PolyPublie
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Title:	Leveraging data from a smart card automatic fare collection system for public transit planning
Auteur: Author:	Ka Kee Chu
Date:	2010
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Chu, K. K. (2010). Leveraging data from a smart card automatic fare collection system for public transit planning [Thèse de doctorat, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/441/

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/441/
Directeurs de recherche: Advisors:	Robert Chapleau
Programme: Program:	Génie civil

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

UNIVERSITÉ DE MONTRÉAL

LEVERAGING DATA FROM A SMART CARD AUTOMATIC FARE COLLECTION SYSTEM FOR PUBLIC TRANSIT PLANNING

KA KEE CHU

DÉPARTEMENT DES GÉNIES CIVIL, GÉOLOGIQUE ET DES MINES

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION

DU DIPLÔME DE PHILOSOPHIÆ DOCTOR

(GÉNIE CIVIL)

DÉCEMBRE 2010

© Ka Kee Chu, 2010.

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée:

LEVERAGING DATA FROM A SMART CARD AUTOMATIC FARE COLLECTION SYSTEM FOR PUBLIC TRANSIT PLANNING

présentée par : CHU Ka Kee

en vue de l'obtention du diplôme de : Philosophiæ Doctor

a été dûment acceptée par le jury d'examen constitué de :

Mme MORENCY Catherine, Ph.D., présidente

M. CHAPLEAU Robert, Ph.D., membre et directeur de recherche

M. TRÉPANIER Martin, Ph.D., membre

M. EL-GENEIDY Ahmed, Ph.D., membre

ACKNOWLEDGMENT

I would like to thank my immediate family, Eric, Cecilia, Garry and family, for their unconditional love and care. I thank them for the sacrifices they made so that I can become who I am today.

I would like to thank Monsieur Robert Chapleau, a visionary and uncompromising professor, my mentor and research director, who taught and educated me on the subject of transport, for his guidance, support and dialogues throughout this multi-year journey. I would like to thanks members of the MADITUC group, especially Mr. Guillaume Bisaillon and Mr. Daniel Piché, for their help throughout the research. Thanks to my fellow students for the stimulating discussions and comradeship.

I would like to thank Professor Martin Trépanier and the Société de Transport de l'Outaouais for providing the smart card dataset in this research. Thanks to the Fonds québécois de la recherche sur la nature et les technologies and École Polytechnique de Montréal for the scholarships. Thanks to the Agence Métropolitaine de Montréal and Mr. Daniel Bergeron for being accommodative and the opportunity to work on the smart card dataset of Montréal. Thanks to Québec and Canada for the great environment.

Thanks to brothers and sisters in Christ for the fellowship and prayers. Thanks to my extended family and friends. They are a blessing to me and enrich my life.

Thanks to God, creator and saviour, for all things he has done:

> *Oh, the depth of the riches of the wisdom and knowledge of God!*
>
> *How unsearchable his judgments, and his paths beyond tracing out!*
>
> *Who has known the mind of the Lord? Or who has been his counselor?*
>
> *Who has ever given to God, that God should repay him?*
>
> *For from him and through him and to him are all things. To him be the glory for ever! Amen.*
>
> (Rom 11:33-36, New International Version)

RÉSUMÉ

Le système de transport en commun est une créature artificielle et complexe. L'interaction spatio-temporelle entre l'offre de service par les opérateurs et la demande des usagers est difficile à mesurer et évolue constamment. C'est dans ce contexte que de nombreux efforts sont mis à la recherche de l'information et de la méthodologie qui peuvent contribuer à révéler et à comprendre cette relation dynamique afin que les services répondent aux besoins des voyageurs.

Récemment, des changements aux paradigmes remodèlent ce processus. D'une part, les opérateurs de transport en commun adoptent une orientation axée sur la performance et le client. Ceci demande des données qui ne sont pas recueillies par des enquêtes traditionnelles. D'autre part, l'avancement des systèmes automatiques de collecte de données et leur adoption par les opérateurs génèrent une abondance de données dans un environnement où la collecte de données était auparavant limitée. Les systèmes d'analyse et de planification sont souvent adaptés à ces réalités et sont inadéquats pour exploiter de nouvelles sources de données. Au confluent de ces évolutions, se retrouvent un défi et une opportunité : apprivoiser les nouvelles technologies informationnelles dans le but de les réconcilier avec les besoins grandissants de données dans le domaine de transport en commun. Cette recherche se fonde sur un jeu de données de validation provenant d'un système de perception par carte à puce (CAP). Le but de la recherche est de développer de nouvelles méthodologies d'exploitation des données, notamment au niveau de leur traitement, de leur enrichissement et de leur analyse afin de mieux connaître la demande de transport en commun, d'améliorer la planification opérationnelle, de raffiner la gestion du système et de comprendre les comportements de déplacement.

Le jeu de données principal provient du système de perception par CAP de la Société de transport de l'Outaouais (STO). Le système est muni de GPS et le jeu contient toutes les validations désagrégées pour le mois de février 2005. Les technologies informationnelles, incluant la base de données relationnelle, le système d'information géographique (SIG), les statistiques spatiales, le data mining et les visualisations, sont des principaux outils de traitement et d'analyse.

Trois principes globaux guident les travaux de recherche : l'approche informationnelle basée sur les données réelles, l'approche totalement désagrégée et l'approche orientée objet. Un nouveau concept, l'approche informationnelle multi-jour, est proposé en jumelant ces principes et des données multi-jour de CAP. Cette approche est le concept de base dans plusieurs procédures de

traitement et d'enrichissement de données. Elle fait l'hypothèse que l'information contenue dans les données d'un jour est partielle et peut même comporter des erreurs. En cumulant et faisant la synthèse de plusieurs jours de données, il est possible de reconstituer l'information complète. Cette dernière peut subséquemment servir comme référence pour la correction de données ainsi que pour l'interprétation de l'information.

En comparant avec les données de l'enquête origine-destination régionale, il est démontré que les données de CAP répondent plus adéquatement aux besoins de la planification de transport en commun en matière de fréquence, de couverture et de résolution. Elles permettent aussi les bénéfices suivants : absence d'effet de fatigue et de refus chez les répondants, absence d'erreurs de transcription, codification uniforme, intégration de données opérationnelles ainsi que précision accrue des réponses. Grâce à ces avantages, les données de CAP ont des usages polyvalents. Par contre, tout comme les autres méthodes passives, certaines dimensions ne sont pas capturées. Des procédures de traitement et d'enrichissement de données sont nécessaires pour combler ces lacunes.

Le jeu de données compte 763 570 transactions de validation provenant de 21 813 cartes. Une stratégie de validation de données est proposée pour améliorer la qualité de celles-ci tout en les gardant cohérentes. Ceci implique la détection d'erreurs et la correction d'erreurs par imputation. Deux raisons principales justifient cette approche : la discontinuité spatio-temporelle causée par les valeurs erronées ainsi que la propagation et l'amplification de l'erreur. La stratégie de détection d'erreurs se base sur les logiques spatio-temporelles et celles du transport en commun. Environ 15% de transactions contiennent des valeurs erronées et suspectes. Les données portant sur le voyage et l'arrêt de montée sont corrigées par imputation en utilisant les concepts de la répétition des services planifiés et de l'historique de montées des individus. Après la validation des données, 98,1% de transactions sont considérées comme valides comparativement à 84,3% avant la procédure.

Les données corrigées satisfont les besoins de nombreuses opérations en planification via l'application de procédures standards de SIG et de statistiques spatiales. Il est possible de décrire les montées selon plusieurs niveaux d'agrégation : arrêt, ligne ainsi que liens et nœuds au niveau du réseau. La modélisation de la demande au niveau du réseau requiert un traitement plus avancé des données. L'affectation totalement désagrégée de la demande selon l'approche MADITUC

nécessite des déplacements désagrégés sous forme d'itinéraires. Pour ce faire, plusieurs étapes d'enrichissement sont effectuées : l'estimation de l'arrêt de descente pour chaque montée selon le concept « chaîne de montées»; l'interpolation de l'heure d'arrivée à l'arrêt pour chaque voyage selon l'information temporelle contenue dans les transactions et l'identification de correspondance selon le concept « coïncidence spatio-temporelle ».

Ces enrichissements permettent de reconstruire l'itinéraire complet à partir de données de montée. Les autres objets liés aux itinéraires, tels que la durée d'activité, la distance de déplacement, la durée de déplacement et la vitesse moyenne de déplacement, sont dérivés grâce à l'approche totalement désagrégée. L'analyse des correspondances montre que durant un jour typique de semaine, le nombre de correspondances selon la définition tarifaire est environ 40% plus élevé que celui estimé selon le concept « coïncidence spatio-temporelle ».

Le profil de charge, un outil classique en planification, est construit à partir de données bonifiées. Il peut être décomposé par déplacement individuel (en fonction de différents attributs) et jumelé avec la trajectoire spatio-temporelle du véhicule pour des analyses de ponctualité. Il permet aussi de calculer des statistiques opérationnelles standards par arrêt, par segment ou par ligne.

L'association entre générateurs de déplacement et arrêts est accomplie par l'approche multi-jour et la consolidation spatio-temporelle des itinéraires. L'approche multi-jour vise à caractériser ou interpréter des attributs d'un déplacement par rapport à l'ensemble des déplacements à l'intérieur de la période d'analyse. Une analyse des points chauds permet d'identifier les lieux d'ancrage de l'usager. Grâce à cette approche, 43% des cartes à tarif étudiant sont associées aux établissements scolaires. Les secteurs de domicile sont aussi dérivés. Chaque déplacement est interprété selon la liste personnalisée de lieux d'ancrage. Cette démarche permet d'abord d'étudier le comportement multi-jour d'un usager. Une table de déplacements et le journal d'activité mensuelle peuvent être reproduits de manière très détaillée pour chaque usager. Il est aussi possible de comparer le comportement d'un sous-groupe d'usagers qui partagent un lieu d'ancrage commun. La caractérisation multi-jour amène les chercheurs à repenser certains aspects fondamentaux dans la description des déplacements. L'application de deux techniques de data mining - les règles d'association et la classification - est proposée pour l'analyse des comportements de déplacement.

Les données provenant du système de la STO sont relativement simples et la recherche examine principalement les données de validation. Afin d'illustrer la complexité et les défis techniques d'exploiter les données d'un système multi-opérateur et multimodal, les données provenant du système de perception par CAP de la région métropolitaine de Montréal, OPUS, sont utilisées. La potentialité des données de vente et de vérification en planification, opération et gestion est démontrée. Étant donné que le montage de chaque système de perception par CAP, le réseau et la structure tarifaire sont uniques, les besoins sur le traitement et l'enrichissement de données doivent être évalués individuellement. Cependant, les principes, les approches méthodologiques et les analyses proposés dans la recherche peuvent être adaptés et transférés à des données similaires. Les perspectives de recherche sont proposées.

ABSTRACT

Public transit system is an artificial and complex creature. The interaction between operators' supply and users' demand is at the same time spatial and temporal. It is also difficult to measure and in constant evolution. There is a continuous quest for information and methodology that can help reveal and facilitate the understanding of this dynamic relationship, so that public transit services can be better organized to suit travelers' needs.

Recent paradigm shifts have contributed the reshaping of this process. On the one hand, public transit service has become more performance-driven and customer-oriented. These require data not covered by traditional survey methods. On the other hand, advances in passive data collection methods and their adoption by transit operators progressively transform the industry from data-poor to data-rich. Traditional analysis and planning tools are adapted to past conditions and are not suited to fully leverage new sources of data. At the confluence of these evolutions lies opportunity and challenge: to embrace the data-rich environment with the view of reconciling with the increasingly demanding data needs in public transit. The research is based on a set of validations data from a smart card automatic fare collection (AFC) system. The goal of the research is to develop new methods in data processing, data enrichment and data analysis in order to better quantify transit demand, enhance operations planning, improve system management and understand travel behaviour.

The primary dataset comes from the smart card AFC of the Société de transport de l'Outaouais (STO). The system is equipped with GPS and the dataset contains all fare validations in a disaggregate form for the month of February 2005. Information technologies, including relational database, geographic information system (GIS), spatial statistics, data mining and visualization are the main data processing and analysis tools.

Three overall principles guide the research: the information-based (data-driven) approach, the totally disaggregated approach and the object-oriented approach. Combined with multi-day smart card data, these principles lead to the multi-day information approach, a new concept used in the proposed data processing and enrichment procedures. The assumption is that each day of data represent partial information of the universe and may contain errors. By synthesizing the correct information from each day, it is possible to reconstruct complete knowledge. The latter is in turn used as a reference to analyze and interpret multi-day data.

When studied as an analogue of the regional origin-destination survey, it is demonstrated that smart card data answer more adequately the needs of transit planning in terms of timeliness, coverage and resolution. Other benefits include absence of non-response and respondent fatigue; absence of transcription error; systematic and uniform coding; more precise values and integration of operations data. These properties allow them to be used as a versatile multi-purpose transit survey. Similar to other passive data, certain dimensions of travel cannot be captured. Data processing and enrichment procedures are therefore required.

The dataset contains 763,570 validation transactions from 21,813 cards. A validation strategy is proposed to improve data quality by assuring their internal coherence. This involves error detection and data correction by imputation. The rationale of this approach is to re-establish spatial-temporal continuity, and to avoid the propagation and amplification of error. The error-detection strategy is based on spatial-temporal and public transit logics. About 15% of transactions contain erroneous or suspect values. Run and stop values are corrected by imputation based on the concepts of repetition of scheduled service and the boarding history of individual cardholders. After the procedure, 98.1% of transactions are considered valid as opposed to 84.3% before the procedure.

Many tasks in transit planning can be served directly by applying standard GIS procedures and spatial statistics to corrected data. It is possible to describe boarding in several levels of aggregation: stop, route, and links and nodes in a network. Demand modeling at the network level requires a more advanced data processing. A MADITUC-style totally disaggregate transit assignment implies a trip in the form of an itinerary. Several data enrichment steps are undertaken: alighting stop estimation for each boarding with the concept of “boarding chain”; interpolation of stop arrival time for each vehicle run according to temporal information embedded in the transactions and transfer identification with the concept of “spatial-temporal coincidence”.

These enrichments allow the reconstruction of complete itineraries from boarding data. Other objects, such as activity duration, distance traveled and average speed of a trip, are derived thanks to the totally disaggregated approach. Transfer analysis shows that on a typical weekday, the number of transfers revealed by the AFC system is about 40% higher than those estimated with the concept of “spatial-temporal coincidence”.

Load profile, a classic tool in transit planning, is synthesized from enriched data. It can be broken down into individual trip and its attributes, merged with spatial-temporal path of vehicle for schedule adherence analysis. Standard operations statistics by stop, segment or route can be calculated.

The association between trip generators and stops are achieved by the multi-day informational approach and spatial-temporal consolidation of itinerary. Multi-day informational approach aims to characterize or interpret trip attributes with respect to all trips within the analysis period. A hotspot analysis reveals anchor points of a cardholder. With this approach, 43% of cards with student fare are assigned to educational establishments. Residence areas are also derived. Each trip is interpreted with respect to a personalized list of anchors. A trip table and a monthly activity schedule can be reconstructed with a lot of details for each cardholder. It also allows travel behaviour comparison between a subgroup of cardholders sharing a common anchor. The multi-day trip characterization leads researchers to rethink some fundamental aspects of trip description. Applications of two data mining techniques, association rules and classification, are proposed for travel behaviour analysis.

Data from the smart card AFC system of the STO are relatively simple and the primary focus of the research is on the validation data. To address this issue, data from a multi-operator and multi-modal AFC system in the Greater Montréal Area, OPUS, are used to illustrate the complexity and technical challenges. They are also used to introduce the potential of other types of data, namely sales and verification data, that are suitable for transit planning, operations and management. Since the setup of each smart card AFC system, the transit network and its fare structure is unique, the needs on data processing and enrichment vary and are specific to each system. However, the principles, the methodological approaches and the analyses proposed in this research can be adapted and transferred to datasets with similar structure. Perspective research topics are also proposed.

CONDENSÉ EN FRANÇAIS

Chapitre 1 Les données dans le domaine du transport en commun : une introduction

Le système de transport en commun est une créature artificielle et complexe. Son existence et son utilité dépendent de l'adéquation entre l'offre de service de l'opérateur et la demande qui représente les besoins de la population en matière de déplacements. L'interaction est à la fois spatiale – car les origines et les destinations de la population doivent coïncider avec celles du service; elle est aussi temporelle – car les déplacements doivent être effectués selon les contraintes du temps. Cette interaction est difficile à mesurer et évolue constamment. C'est dans ce contexte que de nombreux efforts sont faits afin d'identifier l'information et la méthodologie qui puissent contribuer à révéler et à faciliter la compréhension de cette relation dynamique.

Récemment, des changements aux paradigmes remodèlent ce processus. D'une part, les opérateurs de transport en commun adoptent une orientation axée sur la performance et le client. Pour ce faire, ils ont besoin de continuellement surveiller plusieurs aspects du service, dont la performance et la qualité. Ceci nécessite des données qui ne sont pas recueillies par des enquêtes traditionnelles. D'autre part, l'avancement des systèmes automatiques de collecte de données et leur adoption par les opérateurs génèrent une abondance de données dans un environnement où la collecte de données était auparavant limitée par le coût et la main d'œuvre des méthodes manuelles. Les systèmes d'analyse et de planification sont souvent adaptés à ces réalités et sont inadéquats pour exploiter de nouvelles sources de données.

Au confluent de ces évolutions, se retrouvent un défi et en même temps une opportunité : apprivoiser les nouvelles technologies informationnelles dans le but de les réconcilier avec les besoins grandissants de données dans le domaine de transport en commun. Cette recherche se fonde sur un jeu de données de validation provenant d'un système de perception par carte à puce (CAP). Le but de la recherche est de développer de nouvelles méthodologies d'exploitation des données, notamment au niveau de leur traitement, de leur enrichissement et de leur analyse afin de mieux connaître la demande de transport en commun, d'améliorer la planification opérationnelle, de raffiner la gestion du système et de comprendre les comportements de déplacement.

