transferts dans la rue). Les transports londoniens se composent de 9 zones tarifaires, de 11 trains souterrains parcourant 402km de voies avec environ 270 stations, 5 trains aériens avec 78 stations, 8000 bus et 19000 arrêts, plus des tramways et des services fluviaux. Environ 80% des trajets effectués sur le réseau le sont avec une carte.

1.7 Données de Tokyo

Les données utilisées dans [Fuse et al., 2010] proviennent des cartes présentent dans la ville de Kanto au Japon. Les auteurs disposent d'un jeu de données de plus de 38 millions d'enregistrement par mois, enregistrés sur environ 20 millions d'utilisateurs depuis 2007. Pour chaque enregistrement l'identifiant, les informations sur le déplacement (montée ou descente du bus, heure), les informations sur l'arrêt de bus (identifiant arrêt, longitude, latitude), le nombre de personnes qui montent ou descendent et le nombre de personnes entre 2 arrêts sont connus. Cela permet de générer des données OD. Lors de la validation, sont enregistrés, l'identifiant de la carte et l'heure. Sont alors définies comme l'heure d'arrivée l'heure de validation du premier passager, l'heure de départ l'heure de validation du dernier passager. le temps d'arrêt est exclu du calcul du temps de trajet.

2 Potentiel des données billettiques

Dans leurs articles Bagchi and White [2004] et Bagchi and White [2005] s'interrogent sur le rôle des données billettiques comme nouvelle source d'information pour l'analyse des pratiques de voyage et étudient leur potentiel et leur faculté à compléter ou remplacer les données plus traditionnelles (enquêtes, ...). Ils rappellent que les données billettiques sont un nouvel outil permettant de suivre tous les trajets d'une "carte" (sur certains réseaux la carte est transférable) et ainsi d'obtenir un gros volume de données données continues dans le temps et sur de longues périodes (périodes plus longues que pour les enquêtes traditionnelles), et que ces données sont plus riches et contiennent, par exemple, les informations sur les changements de bus (indisponibles auparavant avec les tickets). Les cartes de transport peuvent permettre de faire des estimations du chiffre d'affaire, du taux de voyage par carte émise et des inférences sur le voyages "liés".

Cependant ces cartes possèdent également des limites, comme le fait qu'il n'y ai pas de validation à l'arrivée et donc que la longueur des trajets soit inconnue, que certains traitement requièrent encore les informations issues des sondages pour par exemple expliquer le motif d'un trajet, ou bien qu'elles ne représentent pas l'entièreté de la population et ne soient donc pas nécessairement représentative de celle-ci. De plus la théorie est souvent différente de la pratique, en effet en théorie l'usager va valider sa carte à chaque montée dans un bus. Mais dans la pratique celui-ci peut oublier de valider sa carte, utiliser un ticket ou bien le matériel peut être en panne et les validations non enregistrées.

Les auteurs estiment donc que ces données offrent un rôle complémentaire vis à vis des autres données, rôle qu'ils souhaitent approfondir. Pour cela ils font un bilan des informations données par les données billettiques et les analysent. De plus ils s'intéressent au renouvellement de la population sur une période d'un an. En effet il est parfois difficile de savoir qui est vraiment client d'un opérateur de transports en commun car un usager peut être amené à partir (déménagement) et être immédiatement remplacé par un autre sans que pour autant l'opérateur de transport le sache. Ou bien il peut changer de zone géographique (les étudiant restent environ 3 ans au même endroit). En cela les cartes offrent une solution car elles permettent de suivre un usager et de savoir quand celui-ci a arrêté de l'être.

Pour les besoin de leur études les auteurs ont défini plusieurs concepts. Ainsi un voyage correspond à un déplacement dans un sens, celui-ci est constitué d'étapes de déplacement définies comme changement de modes durant ce déplacement. Les voyages ne sont donc pas identifiables par la carte. Les taux de voyages concernent les voyages non liés, c'est à dire les étapes de déplacement, alors que les transferts concernent les voyages liés. Ils ont également défini la notion d'usager. Un usager est une personne avec un compte actif, c'est à dire qui possède au moins un voyage enregistré sur la période d'un an (permet de prendre en compte

les variation saisonnales).

Plusieurs analyses ont été menées sur les données tirées des jeux de Brandford et Southport (1.1) en utilisant des échantillons de cartes ayant servi sur de courtes périodes (7, 14, 21, 28, et 35 jours). Les auteurs se sont intéressés aux changements de bus. Ils ont défini un trajet lié une séquence composée de 2 ou plus montées dans des bus différents par un même individu, le même jour dans une limite de temps, comme morceau d'un voyage dans un sens d'une activité vers une autre. Par défaut ils ont défini la limite de temps de transfert à 30 minutes. Ils soulèvent le problème des voyages aller retour comptabilisés comme trajets liés par manque d'information sur la descente. Ils notent également que le temps de transfert devrait prendre en compte le groupe cible. Ensuite ils ont comparé le taux de voyages entre les utilisateurs de cartes et les autres usagers. Enfin ils ont étudié le taux de renouvellement des cartes. Celui-ci ne peut être étudié que si la carte est strictement personnelle, or à Brandford la carte est cessible d'une personne à l'autre. De plus certaines cartes ne sont plus actives et certains groupes sont surreprésentés par rapport à d'autres. Les auteurs ont tout de même calculé le nombre de cartes ayant cessé d'être actives durant les mois de l'année précédente et le nombre de cartes ayant été utilisées pour la première fois. Ils ont supposé que les abandons de carte étaient dus à des utilisateurs ayant essayé la carte et l'ayant abandonné pour revenir au titre classique faute d'être satisfait. Cela peut aussi venir d'une mauvaise fusion des cartes.

Plusieurs informations complémentaires seraient nécessaires pour que l'opérateur ai une pleine connaissance de son réseau. Tout d'abord la connaissance de l'origine et de la destination finale n'est pas connue en partie à cause de l'absence de validation en descente du bus. Ensuite aucune information est disponible sur le motif du voyage ou la qualité du transport. Enfin l'analyse des données billettiques est dominée par des processus basés sur des "règles" qui sont souvent définies arbitrairement en l'absence d'informations. Il faudrait donc vérifier ces règles (comme par exemple celle du temps nécessaire à un transfert).

Dans leur article Park et al. [2008] recherchent également le potentiel des données billettiques et souhaitent connaître leur fiabilité. Leur étude porte sur les données du jeu de Seoul (1.2). L'utilisation de ces données pour l'étude des transferts permettrait à terme d'obtenir un paiement basé sur la distance parcourue. Dans l'ensemble on peut considérer que les données sont fiables car la majorité des usagers ont une carte. Cependant les usagers ne disposant pas de carte sont des voyageurs occasionnels qui par conséquent voyagent différemment. De plus les informations de sorties sont disponibles pour les métros mais pas nécessairement pour les bus, particulièrement pour le metropolitan express. En conclusion à partir de ces données les auteurs ont pu calculer le nombre de transferts, les heures de monter, la distribution du nombre de voyageurs par heures et la distribution du temps de voyage. De plus ces données ne diffèrent pas (en nombre d'utilisateurs par station de métro) de celles de SMC. Par la suite il faudrait faire une étude de routine et une autre sur les usagers ne validant pas à la descente du bus.

En conclusion, ces données sont très riches mais l'absence de données sur la descente dans la plupart des cas pose des problèmes. De plus les auteurs ne disposent pas de résultats extérieurs auxquels comparer les leurs. Il faudrait donc une enquête pour pouvoir vérifier les résultats et identifier les voyages non enregistrés. Il faudrait également avoir des informations sur d'autres variables susceptible d'éclairer ce résultat et idéalement que ces informations concorde en temps avec les dates des relevés. L'étude des données billettiques ne peux donc pas remplacer les enquêtes mais peut à terme les réduire. Elles peuvent également être utilisées à des fins de marketing.

3 Estimation des destinations et matrices OD

Comme cela a été soulevé par les travaux de [Bagchi and White, 2004] et [Bagchi and White, 2005], l'absence de validation lors de la descente sur la plupart des réseaux de transport est un grand manque d'information pour l'analyse des données billettiques. Cependant ce type de données offre une possibilité d'estimation de destination qui n'était pas envisageable avec les titres papier.