Chapitre 2 La mise en contexte : une revue de littérature

Afin de bien comprendre le contenu des données de CAP, il est nécessaire de maîtriser les concepts de la planification opérationnelle en transport en commun. La compréhension des objets véhicule, voyage, tournée, arrêt, ligne-arrêt et leurs interrelations est essentielle dans le développement de procédures de traitement et d'enrichissement ainsi que dans les analyses de données. Au cours des dernières années, est apparu un volume grandissant d'études sur l'utilisation de données passives en transport en commun. Les grands thèmes sont les suivants :

- la détection d'erreurs et la validation de données passives;
- le potentiel et les exemples d'usage dans la planification du transport en commun et l'évaluation de la performance des services;
- l'estimation de l'origine et de la destination des déplacements en transport en commun;
- l'analyse sur le comportement des voyageurs;
- l'enrichissement de données en ce qui a trait au mode de transport, aux correspondances, aux extrémités du déplacement et aux motifs de déplacement.

Les technologies informationnelles suivantes sont des principaux outils d'analyse :

- la base de données relationnelle et la requête SQL;
- le système d'information géographique (le SIG);
- les statistiques spatiales;
- le data mining (l'exploration de données);
- les visualisations.

Chapitre 3 Les données et les principes de recherche

Le jeu de données de validation provient du système de perception par CAP de la Société de transport de l'Outaouais (STO). Le système est muni de GPS et le jeu de données contient toutes les validations effectuées pendant le mois de février 2005 sous une forme désagrégée. Quelques dictionnaires sont nécessaires pour le décodage du contenu. La recherche intègre d'autres jeux de données qui ne proviennent pas directement du système. Ce sont :

- l'information dérivée à partir de données de validation telle que les services planifiés;

- les horaires planifiés aux points de contrôle;
- les données spatiales des points d'intérêt.

Trois principes généraux guident les travaux de recherche :

- l'approche informationnelle;
- l'approche totalement désagrégée;
- l'approche orientée objet.

Deux concepts ont également inspiré le développement des procédures de traitement et d'enrichissement de données : la maximisation de l'entropie et le programme d'amorce (bootstrapping).

L'approche informationnelle multi-jour, qui est utilisée dans plusieurs étapes de la recherche, fait l'hypothèse que l'information contenue dans les données d'un jour est partielle et peut même comporter des erreurs. En accumulant plusieurs jours de données validées et en en faisant la synthèse, il est possible de reconstituer l'information complète. Cette dernière peut subséquemment servir comme référence pour la correction de données ainsi que pour l'interprétation de l'information.

Chapitre 4 Les données de CAP comme une enquête origine-destination

Les enquêtes traditionnelles sur le transport sont souvent classifiées selon les caractéristiques non-exclusives suivantes :

- le type « préférences déclarées » ou « préférences observées »;
- le mode d'enquête;
- l'entité « ménage » ou « individu »;
- les modes de transport couverts;
- la durée et la fréquence.

Les données des enquêtes OD régionales sont appropriées pour des analyses de grande tendance, mais ne sont pas adéquates pour la planification opérationnelle de transport en commun en termes de fréquence, d'échantillonnage et de résolution. Les données de CAP captent en continu la totalité (ou presque) des activités de tous les détenteurs de carte dans tout le réseau et ont des

résolutions spatiale et temporelle très fines (à l'arrêt et à la minute près). Elles offrent aussi d'autres avantages :

- l'absence d'effet de fatigue chez les répondants;
- l'absence de refus;
- une codification uniforme, l'absence d'erreur de transcription et une précision accrue des réponses;
- l'absence de l'entrevue et l'épargne des ressources associées;
- l'intégration de données opérationnelles.

Grâce à ces avantages, les données de CAP ont des usages polyvalents et peuvent être exploitées comme une enquête sur la demande de transport en commun, une enquête en continu, une enquête des générateurs de déplacement, une enquête sur l'allocation et la consommation de ressources ainsi qu'une enquête de type « préférences observées ».

Par contre, comme la collecte de données est passive, plusieurs éléments, notamment la dimension socio-économique, les extrémités et le motif de déplacement ne sont pas saisis. De plus, comme la validation de titre se fait à l'entrée du véhicule, l'arrêt de descente n'est pas capté par le système. Des procédures de traitement et d'enrichissement de données sont nécessaires pour combler ces lacunes.

Chapitre 5 Analyses exploratoires et stratégie de validation de données

Le mois de février 2005 compte 28 jours, dont 20 jours de semaine. Au total, il existe 763 570 transactions provenant de 21 813 cartes dans le jeu de données. Le 10 février est choisi en tant que représentant du jour typique de la semaine dans diverses démonstrations. Le type de tarif est un attribut intéressant pour la segmentation des usagers. L'analyse exploratoire démontre qu'il est possible de générer des indicateurs globaux avec des données brutes.

Une stratégie de validation de données est proposée afin d'améliorer la qualité de celles-ci tout en les gardant cohérentes. Ceci implique la détection et la correction d'erreur par imputation. Deux raisons principales justifient cette approche : la discontinuité spatio-temporelle causée par les valeurs erronées ainsi que la propagation et l'amplification de l'erreur. Ces deux lacunes

empêchent l'enrichissement de données et influencent négativement la validité des résultats analytiques.

La stratégie de détection d'erreur est basée sur des logiques spatio-temporelles et des logiques liées à l'organisation du transport en commun. Des règles sont appliquées aux données afin de détecter les erreurs. Environ 15% de transactions contiennent des valeurs erronées ou suspectes. Les transactions erronées forment souvent une séquence chronologique de transactions appartenant à un même véhicule. Sous l'hypothèse que le numéro du véhicule, le numéro de la carte et l'heure de transaction soient toujours valides, les données portant sur le voyage et l'arrêt de montée sont corrigées par imputation selon les deux concepts suivants:

- la répétition de services planifiés : les validations sont réaffectées à un voyage selon l'horaire planifié. Un dictionnaire des services planifiés dérivé des données multi-jour et une association entre les objets tournées et véhicules sont essentiels.
- l'historique des montées des individus : on suppose qu'il existe une régularité dans les déplacements d'un voyageur. Pour chaque transaction erronée, un algorithme prédit le voyage et l'arrêt de montée les plus probables selon l'historique mensuel de montées de la carte.

L'ensemble des prédictions sert à la contre-validation. Le concept de la répétition des services planifiés aide principalement à proposer un voyage tandis que le concept de l'historique des montées des individus sert notamment à suggérer un arrêt de montée. Après la procédure de validation de données, 98.1% de transactions sont considérées comme valides comparativement à 84.3% avant la procédure. D'ailleurs, la validation des données de référence, telles que la géométrie des lignes, est également importante car elle peut influencer de façon significative les résultats des traitements de données.

Chapitre 6 Les méthodes et les analyses pour la planification opérationnelle

Les données corrigées ont un niveau de résolution spatio-temporelle très fin. Elles satisfont les besoins de nombreuses opérations en planification en appliquant des procédures standards de SIG et de statistiques spatiales. Il est possible de décrire les montées selon plusieurs niveaux d'agrégation : arrêt, ligne ainsi que liens et nœuds au niveau du réseau. Par exemple :

- la description des montées à l'arrêt et de leur évolution spatio-temporelle au cours d'une journée par des ellipses écarts-types, des barycentres, la densité sur une grille ou d'autres représentations géolocalisées et multivariées;
- la description des montées par ligne selon l'heure;
- le nombre de montées à l'arrêt par voyage pour une ligne et la progression temporelle des véhicules.

Chapitre 7 Les méthodes et les analyses pour la modélisation de la demande

La modélisation de la demande au niveau du réseau appelle à un traitement plus avancé des données. L'affectation de la demande selon l'approche MADITUC nécessite des déplacements désagrégés sous forme d'itinéraires, où l'itinéraire est défini comme un déplacement-personne unidirectionnel d'une origine à une destination avec un motif spécifique. Pour ce faire, plusieurs étapes d'enrichissement sont effectuées :

- l'estimation de l'arrêt de descente pour chaque montée selon le concept « chaîne de montées »;
- l'interpolation de l'heure d'arrivée à l'arrêt pour chaque voyage selon l'information temporelle contenue dans les transactions. La trajectoire est extrapolée selon le temps de parcours associé à l'horaire planifié;
- l'identification de correspondance selon le concept « coïncidence spatio-temporelle ».

Ces enrichissements permettent de reconstruire l'itinéraire complet à partir de données de montée. Les autres objets liés aux itinéraires tels que la durée d'activité, la distance de déplacement, la durée de déplacement et la vitesse moyenne de déplacement sont dérivés grâce à l'approche totalement désagrégée.

D'autres applications en planification de transport sont étudiées. L'analyse des correspondances permet aux planificateurs d'étudier la distribution spatio-temporelle des correspondances observées dans le réseau. Il est démontré que durant un jour typique de semaine, le nombre de correspondances selon la définition tarifaire est de l'ordre de 40% plus élevé que celui estimé selon le concept « coïncidence spatio-temporelle ». Le temps de correspondance et le temps d'attente lors des correspondances sont examinés.

Le profil de charge, un outil classique en planification, est construit à partir de données bonifiées et a les caractéristiques suivantes :

- le profil de charge peut être décomposé et étudié en fonction de chaque déplacement individuel comportant un arrêt de montée et un arrêt de descente;
- il a une résolution spatiale à l'arrêt;
- il peut être construit pour tous les voyages;
- il peut être jumelé avec la trajectoire spatio-temporelle du véhicule et sert à l'analyse de ponctualité;
- les statistiques opérationnelles standards, telles que les passagers-kilomètres, le facteur de charge, le point de charge maximale, la proportion des usagers provenant des correspondances et le type de tarif, peuvent être calculées par arrêt, par segment ou par ligne.

La construction d'un réseau pour l'affectation des itinéraires amène à une fusion des lignes-arrêts en nœuds de type MADITUC ainsi que la « recodification » des lignes selon ces nouveaux nœuds. L'affectation résulte en des profils de charge sur les liens et en des entrants-sortants aux nœuds. Les attributs des déplacements ou des usagers, par exemple le secteur de la première montée et le type de tarif, peuvent être différenciés dans le profil de charge.

L'évolution spatio-temporelle de l'occupation des lieux d'activité par des usagers de transport en commun peut être modélisée et représentée par des cartes de l'occupation du sol au cours d'une journée. Celle-ci amène naturellement à associer des générateurs de déplacement aux arrêts d'autobus et ensuite à dériver les extrémités pour chaque itinéraire. L'approche multi-jour et la consolidation spatio-temporelle des itinéraires révèlent la régularité dans le comportement de déplacement de l'usager.

Chapitre 8 Les méthodes et les analyses pour l'étude des comportements de déplacement

L'approche multi-jour vise à caractériser ou interpréter les attributs d'un déplacement par rapport à l'ensemble de déplacements réalisés à l'intérieur de la période d'analyse. Une analyse des points chauds permet d'identifier les lieux d'ancrage de l'usager. Ces lieux sont associés à un générateur spécifique, un secteur géographique ou une région vague. Dans le cas des cartes à tarif

étudiant, il est supposé que chaque carte est associée à une institution scolaire. Une affectation est faite selon la régularité spatio-temporelle des lieux de montée. Environ 43% des cartes sont affectées à une institution scolaire. Une surface probabiliste est créée afin de montrer les lieux d'ancrage sans coordonnées spatiales spécifiques tels que le domicile. Avec la liste de lieux d'ancrage de l'usager comme référence, les extrémités de chaque déplacement sont interprétées.

Cette démarche permet d'abord d'étudier le comportement multi-jour d'un usager. Une table de déplacements est compilée. Elle contient notamment les paires origine-destination, leur fréquence, la durée moyenne d'activité, l'heure de montée et sa variance ainsi que la distance parcourue. Le journal d'activité mensuelle est reproduit en détail pour un usager. La démarche permet aussi de comparer le comportement d'un sous-groupe d'usagers qui partagent un lieu d'ancrage commun. Par exemple, l'utilisation de SIG, de statistiques spatiales et d'animation peut illustrer les domiciles dérivés des étudiants fréquentant un même établissement scolaire et l'heure de départ depuis leurs domiciles. La fidélité envers le service de transport en commun à travers les jours représente aussi une analyse multi-jour intéressante.

Deux techniques de data mining sont utilisées pour l'analyse des comportements de déplacement. Les règles d'association permettent de détecter les valeurs des attributs associées fréquemment au même déplacement et par conséquent, la régularité dans le comportement de déplacement. La classification permet de comparer la régularité du comportement de déplacement parmi différents individus en utilisant les taux de prédiction des modèles individuels comme mesure.

Les données de CAP multi-jour permettent aux planificateurs de découvrir des cycles d'activité étendus. Elles amènent les chercheurs à repenser certains aspects fondamentaux associés à la description des déplacements. Avec la caractérisation multi-jour des déplacements, il n'est peut-être plus nécessaire d'identifier le motif de déplacement. La procédure de l'estimation de l'arrêt de descente peut être bonifiée avec le concept « multi-jour ».

Chapitre 9 Les limitations et la généralisation des conclusions

Les données provenant du système de la STO sont relativement simples et la recherche examine principalement les données de validation. Afin d'illustrer la complexité et les défis techniques d'exploiter les données d'un système multi-opérateur et multimodal, les données provenant du système de perception par CAP de la région métropolitaine de Montréal, OPUS, sont utilisées. La

potentialité des données de vente et de vérification en planification, opération et gestion est démontrée. Étant donné que le montage de chaque système de perception par CAP, le réseau et la structure tarifaire sont uniques, les besoins sur le traitement et l'enrichissement de données doivent être évalués individuellement. Cependant, les principes, les approches méthodologiques et les analyses proposés dans la recherche peuvent être adaptés et transférés à des données similaires.

Chapitre 10 Les contributions et les perspectives

La recherche contribue sur les plans conceptuel, théorique, méthodologique et analytique. Les méthodologies proposées s'appuient sur des données réelles et des résultats expérimentaux. La recherche témoigne de l'utilité des données de CAP dans la planification du transport en commun et l'étude du comportement des voyageurs. Elle justifie le traitement et l'enrichissement de données. Les perspectives de recherche incluent :

- l'exploitation systématique de données provenant d'un système de perception multimodal et multi-opérateur par CAP;
- la synthèse de données passives pour la planification et l'exploitation du transport en commun;
- l'étude du choix d'itinéraire en transport en commun avec des données de CAP;
- le développement d'un algorithme d'affectation des itinéraires multi-jour dérivés à partir de données de CAP.

Certaines conclusions de cette recherche sont déjà diffusées via la participation aux congrès régionaux et internationaux (AQTR, WCTR, TRB, ISCTSC et IATBR) ainsi que publiées dans trois articles dans une revue à comité de lecture (TRR).

TABLE OF CONTENTS

ACKNOWLEDGMENT III

RÉSUMÉ IV

ABSTRACT VIII

CONDENSÉ EN FRANÇAIS XI

TABLE OF CONTENTS XX

LIST OF TABLES XXVII

LIST OF FIGURES XXIX

LIST OF ACRONYMES AND ABREVIATIONS XXXIV

CHAPTER 1 DATA IN PUBLIC TRANSIT: AN INTRODUCTION 1

1.1 Motivation of Research 1

1.2 Data Needs and Data Collection in Public Transit 3

1.3 Structure of the Thesis 8

CHAPTER 2 THEORETICAL BACKGROUND: A LITERATURE REVIEW OF PREVIOUS WORKS ON PASSIVE DATA 10

2.1 Transit Planning 10

2.1.1 Concepts in Operations Planning 10

2.1.2 Transit Assignment 11

2.2 Processing and Applications of Passive Data in Public Transit and Travel Behaviour 12

2.2.1 Data Validation of Passive Data in Public Transit 12

2.2.2 Comparing Smart Card Data with Survey Data 15

2.2.3 Potential Use of Smart Card Validation Data in Transit Planning 16

2.2.4 Transit Service Performance Analysis and Monitoring 17

2.2.5 Deriving Origin-Destination Information of Transit Trips 18
2.2.6 Analyzing Travel Behaviour of Transit Users 20
2.2.7 Enriching Passive Data with Trip Details 22
2.3 Travel Behaviour Studies 26
2.3.1 Cross-sectional Analysis 26
2.3.2 Multi-day Analysis 26
2.4 Information Technologies 27
2.4.1 Data Storage, Retrieval and Processing 27
2.4.2 Geographic Information System 28
2.4.3 Spatial Statistics 28
2.4.4 Data Mining 29
2.4.5 Visualizations 30
2.5 Recapitulation on Theoretical Background 32
CHAPTER 3 DATA AND GUIDING RESEARCH PRINCIPLES 34
3.1 Data Source 34
3.1.1 Data from the Smart Card AFC system of the STO 36
3.1.2 Data Derived from Validation Data 44
3.1.3 Timetable Data from User Guide 44
3.1.4 Spatial Data from External Sources 44
3.1.5 Data from the OPUS Smart Card AFC System 45
3.2 Making Sense of the Data 46
3.2.1 Understanding the Organization of Transit Service 46
3.2.2 Understanding the Sources of Error 47
3.3 Guiding Research Principles 48

3.3.1 Generalized Approaches 48
3.3.2 Concepts Used in Data Processing 50
3.3.3 Multi-day Information Approach 50
CHAPTER 4 SMART CARD FARE VALIDATION DATA AS A TRAVEL SURVEY 53
4.1 Travel Survey Type and Data Quality 53
4.1.1 Stated Preference vs Revealed Preference 53
4.1.2 Survey Method 53
4.1.3 Household-based vs Non Household-based 54
4.1.4 Travel Mode Coverage 55
4.1.5 Timing and Duration 55
4.2 Drawbacks of Traditional Survey Methods 55
4.3 Uniqueness of Smart Card Data 56
4.4 Smart Card Data Quality Issues 58
4.4.1 On Respondents 58
4.4.2 On Response Rate 58
4.4.3 On Coding 58
4.4.4 On Sampling 59
4.4.5 On Survey Universe 60
4.5 A Multi-use Survey 61
CHAPTER 5 EXPLORATORY DATA ANALYSIS AND DATA VALIDATION STRATEGY 62
5.1 Exploratory Data Analysis 62
5.2 Data Validation Strategy 65
5.2.1 Rationale of the Data Validation Process 66
5.3 Error Detection by Logical Rules 70

5.3.1 Public Transit Logics 70
5.3.2 Spatial-temporal Logics 71
5.3.3 Rule-based Detection 73
5.3.4 Examples of Logical Rules and Errors 74
5.3.5 Results of Error Detection 78
5.4 Data Replenishment by Imputation 80
5.4.1 Concept 1: Repetition of Transit Service 80
5.4.2 Concept 2: Boarding History of Individual Cardholders 83
5.4.3 Analysis of Results from Data Validation 89
5.5 Assessing the Integrity of Databases: a Case study 93
5.5.1 Case Study: School Routes 93
CHAPTER 6 METHODS AND ANALYSES FOR OPERATIONS PLANNING 95
6.1 Spatial-temporal Description of Boardings 95
6.1.1 Levels of Aggregation 95
6.1.2 Spatial-temporal Distribution of Boardings 96
6.1.3 Stop-level Boarding 99
6.1.4 Route-level Boarding 100
6.1.5 Run-level Boarding 103
CHAPTER 7 METHODS AND ANALYSES FOR TRANSIT DEMAND MODELING 105
7.1 Reconstructing an Itinerary 105
7.1.1 Definition of an Itinerary 105
7.1.2 Determining the Most Probable Alighting Stop 106
7.2 Estimating Vehicle Spatial-temporal Path with Timetable Data 107
7.2.1 Running Time Estimation 107

7.2.2 Spatial-temporal paths of vehicles 108
7.3 Identifying Transfer Activities 109
7.3.1 Algorithm to Detect Transfer Coincidence 109
7.4 Cardholder Itinerary with Trip Details 111
7.4.1 Reconstruction of Cardholder Itinerary 111
7.4.2 Activity Duration 111
7.4.3 Distance traveled and Trip Duration 111
7.4.4 Example of Itineraries from a Cardholder 111
7.5 Applications for Transit Planning 113
7.5.1 Transfer Analysis 113
7.5.2 Load profiles 117
7.5.3 Detailed Analysis of a Route 121
7.6 Transit Network Analysis 124
7.6.1 Transit Assignment with a Totally Disaggregated Approach 124
7.6.2 Examples of Transit Network Assignments 126
7.7 Deriving Activity-space Profile 131
7.7.1 Modeling Activity Space 131
7.7.2 Activity Space with Trip Generators 134
7.7.3 Analysis by Trip Generator 136
7.8 Deriving Trip Ends for Individual Itinerary 138
7.8.1 Travel Pattern Consolidation by Aggregations 139
7.8.2 Example: Linking Student Boardings to Schools 139
CHAPTER 8 METHODS AND ANALYSES FOR THE STUDY OF TRAVEL BEHAVIOUR 141
8.1 Framework for Analyzing Multi-day Data 142

8.2 Characterizing Trips with Multi-day Data 144
8.2.1 Identifying Anchor Points for Each Cardholder 144
8.2.2 Discovering a Card's Anchor Points 144
8.2.3 Linking Trip Ends to Anchors and Nodes 148
8.3 Multi-day Travel Behaviour Analysis 148
8.3.1 Analyzing Travel Behaviour of Cards Tied to a Specific Anchor 148
8.3.2 Analyzing Travel Behaviour of a Specific Card 153
8.3.3 Data Mining as a Tool for Travel Behaviour Analysis 156
8.4 Analyzing Travel Behaviour of Cards Tied to a Specific Transit Service 159
8.5 Observations from Travel Behaviour Analysis 160
CHAPTER 9 LIMITATIONS AND GENERALIZATION OF FINDINGS 162
9.1 Limitations 162
9.2 A Multi-modal and Multi-operator AFC System 163
9.3 Potential Applications of Sales Data 164
9.3.1 Sales Volume and Spatial Distribution 165
9.3.2 Purchase Behaviour of Transit Users 167
9.3.3 Usage Pattern of Sales Equipments 168
9.4 Applications of Validation Data 169
9.4.1 Macroscopic Analysis of Validation Data 169
9.4.2 Microscopic Analysis of Validation Data 171
9.5 Applications of Verification Data 173
9.5.1 Estimating Fare Evasion Rate 173
CHAPTER 10 CONTRIBUTIONS AND PERSPECTIVES 177
10.1 Summary of Key Findings 177

10.1.1 Contributions to Methodology 177

10.1.2 Experimental Results from the Specific Datasets 179

10.1.3 Implications of the Research 181

10.2 Future Research Directions 183

10.2.1 Large-scale Applications of Data from a Multi-operator and Multi-modal Smart Card Automatic Fare Collection System 183

10.2.2 Synthesis of Passive Data from Different Systems 184

10.2.3 Transit route-choice behaviour 185

10.2.4 Transit Assignment with Multi-day Smart Card Itineraries 185

CONCLUSION 186

REFERENCES 188

LIST OF TABLES

Table 1.1 Data needs of various actors in public transit (p.t.) (taken from Sammer, 2009). 3

Table 3.1 Excerpt of smart card fare validation transaction records. 39

Table 3.2 Excerpt of fare type data. 40

Table 3.3 Error message data. 41

Table 3.4 Excerpt of route geometry data. 42

Table 3.5 Excerpt of bus stop inventory data. 43

Table 3.6 Excerpt of point of interest data of Gatineau. 45

Table 5.1 Statistics on various objects. (Potentially sensitive statistics are greyed out.) 63

Table 5.2 Monthly statistics on various fare types. (Potentially sensitive statistics are greyed out.) 64

Table 5.3 Monthly statistics on aggregated fare types. (Potentially sensitive statistics are greyed out.) 65

Table 5.4 Transactions illustrating a case where a driver failed to initialize a service run after a deadheading run. 75

Table 5.5 Transactions with departure time “00:00”. 75

Table 5.6 Transactions illustrating a case where the driver fails to initialize a service run after another service run. 76

Table 5.7 Transactions illustrating a case where the dwell time at boarding stop 5006. 77

Table 5.8 Transactions illustrating a case where all boardings of a vehicle from the whole day are tied to the same stop. 77

Table 5.9 The order of stops within a run is not respected in the transactions. 78

Table 5.10 Distribution of transactions by flag type for the month of February 2005. The same transaction can be flagged more than once by different rules (Chu & Chapleau, 2007). 79

Table 5.11 A typical vehicle block for a weekday derived from transactions taken from the enriched operations dictionary. 82

Table 7.1 Results of running time estimation. Bold values denote times from actual boarding validations (Chu & Chapleau, 2008). 108

Table 7.2 Derived trip and activity details (Chu & Chapleau, 2008). 113

Table 7.3 Performance indicators for inbound route 44. 122

Table 8.1 Results of the school assignment (Chu & Chapleau, 2010). 147

Table 8.2 Origin-destination trip table with multi-day attributes (Chu & Chapleau, 2010). 155

Table 9.1 Evasion rate by day of week (Chu & Bergeron, 2010). 175

Table 9.2 Evasion rate by fare zone and station on the Montréal / Deux-Montagnes train line (Chu & Bergeron, 2010). 175

LIST OF FIGURES

Figure 1.1 Opportunity and challenge in data needs fed by a constant quest for data and paradigm changes. 2

Figure 1.2 Representations of a load profile, a load profile incorporating time and a load surface by synthesizing multiple load profiles and time (taken from Vuchic, 2005). 4

Figure 1.3 Structure of the thesis. 9

Figure 2.1 Object-oriented modeling of the smart card AFC system of the STO (taken from Trépanier et al., 2004). 17

Figure 2.2 Enriching GPS data with land use information and GIS (taken from Wolf et al., 2004). 25

Figure 2.3 An example of "false-GIS" with GPS data (taken from Dionne, 2006). 31

Figure 2.4 An animation based on OD survey data of the Greater Montréal Area (taken from Morency, 2004). 32

Figure 3.1 Simplified diagram of the data flow in a smart card AFC system (Chu & Bergeron, 2010). 35

Figure 3.2 Interactions among various data from the smart card AFC system. 36

Figure 3.3 Bus stop locations of the STO network. 43

Figure 3.4 Atlas of Gatineau in the City of Gatineau website. 45

Figure 3.5 Illustration of a vehicle block composed of 3 consecutive runs (route-direction-departure time) and the related operations information (Chapleau & Chu, 2007). 47

Figure 3.6 An individual trip defined with the totally disaggregated approach (based on Chapleau, 1992). 49

Figure 3.7 Schematic definitions and relations of various objects (Chu et al., 2009). 50

Figure 3.8 Schematic representation of the multi-day information approach. 51

Figure 4.1 The concept of Driver-Assister Bus Interview (Chu et al., 2009). 57

Figure 5.1 Number of transactions by fare type. 63

Figure 5.2 A 3-dimensional time-space diagram showing the temporal movement of the object cardholder before the data validation process (Chu & Chapleau, 2007b). 69

Figure 5.3 Boarding transactions associated with vehicle block 140 for 20 weekdays (Chu & Chapleau, 2007b). 71

Figure 5.4 The use of a time-space diagram to detect error in run assignment (Chu et al., 2009).73

Figure 5.5 Results from the validation procedure on smart card transactions in February 2005 (Chu & Chapleau, 2007). 79

Figure 5.6 The transaction history of a card in the month of February 2005 (Chu & Chapleau, 2007).....85

Figure 5.7 Summary of the data validation strategy (Chu & Chapleau, 2007). 88

Figure 5.8 Time-space diagrams of bus objects used to evaluate the imputation procedure (Chu & Chapleau, 2007). 90

Figure 5.9 A 3-dimensional time-space diagram showing the re-established temporal movement of the cardholder object after the data validation process (Chu & Chapleau, 2007). 92

Figure 6.1 Standard deviation ellipses of transactions during the course of a typical day (Chapleau & Chu, 2007). 97

Figure 6.2 Grid analysis of boardings by time of day (Chapleau & Chu, 2007). 98

Figure 6.3 Temporal boarding intensity by 15-minute interval (Chapleau & Chu, 2007).....99

Figure 6.4 Stop-level boardings and derived alightings for a typical AM peak (Chu et al., 2009). 100

Figure 6.5 Stop locations and hourly route-level boardings of route 37 (Chapleau & Chu, 2007). 102

Figure 6.6 Space-time diagrams of bus route 37 illustrating vehicle location and boarding intensity during the AM and PM peaks (based on Chapleau & Chu, 2007). 104

Figure 7.1 The schematic definition of an itinerary of a linked trip (Chu & Chapleau, 2008). ...105

Figure 7.2 An example of a transfer coincidence (Chu & Chapleau, 2008). 110

Figure 7.3 A three-dimensional representation of the itineraries of a cardholder (Chu & Chapleau, 2008).....112

Figure 7.4 The temporal distribution and the cumulative percentage of the derived transfer times (Chu & Chapleau, 2008). 115

Figure 7.5 The ratio between derived transfer wait time and planned headway (Chu & Chapleau, 2008).....116

Figure 7.6 The spatial distribution of the first and transfer boardings in a typical day (Chu & Chapleau, 2008). 117

Figure 7.7 Load profile of a run showing ridership at both aggregate and disaggregate levels (Chu & Chapleau, 2008). 118

Figure 7.8 Three-dimensional stop-level spatial-temporal load profiles of inbound route 44 (Chu & Chapleau, 2008). 120

Figure 7.9 Three-dimensional stop-level spatial-temporal load profiles of route 64 direction Gabrielle-Roy. 121

Figure 7.10 An intersection with several transit routes providing multiple transfer possibilities (Chapleau & Chu, 2007). 125

Figure 7.11 Example of MADITUC nodes. 126

Figure 7.12 Three-dimensional representation of the result of a transit assignment (Chapleau & Chu, 2007). 128

Figure 7.13 Assigning smart card itineraries onto a MADITUC transit network (Chu et al., 2008). 129

Figure 7.14 Load profile from transit assignment according to fare type. 130

Figure 7.15 Load profiles and passenger movements from transit assignment. 131

Figure 7.16 Assumptions used to derive land occupation profile. 132

Figure 7.17 Derived land occupation profile of smart cardholders on a typical weekday (Chu et al., 2009). 134

Figure 7.18 Trip generators are revealed by activities in the transit network (Chu et al., 2009). 135

Figure 7.19 Various types of trip generators located along the inbound route 44. 136

Figure 7.20 Functional analysis of transit stop with ridership and trip generator data (Chu et al., 2009).....138

Figure 8.1 Schematic representation of the multi-day travel behaviour analysis framework.....143

Figure 8.2 Two anchor points for card A and three for card B (Chu & Chapleau, 2010). 145

Figure 8.3 Kernel density showing the derived residence of cardholders attending École de l'Île High School (Chu & Chapleau, 2010). 149

Figure 8.4 Various spatial measures describing the derived residence of cardholders at node level (Chu & Chapleau, 2010). 150

Figure 8.5 Origins of school-bounded trips at 15-minute interval (Chu & Chapleau, 2010). 151

Figure 8.6 Distribution of departure time and on-board distance for trips heading to École de l'Île High School (Chu & Chapleau, 2010). 152

Figure 8.7 Distribution of departure time from École de l'Île High School by the minute (Chu & Chapleau, 2010). 153

Figure 8.8 Trips details of card B (Chu & Chapleau, 2010). 154

Figure 8.9 Derived activity schedule of card B (Chu & Chapleau, 2010). 156

Figure 8.10 A subset of rules describing card B's travel pattern (Chu & Chapleau, 2010).....157

Figure 8.11 A decision tree generated by the C4.5 algorithm along with the confusion matrix (Chu & Chapleau, 2010). 159

Figure 8.12 Cardholders' loyalty towards a route and a run (Chapleau & Chu, 2007). 160

Figure 9.1 Spatial distribution of sales of monthly integrated fare products (Chu & Bergeron, 2010).....166

Figure 9.2 Monthly pass sales by day of month for purchase behaviour study (Chu & Bergeron, 2010).....168

Figure 9.3 Fare product sales pattern of automatic vending machines for level of service study (Chu & Bergeron, 2010).....169

Figure 9.4 Monthly validation (November, 2009) in the métro network by fare type (Chu & Bergeron, 2010).....170

Figure 9.5 Temporal distribution of entry validations in three métro stations (Chu & Bergeron, 2010).....172

Figure 9.6 Verification and fare evasion rate in November 2009 (Chu & Bergeron, 2010).....174

LIST OF ACRONYMES AND ABREVIATIONS

The following lists, in alphabetical order, the acronyms and abbreviations used in the thesis:

AFC	Automatic Fare Collection
AMT	Agence métropolitaine de transport
APC	Automated Passenger Counter
AVL	Automatic Vehicle Location
AWD	Average Weekday
CAP	Carte à puce (Smart card)
CATI	Computer-Assisted Telephone Interview
DABI	Driver-Assisted Bus Interview
EFC	Electronic Fare Collection
ERF	Electronic Registering Farebox
GIS	Geographic Information System
GPS	Global Positioning System
KDD	Knowledge Discovery in Databases
LAD	Location-Aware Device
MADITUC	Modèle d'Analyse Désagrégée des Itinéraires de Transport Urbain Collectif
OD	Origin-destination
POI	Point of interest
SQL	Structured Query Language
STO	Société de transport de l'Outaouais
TCRP	Transit Cooperative Research Program
WRT	With respect to

CHAPTER 1 DATA IN PUBLIC TRANSIT: AN INTRODUCTION

1.1 Motivation of Research

Public transit system is an artificial and complex creature. Its existence and usefulness are attributed to the collaboration between supply and demand – with the supply being the service offered by transit operators and the demand representing the travel needs of the population. The interaction is spatial because it requires common origin and destination pairs between the population and the service. At the same time, it is temporal because those trips need to be carried out within a specific time window. Transit demand is difficult to measure and its interaction with service is in constant evolution. Therefore, there is a continuous quest for information and methodology that can help reveal and facilitate the understanding of this dynamic relationship, so that operators can plan their services according to the demand and allocate their resources in an efficient manner.

Recent paradigm shifts have contributed to the reshaping of this process. On the one hand, public transit service has become more performance-driven and customer-oriented. To ensure its competitiveness against automobile, improve customer satisfaction and reduce operating costs, transit operators need to constantly monitor the quality and performance of its services. These include aspects such as network coverage, connectivity, comfort, frequency, reliability and fare structure, and would require data that are not covered by conventional survey methods.

On the other hand, transit agencies traditionally rely heavily on manual data collection methods to collect ridership data for planning and finance purposes. In many cases, they remain the most complete and reliable source of information. Advances in passive data collection methods and computing power along with the low cost of data storage and communication devices have prompted another paradigm shift: transit agencies that have adopted the passive data collection technologies move progressively from a data-poor to a data-rich environment. Transit agencies are used to be limited by the high cost and labour-intensive nature of manual data collection methods, resulting in a data-poor environment. Naturally, analysis and planning tools are adapted to the conditions, using methods that are often manual, aggregated or relying on algebraic model to reduce data requirement. As a consequent, they are not suited to fully leverage new sources of data.

At the confluence of all these evolutions lie opportunity and challenge: to embrace the data-rich environment with the purpose of reconciling with the ever more demanding data needs in public transit (Figure 1.1).