Parmi les premiers à s'être intéressés à l'estimation des origines et destinations par les données billettiques, Barry et al. [2002] ont développé une méthodologie qui estime des tables de voyage OD station par station, pour toutes les stations de métro de la ville de New-York. Pour cela, ils déterminent une séquence de déplacement effectuée tout au long de la journée pour chaque carte de métro. Les cartes sont préalablement triées par numéro de série et heure puis la séquence de déplacements, et les stations utilisées à l'origine de chaque déplacement sont extraites. Une série d'algorithme est alors appliquée à chaque ensemble de carte de métro afin d'estimer la station de destination pour chaque station d'origine. Ces algorithmes sont basés sur deux hypothèses, tout d'abord qu'un grand pourcentage des usagers retournent à la station de destination de leur déplacement précédent pour effectuer leur déplacement suivant, deuxièmement qu'un grand pourcentage d'usagers finissent leur dernier voyage de la journée à la station où ils ont commencé leur premier déplacement du jour. Ces hypothèses ont été confirmé pour au moins 90% de la population des usagers du métro à l'aide des informations des journaux de voyage collectés auprès du conseil des transports de la métropole de New-York (NYMTC). Les sorties ont été de plus validées en les comparant aux comptages des sorties de stations.

Pour leur étude les auteurs ont obtenus les fichiers de transactions de cartes de métro (6 millions d'enregistrements par jour de semaine) de la ville de New-York, ceux-ci déjà triés par numéro de série et heure. Les informations sur les usagers ne disposant pas de cartes sont aussi obtenues, leurs motifs de déplacement sont supposés être semblable aux usagers disposant d'une carte dans la même zone et la même période.

La méthodologie est implémentée sur ordinateur en utilisant Microsoft Access et Visual Basic. Un premier traitement consiste à déterminé les transactions utilisables, puis l'algorithme est appliqué à ces transactions afin d'estimer les destinations. Les transactions non utilisables peuvent être les voyages constitués d'un unique déplacement dans la journée, les voyages pour lesquels la destination du premier voyage ne peut pas être inférée car deux transactions arrivent à la même station, ce qui peut arriver quand deux personnes voyagent ensemble en utilisant un seul titre et les voyages pour lesquels la destination du dernier voyage ne peut pas être inférée car la première et la dernière validation du jour ont lieu au même endroit. Les auteurs ont ainsi pu inférés 83% des stations de destination.

Les auteurs ont ensuite utilisé trois méthodes différentes pour comparer les voyages OD au comptages des usagers. Une première méthode consiste en la comparaison des sorties comptées par les systèmes de collecte des données billettiques avec celles estimées par le modèle, ces comparaisons ont été utilisées afin de détecter des anomalies dans les données (sauf dans le cas ou les portails de contrôle sont plus anciens et ne fournissent pas d'information comparable). La deuxième méthode consiste en un comptage manuel des paires OD qui vont contribuer au total des passagers lorsque le train rentre et quitte la station. Enfin la troisième méthode utilise un modèle de réseau et une procédure d'affectation de voyages pour affecter les voyages de la période de pic du matin à des lignes.

Ces données peuvent être utilisées afin d'adapter le réseau lorsque par exemple une station de métro est fermée pour cause de travaux, ou bien encore pour améliorer le réseau en ajoutant des lignes là ou cela est nécessaire.

Dans la suite de ces travaux, Zhao et al. [2007] ont présenté le potentiel des systèmes de collecte des données billettiques, les différences entre ce qu'ils offrent et ce dont nous avons besoin et ont développé une méthode permettant de déduire la destination avec l'origine uniquement. Les auteurs utilisent les données provenant du CTA (1.3). La différence avec les données présentées précédemment est qu'ils disposent des AVL (Automatic Vehicle Location) c'est à dire de la localisation des bus au moment de la validation. Les auteurs mettent l'accent sur le grand potentiel de ces données. En effet elles permettent une couverture spatiale et temporelle plus importantes que les données classiques car elles sont continues dans le temps. De plus leur collecte et leur stockage est automatique ce qui permet d'avoir un traitement plus facile et des données sans biais introduits par l'utilisateur (enquête voyageur). En revanche la mise en place d'un tel système de collecte nécessite un investissement initial assez important. Les auteurs souligner, à juste titre que la plupart de ces systèmes n'étant pas prévu initialement pour la collecte de données, il arrive que les données ne soient pas au bon format ou avec une structure incompatible et de ce fait ces systèmes ne fournissent pas toujours toutes les informations nécessaires à l'étude de la mobilité.

Avant de pouvoir estimer les destinations, les auteurs ont enrichi leurs données à l'aide des AVL. Ils ont tout d'abord obtenu les arrêts à partir des lignes de bus pour en déduire ensuite le numéro de bus, et enfin ils ont fusionné les données de validation avec celles de localisation des bus. Ils ont également utilisé des GIS (Geographic Informations Systems), leur permettant d'étudier les proximités entre les différentes arrêts. Une fois les données complétées, un certain nombre d'hypothèse est utilisé afin d'estimer les destinations :

— Une grande partie des usagers repartent de la station à laquelle leur précédent voyage s'est terminé
— les usagers n'utilisent pas de transport privé entre 2 segments dans une séquence quotidienne
— les passagers ne vont pas marcher jusqu'à une autre station que celle où ils sont descendus
— les usagers finissent leur dernier voyage là où ils ont commencé le premier

L'étude portant sur le réseau ferré, les destinations ne sont estimées que dans le cas des correspondances métro – bus et métro – métro. Trois cas sont distingués : lorsqu'il n'y a ni données GIS ni données AVL disponibles le cas n'est pas étudié, s'il y a les données GIS mais pas des données AVL le cas peut être étudié dans certains cas, enfin si les données GIS et AVL sont disponibles le cas peut toujours être étudié. Lorsque les données de GIS sont disponibles et pas celles de l'AVL, un cercle est tracé autour de l'arrêt de métro. Si ce cercle ne croise aucune ou plusieurs lignes de bus alors l'estimation n'est pas possible. En revanche s'il en croise une seule, le voyage peut être utilisé. Si les données GIS et AVL sont disponibles on peut alors déduire l'arrêt de bus et si celui-ci est dans la bonne distance on peut utiliser le voyage. Pour estimer les destinations les auteurs ont utilisé des chaînes symétriques. Si on a une symétrie par exemple métro – bus – métro – bus alors on vérifie les numéros de ligne et si ceux-ci sont identiques, la symétrie est confirmée, on a donc l'origine et la destination.

Les auteurs ont alors développé un logiciel permettant, une fois les données AVL et de validations disponibles, l'estimation de la matrice OD. Pour cela ils utilisent un algorithme constitué de 3 boucles imbriquées tournant sur chaque carte, jour et segment de journée. Cependant cet algorithme peut mener à différentes erreurs. En effet la matrice OD est ainsi estimée sur des utilisateurs de cartes qui ne sont pas forcément représentatifs de tous les usagers (différences socio-économiques ou géographiques), certaines destinations ont été mal estimées pour le rail et à certaines stations le nombre d'entrées est différent du nombre de sorties (ex : aéroport).

Trépanier et al. [2007] présentent également un modèle permettant d'estimer le lieu de destination pour chaque montée dans un bus validée avec une carte. Ils mettent en avant le fait que les données doivent être attentivement corrigées avant estimation pour éviter des erreurs.

Leur étude s'est basée sur les données du STO 1.4. Le lieu de descente est nécessaire pour calculer par exemple le profil de charge, d'où la nécessité de cette étude. Ils ont ainsi obtenu 60% de réussite et 80% en heure de pic.

La méthode développée ici afin de prédire la destination des usagers est basée sur 3 éléments. La première chose à faire est d'étudier l'architecture du système de collection des données afin d'en tirer les informations les plus intéressantes. Il faut ensuite identifier les objets nécessitant d'être analysés dans le modèle final. Enfin, le modèle analytique est développé pour estimer les lieux.