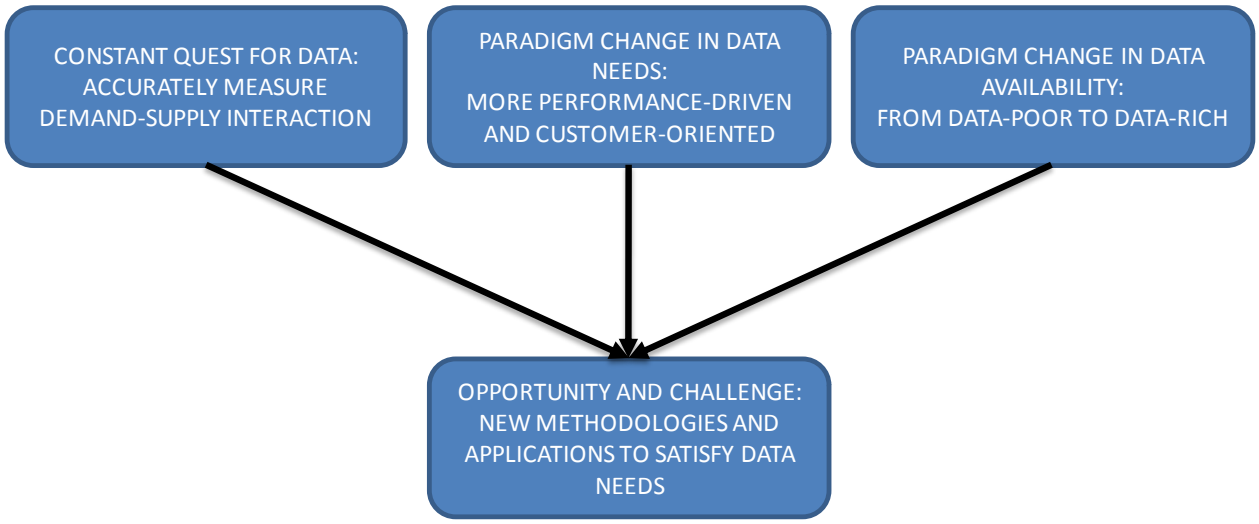


Figure 1.1 Opportunity and challenge in data needs fed by a constant quest for data and paradigm changes.

Over the years, passive data collection technologies have gradually made their way into transit agencies across the world and into the Québec society – one of them being the smart card automatic fare collection (AFC) system. The data generated by a smart card AFC system record, in a very precise manner, many interactions between the transit network and travelers, and can potentially be used to accurately measure the demand-supply interaction as well as to generate performance indicators. With new methods in data processing, data enrichment and data analysis, the knowledge drawn from these data can help operators to better quantify transit demand, enhance operations planning, improve system management, understand travel behaviour and ultimately, provide better transit services with more efficient use of resources. Based on data from an operating smart card AFC system, this research establishes the theoretical framework for using smart card fare validation data as a transit survey. It proposes methodology for data processing and enrichment as well as analyses of operation planning, transit demand modeling and the study of travel behaviour. It demonstrates their applicability using experimental results and specific illustrations.

1.2 Data Needs and Data Collection in Public Transit

Sammer (2009) provides an up-to-date overview of data needs of various actors in public transit and the data collection methods (Table 1.1). Ideally, data such as ridership, trip details, network supply and usage, service quality and performance, and customer satisfaction are collected for public transit planning and operations purposes. However, the precision, the accuracy and the timeliness of the data vary greatly according to the data collection methods.

Table 1.1 Data needs of various actors in public transit (p.t.) (taken from Sammer, 2009).

Target activity for p.t. planning and operation	Specific data needs	Users of data
Transport modelling for operational and infrastructure decision	Detailed information about trip stages and intermodality	Transport authorities, modellers, researchers
Counting p.t. ridership correctly	Accurate figures of p.t. ridership and compatibility of different data resources	Transport authorities, p.t. operators, transport decision-makers and stakeholders
Providing adequate p.t. service for mobility-impaired people	Definition and identification of these specific groups and their needs for p.t.	Transport authorities, transport decision makers, transport planners, researchers
Revenue distribution to operators	Accurate allocation of passengers mileage of p.t. operators over the year, error estimation of sample	Transport associations, transport authorities, p.t. operators
Measurement of service quality of p.t.	Usability for improvement of service quality and marketing	P.t. operators, transport associations
Benchmarking of p.t. operators	Comparability of data collected for benchmarking indicators	P.t. operators
Measurements of customer satisfaction for p.t.	Identification of potential of new customers and loss of current customers of p.t. for marketing	P.t. operators, transport associations, transport authorities
Decision-making process for p.t. measures	Identification of stakeholders' attitudes to identify barriers and drivers of the decision-making process	Transport authorities, transport planners
Marketing and information campaigns for p.t.	Perceptive customer satisfaction data for different target groups, relevant for marketing	P.t. operators, transport associations
Automation of data collection for user counts and travel behaviour	Highly accurate and relliable data covering the whole year: cost-effective data management	Transport associations, transport authorities, p.t. operators, researchers

The most basic information for operation planning of public transit remains the load profile. It combines boarding and alighting information to calculate the number of passengers, or load, inside a vehicle at a series of reference points or check points. Depending on the level of resolution of the data, the most common units are route segment, stop, linear route distance and travel time. The maximum load point identifies the section of a transit route with the heaviest load. Vuchic (2005) provides several representations of load profile (Figure 1.2). With an additional dimension, successive load profiles of a route can be combined to generate a load surface. Operation planning aims to organize services following these load profiles. Since the demand and the composition of clientele vary with time (time of day, day of week and season), weather, economic and other conditions, the more observations and knowledge on these patterns are gathered, the more accurately can the operators adjust their services to the demand. The same holds true for service performance. The types of indicators that can be calculated depend on the level of resolution of the data gathered. For example, the disposal of running times of a route segment at various time of day allows operators to fine tune the schedule.

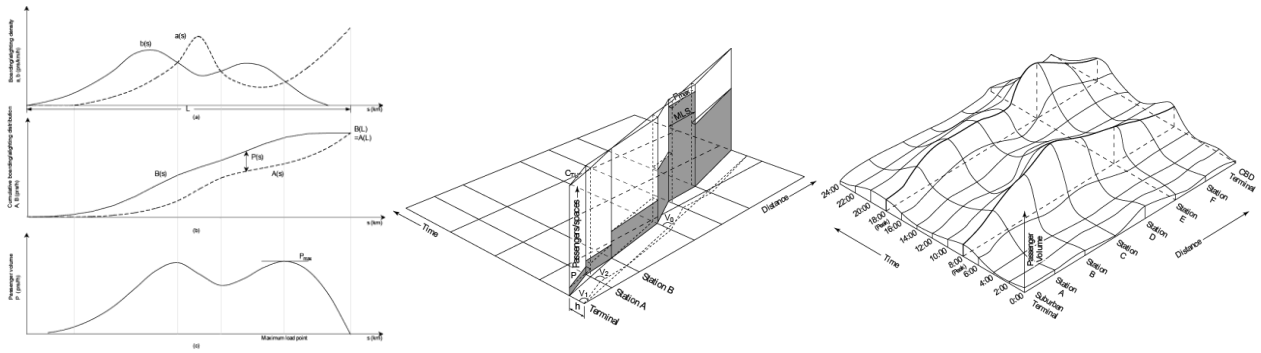


Figure 1.2 Representations of a load profile, a load profile incorporating time and a load surface by synthesizing multiple load profiles and time (taken from Vuchic, 2005).

Traditionally, much of these data are acquired by active data collection methods or are simply lacking. Active data collection methods necessitate the active participation of a respondent to provide information and/or the intervention of a human agent to record information. Since they are labour-intensive, time-consuming and costly, data collections are performed infrequently. For example, a pen-and-paper on-board travel survey involves a questionnaire handed to a transit rider by a distributor. The rider fills up the questionnaire and returns it to a collector. The survey manager needs to assure that the target group are correctly sampled and that riders have enough time and room to complete the survey. Once completed, the survey has to be manually

transcribed into digital format for analysis. Illegible writing, inaccurate or conflicting information by the respondents and errors in transcription represents major sources of uncertainty in the data. At the same time, ride check and stop check involve counting people manually in environments that can be crowded and disorganized, causing inaccuracy in data. A computer-assisted origin-destination telephone survey requires the interviewee to explicitly declare socio-economic and trip details to the interviewer. Although the interview is assisted by real-time on-screen data validation, significant amount of effort is still needed to validate and correct the data in the processing stage.

Passive data collection methods distinguish themselves by their ability to automatically gather or generate data through a mechanical or an electronic device. Although qualified as passive, the methods may still require some manual operations, usually for inputting reference data. Passive data collection methods also suffer from errors and missing data. However, they are generally more systematic and predictable than human manipulation errors. Common passive data collection systems in public transit include automated passenger counter (APC), automatic fare collection (AFC), automatic vehicle location (AVL) and logs of computerized systems.

An automated passenger counter is "an automated system that counts the number of passengers boarding and alighting a transit vehicle" (Kittleson & Associates, Inc. et al., 2003). This technology dates back to the 1970s but is still widely used by transit agencies to gather ridership count and route-based boarding-alighting information. The system uses treadle mats, infrared, laser or light emitting diode (LED) at the entrance and exit doors to detect passenger movement, and sometimes, its direction. The physical characteristics of the vehicle influence the choice of technology. For example, a low-floor bus cannot use treadle mats as it does not have any step. The level of accuracy differs but usually adjustments are needed to balance the on-off count since an APC can capture objects that are not passengers and movements that are not boarding or alighting. The on-off counts acquired by an APC are aggregate data. Thus, it is not possible to link a boarding passenger to an alighting passenger. The data need to be combined with operational, spatial or temporal reference data in order to be meaningful. That unit can be coarsely divided, such as by vehicle-tour or vehicle-trip, or can be refined, such as by route-segment or by stop. The level of resolution is a key determinant on the type of analyses that can be performed with the data. Similarly, turnstiles in station, with a fixed spatial reference, provide the number of passengers passing through a gate.

An AFC system, also known as an electronic fare collection system or an electronic registering farebox (ERF), is a device that counts the money, processes fare media and electronically records the fare information (Kittleson & Associates, Inc. et al., 2003). It records the time and the fare type of the transaction. Similar to APC, it requires a spatial reference point in order to determine where transaction occurs. Depending on the fare media, which includes coin, token, magnetic card and smart card, AFC can provide either aggregate or disaggregate data. Transaction made with a single-use fare medium in an entry-only system provides count data only. Transactions made with a multiple-use fare medium, such as a magnetic fare card with a unique identification number, can relate to each other. The life extent however is generally less than a month.

The smart card fare management system is an AFC using smart card as the fare medium. The smart card technology originated in the 1970s and made its entrance to public transit as a ticket system in the 1990s. Attoh-Okine et al. (1995) and the TCRP Report 94 (Multisystems Inc. et al., 2003) provide some background on the technology. Smart card technology in the transit industry is described in Bagchi and White (2004). The term "smart card" is a generic term used to describe a wide range of technologies and applications. The common feature of all smart cards is the embedded microchip. It can be a computer chip, which can store and process information, or a memory chip, which can only access data already stored in the chip. The point of interaction between a card and a reader is called interface. The interface can either be with contact or contactless. Contactless cards, which do not require insertion or physical contact to communicate with the reader, can be further divided into two standards: "proximity" and "vicinity". Proximity card can communicate within a range of 10 centimetres and conforms to the international card standard ISO 14443 while vicinity card has a range of 70 centimetres and conforms to ISO 15693. Public transit agencies usually choose proximity card primarily to avoid unintended communication near the reader. For transit applications, smart card can be "loaded with monetary value which is decremented for each ride, in flat amount or, with exit checks, for distance-based fares" (Kittleson & Associates, Inc. et al., 2003), loaded with tokens or used as a pass. The main differences between smart card and magnetic fare card are the amount of information that can be stored and processed, and the life extent of the media.

Smart card AFC provides disaggregated data for each individual transaction. Depending on system specifications, data usually include transaction details such as card ID, equipment ID, transaction time, fare product and some operations details. However, it lacks spatial reference,

unless the location of the equipment is known. Since smart cards contain a unique identification number and some cards are personalized, trips made by the same card can be linked over a long period of time.

An AVL system "determines the location of vehicles carrying special electronic equipment that communicates a signal back to a central control facility" (Kittleson & Associates, Inc. et al., 2003). Strictly speaking, it only includes systems that communicate in real-time between vehicles and the central control facility. AVL by the global positioning system (GPS) receives signals from a network of satellites and compute the position by time differential. The drawback is that signal can be weakened in densely built areas, a phenomenon known as urban canyon, and absent in covered areas. GPS provides high-resolution spatial-temporal data of an object. Position can be updated at each second. However, due to the high bandwidth required for data transmission, real-time applications usually have a lower update frequency. Data archival practice varies: some systems are event-based, meaning that data is only recorded when a specific event occurs; some record at pre-defined time interval and some combine both. Raw coordinates need to be matched against a reference, such as a street network or transit stops, in order to be meaningful. This procedure is called map matching. AVL by radio frequency identification (RFID) uses a network of beacons (signposts) installed along a bus route and radio signal to locate a vehicle. The system does not have the drawback of GPS but requires an important amount of capital investment. Other possible tracking methods include radio triangulation with the cellular phone network. AVL can provide a spatial reference to APC and AFC data.

Other computerized systems also passively gather data on transit service. Trip planner is an advanced web tool that proposes transit itineraries based on user input of an origin and destination pair (Schaller, 2002). Examples include Tout Azimut of the MADITUC Group (Chapleau, Allard & Trépanier, 1996) and Google Transit. Trépanier, Chapleau & Allard (2005) explore the usefulness of trip planner log in transit planning. Data from para-transit reservation system and service log can be used to analyse demand pattern and user characteristics (Chapleau & Allard, 2007; Desharnais & Chapleau, 2010). The same can be done with data from the reservation system of car-sharing services (Morency, Trépanier & Martin, 2008).

1.3 Structure of the Thesis

This thesis is based on research performed on data from two operating smart card AFC systems. It elaborates on the findings of the research in three principal aspects: conceptual and theoretical, methodological, and analytical. Many of the research materials have previously been presented in academic conferences or published in peer-reviewed journals. Figure 1.3 shows the outline of the thesis.

Data need to be analyzed in valid and pertinent ways in order to clarify complex issues. To achieve this goal, one needs sound methodological and relevant analytical frameworks. Since smart card AFC systems are emerging technologies and the data have not been studied extensively, the first assignment is to understand the data and to study their ontology, structure, properties and validity. This establishes the theoretical framework of the research. Chapter 2 examines previous works on passively-collected data within the public transit context as well as the analytical tools that are used. Chapter 3 presents the data sets used in this research, the way they are generated and the inter-relationship among objects in the transit system. It also outlines the guiding research principles which underlie various topics of the research. Chapter 4 compares smart card validation data with traditional travel survey. It describes the properties of the data and proposes potential applications. An exploratory data analysis in Chapter 5 looks at the validity of the data.

With an understanding of the data, the subsequent chapters propose methodologies to process and analyze the data. The second part of Chapter 5 puts forward a data validation procedure with the aim to improve the quality of the data. Chapters 6 and 7 propose and apply enrichment techniques to the validated data in order to perform analyses on transit performance and demand. Chapter 8 looks into the multi-day aspect of the data and derives a new approach to characterize transit trip and to study the travel behaviour of transit users.

To conclude, chapter 9 presents the limitations of the research by illustrating the complexity of data from a multi-operator and multi-modal smart card AFC system and the potential uses of other types of data. It also discusses the generalization of findings from this research. Chapter 10 summarizes the key findings and contributions of this research, and provides perspectives on further research.

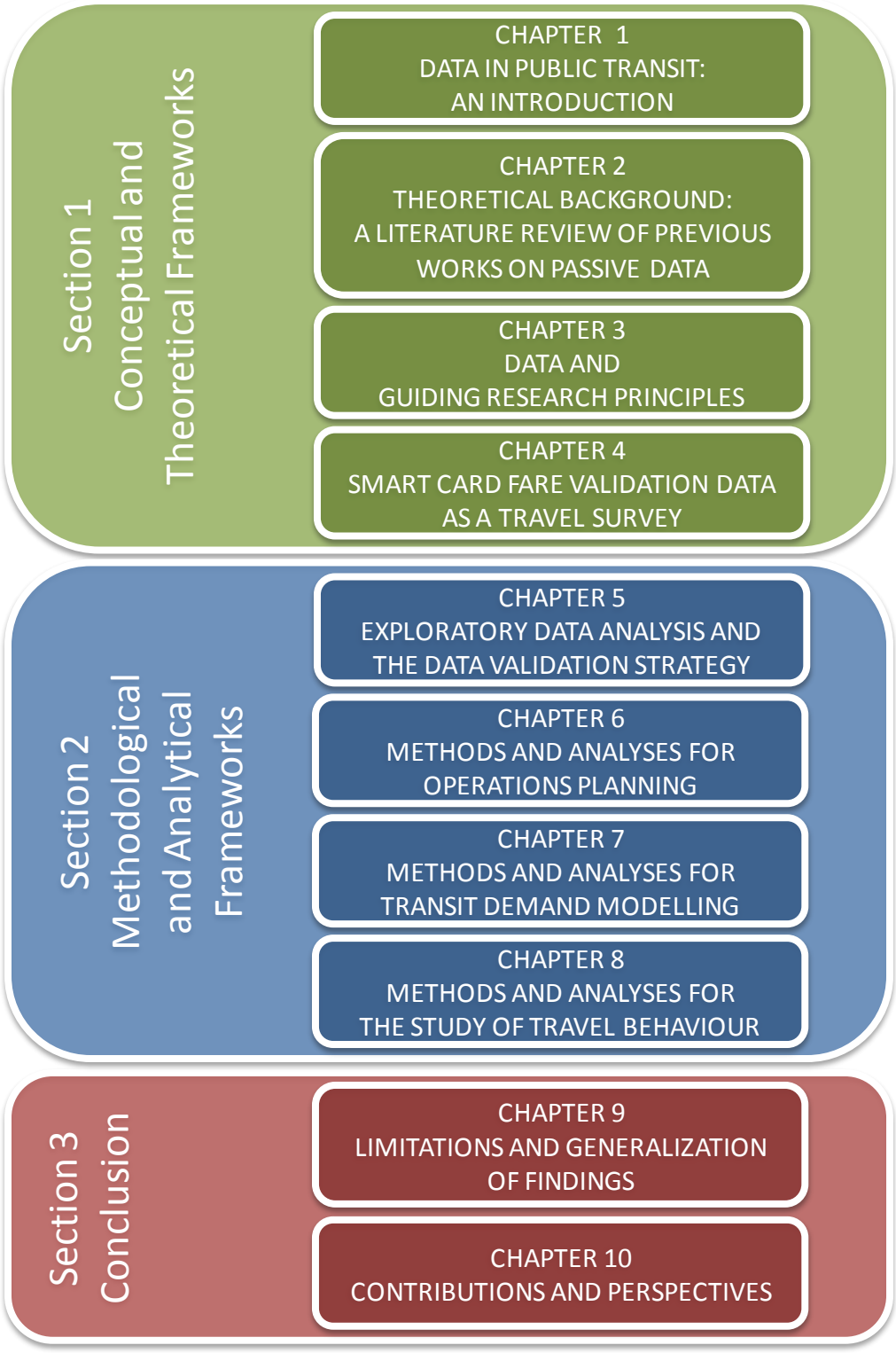


Figure 1.3 Structure of the thesis.

CHAPTER 2 THEORETICAL BACKGROUND: A LITERATURE REVIEW OF PREVIOUS WORKS ON PASSIVE DATA

This chapter provides a theoretical background on the use of passive data in public transit, with an emphasis on smart card data. It starts with a brief presentation of the public transit planning mechanism, followed by a literature review of recent research on the subject and an introduction of common tools used to analyze passive data.

2.1 Transit Planning

Transit planning can be classified according to planning horizon. The medium- to long-range planning relates to network geometry while the short-term planning focuses to day-to-day service operations. Smart card validation data are intricately linked to both of them: the data can be analyzed in details for fine-tuning of operations and when aggregated, they reveal the overall trend and variation. For this reason, an understanding of the concepts used in operations planning is essential to put the data into perspective.

2.1.1 Concepts in Operations Planning

A schedule or timetable is usually established for a fixed-route network, which is defined in transit manager's handbook as "transit services where vehicles run on regular, pre-designated, pre-scheduled routes, with no deviation. Typically, fixed-route service is characterized by features such as printed schedules or timetables, designated bus stops where passengers board and alight and the use of larger transit vehicles" (Iowa Department of Transportation, 2005). Although it is well documented that transit demand varies according to the day of the week, the common practice within the public transit industry is to conduct service planning around the concept of an average week day, for which a schedule is developed. Other sets of schedules are often constructed for Saturday, Sunday and holidays. This facilitates vehicle and crew assignments and makes the timetable easier for travelers to follow. The timetable can be frequency-based or run-based. It is valid for a period of time and is modified in order to follow the seasonal variation of demand. For example, in Canada, the Toronto Transit Commission (TTC) makes about ten changes in timetable per year, with each of them lasts 4 to 6 weeks; the Société de Transport de Montreal (STM) has 4 timetables a year along with minor modifications

for major holidays and the Société de transport de l'Outaouais (STO) alters its timetable twice a year. The main goal of operations planning is to dress the demand within a service envelop and to allocate resources to perform the service. Ceder (2001; 2007) provides thorough explanations on operations planning.

To construct timetables, the demand of a route at the segment or stop level, which can either be revealed by ridership data or obtained from transit assignment model, is brought back to the point of injection using estimated travel time. Travel time and its variation within the day are estimated with average commercial speed or more accurately, with AVL data. The number of runs and vehicles required are calculated using either the even-load or the even-headway approach with a specified load factor. To optimize resource, one can minimize the number of vehicles required to carry out the schedule by tolerating minor shifts in the timetable or minimizing unproductive journey of the fleet, known as deadheading, with interlines (Ceder, 2001).

Vehicle blocks are created by combining runs. During crew scheduling, drivers are assigned to one or more blocks, called a work piece. All of these procedures are constrained by collective agreement governing the working conditions of the drivers. A roster is "a periodic duty assignment which guarantees that all the trips are covered for a certain number of consecutive days" (Ceder, 2001) and each roster is usually maintained for several weeks.

2.1.2 Transit Assignment

The objective of a transit assignment model is to "predict passenger flows and levels of service on a given transit network that consists of a set of fixed lines" (De Cea & Fernández, 2000). In its early stage of development, many models are simplification or modification of road network assignment. Nielsen (2000) presents five characteristics of transit assignment that differ from traffic assignment and are summarized below:

- Public transit network often consist of "common lines", segments served by more than one routes with the same or different frequencies. The determination of relative weight of wait time and in-vehicle time is often a concern.
- Depending on the passenger, the objective may be to minimize travel time or the number of transfers.

- The deterministic travel time must be used within a utility function. The weights of different terms in the function can vary across public transit sub-modes.
- The choice of links is not independent as it depends on the preceding mode.
- Passengers are not aware of all feasible routes within a complex public transit network.

Transit assignment models differentiate themselves by "the hypotheses made, either explicitly or implicitly, on the users' behaviour when faced with route-choice decisions" (De Cea & Fernández, 2000). Assumptions are made to handle common lines, capacity and schedule.

2.2 Processing and Applications of Passive Data in Public Transit and Travel Behaviour

The use of passively collected data in public transit is not new. Many public transit operators have been using passive data in one form or another, though recent advances in information technology create new opportunities and challenges. The amount and level of details provided by passive data offer potential for new applications but at the same time demand new data processing and integration techniques in order to generate useful knowledge. This section reviews recent advances in the processing and applications of passive data within the public transit and travel behaviour context.

2.2.1 Data Validation of Passive Data in Public Transit

The accuracy of APC, AFC and AVL data is a primary concern in the transit industry as it affects the operations planning of transit services and transit performance measures. The Transit Cooperative Research Program (TCRP), with a principal audience of practitioners, has publications discussing issues on data validation:

- Synthesis of Transit Practice 29: Passenger Counting Technologies and Procedures (Boyle, 1998);
- Report 113: Using Archived AVL-APC Data to Improve Transit Performance and Management (Furth, Hemily, Muller & Strathman, 2006);
- Report 126: Leveraging ITS Data for Transit Market Research: a Practitioner's Guidebook (Strathman et al., 2008).

2.2.1.1 The Need of an Automated Data Validation Procedure

A number of authors stress the importance of a validation process on passive data. Utsunomiya Attanucci & Wilson (2006) state that:

> *Clearly, data from automatic data collection systems such as AFC, automatic vehicle location, and automated passenger count systems will contain some errors, and it is essential to identify and exclude or correct such errors if these data are to be used to support planning and management decisions. This paper has only begun the process of error detection and correction, and much more work in this area is required.*

Trépanie, Barj, Dufour & Poipré (2004) write that "as far as the present case study is concerned, previous research had identified errors to be corrected in the smart card automatic fare collection before any data were treated". "Some errors are particularly worrying because of the impacts they could have on the analysis. For example, assigning wrong route numbers to boarding information could completely change the statistics on the route's load profiles" (Trépanier, Chapleau & Tranchant, 2007).

One critical element in the data validation process is the difficulty of measuring the "ground truth" (Furth et al., 2006), which would be very useful to access the level of errors in the database and the effectiveness of validation and enrichment processes. However, it would be difficult to obtain ground truth using manual methods without itself incurring certain amount of errors. The amount of data also makes manual validation impractical. "An immense amount of effort is needed to validate ridership data at the trip level, making automated validation components extremely useful" (Boyle, 1998). Chapleau & Allard (2010) see redundancy in data collected from different passive devices as an opportunity to cross-validate their data and to derive reliable information.

2.2.1.2 Common Types of Errors

The identification of underlying causes of the error is important as it allows researchers to derive proper imputation techniques. Furth et al. (2006) finds that "accuracy of automated passenger counts may be reduced by many types of errors, including counting error, location error, attribution error (i.e., attributing counts to the wrong trip), modeling error (e.g., assumptions

about loops), and sampling error". Trépanier et al. (2004) examine location-stamped smart card AFC data and describe four main causes of error in the system:

- Computer and database-related errors;
- Errors due to the mishandling of the AFC system by drivers and operators;
- Card-reading and card-validating errors due to the on-board equipment, and location errors by AVL;
- Errors due to the desynchronizing of planned service data and operating service data (for example, emergency detour on a route).

One specific difficulty is identified for ERF of bus transit. According to Boyle (1998):

> *Bus operator compliance and attitudes are key issues in ensuring the usefulness of farebox data. Operators must enter specific codes at the beginning of their shifts and at the start of each new trip to tie fares to specific blocks and trips. Operators also need to record non-cash boardings using specific keys on the keypad. These additional operator duties frequently must be agreed to in negotiation with their union representatives. Lack of compliance can render much of the data useless. A contributing factor is that problems with operator compliance are non-random, i.e., data on specific trips driven by specific operators are consistently missing or inaccurate. One transit property noted that the first operator to take a bus out in the morning generally enters correct data, but that the level of compliance declines with subsequent reliefs. At another agency, Route 0 often has the highest ridership on daily printouts of ERF data.*

Following the same line, in a paper exploring the potential and the use of smart card data, Utsunomiya et al. (2006) identifies two types of error: missing transactions and incorrect bus routes. A missing transaction error occurred when a boarding "was not recorded because of equipment malfunction or a customer entering a station or boarding a bus without touching the reader" (Utsunomiya et al., 2006). A missing transaction is difficult to discern from a trip that is simply not made. An incorrect bus route error occurred when the bus operator entered an incorrect route number in the AFC system.

Even for vehicles equipped with AVL system, Furth et al. (2006) state that "the rate at which data is rejected for inability to match it to a route can be substantial, reaching 40% at agencies that

were interviewed. Data matching was cited by many agencies as the single greatest challenge faced in making their AVL-APC data useful". In real-life situation, "end-of-line operations can be both complex and unpredictable, making a trip's start and end times difficult to identify" (Furth et al., 2006). As a consequence, data associated to end-of-line, or to the beginning-of-line of the next route, are more prone to errors.

In sum, a widespread issue encountered by transit agencies is the uncertainty and inability to match AFC transactions, AVL or APC data with the correct run. "Success in matching depends to a large extent on the data captured. If the AVL system is integrated with a radio, operator sign-in including route-run number can be captured, which aids matching" (Furth et al., 2006). Although it is less problematic in transit agencies where the sign-in procedure is enforced towards the operators, data matching remains a serious limiting factor to fully leverage the data. "Systems in which sign-in errors are not detected until off-line processing cannot benefit from operators correcting their own input errors. As an example, farebox data systems often have very high rates of sign-in error, making boarding counts and revenue difficult for agencies to attribute to route" (Furth et al., 2006).

Furth et al. (2006) mention some transit agencies that have put in place mechanisms to tackle the data matching issue:

- Data from New Jersey Transit's APC/event recorder system are tied to route/run inferred from scheduled runs
- Metro Transit provide automated sign-in based on vehicle-block assignment
- Houston Metro validates route/run data by comparing sign-in data with payroll data in post-processing

However, details on how and the extent to which these practices are successful in validating and correcting data are not discussed.

2.2.2 Comparing Smart Card Data with Survey Data

Previous research has looked at the comparability between smart card data and survey data. Numerous aspects including sampling rate and accuracy of the declared value may contribute to the discord between the data. Farzin (2008) does a brief comparison between data collected from

traditional household survey and those from AFC with AVL in São Paulo, where the bus network is complex. Park, Kim & Lim (2008) investigate whether smart card data from Seoul, South Korea, are consistent with other survey data in terms of ridership and the number of transfers. Trépanier, Morency & Blanchette (2009) compare transit use indicators derived from smart card transactions and household travel survey data of the Ottawa-Outaouais region to evaluate the potential of data fusion. They identify comparable variables from the two datasets and find that there are significant differences in terms of ridership, especially when examined at the route level.

2.2.3 Potential Use of Smart Card Validation Data in Transit Planning

The potential applications of smart card validation data in transit planning have been extensively acknowledged in recent works. After the first field operation test of a smart card AFC system with GPS in California, Chira-Chavala & Coifman (1996) reports one of the benefits is the enhancement of collection and quality of transit data. They include "bus users by time of day, day of week, date, and week; bus users at individual bus stops; load profiles; on-time performance statistics; and average speed by segment and route". Lehtonen et al. (2002) define various methods of using public transit payment system data. The authors suggest by combining interconnected information systems, transit routes can be described by travel time, number of boarding, load, transfer area, boarding and exit stops. The data can also provide transport statistics. Wofinden (2003) acknowledges the smart card as a new and relevant technology for non-household survey. Trépanier et al. (2004) study the potential of smart card data using the totally disaggregated approach and object-oriented modeling (Figure 2.1). The following sections review this emerging research area with examples of passive data from actual systems. They are grouped by the types of application. Pelletier, Trépanier & Morency (2009) and Trépanier (2010) also provide literature reviews on the use of smart card data in transit planning.

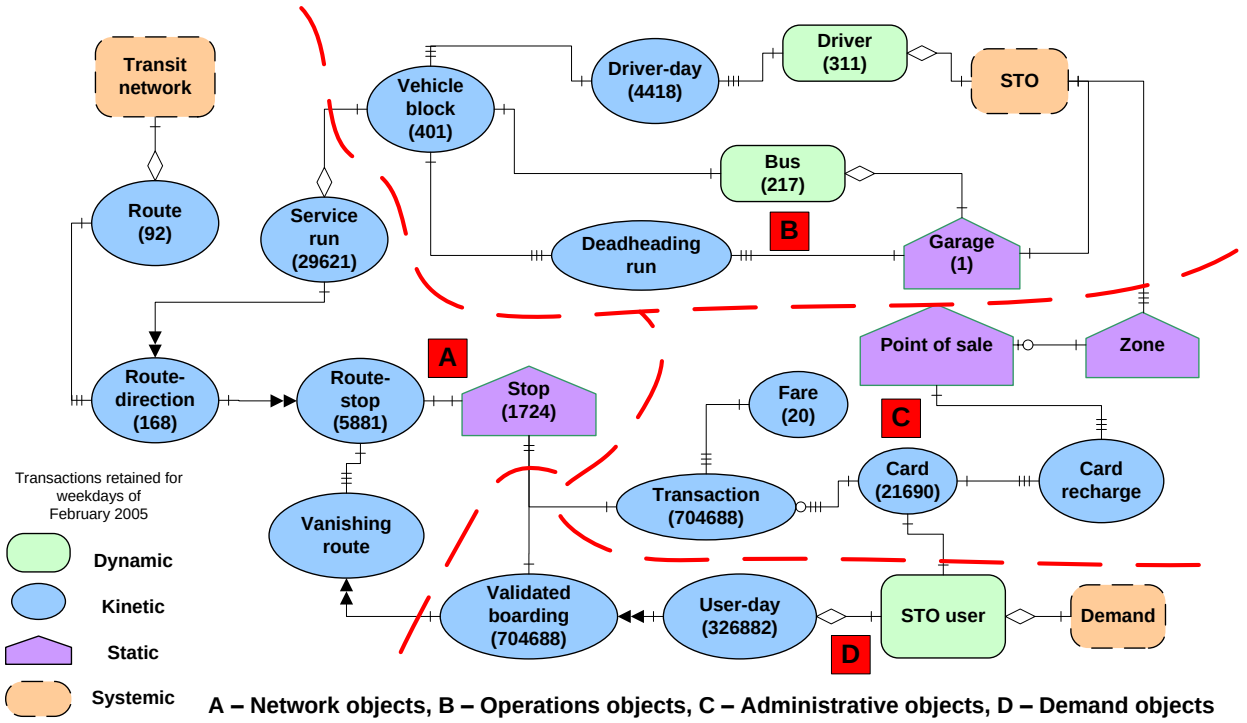


Figure 2.1 Object-oriented modeling of the smart card AFC system of the STO (taken from Trépanier et al., 2004).

2.2.4 Transit Service Performance Analysis and Monitoring

The most intuitive applications of passive data spawn from the needs of monitoring the performance of transit services by the operators. Bertini & El-Geneidy (2003) use data from TriMet, which are mainly gathered from the APC and GPS systems, to produce transit performance measures. The procedure first integrates operations data, vehicle locations from GPS, door activations and passenger movements at the stop level. Vehicle speed, schedule adherence, dwell time and service area analyses with GIS are subsequently performed using the data. Bullock, Jiang & Stopher (2005) use GPS technology to measure on-time running of scheduled bus services. Archived continuous GPS logs are broken into basic trips. On-time running assessment is done by comparing the recorded time against scheduled time at time points. Hammerle, Hayes & McNeil (2005) analyze a sample of AVL-APC data from the Chicago Transit Authority (CTA). The system records stop-level events, driver sign-ins, time points and distance traveled. The data are used for schedule adherence analysis, examining headway regularity and bus-bunching with time-space diagram. Golani (2007) uses GIS to visualize AVL-APC data of a route of the Champaign-Urbana Mass Transit District to look at schedule deviation.

The author also computes statistics on passenger boarding and on-time performance. Berkow, El-Geneidy, Bertini & Crout (2009) apply statistical models and visualizations on historical data from a bus dispatch system that includes AVL and APC. Joining 1.6 million ERF data and weather data in an Irish city, Hofmann & O'Mahony (2005a) examine the impact of adverse weather, mainly rain or no rain, on ridership, average frequency of service, headway regularity and bus bunching and travel time variability. However, the paper does not clearly indicate how headway and bunching are measured. Trépanier & Vassivière (2008) present an intranet interface containing multiple indicators generated from data from a smart card AFC with GPS. The indicators, intended for monitoring transit network use, include the number of boardings and transfers by fare type, load profile, passenger-kilometres and operations statistics for each run. Park et al. (2008) use smart card data of a bus and metro system from Seoul to calculate indicators describing transit trips: ridership, number of transfers, boarding time, trip time distribution and travel time distribution. The AFC system has a very high market penetration rate and in most cases collects both entry and exit data. Trépanier, Morency et Agard (2009) propose the use of location-stamped smart card transaction data to calculate transit performance measures such as schedule adherence, network supply and usage, and travel speed. Uniman, Attanucci, Mishalani & Wilson (2010) study service reliability of London's underground with smart card transaction data.

Passive data are also used to model dwell time and travel time. Rajbhandari, Chien & Daniel (2003) use APC data from New Jersey and regression models to estimate dwell time at bus stops. Farhan, Shalaby & Sayed (2002) model bus travel time using AVL and APC data.