Les systèmes de collecte de données billettiques regroupent plusieurs objets qui peuvent être identifiés et reliés à la base. Pour cela les auteurs ont utilisé TOOM (Transportation Object-Oriented Modeling, Trépanier and Chapleau 2001). Ce système classe les données en 4 méta-classes d'objets : statiques, cinétiques, dynamiques et systémiques. Quatre grands groupes d'objets peuvent être alors décrits : Les objets du réseau (éléments du réseau STO : conducteurs, bus, lignes, arrêts), les objets de l'opération (conducteurs, bus, pièces, garages), les objets administratifs (utilisé pour la carte elle même) et les voyages et chaînes de déplacements qui caractérisent la demande. Le modèle objet permet d'exprimer les relations entre les éléments des données disponibles.

Le modèle d'estimation de la destination du voyage a pour but d'estimer le lieu de descente des usagers pour chaque montée et chaque usager. Il se concentre sur les objets usager, voyage, ligne et arrêt. On notera i l'usager, j le nombre séquence des arrêts sur une ligne, r le nombre séquence de la ligne dans la journée de l'usager et enfin k le jour. Une ligne de transit R (une seule direction) est définie par un ensemble ordonné d'arrêts $s : R = \{s^j\}$. ... Lors d'un transfert, le point de descente sera le point de transfert. Pour choisir l'arrêt de descente on va se baser sur la distance euclidienne $d(a, b)$ entre le lieu a et le lieu b. Le modèle doit estimer le meilleur lieu en fonction des différentes valeurs de r dans la journée. Le but est d'estimer l'arrêt de descente d_{rik} de la route, de telle sorte que la distance avec l'arrêt de montée suivant $s^B_{r+1)ik}$ soit la plus petite : $d_{rik} = z \rightarrow \min_z(s^B_{r+1)ik}, z) \quad z \in \{V_{rik}\}$ avec $r < N_k, d(s^B_{r+1)ik}, z) < M$. Une valeur de tolérance M est définie. Pour le dernier voyage de la journée, on suppose que c'est le même arrêt qu'en début de journée s'il est relié à la ligne prise ($r = N_k$) et le même arrêt que le lendemain si $d(s^B_{r+1)ik}, z) > M$.

Pour pouvoir appliquer ce modèle, il faut au préalable effectuer certains ajustements sur les données. Il faut d'abord corriger certaines erreurs avant la première analyse, puis analyser le transit global sur le réseau, analyser les comportements des usagers, effectuer les opérations de reconstruction et enfin appliquer le modèle d'estimation des destinations. Ce modèle produit des statistiques utiles, mais une attention particulière doit être portée au pré-traitement des données afin d'éviter de futurs erreurs dues par exemple à des validations enregistrées sur la mauvaise ligne. La plupart des destinations non estimées proviennent de voyages atypiques hors des heures de pic.

Enfin Munizaga and Palma [2012] ont travaillé sur l'estimation d'une matrice OD. Dans cet article ils ont repris les travaux de [Barry et al., 2002], [Trépanier et al., 2007] et [Zhao et al., 2007] pour développer à leur tour une méthodologie pour l'estimation d'une matrice OD pour les transports en commun grâce aux données billettiques et GPS. Les auteurs utilisent ici le jeu de données provenant de Santiago (1.5). L'étude s'est déroulée sur 2 semaines différentes en mars 2009 et en juin 2010. En moyenne on a 6 millions de validation par jour en semaine, moins de 4 millions le samedi et moins de 2.5 millions le dimanche. En semaine 60% des transactions se font dans le bus, 33% dans les métros et 7% dans les stations de bus. Les motifs globalement observés correspondent à un motif avec deux pics d'utilisation pour la semaine, un le matin et l'autre le soir, alors que le week-end le motif observée est de la forme d'une cloche sur toute la journée.

Le but étant de construire une matrice OD, les auteurs définissent un voyage comme étant le trajet menant d'une origine à une destination, ce trajet pouvant être décomposé en plusieurs segments. Ils disposent des informations sur les transferts bus–bus et bus–métro mais pas des métro–métro. Ils effectuent donc une reconstruction de la chaîne du voyage par estimation du point d'arrivée. En tout, trois bases de données sont disponibles : l'une contenant les transactions, la deuxième la position du véhicule et la troisième la définition géo-codée du réseau de transport. À partir de ces trois bases, on peut alors estimer les coordonnées spatio-temporelles du point de descente. De plus ils disposent d'un module différenciant transferts et destination. Pour l'estimation du point de descente les auteurs se sont appuyés sur les travaux de [Barry et al., 2002], [Trépanier et al., 2007] et [Zhao et al., 2007]. Ils ont repris leurs hypothèses c'est-à-dire qu'on suppose que l'usager retourne en fin de journée au point d'où il est parti en début de journée et que dans la journée, il repartira de la station où il est sorti. Ils supposent également que l'usager ne parcourra pas plus de 400 mètres et n'attendra pas plus de 5 minutes. L'information du jours suivant est également utilisée pour l'estimation. La seule variante est qu'au lieu de prendre le lieu de départ du matin comme lieu d'arrivée on prend l'arrêt le plus proche du lieu de départ du lendemain. Lorsqu'il n'y a qu'un seul voyage aucune inférence n'est possible. L'estimation des arrêts de descente pour le bus a posé des problèmes car l'arrêt le plus proche en distance n'était pas nécessairement le plus proche en temps. En effet, on peut imaginer un bus qui fait une boucle autour d'un arrêt de métro et passe à proximité deux fois. Cependant, l'arrêt de bus le plus proche sera peut être celui se trouvant à la fin de boucle, l'usager descendra plus probablement à l'arrêt le plus éloigné se trouvant en début de boucle. Les auteurs minimisent donc le temps général défini par :

$$T_{gi} = t_i + f_w.\frac{d_{i-post}}{s_w}$$

avec t_i le minimum du temps d'embarquement, f_w la pénalisation, s_w la vitesse de marche moyenne, d_{i-post} la distance entre la position i et celle de l'embarquement, la distance maximale étant de 1000. Le temps général est donc défini comme le temps associé à la position i t_i plus une variable qui représente une estimation du temps de marche entre la descente et la montée suivante, multipliée par un facteur de pénalisation. En minimisant le temps général, on peut donc identifier un arrêt assez proche pour que l'usager descende et marche. On peut aussi l'écrire :

$$\min_i T_{g_i}, d_{i-post} < d$$

Dans le cas du métro, on estime le trajet le plus court en utilisant le temps de trajet entre les stations, le temps d'attente à la station et le temps de marche entre les stations. Les stations de bus posent un problème car on ne sait pas dans quel bus va monter l'usager. Les auteurs supposent donc que l'usager va prendre le bus qui minimisera son trajet. Les auteurs supposent qu'un voyageur restant plus de 30 minutes à un arrêt est arrivé à destination. Ils déduisent les destinations des voyages sans destination des autres voyages pour lesquels ils ont pu en estimer une. Pour cela ils utilisent le facteur d'expansion défini par :

$$f_{it} = \frac{\sum_j Trips_{ijt}}{\sum_{j \neq null} Trips_{ijt}}$$

avec i l'origine, j la destination et t la période.

Si ni l'origine ni la destination ne sont connues (comme pour les fraudes), alors on utilisera :

$$f_{it} = \frac{\sum_j Trips_{ijt}}{\sum_{j \neq null, i \neq null} Trips_{ijt}.f_{ijt}}$$

Ainsi plus de 80% des points de descente ont pu être estimés.

4 Étude des distances et fréquence d'utilisation

Dans leur article, Utsunomiya et al. [2006] décrivent les premières étapes nécessaires pouvant être utilisées à des fins de planification. Pour cela ils utilisent un jeu de données billettiques de la CTA (1.3). Ils mettent aussi en avant le fait que contrairement aux données traditionnelles (comptage manuel du nombre de voyageurs, enquête voyageurs), les données billettiques n'introduisent pas de biais. À partir des données , les auteurs se sont intéressés plus particulièrement aux adresses enregistrées des utilisateurs et ont étudié leurs pratiques de déplacement, le lien entre adresse et lieu de payement. Cependant il est indiqué qu'une telle étude n'est pas représentative de la population de Chicago car les utilisateurs du réseau sont principalement des habitants aisés.

Les auteurs ont étudié la distance des usagers au transports, c'est à dire la relation entre l'adresse enregistrée dans la base de données du CTA et le premier voyage sur le réseau des usagers sur une période de 7 jours (du 12 au 18 septembre 2004). Pendant cette période 62 351 cartes étaient actives, 521 630 transactions de validation ont été effectuées par les propriétaires de Chicago card (314 851 sur le réseau ferré et 206 779 sur le réseau routier). L'étude s'est portée sur les 91% des usagers (56894) dont l'adresse était géo-codée. Sur l'ensemble de ces usagers, le calcul de la distance entre le foyer et le premier lieu de validation montré qu'un grand nombre de ces distances était supérieur à 2 miles (3.22 km), ce qui suppose une mauvaise adresse, un déplacement en voiture avant de prendre les transports en commun ou bien une erreur. Seuls les usagers ayant marché moins d'un miles (1.61 km) ont alors été conservés. Il en est ressorti que les usagers étaient prêts à marcher plus pour avoir un meilleur service car les distances au métro étaient plus élevées que celles au bus.

L'étude de l'usage quotidien des cartes a mis au jour un grand nombre de voyages unique (c'est à dire pas d'aller-retour), ce qui implique qu'il s'agit essentiellement de voyageurs occasionnels. De plus, l'utilisation en week-end est plus faible. En ce qui concerne la constance dans l'utilisation des transports, 41852 usagers utilisent leur carte 3 jours ou plus par semaine. Il a aussi été observé que si la zone était bien desservie et que plusieurs chemins étaient possibles, l'utilisation des transports variait selon le jour.

Le problème du grand nombre de validation (au moins une fois par semaine) à une distance supérieure à 1 mile du lieu d'habitation sur les données d'utilisateurs réguliers peut provenir d'une erreur de bus, de validation ou bien de l'usager qui a séjourné à l'extérieur de son domicile habituel. Pour pallier à ce problème, les auteurs suggèrent d'ajouter des données GPS au bus ou bien de passer au mois.

Les auteurs concluent en énonçant un certain nombre de pistes pour de futurs travaux, notamment l'influence du choix de payement (mensuel ou à l'unité), l'influence de critères qualitatifs et quantitatifs sur l'usage des transports (fréquence d'utilisation, sécurité et confort), la comparaison des données billettiques avec des données socio-économiques. Ils envisagent également de demander adresse de facturation et adresse de domiciliation ou bien de demander à l'usager de tenir un journal de ses déplacements afin de détecter les erreurs de validation.

5 Utilisation de méthodes de fouille de données (ou data mining) sur les données billettiques

Dans leur article Agard et al. [2006] se sont intéressés à l'utilisation des méthodes de data mining pour l'étude des données billettiques et à l'utilisation des modèles de planification des transports. L'objectif était d'obtenir une représentation améliorée des utilisateurs dans les

transports publiques avec une attention particulière portée sur les habitudes durant les jours de semaine. Les données proviennent du jeu du STO (1.4). Le jeu de données se compose de 2 147 049 validations à la montée, provenant de 25 452 usagers possédant une carte de transport, entre le 10 janvier et le 1er avril 2005. Les transactions ont été résumées en 12 enregistrements d'une semaine avec environ 238 895 usagers par semaine. Chaque enregistrement est divisé en 20 variables binaires représentant les 5 jours de semaine divisés en 4 périodes par jour (mati, midi, après-midi et soir). De plus les usagers sont classés selon trois groupes en fonction de leur carte : adulte, jeunes, personnes âgées.

Le protocole appliqué aux données afin de caractériser les habitudes des usagers est constitué de 3 étapes. Tout d'abord les auteurs ont fait des groupes selon les motifs des voyages à l'aide de la méthode HAC (Hierarchical Ascending Clustering). Pour accélérer celle-ci, ils ont auparavant fait un pré-clustering de 20 groupes à l'aide de l'algorithme des K-means et ont renvoyé son résultat au HAC. Une fois les groupes obtenus, les auteurs ont analysé leur composition et les ont comparés aux types de cartes. Ils ont également étudié la variabilité des groupes sur les 12 semaines d'observation. À la suite de cette étude, 4 groupes ont été obtenus après segmentation, dont 2 facilement interprétables. Un groupe d'utilisateurs réguliers (qui voyagent pendant les heures de pic), une groupe d'utilisateurs réguliers le matin (qui voyagent pendant les heures de pic le matin), un groupe sans motifs particuliers et enfin un groupe d'utilisateurs occasionnels. L'étude de la variabilité sur les 12 semaines fait ressortir la semaine de vacances scolaires où les habitudes des jeunes usagers sont modifiées.

Morency et al. [2006] développent dans leur article une étude sur 10 mois de données en utilisant des méthodes de data mining pour classifier les jours de voyage en fonction des périodes de montée. L'étude s'est portée sur les données du STO (1.4). Au total, 6 200 000 transactions de 43 248 cartes ont été enregistrées sur la période du 1er janvier au 4 octobre 2005, soit en moyenne 238 895 cartes par semaine. Chaque transaction est découpée en 17 variables comprenant l'identifiant de la carte, la date, le type de jour (D1,... D7) et le nombre de validation pour chaque heure du jour (H08=1 équivaut à une validation entre 8H et 9H). Les auteurs ont alors appliqué à ces données plusieurs outils de data mining : filtrage, segmentation (K-means) et visualisation. Ils ont construit des segmentations de jours présentant des motifs temporels similaires d'embarquement dans les réseaux de transport. Le but étant de comprendre si les voyageurs avaient des trajets réguliers et si les jours différaient significativement les uns des autres.

L'analyse longitudinale des données et l'analyse des pratiques de voyage s'est faite en calculant un taux d'activité $AR = D_B/D_O$ avec D_B le nombre de jours où la carte a été validée et D_O le nombre de jours observés. Le nombre de validation par jour a également été calculé. De plus l'énumération des stations utilisées a permis d'observer un effet de la saison sur celles-ci. Pour étudier la variabilité spatio-temporelle de l'usage du réseau, les auteurs ont utilisé deux cartes avec un taux d'activité de 100%, l'une d'un adulte, et l'autre d'une personne âgée. Ils ont ensuite comparé leur nombre de validation ainsi que les heures de celles-ci.

Enfin Agard et al. [2009] ont également montré le potentiel des méthodes data mining pour fournir des informations utiles et nouvelles sur les pratiques des usagers sur le réseau. Une hypothèse employée couramment est de considérer un comportement similaire quelque soit la semaine, est remise en cause dans la mesure à l'analyse de ce type de données a mis en évidence une variabilité dans l'espace et dans le temps. De plus il est difficile d'évaluer la variabilité car les données disponibles ne le sont souvent que sur une journée. Cette étude utilise les données du STO 1.4.

Chaque client est représenté par un vecteur de longueur fixe de façon à pouvoir extraire ses pratiques de déplacement en terme temporel. Deux transformations sont alors possibles :

un jour est échantillonné en 24 périodes (heures) ou bien à une échelle moins fine une semaine est divisée en 24 périodes, soit 4 périodes par jour (sur 7 jours) On dispose également d'un identifiant de carte et du jour ou de la semaine en question.

Les auteurs ont alors cherché à identifier différents groupes d'utilisateurs en terme de pratiques temporelles. Le nombre de groupe était alors choisi en fonction de la précision voulue et les groupes sont obtenus à l'aide de l'algorithme des K-means. L'étude des groupes obtenus a permis de reconnaître leur composition : étudiants, personnes âgées, travailleurs, ... Une grosse variabilité a également été observée pendant le spring break. L'étude de chaque usager individuellement permet la construction de maps pour chacun d'eux. De plus le type d'un usager est facilement déductible de ses habitudes.

En conclusion, les auteurs présentent une définition des clients typiques et une mesure de leurs déplacements. Ils analysent également la variabilité de l'usage du système en fonction du jour de la semaine ou de la saison. Leur étude peut permettre à terme d'améliorer les moyens de transport en ajustant la balance entre les pratiques des usagers et le niveau de service fourni sur chaque ligne. Elle peut également permettre de proposer de nouveaux titres de transport qui encouragent des shifts temporels ou spatials de manière à éviter la congestion du réseau.

6 Enrichissement des données billettiques

Dans leur article Chu and Chapleau [2008] ont développé une méthodologie visant à enrichir les données billettiques et ainsi à obtenir des informations plus complètes à des fins de planification. Les données proviennent du STO (1.4). Ici les auteurs s'intéressent à un jeu de données contenant 37 781 transactions effectuées par 17 434 usagers le 10 février 2005. Cette journée a été choisie car elle représente un jour de semaine typique.