2.2.5 Deriving Origin-Destination Information of Transit Trips

Another priority for transit operators is to measure the transit demand. Origin-destination pairs can be quantified at the route-level but should ideally be expressed with respect to the transit network, or the entire study area. AFC systems generate fare validation records which provide information on the trips made. In general, the data content depends primarily on three characteristics. The first concerns the location of the fare validation. It can either be recorded at a fixed and known location, such as a train station, or at a moving location, such as a vehicle, where the position can only be determined with an AVL system. In the latter, the referencing of the fare validations to a location is more susceptible to error because of vehicle movement. The

second characteristic relates to the fare control strategy. Some fare structures, such as distance-based fares, require users to perform fare validations at both ends of the trip (entry and exit), while others require users to perform only once, usually at entry. In the second case, an extra procedure is required to infer the location of the other trip end. The third characteristic involves the level of aggregation of the data that the AFC system provides. They range from an aggregated ridership count to disaggregate transactional records where individual validation contains complete details and is tied to a specific card. The following paragraphs summarize recent works that aim to derive origin-destination information from APC or AFC data.

2.2.5.1 Aggregate data

Although many AFC systems do not capture alighting information, there has been research on deriving origin-destination pattern. Navick & Furth (2002) use location-stamped farebox data from buses in Los Angeles to estimate alighting patterns based on the assumption of origin-destination symmetry. Due to the aggregate nature of the data, only travel patterns on a single bus line can be obtained. Richardson (2003) estimates alighting pattern of a bus route in Australia assuming symmetry between the boarding count in one direction and the alighting count in the opposite direction. Both works focus on the aggregate travel pattern of a bus route and use symmetry as assumption.

2.2.5.2 Disaggregate Data

Farecards such as magnetic cards and smart cards are the main sources of disaggregate ridership data. Barry, Newhouser, Rahbee & Sayeda (2002) use data from an entry-only AFC system to estimate origin-destination matrices for the New York subway system. The alighting station is deduced by assuming that an individual starts a trip at the station where the previous trip ends and returns to the entry station of the first trip in the last trip of the day. Using a similar concept, Zhao (2004) infers the destination station for individual trips and models the path choice in rail-to-rail sequences using data from the smart card AFC system in Chicago. Zhao & Rahbee (2007) describe the procedure from a programming perspective and estimate a rail passenger trip origin-destination matrix. These two applications are also discussed in Wilson, Zhao & Rahbee (2008). Cui (2006) applies the inference concept to estimate an origin-destination matrix of a bus route where smart card boarding data are available. Trépanier et al. (2007), assuming that individual alights at a downstream stop closest to the subsequent recorded boarding stop, derive the

alighting stop for bus trips with entry-only smart card validation data. These works take advantage the disaggregate property of the data and focus on the continuity of boarding sequence of an individual card, sometimes referred as trip chain or more preferably "boarding chain", revealed by the card's identification number. Zhang, Zhao, Zhu, Y. & Zhu, Z. (2007) estimate a bus-stop origin-destination matrix using smart card boarding validation data without AVL. The transactions are joined with time check data filled out by drivers at each stop. Only two peak periods (lasting an hour each) are considered. The alighting stop of a trip made in the AM peak is assumed to be the boarding stop in PM peak and vice versa. Using the Furness method, the partial OD matrix is expanded to represent the whole region of Changchun, China. In order to synthesize a zonal origin-destination matrix from smart card AFC-AVL data in São Paulo, Farzin (2008) proceeds with the following processing steps:

- Cleaning datasets, and matching raw GPS data from the vehicles to bus stops and zones;
- Associating smart card boarding to zones according to transaction time;
- Inferring destination zone.

The last step follows the same concept as the previous works on destination inference. For unknown destinations, a proportional model is used. The final matrix contains more than 500,000 records. Seaborn, Attanucci & Wilson (2009) analyze multi-modal journeys in London, United Kingdom, with eight millon smart card fare validations from a representative day. They identify complete journey including transfers under some transfer time assumptions. Jang (2010) derives estimated travel time for stop-level origin-destination pairs with more than 100 million smart card records from Seoul, South Korea. The author also analyzes transfer trip patterns and transfer location choice. Munizaga, Palma & Mora (2010) estimate an OD matrix for Santiago, Chile, from smart card validation data.

2.2.6 Analyzing Travel Behaviour of Transit Users

Passive data can be a valuable source to study the travel behaviour of transit users. The understanding of travel behaviour allows researchers to improve planning tools and transit operators to better organize operations and service. Bagchi & White (2004; 2005) examine samples of public transit smart card data from Southport, Merseyside and Bradford, in the United Kingdom. The authors acknowledge that through smart card system, transit agencies:

- Have access to larger volumes of personal travel data;
- Are able to link those data to the individual card and/or traveler;
- Have access to continuous trip data covering longer periods of time than it is possible to obtain using existing transport data sources;
- Know who their most frequent customers are.

A random sample of 10% of cards is drawn following data validation to estimate trip rates per card, the proportion of linked trips and turnover rates using rule-based processing. The authors suggest the use of "verificatory" or complementary surveys to supplement information not collected by the system, such as trip ends and trip purpose, and to validate modeling assumptions and inferred information against actual behaviour. Bryan & Blythe (2007) use data from concessionary customers in Nottinghamshire County, United Kingdom, to analyze trip rates by user subgroups: elderly, disabled and school children. Boarding by route, stop and time of day are also examined. Limitations include large temporal and spatial gap between successive uses and the lack of alighting information. Utsunomiya et al. (2006) present several types of analysis that can be done using smart card registration and transaction data from the CTA, Chicago. More than 520,000 boarding transactions from more than 62,000 cards were made over the seven-day analysis period. Analyses include walk access distance from billing address to the first transit entry point, frequency of use, transfer trips, daily travel patterns, variation in transit mode, route and stop choice, and comparison of usage by area of residence. Mojica (2008) examines changes in travel behaviour under deteriorated service conditions with smart card data and subsequently provides a modal shift model using a binary logit formulation.

Data mining is a useful technique to examine travel behaviour of transit users from a large amount of data. Okamura, Zhang & Akimasa (2004) study travel behaviour using one month of boarding records from an AFC system in Hiroshima, Japan. They analyze passenger boarding time and transfer behaviour. When compared to a questionnaire-based person trip survey, it is noticed there is significant rounding in the surveyed boarding time. The location, quantity of transfer trips and transfer wait time at major transfer points are examined. The authors also use data mining technique to classify cardholders based on similarity in monthly boarding frequency, boarding time period, average on-board time and average payment per ride. Tseytin, Hofmann, O'Mahony & Lyons (2006) use the market basket data mining technique to extend the temporal

coverage of travelers using weekly magnetic card by matching cards based on a similarity function between the travel patterns of two cards. Morency, Trépanier & Agard (2007) examine a complete dataset of 2 million boarding transactions by about 25,000 smart cards made within an 11-week period. The boardings of each card are transformed into binary attributes by day of week and period of day and are fed into a data mining algorithm to form clusters. Travel behaviours are analyzed by frequency and time of transit use, fare type, number of boardings, and number of stops used. Two cards, one adult and one senior, are examined in more detail over a period of 9 months. The number of boardings and its variation are studied, and clusters are formed according to the presence of transaction by each hour. The research provides a framework to study the variability of travel behaviour over a long time horizon. Morency & Trépanier (2010) use smart card validation data from an extended period to assess loyalty of different segments of cardholders. Asakura, Iryo, Nakajima & Kusakabe (2009) estimate change in behaviours of train users after a schedule change with smart card data. The authors take advantages of the spatial-temporal constraint and the temporal resolution of the data to assign users to the most probable train itinerary. The main behavioural aspects examined are arrival time, exit time and travel time distribution.

2.2.7 Enriching Passive Data with Trip Details

Data enrichment techniques are often applied in data processing to derive information that is not directly captured during the data collection phase. Household travel surveys that require interviewees' direct participation usually solicit all essential information and require little enrichment. When done, it is often direct and has a high level of certainty. For example, activity duration can be derived from the time gap between the departure times of two successive trips. The process does not involve external information and complex algorithm. Meanwhile, the information included in passive data is limited by the system setup and may not contain all the necessary data for transport analyses. Some elements related to the person, the household or the trips cannot be recorded without active input. Because of this, surveys using passive data collection methods are often followed up with complementary questionnaires. As an alternative, various techniques of data enrichment aiming to obtain the sought-after information have been developed. Data enrichment is often linked to terminology such as "to derive", "to identify", "to infer", "to deduce", "to estimate" and "to interpret". Although they are not associated with

specific scientific definition, they suggest varying levels of confidence regarding the technique and the result.

Other than deriving the alighting stop of a fare validation, which is reviewed in the previous section, enrichment techniques on passively gathered data mostly focus on the following information:

- Travel mode;
- Transfer information for transit trips;
- Trip ends and trip purpose of automobile trips.

2.2.7.1 Travel Mode

Since the advent of portable GPS units, the GPS has become a viable alternative to traditional travel survey. Research has been done to test whether missing information can be accurately reproduced using data enrichment techniques. Chung & Shalaby (2005) develop a software that performs map-matching and assigns trips into one of the four modes: walk, bicycle, bus and passenger car. The tool is tested against trips in a multi-modal origin-destination survey replicated by participants wearing GPS. Tsui & Shalaby (2006) use algorithm to automate the post-processing of GPS log, with or without GIS. The GPS-alone process identifies activity and mode. The GPS-GIS process includes link identification with street network data.

2.2.7.2 Transfer Information for Transit Trips

Linked trips and transfer activities can only be studied if the passive systems recognize the fare media by their identification number. This is the case for AFC with multi-use magnetic cards or smart cards. In some countries, like in Canada, a linked trip is the basic measure unit of transit use.

There are efforts to identify and analyze linked-trips with AFC data. Apart from comparing the routes taken, the identification of linked trips in previous studies is solely based on a fixed temporal threshold between boardings, which is identical to the concept used in fare policy by many transit agencies. The concept, however, does not aim to accurately determine linked trips but to allow users to transfer between routes for free or with a supplement. Okamura et al. (2004) analyze transfer wait time at major transit hubs. They define transfer as a trip which involves two

different operators and has a wait time less than 60 minutes at the transit hub. Hofmann & O'Mahony (2005b) describe a procedure that classifies unlinked and linked trips within a huge database of magnetic card transactions. The iterative classification algorithm is based on the assumption of time elapsed between successive boarding transactions, a 90-minute threshold chosen according to preliminary analysis, and performs a comparison on routes taken. Bagchi and White (2004; 2005) examine bus-to-bus linked trips using smart card transaction data. The algorithm compares the routes taken and the default temporal threshold is 30 minutes between the successive boardings. Seaborn et al. (2009) also use temporal thresholds to identify transfer trips in London. They establish different time assumptions for underground-to-bus, bus-to-underground and bus-to-bus transfers in order to reconstruct complete trips.

Although it is computationally simple to handle, the weakness of applying a threshold value is that it can be seen as arbitrary even with considerations of network size and specific user groups. A fixed value that does not take into account in-vehicle time and route headway would invariably classify all boardings that are carried out within the threshold as linked trips. Trips with a short activity duration or trip chains might be masked and return trips might be counted as transfers. It does not consider the spatial-temporal coincidence of a transfer.

2.2.7.3 Trip Ends and Trip Purpose of Automobile Trips

Two of most important pieces of information in a transport survey are trip ends and trip purpose. In passive automobile GPS survey, trip ends are usually detected by engine start up and shut off or by the absence of movement at a particular location. However, the exact activity location and trip purpose cannot be known without the respondent input. Similarly, trip ends in transit trip, as opposed to boarding and alighting stops, and trip purpose are not contained in AFC data. In some cases, cardholder address is available through card registration. However, the address does not necessarily represent a home address and may not be related to trip ends.

Trip ends and trip purpose are closely linked. Both elements are not explicitly collected by passive methods and require inference by post-possessing. Wolf, Guensler & Bachman (2001) examine the possibility of deriving trip purpose in a view to eliminate travel diary. In a further study, Wolf, Schönfelder, Samaga, Oliveira & Axhausen (2004) segment a continuous GPS track into trips based on halt in movement. Using an external database of land use and GIS, based on activity duration and the proximity from parking location, an algorithm automatically selects a

trip end and a trip purpose (Figure 2.2). Kuhnimhof & Wassmuth (2002) work on refining broad categories of trip purpose into more specific sub-categories in a survey. The use of trip context, or less scientifically called “common-sense”, is invaluable in choosing the correct sub-category. Bohte & Maat (2009) derive trip purpose and travel modes from a multi-day GPS survey in the Netherlands. The GPS logs are interpreted according spatial data and individual characteristics. The results are then validated and adjusted through interactive web-based applications by respondents.

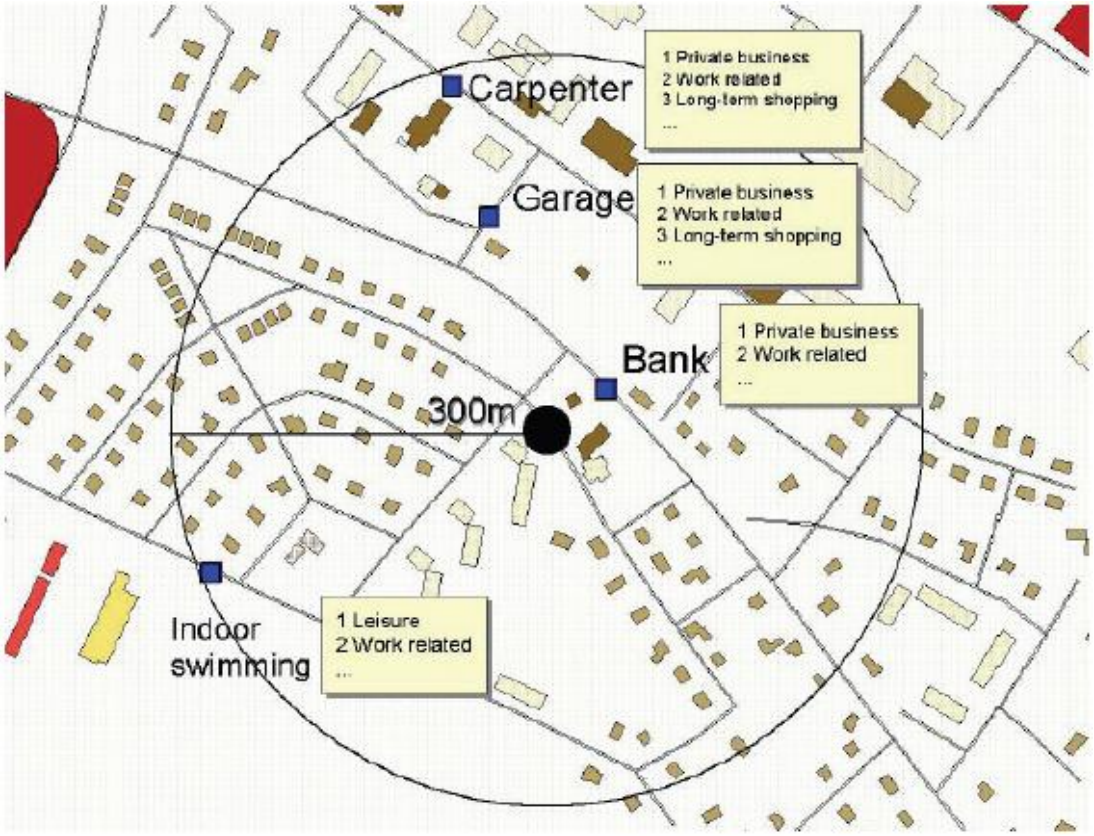


Figure 2.2 Enriching GPS data with land use information and GIS (taken from Wolf et al., 2004).

In summary, data enrichment techniques on travel mode, trip ends and trip purpose focus on the post-processing of data. It also requires the input of external data, especially spatial data, to suggest candidates for trip ends. Many algorithms involve judgment and interpretation, whether programmed into rules and algorithms or remained manual, in determining the trip ends and trip purpose. In earlier studies, the number of records in the experiment is small while recent developments are applied to large-scale dataset. Schuessler & Axhausen (2009) propose a data

processing procedure where trips, activities and travel modes are identified from raw GPS data without additional information.

2.3 Travel Behaviour Studies

Travel demand is derived from the need to participate in activities. In order to estimate and predict travel demand, it is necessary have a good understanding of the underlying travel behaviour that drives it. Travel behaviour studies often require cross-sectional data or multi-day data.

2.3.1 Cross-sectional Analysis

Morency (2004) studies mobility, activity location and the characterization of activity rhythm using the Montréal regional travel surveys and the Canadian census. Cross-section data are useful for determining the travel behaviour on an average day but lack the ability to provide information on variability over time.

2.3.2 Multi-day Analysis

Analysis on multi-day data seeks to address the varying travel needs over time. In general, it involves measuring similarity or variability in travel of an individual or a group. The way that similarity is defined depends on the perspective of the researcher (Huff & Hanson, 1990). The main target aspects are:

- the number of daily trips or trip rate;
- the number of visited locations;
- the number of different trips, defined by two or more attributes;
- travel and activity time budget;
- travel time and distance;
- activity sequence.

Schönfelder (2006) uses recently available longitudinal data, such as the Mobi*drive*, to investigate urban rhythm and model intra-personal travel behaviour. The research provides several spatial measures of activity location. Elango, Guensler & Ogle (2007) explore the

significance of demographic attributes on the day-to-day variability of the total number of daily household trips. Evolution of travel behaviour can also be examined using panel data (Stopher, Clifford & Montes, 2008).

Analysis outcome can be influenced by the way a trip is measured. Hanson & Huff (1988) "generally notice that the more detailed a measuring procedure is and the more attributes it covers, the smaller are the observed similarities". This leads to the idea of aggregating trip details to reveal travel pattern. Schlich & Axhausen (2003) provide a comparison among various methods to measure similarity in travel behaviour with multi-day data. Most often, the methods take on an algebraic approach. Susilo & Axhausen (2007) describe the use of Herfindahl-Hirschman Index, an indicator that measures market concentration, to examine the stability of individuals' choices of their daily "activity – travel location" combinations with longitudinal data. Bayarma, Kitamura & Susilo (2007) use Markov chain to analyze transition among travel patterns.

The duration of observation period is another factor that influences analysis outcome. Traditional multi-day surveys usually have a short duration of observation because of the level of effort required on the respondent to complete the survey. Passive data collection methods, especially those that do not require follow-up, can have a longer duration. Based on different measures, Schlich & Axhausen (2003) conclude that similarity analysis of travel behaviour on weekdays should covers at least two weeks. This provides a guideline on the minimum amount data for multi-day analysis.

2.4 Information Technologies

The recent works on passive data make use of various information technologies. Data in their raw form provide little value. In order to analyze data and generate useful information in support systems, they need to be stored, retrieved and processed. Some analytical tools are particularly suited to treat the amount and the spatial-temporal properties of passive data.

2.4.1 Data Storage, Retrieval and Processing

A common tool to store, retrieve and process data is the relational database. A relational database contains one or more tables and is structured in a way where tables are associated to each other in an explicit relationship via a shared attribute. Through queries, which can be expressed in the

Structured Query Language (SQL), data from different tables can be selectively combined. This eliminates the need to store redundant data and allows users to combined different data sources. SQL can also be used to program processing procedure and to automate tasks. On the other hand, spreadsheets have some of the properties of a relationship database. They provide a simple interface to visualize and manipulate data. They are customizable and programmable with the built-in language. The latest version of spreadsheet can handle more than one million records on each sheet.

2.4.2 Geographic Information System

Geographic information system (GIS) is a computer system for capturing, storing, checking, integrating, manipulating, analyzing and displaying data related to positions on the Earth's surface (Stanford, 2010). It behaves like a database but allows users to visualize, analyze data with a spatial component and to uncover spatial patterns. Data are organized by layers on which operations are performed. Data format and operations can be primarily divided into two categories: vector and raster. Data in vector format are classified into point, line and polygon features. Each feature is associated with one or more attributes. Data in raster format are organized by grids, or cells, of a particular size. Each cell is associated with one numeric value.

A transportation network is more easily represented in the vector format, while aggregation and arithmetic calculations are more easily done on raster data. Spatial data in Québec uses coordinates in MTM format (Modified Transverse Mercator, Zone 9).

2.4.3 Spatial Statistics

Spatial statistics examine data with a location component. The most basic type of spatial distribution descriptors are centrographic statistics (CrimeStats, 2005). They are two dimensional correlates of the traditional statistics that describe a univariate distribution, such as the mean and standard deviation. Examples of spatial descriptors include the mean centre and the standard deviational ellipse. The mean centre, also known as the centre of gravity or barycentre, is the mean of the x and y coordinates. In two dimensions, distributions are frequently skewed in one direction or another, a condition called anisotropy. The standard deviational ellipse can be used to measure dispersion in two dimensions. These statistics are often used to describe or compare the spatial distribution of different populations of incidents.

2.4.4 Data Mining

Data mining, sometimes referred as knowledge discovery in databases (KDD), is a broad term described as “the nontrivial extraction of implicit, previously unknown, and potentially useful information from data” (Frawley, Piatetsky-Shapiro & Matheus, 1992) and “the science of extracting useful information from large data sets or databases” (Hand, Mannila & Smyth, 2001). The purposes of data mining are about “solving problems by analyzing data already present in databases”, “discovering patterns in data” as well as data cleaning (Witten & Frank, 1999), which is usually defined as the process of modifying the form or content, such as filling in missing values and correcting erroneous data, in order to improve the accuracy of the data. Common data mining techniques include classification, cluster analysis and association rules. There are many algorithms developed for each technique. A set of parameters needs to be defined depending on the algorithm. The following is a brief description of each technique:

- Classification tries to build a model using observations (instances) with several attributes and one target category, known as class. The use of a predefined class makes it a supervised method. Observations are usually divided into the training set and the test set. An algorithm uses the training set to build a model and then try the model with the test set. The goodness-of-fit of the model can be evaluated by how well the model predicts the class in the test set. The goodness-of-fit can be shown as a percentage of correct predictions or as a confusion matrix.
- Cluster analysis groups similar observations into categories without the need of a predefined class, thus considered an unsupervised method. It calculates distances, usually Euclidean distances, among the instances in a multi-dimensional space representing the attributes. It minimizes the variance of the instances within the same cluster and maximizes variance across clusters. It uses all the instances and there is no common measure for goodness-of-fit.
- Association rules reveal how often various binary attributes appear within the same instance and how strongly the rules are held across instances.

In all cases, data pre-processing and transformation are essential in the knowledge discovery process. Steps such as deriving additional attributes from the data, removing correlated attributes and transforming data type can improve results.

2.4.5 Visualizations

Data by themselves are hard to comprehend. This is especially true with dealing with the amount of passive data. Visualization helps to present concepts, summarize data, detect inconsistencies and errors as well as reveal patterns in data. A well-thought visualization is coveted. The following types of visualization are frequently used to analyze multi-dimensional data:

- A scattered plot shows a data point by two variables, which can be numeric or categorical. A spatial location can generally be represented by X-Y coordinates, namely the longitude and latitude. A three-dimensional scattered plot shows a data point by three variables. Multivariate visualizations, such as a bubble diagram, extend the three spatial dimensions with the use of colour, size and shape in order to represent complex data. Data reduction techniques can also be used to reduce the dimensionality of the data.
- GIS provides powerful multivariate visualizations of spatial data. However, the GIS interface can be manipulated to illustrate any numeric or categorical variables instead of using the real x-y coordinates of a spatial object. It can be called false-GIS. It takes advantage of the spatial analysis and visualization capability of GIS to study multi-dimensional data. Figure 2.3 shows an example of false-GIS in a public transit context (Dionne, 2006).

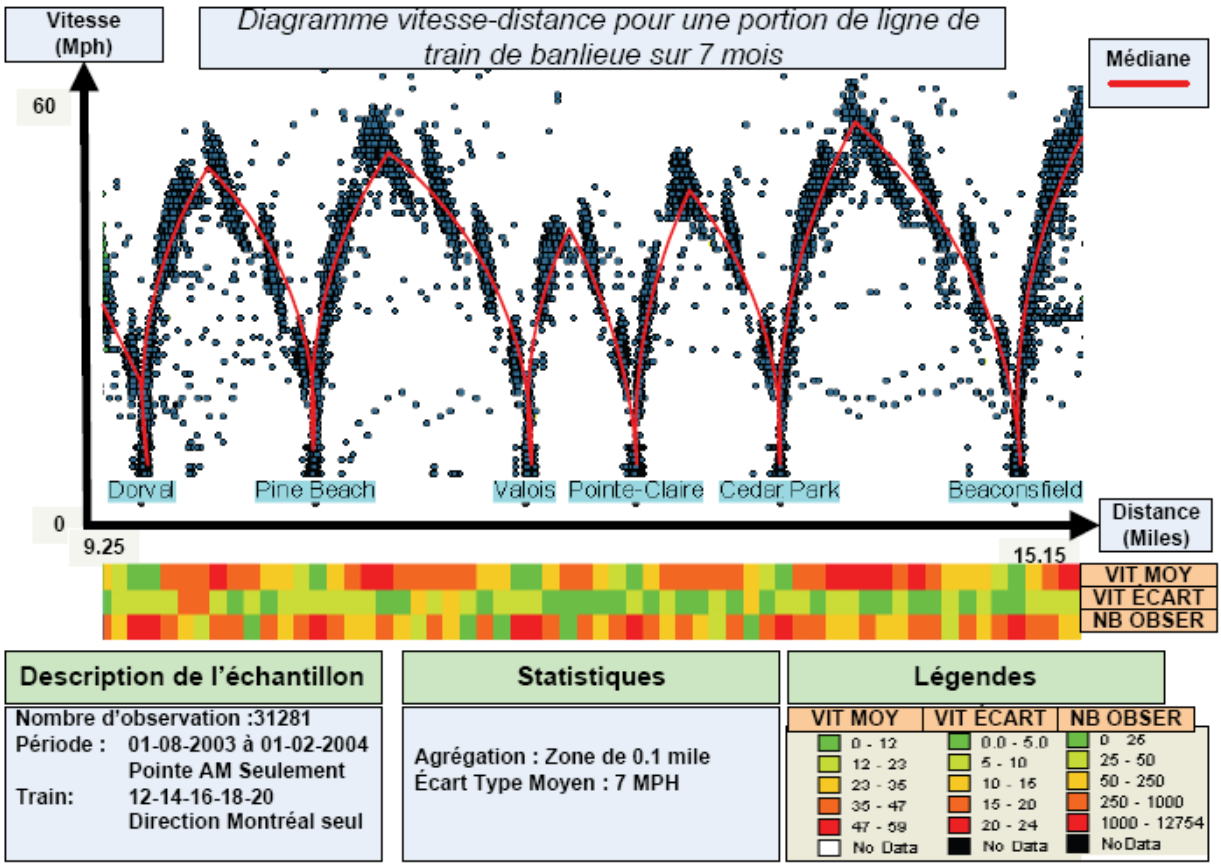


Figure 2.3 An example of "false-GIS" with GPS data (taken from Dionne, 2006).

- Time-space diagram is a type of X-Y plot where one axis represents time and the other, linear distance. It is often used in transportation to study the movement of people and goods in space and time.
- Animation consists of successive and overlapping figures illustrating the evolution of objects at different time. It is not possible to show animation on printed material. Instead, it is illustrated by successive snapshots. An example shows the within-day spatial-temporal evolution of the population with OD survey data (Figure 2.4).

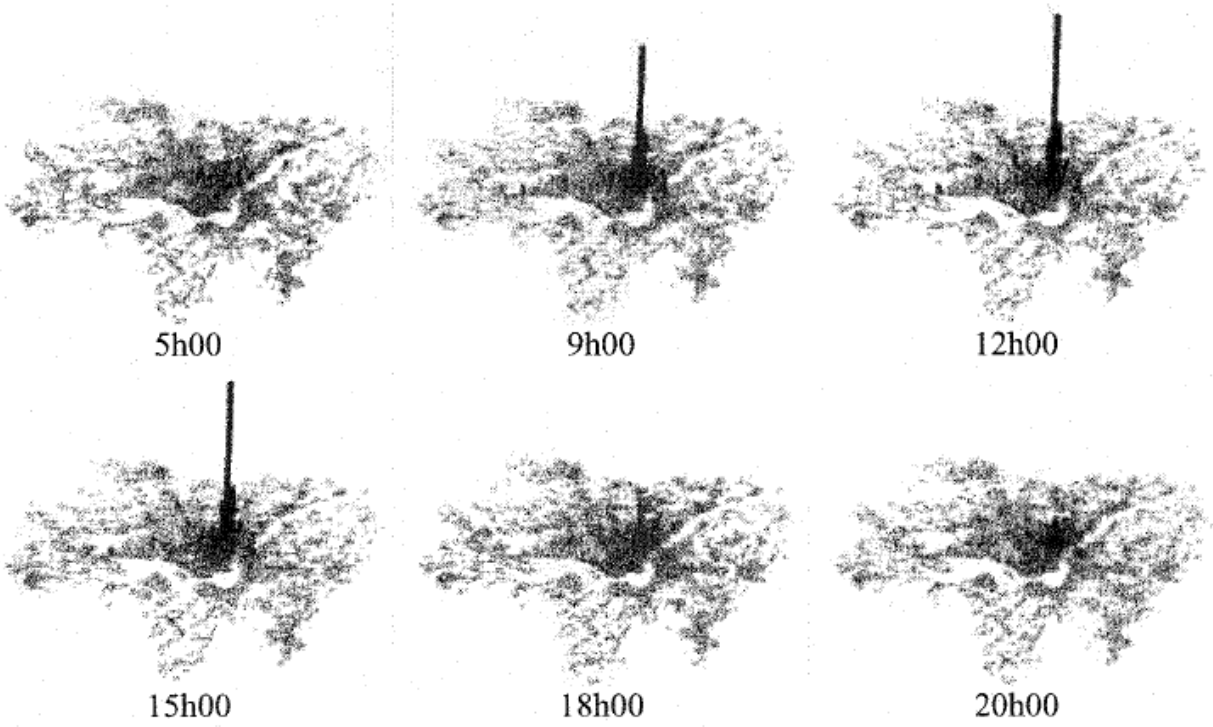


Figure 2.4 An animation based on OD survey data of the Greater Montréal Area (taken from Morency, 2004).

2.5 Recapitulation on Theoretical Background

The study of passive data in public transit, more specifically those from a smart card AFC system, comes across a broad range of subjects. The literature review provides a theoretical foundation upon which the research is built. The methodological aspect of the research would:

- Take into consideration related subjects;
- Evaluate existing methods, identify gaps in the existing literature and propose innovative methodology in data processing and analysis;
- Generalize the methodology for applications on similar data.

Meanwhile, the analytical aspect of the research would:

- Contribute to the improvement of public transit planning by clarifying issues on demand and supply of the network;

- Contribute to the understanding of travel behaviour by revealing travel and activity patterns of transit users.

CHAPTER 3 DATA AND GUIDING RESEARCH PRINCIPLES

One of the unique features of this research is having the privilege to access data from two distinctive operational smart card AFC systems. Another one is to inherit the knowledge in transit planning and the tradition from the MADITUC Group. In this chapter, the datasets used in the research will be presented in details. The guiding research principles of the research, representing the school of thought of the MADITUC Group, are also explained.

3.1 Data Source

In general, a smart card AFC system typically interacts with the transit clientele in three different situations:

- During the advanced purchase of a fare product from a ticket booth, an automatic ticket vending machine, an authorized ticket vendor or by subscription. This involves a fare product being loaded onto a fare medium.
- During fare validation, when a fare product is consumed for a transit service. This can take place at a validation machine (validator) located at a turnstile inside a subway station, in a bus vehicle, on a station platform or using a handheld validator machine. The system distinguishes first boarding, transfer boarding or refused validation. For some distance-based or zone-based fare system, an exit reading may be required.
- During fare verification, when fare product and fare medium are checked for their validity inside the controlled area of the transit network by an inspector using a handheld card reader (verifier).

In each of these events, including unsuccessful attempt, a transactional record is generated, stored and subsequently uploaded to the central server. They can be classified as sales, validation and verification events respectively. Each record represents the communication between a fare medium (thus its holder) and an equipment at a specific time and place. The concept of a transactional record is therefore based on the serial number (unique ID) of the fare medium, the unique code of the equipment, a transaction time and a fare product. Depending on the host of equipment, the location of the equipment can be fixed (e.g. inside a station) or variable (e.g. inside a vehicle). A simplified diagram of the data flow in a typical multi-modal and multi-operator smart card AFC system is illustrated in Figure 3.1.

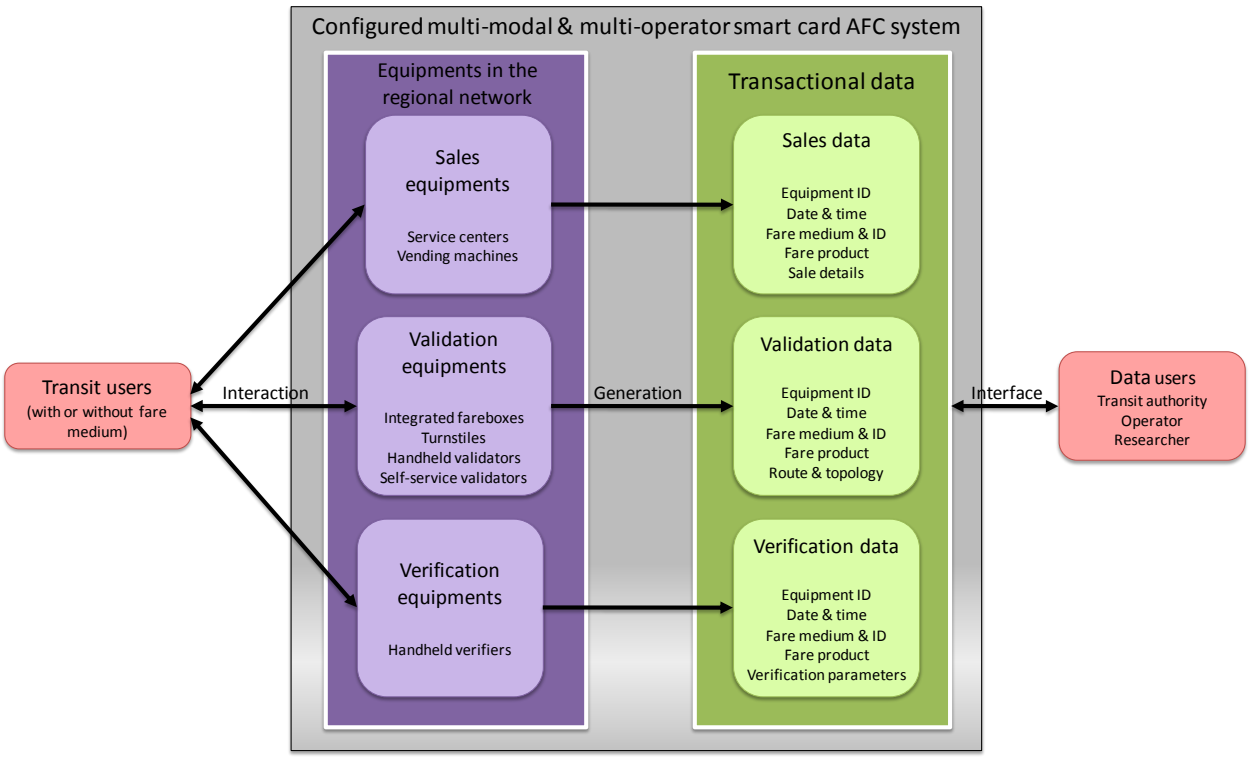


Figure 3.1 Simplified diagram of the data flow in a smart card AFC system (Chu & Bergeron, 2010).

Other than the transactional validation data of the smart card AFC system, complementary datasets are necessary to analyze transit demand and to understand travel behaviour:

- Data dictionaries from the AFC system which are used to decode values in the data tables;
- Information synthesized from the smart card AFC data which is used for data validation and enrichment;
- Spatial data from external sources.

Figure 3.2 illustrates the lineage and interaction of complementary data from the smart card AFC system. All the tables, presented in the following sections, are valid for the period of analysis.