La première étape pour enrichir les données est de reconstruire l'itinéraire à partir des différents segments d'un même voyage. Le choix est fait de reconstituer l'itinéraire et pas uniquement d'étudier l'origine et la destination car l'information contenue dans celui-ci est plus riche. Une première étape consiste à déterminer si l'usager va être dans un transfert (correspondance) ou non, les auteurs utilisent à cette fin un seuillage avec une valeur fixée a priori. Des travaux pourraient être menés sur le choix du seuil.

Pour mieux identifier les transferts et afin de construire des profils spatio-temporels de charge des bus, le chemin spatio-temporel des bus doit être estimé. Pour cela, les auteurs prennent en compte toutes les contraintes spatio-temporelles disponibles pour produire un estimateur : heures de départ et d'arrivée, distance entre les arrêts. Ils posent alors plusieurs hypothèses :

— Le départ des véhicules du terminus est celui prévu, à moins que la transaction l'indique.
— L'heure d'embarquement des passagers à l'arrêt suivant sert de borne supérieure à l'estimation de l'arrêt.

En tenant compte des contraintes, une interpolation est utilisée pour estimer les heures d'arrivée aux arrêts de descente où il n'y a pas de montée. On prend donc en compte la première et la dernière transaction à chaque arrêt ainsi que la distance entre les arrêts. La même méthodologie est employée pour estimer la position du véhicule entre la dernière validation et le terminus, en supposant que le bus arrive à l'heure prévue au terminus. En revanche si la vitesse est trop faible ou trop élevée, une extrapolation linéaire est utilisée avec une vitesse moyenne. Les auteurs notent que cette extrapolation reste limitée par l'heure de départ de la prochaine course du véhicule et l'heure de montée du passager lors de l'arrêt suivant. Il est important de noter que ce système peut être amélioré si on dispose des coordonnées GPS des véhicules.

Pour détecter les transferts, chaque montée est associée à une course et un arrêt. L'heure estimée de descente est alors déduite de celle de la course puis comparée avec les heures de montée de ce même utilisateur. Le temps de transfert pris en compte est celui d'une marche à 1.2m/s entre les arrêts. Un algorithme est également utilisé pour tester si le bus pris était le

premier ou non, pour cela de la variabilité à l'heure d'arrivée est ajoutée (plus 5 minutes) afin de prendre en compte les bouchons ou la vitesse de marche. De plus les auteurs font l'hypothèse qu'il y aura toujours de la place dans le bus et qu'il y aura une transaction si et seulement si le bus n'est pas dans la même direction. L'analyse de ces transferts a fait apparaître que 50% des transferts se font dans une durée inférieure à 7 minutes et 80% dans une durée inférieure à 18 minutes.

L'analyse du profil de charge des bus permet d'étudier la variabilité dans l'utilisation d'une ligne. De plus, si on combine cette information avec celle des validations des voyageurs on peut obtenir le nombre de km parcourus par usager.

À la suite de leur article précédent [Chu and Chapleau, 2008], Chu et al. [2009] s'intéressent à l'enrichissement des données billettiques. Ils présentent donc la collection automatique d'informations sur la validation des usagers avec une couverture géographique et temporelle. Ils présentent ensuite une méthode orientée objet pour comprendre, valider, corriger et enrichir les données. Pour cela ils travaillent sur un mois de données. Ils soulignent également le fait que les outils habituels (enquête ménage/habitation) sont inadéquates pour la planification des déplacements. En effet ce type d'enquêtes est coûteuse et n'est donc pas menée plus d'une fois par an alors qu'elle nécessiterait des ajustements tout au long de l'année. De plus ces enquêtes concernent des échantillons insuffisants de la population et présentent une surreprésentation des utilisateurs d'automobiles. Enfin elles ne disposent pas d'une assez bonne résolution (détails spatio-temporels).

Les données utilisées pour cette étude sont générées automatiquement et d'une haute fiabilité spatiale et temporel. Elles permettent un suivi dans le temps et une couverture étendue de tout le réseau. Enfin ces données sont individuelles. Ils disposent de plusieurs informations. Tous d'abord des informations sur la carte, c'est à dire le numéro de carte, le type et enfin le tire de transport. Ils disposent ensuite d'informations sur le voyage : heure de montée, localisation et transfert. Enfin ils ont des informations sur l'opération : données sur la course (route, direction, heure de départ prévue), le bloc du véhicule, son numéro et le numéro du chauffeur. Cependant certaines informations ne sont pas collectées comme l'heure et le lieu de descente, la destination finale, la raison du voyage, le statut socio-économique de l'usager, les compagnons de voyages et la localisation du véhicule quand personne ne monte. Ces données ont pour avantage de ne pas être intrusives et de ne pas demander à l'usager de se rappeler de ces déplacements. Elles sont donc sans biais et on dispose facilement d'un gros volume de données. De plus les heures de départ ne sont pas arrondies (à 15 minutes près) comme cela peut être le cas lors d'une enquête, mais la localisation précise de l'usager n'est pas connue. Enfin pour les données issues d'enquête on ne dispose pas d'observations d'un jour sur l'autre et donc on n'a pas d'effet de la saisonnalité ou de la météo.

Les données ont été analysées avec 2 processus de modélisation des données, le premier avec une approche orientée objet, c'est à dire que chaque objet peut être analysé en détail et utilisé pour différentes applications de planifications, et une approche désagrégée, c'est à dire mettant en avant la préservation du voyage et des composantes du voyage liés à chaque utilisateur, cela permet une analyse multivariée. Les données originales ont plusieurs propriétés. Le service va s'organiser en suivant le concept de jour ou de semaine moyenne. On aura les même répétitions tous les jours de la semaine. Il y a plusieurs sources d'erreurs dans ces données, comme par exemple une inversion des coordonnées GPS qui peuvent mener à la déduction d'un mauvais arrêt. Cependant les données GPS n'étant pas conservé, la qualité des données va dépendre du chauffeur.

Les données de validations sont organisées par bloc selon le véhicule dont elles proviennent. On cherche ensuite les erreurs en comparant à la course théorique. Les erreurs peuvent être une montée au terminus, ou une montée pendant un non service, un chemin spatio-temporel

incohérent ou bien une opération manquante ou incomplète. Au total 80 % des opérations n'ont pas d'erreurs. Les erreurs trouvés ne sont pas supprimés mais corrigées, cela afin d'éviter des erreurs futures lors de l'enrichissement. 3 des champs sont connus de manière absolue, le numéro du véhicule, le numéro de carte, et l'heure de la transaction. On peut donc récupérer une course correct en mettant en relation le numéro du véhicule avec le numéro du bloc un jour donné.

L'enrichissement se fait en localisant l'arrêt à l'aide des données GPS puis en mettant en relation arrêt et but du voyage (éducation, hôpitaux, centre commerciaux, parcs, etc...).

Pour mieux connaître les lieux avec une forte demande les auteurs ont effectué une agrégation par arrêts et ont observés de nombreuses descentes près des écoles pour les étudiants de moins de 21 ans. Ils ont également noté que cela était dominé par les adultes sauf près de l'université où c'était les étudiants. L'agrégation par route a permis d'étudier le nombre de montées et de descentes, ainsi que le taux d'occupation. Enfin l'agrégation par liens et nœuds s'est faite en agrégeant les arrêts à proximité (MADITUC – Modèle d'Analyse Désagrégée des Itinéraires de Transports Urbains Collectifs). Les routes enregistrées avec ces nouveaux nœuds et itinéraires chargés avec les données. Les auteurs ont fait une cartes des zones d'occupation des utilisateurs en fonction de leur lieu de validation et de descente estimée lorsqu'il était connu. Ils ont également étudié le nombre de montées et de véhicules au cours du temps Pour assigner des points d'encrage aux voyageurs les auteurs ont commencé par agréger les transactions en itinéraire. Ils ont ensuite agrégé les profils temporels selon différentes périodes du jour (pic du matin, midi, ...). Enfin ils ont effectué une agrégation spatiale, par exemple en assignant un établissement scolaire à des cartes de type scolaire. De l'étude des données il est ressortit que réduite à un jour de semaine moyen les données sont plus significatives et représentatives. Les ajustements saisonnale demandant de pouvoir regarder la demande en fonction du temps ce qui est permis par ces données. Enfin au lieu d'être uniforme sur tous les jours de la semaine on pourrait adapter en fonction des activités (parc exemple le cinéma).

Dans [Chu and Chapleau, 2010], les auteurs ont développé une méthodologie pour améliorer la caractérisation des voyages en ajoutant une dimension "multiday" d'un mois aux transactions des cartes. Ils ont estimés des points d'encrage individuels (liés à l'adresse) et en ont déduit les lieux de monté et de descente. Les données utilisées proviennent du jeu STO (1.4). Le jeu de données contient 763 570 validations enregistrées en 2005.

Les auteurs ont défini un point d'encrage comme un point où un usager va aller souvent. Pour le trouver ils ont agrégé les arrêts dans un rayon de 50m, et les séquences de déplacement en un unique voyage, car sinon cela risquait de créer des ancrages artificielles (par exemple aux stations regroupant un grand nombre de correspondances). Une station ne pouvait être définie comme ancre qu'aux conditions que l'usager ai validé au moins 5 fois dans cette station et qu'elle concentre au moins 20 % des validations totales. La définition de l'ancre se fait sur la supposition que les usagers vont avoir tendance à minimiser la distance d'accès au bus. Trois types d'ancrages sont alors définis : le point d'encrage qui consiste en un ensemble fini de lieux avec des coordonnées spatiales connues (par exemple les écoles), les airs d'ancrage qui sont des aires prédéfinies qui peuvent être une grille virtuelle ou une région géopolitique et enfin les "Fuzzy area" qui sont des zones pour lesquelles on aimerait avoir dans l'idéale les coordonnées précises mais en leur absence on défini une air (comme par exemple pour les domiciles). Pour assigner une ancre le titre de transport est utilisé (par exemple pour lier un étudiant à une école). On va alors assigner un étudiant à une école en fonction des horaires de celle-ci (entre 14 et 17 heures). Puis on compare une carte des écoles avec les validations et on assigne une école à chaque carte. Les cartes non assignées concernent les étudiants d'Ottawa, les étudiants de moins de 21 ans inscrit à l'université ou à des cours du soir ou bien les cartes ayant enregistré trop peu de transactions dans le mois. Lorsque aucune coordonnées n'est disponible les auteurs ont utilisé une approche probabiliste à partir de l'heure, du jour, de la distance et de la fréquence.

Ils ont effectué une analyse de la densité des noyaux pour associer les ancres à des activités et ont défini le domicile comme la première validation de la journée. Lorsque le lieu de descente ou un arrêt se trouve à moins de 500 mètres d'une ancre alors la destination est définie comme étant l'ancre.

L'étude des résultats obtenus a montré que plus les usagers habitaient loin de leur lieu de travail, plus ils partaient tôt. On peut également observer des départs groupés aux heures de fin de cours. Ils ont également noté plus de départs que d'arrivées.

Dans leur article les auteurs ont utilisé un algorithme de règles d'association pour analyser les pratiques de voyage. Ils ont également appliqué un algorithme de classification pour mesurer la régularité des motifs de déplacement. Cependant il faut faire attention à ne pas trop ajuster cet algorithme afin de ne pas devenir trop spécifique.

7 Étude des pratiques et des motifs de déplacement

Dans leur article Seaborn et al. [2009] cherchent à développer une méthodologie pour décrire les pratiques des utilisateurs depuis et vers le réseau de bus en utilisant les données billettiques pour identifier les motifs des voyages, puis en les comparant aux données issues des enquêtes. Le but est donc d'utiliser les données issues du bus pour améliorer le réseau. L'étude utilise les données des TFL (1.6). Il y a une bonne évaluation de la demande entre les arrêts et sur les routes parallèles du bus mais pas pour le métro et les origines-destinations. Ils souhaitaient donc étendre les informations sur les passagers, c'est à dire étudier : les flux de passagers entre les intersections de routes pour fournir des lignes directes réduisant ainsi les transferts, les volumes de transferts du bus vers le métro pour voir quelles entrées ont le plus d'accès et ainsi améliorer le design du bus et du métro. Ils souhaitent également comparer les temps de transfert bus et métro et identifier les voyages journaliers individuels répétés sur une ligne pour indiquer un lien fort à ce service.

Dans leur article Fuse et al. [2010] souhaitent développer des méthodes d'analyse avec les pratiques détaillées des utilisateurs. Pour cela ils convertissent les données de bus en données de temps, analysent les origine-destination, estiment le temps moyen de trajet en bus, et cherchent le lien entre temps de trajet et congestion du trafic (heure du trajet). Les données proviennent du jeu de Tokyo (1.7). Les auteurs ont comparé leurs données aux données de météo (pour par exemple voir l'effet de la pluie). Ils ont étudié la congestion pour prioriser la maintenance des routes.

Lathia et al. [2010] analysent les données de déplacement individuelles dans le métro londonien (offre de service personnalisé) afin d'estimer les voyages personnels. Ils présentent une méthode de prédiction des heures de voyage personnalisées pour les usagers et font un classement des stations basé sur les futurs motifs de mobilité en identifiant des sous ensemble de grand intérêt pour ainsi fournir aux usagers des informations utiles. Les auteurs font la constatation que seulement 46 à 62% passé dans le métro l'est dans un wagon. Le reste est du temps passé dans les changements à marcher ou à attendre. Il y a alors deux aspects de la personnalisation qui sont chacun liés à des problèmes de prédiction : la prédiction du temps de trajet personnalisé entre une origine et une destination pour fournir à l'usager son temps de parcours et la prédiction et le classement de l'intérêt que les voyageurs individuels vont avoir pour les notifications et alertes sur des stations particulières basés sur l'historique de leurs voyages précédents. Le jeu de données utilisé provient du TFL (1.6). Deux jeux de données de 83 jours sont utilisés, le premier de mai à juillet 2009 et le deuxième d'octobre 2009 à janvier 2010. Les auteurs travaillent sur un sous-ensemble de 5% des usagers enregistrés avec des tuples définis par $< u, (o, d), t_o, t_d >$.

Les auteurs mettent en avant des motifs systématiques qui donnent une large perspective

du système et impactent sur la faculté à prédire les heures de voyages et stations d'intérêt. Il en ressort un motif temporel constitué de 2 pics en semaine et d'aucun le week-end. De plus le temps de parcours est en moyenne de 30 minutes. L'étude des données montrent que 60% des couplets OD sont des répétitions et que 88% font une boucle dans leur trajet (Attention ! pas de multimodal). Les auteurs effectuent donc un clustering de l'activité des usagers en formant des groupes basés sur les moments de la semaine où l'usager voyage (semaine, week-end ou les deux). Ils séparent une journée de 24 heures en 5 segments par rapport aux heures de pics puis construisent un vecteur de fréquence pour les utilisateurs ayant fait plus d'un voyage. Ils comparent ensuite ces vecteurs pour les différents utilisateurs en calculant une distance

$$d_{a,b} = \frac{1}{s} \sum_i \left| \frac{a_i}{A} - \frac{b_i}{B} \right|$$

avec A et B le nombre de voyage des utilisateurs a et b. La clustering est effectué sur 10 sous groupes de 1000 usagers. Le temps de trajet en min $u_{o,d}$ pour chaque utilisateur u est également calculé à l'aide du temps moyen $\bar{m}_{o,d}$ et des résidus $r_{o,d}$.

Les auteurs souhaitent alors estimer le temps personnel de voyage et évaluer la prédiction. Ils commencent par calculer des statistiques de base comme le temps moyen de trajet (environ 30 minutes), le temps de transite entre zones $z_{o,d}$ et le temps moyen de trajets entre la station d'origine et celle de destination. Ensuite ils posent l'hypothèse que si un usager répète le même voyage il va avoir tendance à suivre le même chemin et avoir le même temps de trajet $U_{o,d} \in T_{o,d}$: l'ensemble de taille M des $x_{u,t}$ des voyages entre o et d, $u_{o,d}$ le temps moyen, moyenne géométrique des temps observés

$$u_{o,d} = \left(\prod_{o,d} x_{u,t} \right)^{1/M} = exp \left(\frac{1}{M} \sum_{U_{o,d}} lnx_{u,t} \right).$$

Ils posent également $F_{o,d}$ l'ensemble (des usagers) des temps moyens $\bar{u}_{o,d}$ de tous les usagers ayant une familiarité f_u qui est au moins M. La prédiction personnelle s'écrit alors

$$\hat{p}_{u,o,d} = \frac{\sum\limits_{F_{o,d}} (\bar{u}_{o,d} \times f_u)}{\sum\limits_{F_{o,d}} f_u}$$

Les voyageurs sont alors séparés en deux groupes et seuls les plus pertinents pour le calcul du trajet sont conservés. Ils définissent également la notion de contexte temporel, c'est à dire que les usagers voyageant dans la même fenêtre temporelle $t \pm w$ sont dans le même contexte temporel. Ils définissent alors un modèle combiné en chaînant les méthodes les unes à la suite des autres : transfert de zone $\rightarrow$ moyenne de voyage $\bar{m}_{z_{o,d}} \rightarrow$ contexte du voyage $\rightarrow$ similarité de l'usager. Pour évaluer le résultats, les auteurs ont utilisé les 9 derniers jours de données, soit environ un ratio de 90% pour 10%. Ils ont ensuite utilisé du MAE (Mean Absolute Error) et MAPE (Mean Absolute Percentage Error).

Un classement de l'intérêt des stations est créé pour chaque utilisateur, celui-ci reflétant leur futur déplacements. Trois méthodes différentes existent : la première consiste à chercher les stations les plus populaires de manière générale, c'est la méthode baseline, la deuxième méthode, user history, cherchent les stations les plus utilisées par l'utilisateur, enfin la dernière méthode, similarity model, crée une matrice de co-occurence C avec $c_{i,j}$ la fréquence à laquelle on a la station i et j comme point final, on normalise ensuite celle ci qui devient W avec $w_{i,j}$. Le rang s'écrit alors

$$\hat{r}_{u,s} = b_s + \beta h_{u,s} + \sum_{s \in S_u} \left(\sum_{n \in N_s} w_{s,n} \right)$$

avec $h_{u,s}$ l'historique. Le rang moyen est alors défini comme

$$\bar{rank} = \frac{\sum_{u,s} interest_{u,s} \times renk_{u,s}}{\sum_{u,s} interest_{us}}.$$

Encore une fois dans cet article Lathia and Capra [2011] cherchent à révéler les pratiques individuelles mais ils s'intéressent ici plus particulièrement aux réponses des usagers vis à vis des incitations de l'opérateur de transport. Ils étudient notamment la différence entre les déplacements et la perception que les usagers en ont et s'interrogent sur l'effet ou non des incitations des opérateurs.

Deux points sont particulièrement intéressants, à savoir, comment les usagers perçoivent leurs déplacement et comment ils perçoivent leurs habitudes d'achat. Les auteurs ont donc étudié les informations sur les jours de semaine, les habitudes de déplacement, le nombre de déplacement par jour, heures et les modalités de choix. Ils ont également regardé les habitudes de choix en séparant les deux types d'achat et en s'interrogeant sur la cause de l'achat et les habitudes. Une enquête a été menée, pour laquelle 119 réponses ont été collectées dont 85 répondant sont accepté l'accès à leurs données. ils ne disposent pas des informations démographiques mais du type d'abonnement (jeune, adulte, personne âgée). Un mois de données billettiques est également disponible sur toutes les cartes et de deux échantillons de 83 jours représentant des échantillons de 5% des voyageurs soit environ 300 000 voyageurs. Pour comparer pratiques réelles et perçues, les auteurs comparent le nombre de voyages par jour, les heures de voyages et heures de points, les modalités de trajet, les origines, destinations, statistiques atypiques et les habitudes d'achat.

Les auteurs étudient la corrélation entre incitation et pratiques. L'étude montre que l'achat d'un pass incite à un plus grand nombre de voyages sur les bus. Ils étudient également l'effet du titre plus cher aux heures de pointes, l'effet du passage au titre de voyage illimités pour la journée ou encore l'absence d'effets de la réduction sur les pratiques des usagers.

Ceapa et al. [2012] souhaitent révéler les motifs afin de pouvoir prévoir quand les transports seront bondés. Cela est un phénomène régulier en semaine et réparti sur une très courte durée.

Le jeu de données utilisé provient des TFL (1.6) et se compose de 31 jours en mars 2010, chaque donnée étant représentée par le tuple $< u, (o, d), t_o, t_d >$ dont on extrait les couplets $< o, t_o >$ et $< d, t_d >$. Pour chaque station les validations sont agrégées par 2 minutes.

Pour analyser la sur-affluence les auteurs ont mené une étude temporelle et une autre spatio-temporelle. L'analyse de l'activité montre que celle-ci n'est pas intense le week-end et ne sera par conséquent pas étudiée. Les auteurs ont aussi remarqué 3 petits pics pics dans l'heure de pointe du soir, l'heure de départ est plus étalée le soir. Les auteurs ont étudié 3 stations dans des secteurs différents (résidentiel, travail, transition). Ils ont fait un clustering hiérarchique par agglomération pour grouper les stations ayant les même motifs d'usage. dynamic time FastDTW et similarités entre clusters $D_{AB} = \frac{1}{n_A n_B} \sum_{a \in A} \sum_{b \in B} FastDTW(a, b)$.

La méthodologie développée afin de prédire la sur-affluence se décompose comme suit. Le jeux de données est séparé en deux jeux, l'un de calibrage et l'autre de test.On travaille sur des intervalles de 10 minutes et on fait un choix de seuil pour déterminer si le réseau est bondé ou non. Trois choix sont possibles : 0.6,0.5 et 0.8. Un choix de λ trop grand donnerait des résultats moins bons. Le but est d'éviter de prédire qu'il n'y aura pas trop de monde alors que si. Pour cela on défini la sensibilité $sensitivity = \frac{t_p}{t_p + f_n}$ avec t_p la proportion de résultats bien identifiés dans les résultats positifs.

Trois algorithmes de prédiction sont développés. Ceux-ci utilisent les termes $Train[t]$ le niveau d'affluence moyen à l'intervalle t sur le jeu d'entraînement, $Test[t]$ les niveaux d'affluence

observée sur le jeu de test et $Train[\bar{t}_1 - t_2]$ la moyenne du niveau d'influence de t_1 à t_2. On aura alors $HistoricValue(t, PW) = Train[t + PW]$, $HistoricMean(t, PW) = Train[t - \bar{(t + PW)}]$ et $HistoricMean(t, PW) = Train[t - \bar{(t + PW)}] - Train[t] + test[t]$.

Les résultats montrent qu'au niveau de la régularité $historicValue$ a de bons résultats, que le choix du λ est important. Cependant l'étude ne portait que sur un mois. période "classique"

8 Croisement avec des données socio-économiques

Lathia et al. [2012] ont testé si la mobilité urbaine permet de rendre compte des communautés existantes dans une ville. La validation est effectuée en utilisant les données de la ville de Londres (1.6) et en étudiant les flux d'indices de bien être (IMD). Les auteurs ont estimé à quelle communauté appartenait chaque utilisateur et ont ainsi obtenu une matrice de flux des motifs de visite entre les communautés. Les stations sont ensuite associées à la zone géographique la plus proche de l'IMD associé. La méthodologie nécessite l'estimation du lieu d'habitation de l'usager. Pour chaque usager, le nombre d'entrées et de sorties est compté et un classement des stations est ensuite effectué. Les deux stations les plus fréquentées pour chaque voyageur ne sont ensuite retenues que si elles contenaient au moins 2 voyages dans les 31 jours et n'étaient pas des stations importantes de correspondance. La création d'une matrice de visite des usagers s'est faite en comptant le nombre de visites du voyageur aux autres stations. Enfin la création d'une matrice de flux station par station est définie comme $F_{ij} = nombre\ de\ personnes\ vivant\ dans\ i\ et\ visitant\ j$. Cette matrice ne prend cependant pas en compte la fréquence. On peut alors corréler les IMD au flux.

Devillaine et al. [2012] ont cherché à obtenir une méthode de détection et d'estimation du lieu, de l'heure, de la durée et du motif de l'activité entreprise par l'usager. Deux jeux de données sont utilisés Santiago (1.5) et Gatineau (1.4). Pour le premier, le jeu de données court de la semaine du 28 juin 2010 au 4 juillet 2010 et contient environ 38 millions de transactions. Environ 3 millions de cartes sont enregistrées, il y a 300 lignes de bus 10 000 arrêts, 6000 bus et 85 km de rails. La validation ne se fait pas à la descente, ces lieux sont donc estimés. Enfin chaque validation génère une ligne de données contenant l'identifiant de la carte, la catégorie de la carte (étudiant, régulier), la date et le lieu de la transaction, les lignes et sens de validation, et le lieu de validation (bus, métro, station de bus). Le deuxième jeu de données est celui de Gatineau (1.4). Là encore il n'y a pas de données de validation à la descente, celle-ci est donc estimée. Chaque enregistrement contient l'identifiant de la carte, la catégorie de la carte, l'heure et le lieu de la transaction, l'arrêt de montée, la ligne et le sens de validation, le numéro du bus et du conducteur et l'heure de départ du bus.

Pour détecter les activités la même méthode est appliquée sur les deux villes, seuls les critères sont différents. La première étape consiste en l'estimation des arrêts de descente (utilisation de chaque méthode pour chaque ville). Puis les différentes étapes d'une même voyage sont regroupées dans le même voyage. Pour Gatineau si le transfert est supérieur à 30 minutes alors on considère que l'usager a effectué une activité et que ce n'est donc pas un transfert. Pour Santiago on fait la même chose sauf si les deux validations sont dans le métro ou sur la même ligne, dans ce cas on dira que c'est des activités. Si on n'a pas d'arrêt de descente on aura le même trajet dans les 2h qui suivent. En récupérant l'activité, l'identifiant d'utilisateur, l'arrêt de bus, l'heure de début, la date et la duré, un module d'affectation des motifs qui prend également en considération les caractéristiques de la ville. Les résultats montrent que les pics se concentrent sur les trajets domicile travail et qu'on constate une différence des heures d'influence selon la ville.

9 Étude des motifs de dynamiques de déplacement

Les travaux menés dans Tao et al. [2014] s'attachent à rendre plus pertinents les motifs de déplacement spatio-temporels des usagers en combinant le GTFS (General Transit Specification Feed) et la technologie GIS (Geographical Information System). Les auteurs ont ainsi créé des cartes représentant les flux conditionnelles (flow-comaps) pour pouvoir visualiser les motifs de flux agrégés sur le temps et l'espace et ce, de façon à faire apparaître la dynamique spatio-temporelle des pratiques des usagers. Ils ont également représenté des flux pondérés afin de faire apparaître les axes majeurs. On peut noter que ces travaux sont les premiers à s'intéresser aux motifs de mouvement et non pas aux motifs de déplacement au niveau des arrêts ou des zones géographiques restreintes et de mettre ainsi en évidence les dynamiques des usagers.

Cette étude utilise une journée de données billettiques fournies par l'agence du transport de Brisbane. Ce réseau est composé de bus, réseaux ferrés et ferry. Au total 515 435 transactions ont été enregistrées sur la journée dans le sud-est du Queensland. Une étude préliminaire ayant permis de détecter environ 3500 (0.06%) enregistrements avec des informations manquantes et ceux-ci n'étant pas très nombreux, ils ont été retirés de l'étude. Le système de données billettiques de Brisbane (appelé Go Card) est utilisé par plus de 80% des usagers lesquels doivent valider leur carte à la montée et à la descente. Pour chaque transaction l'identifiant de la route, la direction, l'identifiant de la carte, le type de la carte et l'identifiant du voyage (1 à n en fonction de la position dans la chaîne de déplacements) sont enregistrés. L'étude se concentre sur les voyages en bus soit un total de 245 549 validations enregistrées.

Pour cela la première étape consiste en l'extraction des motifs de service des bus à l'aide des fichiers GTFS. Cependant cette méthode ne permet pas d'identifier les arrêts auxquels le bus s'est arrêté, notamment pour les lignes express qui ne s'arrêtent pas à tous les arrêts. Une fois les motifs extraits, il s'agit de reconstruire les trajectoires en combinant le fichier GTFS avec les données de validation des cartes. Cette étape sépare chaque enregistrement de voyage d'utilisateur en un certain nombre de petit voyages pour lesquels chaque arrêt est la fin d'un voyage et le début d'un autre. Enfin la dernière étape consiste en la création d'une carte des flux conditionnelle, celle-ci intégrant une carte des flux mais également un graphique conditionnel.

10 Conclusion

En guise d'introduction aux problématiques qui seront abordées dans le projet Mobilletic, ce livrable dresse un état de l'art sur les données billettiques pour l'analyse des mobilités, notamment en détaillant les principales questions abordées dans la littérature sur ce sujet, qui passent par l'enrichissement de ces données, leur analyse ou encore leur croisement avec d'autres sources de données socio-économiques en particulier.

Références

B. Agard, C. Morency, and M. Trépanier. Mining public transport user behaviour from smart card data. In *The 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)*, pages 17–19, 2006.

B. Agard, C. Morency, and M. Trépanier. Mining smart card data from an urban transit network. In *IGI Global*, pages 1–11, 2009.

M. Bagchi and P. R. White. What role for smart card data from bus systems. In *Proceedings of the Institution of Civil Engineers. Municipal Engineer 157, 39–46. March Issue ME1*, 2004.

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

J.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*, 1817 :183–187, 2002.

I. Ceapa, C. Smith, and L. Capra. Avoiding the crowds : Understanding tube station congestion patterns from trip data. In *Proceeding of the 1st ACM SIGKDD International Workshop on Urban Computing. ACM press*, pages 134–141, 2012.

K. Chu and R. Chapleau. Enriching archived smart card transaction data for transit demand modeling. *Transportation Research Record : Journal of the Transportation Research Board*, 2063 :63–72, 2008.

K. Chu and R. Chapleau. Augmenting transit characterization and travel behaviour comprehension. multiday location stamped card transactions. *Transportation Research Record : Journal of the Transportation Research Board*, 2183 :29–40, 2010.

K. Chu, R. Chapleau, and M. Trépanier. Driver-assisted bus interview. passive transit travel survey with smart card automatic fare collection system and applications. *Transportation Research Record : Journal of the Transportation Research Board*, 2105 :1–10, 2009.

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 :48–55, 2012.

T. Fuse, K. Makimura, and T. Nakamura. Observation of travel behavior by ic card data and application to transportation planning. In *Special Joint Symposium of ISPRS Commission IV and AutoCarto 2010*, 2010.

N. Lathia and L. Capra. How smart is your smartcard ? measuring travel behaviours, perceptions, and incentives. In *ACM International Conference on Ubiquitous Computing. Beijing, China*, 2011.

N. Lathia, J. Froehlich, and L. Capra. Mining public transport usage for personalised intelligent transport systems. In *IEEE International Conference on Data Mining. Sydney, Australia*, 2010.

N. Lathia, D. Quercia, and J. Crowcroft. The hidden image of the city : sensing community well-being from urban mobility. In *Proceedings of the 10th international conference on Pervasive Computing, Newcastle, UK*, 2012.

C. Morency, M. Trépanier, and B. Agard. Analysing the variability of transit users behaviour with smart card data. In *The Ninth International IEEE Conference on Intelligent Transportation Systems, Toronto, Canada, September*, 2006.

M. A. Munizaga and C. Palma. Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smart card data from santiago, chile. *Transportation Research Part C : Emerging Technologies*, 24 :9–18, 2012.

J. Y. Park, D.-J. Kim, and Y. Lim. Use of smart card data to define public transit use in seoul, south korea. *Transportation Research Record : Journal of the Transportation Research Board*, 2063 :3–9, 2008.

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 :55–62, 2009.

S. Tao, D. Rohde, and J. Corcoran. Examining the spatial-temporal dynamics of bus passenger travel behaviour using smart card data and the flow co-map. *Journal of transport geography*, 41 :21–36, 2014.

M. Trépanier, N. Tranchant, and R. Chapleau. Individual trip destination estimation in a transit smart card automated fare collection system. *Intelligent Transportation Systems*, 11 : 1–14, 2007.

M. Utsunomiya, J. Attanucci, and N. Wilson. Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation Research Record*, 1971 : 119–126, 2006.

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