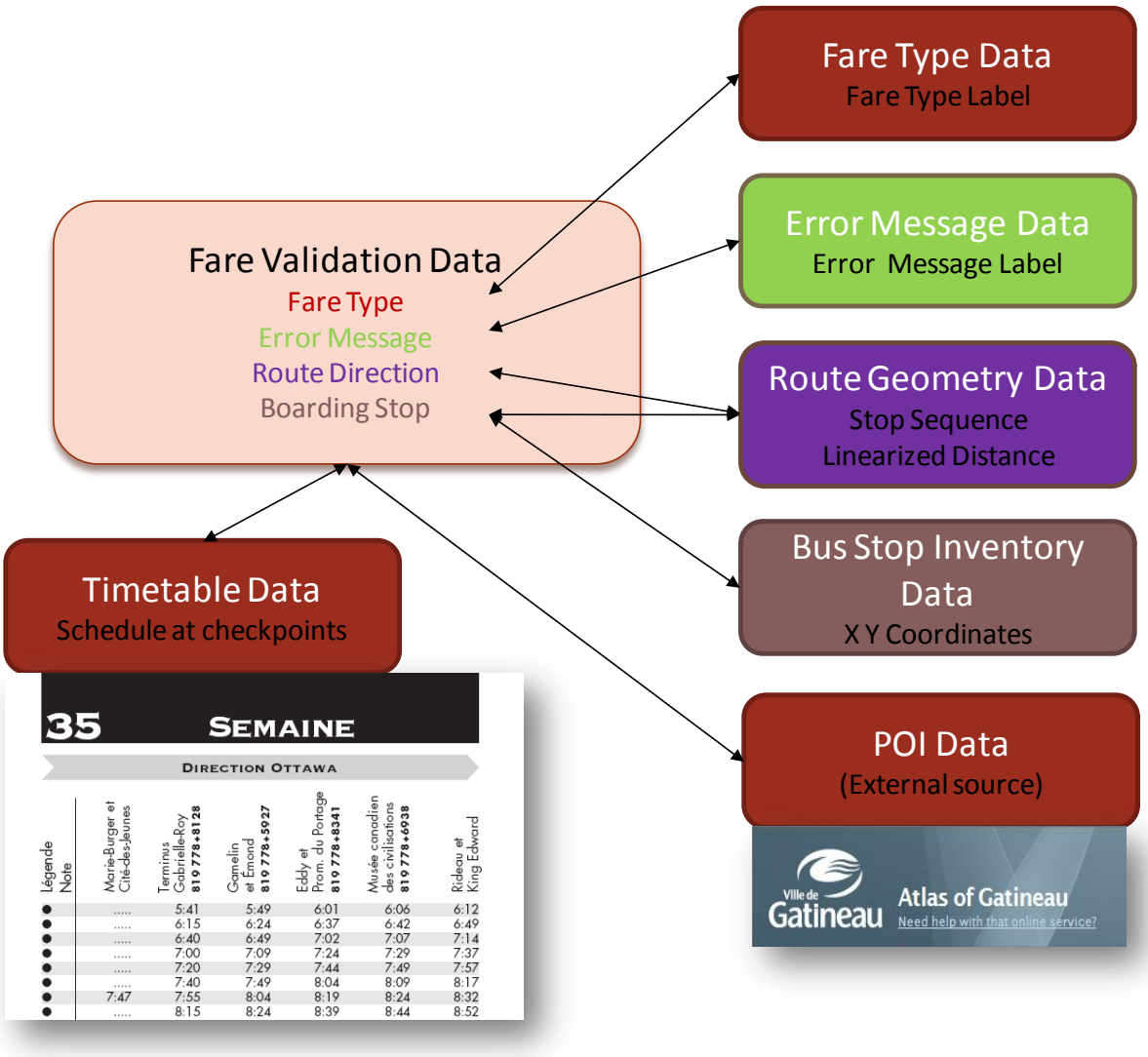


Figure 3.2 Interactions among various data from the smart card AFC system.

3.1.1 Data from the Smart Card AFC system of the STO

The primary data source of the research comes from the smart card AFC system of the STO. It operates a bus system in the National Capital Region of Canada, with most routes connecting Gatineau, Québec with the central business district (CBD) of Ottawa, Ontario. It transports about 18 million passengers annually and covers 11 million vehicle-kilometres. Several features of the system make this particular dataset valuable. First, the whole regular bus fleet is equipped with GPS for location identification, which allows each validation record to be associated with a

boarding stop. Second, each smart card is personalized and bears the photo of the cardholder, thus limited to one cardholder. No personal information is made available to the researchers other than the card number. Third, the system has a high market penetration rate. According to the STO, about 80% of its clients hold a smart card in 2005. Other studies reveal that market penetration varies greatly according to the practice of transit agencies. For instance, payment made with the Chicago Card represent 8.4% of boarding transactions for rail and bus during a 7-day period in September 2004 (Utsunomiya et al., 2006). In contrast, 90% of bus users and 72% of metro users use smart card for fare payment in 2005 in Seoul (Park et al., 2008). A one-day survey of passengers in São Paulo in 2006 revealed that over 75% of passengers use the smart card (Farzin, 2008). The variation can be attributed to the number of available fare options, such as limiting monthly passes to smart card users, and incentives, such as offering free transfers for smart card holders. The high market penetration rate ensures that the majority of demand from regular customers is captured by the smart card AFC system.

3.1.1.1 Fare Validation Data

The fare validation data table contains all the on-board fare validations for the month of February in 2005. The month has exactly four complete weeks with 20 weekdays, 4 Saturdays and 4 Sundays. In total, the dataset includes 763,570 boarding validations of which 713,276 were made on weekdays. The Thursday, February 10, is selected to represent a typical weekday. It contains 37,781 validations.

A smart card fare validation is a transactional event recorded by the AFC system. From the bus transit perspective, a smart card transaction occurs within a moving vehicle along a route. Since the STO bus transit system requires payment immediately upon boarding, it can safely be assumed that the validation location corresponds to the boarding stop. Whereas the boarding location is captured by the system, the alighting location is not as there is no exit reading. Each validation generates a transactional record which is temporarily stored in the on-board computer. Upon return to the garage, the data are uploaded wirelessly to the server from the vehicle, which are then compiled into a database. The temporal and spatial information in the database does not only provide the boarding locations of the cardholders, they also reveal the locations of the vehicles because by definition, a validation is the spatial-temporal coincidence between two objects: the cardholder and the vehicle. Operations data are integrated into each transaction.

Table 3.1 shows a typical sequence of validations generated by the system. Each record is associated with a unique identification number (Record ID) and contains information on:

- The smart card
    - *Card Number*: the serial number of the smart card involved in the validation
    - *Fare Category*: the fare product stored in the smart card involved in the validation
    - *Event Number*: the order of events of each smart card
- The boarding
    - *Validation Time*: recorded to the nearest minute and assumed to be the boarding time
    - *Validation Type*: indicating whether the validation is a first boarding or a transfer boarding
    - *Boarding Stop*: the validation location, represented by a stop number, determined by the on-board computer from GPS data and operations data, and assumed to be the boarding stop
    - *Error Message*: a message indicating the reason why the validation is invalid
- The vehicle
    - *Vehicle Number*: the vehicle associated with the on-board smart card validator. It is assumed that the equipment is permanently fixed to a specific vehicle
    - *Event Sequence Number:* incrementing by one for each event in the on-board computer, such as change of location or validation
- The operations
    - *Vehicle Block Number*: referring to a vehicle-trip
    - *Route Number*: referring to a route in the fixed-route bus network
    - *Direction*: indicating the direction of a route. Inbound is denoted by 0 and usually means towards the CBD of Ottawa. Outbound is denoted by 1 and usually means towards Gatineau

- *Planned Departure Time*: referring to the scheduled time when the vehicle departs from the terminus. Since the service is schedule-based, each run has a planned departure time at the point of injection
- *Driver Number*: the unique identification number representing the driver who is signed in when the transaction occurs

Less relevant fields in the data are not shown and omitted from the study.

Table 3.1 Excerpt of smart card fare validation transaction records.

Record ID	Card Number	Fare Category	Transaction Date	Transaction Time	Transaction Type	Error Message	Vehicle Blcok Number	Route Number	Direction	Departure Time	Boarding Stop	Bus Number	Driver Number	Transaction Number	Event Sequence Number
23080308	1454633320	2	Feb-01	06:50	1		207	83	0	06:48	4412	9217	3283	202	5
23080309	2187852747	2	Feb-01	06:50	1		207	83	0	06:48	4412	9217	3283	108	6
23080310	2173486001	16	Feb-01	06:51	1		207	83	0	06:48	4410	9217	3283	205	7
23080311	1381629131	2	Feb-01	06:53	1		207	83	0	06:48	4406	9217	3283	195	12
23080312	3325226344	2	Feb-01	06:53	1		207	83	0	06:48	4406	9217	3283	200	13

As shown in Table 3.1, some information that is usually collected in traditional survey is not recorded in the validations. It includes alighting stop and time, trip ends (as opposed to entry and exit points in the transit network), trip purpose, cardholder socio-economic status, travel companion and the location of the vehicle when there is no boarding transaction. Therefore, the post-possessing of the data, which includes data validation and enrichment, is of critical importance. In order to derive information from the data, the fare validation data table must be related to explanatory tables, such as the fare type and error message data tables. Route geometry and bus stop inventory data tables are essential to spatially locate the bus stops, to perform spatial analysis and to calculate statistics.

3.1.1.2 Fare Type Data

Fare type is represented by a numeric value in the validations. It requires a descriptive label to be understood. The fare type data table (Table 3.2) provides a description for each fare type. There

are 46 different fare types. For analysis purposes, they are aggregated into more meaningful entities such as adults, seniors, students under the age of 21, students 21 and over, and STO employees.

Table 3.2 Excerpt of fare type data.

Fare Code	Long Label	Short Label	Replacement Type	Beginning of Validity	End of Validity	Active
1	ADULTE REGULIER	AD_REG	FALSE	01/01/1900	01/01/1900	TRUE
2	ADULTE EXPRESS	AD_EXP	FALSE	01/01/1900	01/01/1900	TRUE
3	ADULTE INTERZONE	AD_IZN	FALSE	01/01/1900	01/01/1900	TRUE
4	ET REGULIER 1999	ET_99R	FALSE	01/01/1900	01/01/1900	FALSE
5	ET EXPRESS 1999	ET_99E	FALSE	01/01/1900	01/01/1900	FALSE
6	ET INTERZN. 1999	ET_99I	FALSE	01/01/1900	01/01/1900	FALSE
8	AINE	LP_AIN	FALSE	01/01/1900	01/01/1900	TRUE
9	CARTE A VALEUR	CR_VAL	FALSE	01/01/1900	01/01/1900	FALSE
10	EMPLOYE STO	EM_STO	TRUE	01/01/1900	01/01/1900	TRUE
12	ET REGULIER 2000	ET_00R	FALSE	01/01/1900	01/01/1900	FALSE

3.1.1.3 Error Message Data

The fare validation table contains fields with numeric value representing the message displayed on the screen of an equipment after a fare validation. A value of 0 or no value mean the fare is validated while other values are associated with messages that explain why a fare validation has been refused. The complete list is shown in Table 3.3.

Table 3.3 Error message data.

Reject Code	Reject Message
0	Aucun message
20	Déjà validé
21	Carte refusée LN
22	Carte refusée LN
23	Carte invalide
25	Supplement exigé
26	Carte expirée
27	Titre invalide
28	Jour invalide
29	Ligne invalide
30	Periode invalide

3.1.1.4 Route Geometry Data

The route geometry data table (Table 3.4) contains the order of stops of each route, defined by a route number and a direction. The cumulative linearized distance from the departure terminus, which takes into account the street network geometry, is associated for each stop of a route. They are used to calculate indicators such as vehicle-kilometres, passenger-kilometres and on-board distance traveled. Stops can be used by more than one routes.

Table 3.4 Excerpt of route geometry data.

Beginning of Period	End of Period	Route Number	Direction	Stop Number	Stop Order	Location Code	Linear Distance (m)
06/11/2000	27/03/2005	1	0	5532	0	FP	0
06/11/2000	27/03/2005	1	0	5520	1	FP	3321
06/11/2000	27/03/2005	1	0	5516	2	S105	7728
06/11/2000	27/03/2005	1	0	5528	3	S105	13580
06/11/2000	27/03/2005	1	0	5512	4	S105	15915
06/11/2000	27/03/2005	1	0	5510	5	OLCH	18508
06/11/2000	27/03/2005	1	0	5511	6	OLCH	19138
06/11/2000	27/03/2005	1	0	5506	7	OLCH	20620
06/11/2000	27/03/2005	1	0	6090	8	DHUL	24351
06/11/2000	27/03/2005	1	0	5502	9	DHUL	24431

The route geometry data contain 169 route-directions and 7,337 route-stops. In addition, there are two dummy route-directions and four dummy route-stops. Variants of routes, including those serving educational institutions are coded using a number prefix. For example, variants of route 39 include routes 239, 439, 639, 739 and 839. The number of stops in a route varies from 3 to 102 and the length varies from 2,643 to 53,329 metres.

3.1.1.5 Bus Stop Inventory Data

Each physical stop located within the STO service area is georeferenced by X and Y in the MTM coordinates (Table 3.5). They can be joined with other data tables using the stop number as the primary key. Figure 3.3 shows the spatial distribution of all 1,869 bus stop panels in the transit network.

Table 3.5 Excerpt of bus stop inventory data.

Beginning of Period	End of Period	Stop Number	Zone Number	X in MTM	Y in MTM	Stop Label
06/11/2000	27/03/2005	1	1	0	0	DUMMY 1
06/11/2000	27/03/2005	2	1	0	0	DUMMY 2
06/11/2000	27/03/2005	1000	1	18433093	5028191	FRONT/CORMIER
06/11/2000	27/03/2005	1001	1	18432986	5028635	FRONT/DE LA TERRASSE-EARDLEY
06/11/2000	27/03/2005	1002	1	18433105	5028185	FRONT/CORMIER
06/11/2000	27/03/2005	1003	1	18432975	5028568	FRONT/DE LA TERRASSE-EARDLEY
06/11/2000	27/03/2005	1004	1	18433083	5028046	FRONT/PEARSON
06/11/2000	27/03/2005	1005	1	18433080	5028429	FRONT/DE LA TERRASSE-EARDLEY
06/11/2000	27/03/2005	1006	1	18433071	5028067	FRONT/PEARSON
06/11/2000	27/03/2005	1007	1	18433563	5028805	MCCONNELL-LARAMÉE/WILFRID LAVI

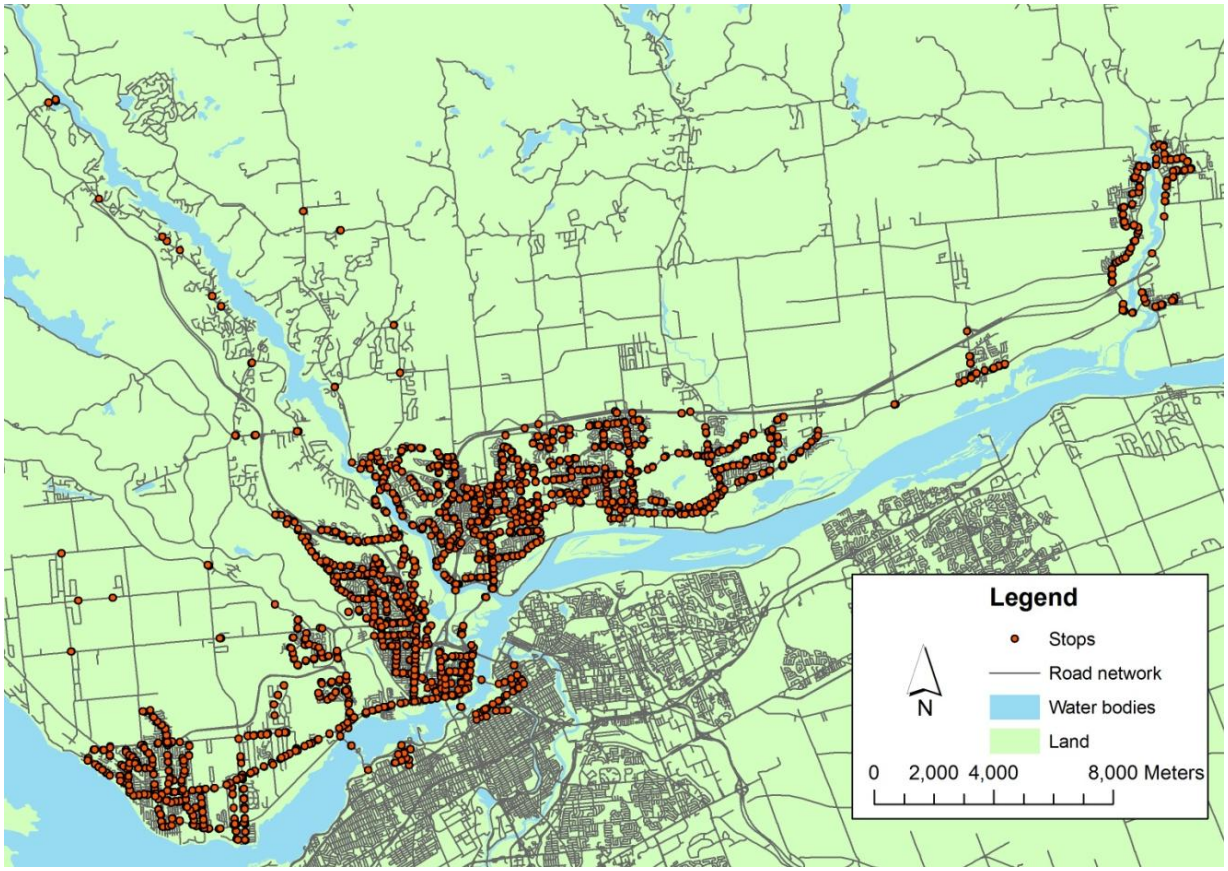


Figure 3.3 Bus stop locations of the STO network.

3.1.2 Data Derived from Validation Data

The boarding transaction data contain operations data. With a long enough time span, one can synthesize the complete scheduled service from the information recorded in boarding validations. The assumption is that the service is constant within the time span and that every route-direction offered has at least one validation associated with it. The synthesized information from multi-day data, labelled as “dictionaries”, serves as a reference in the validation process. They will be described in more details in Chapter 5.

3.1.3 Timetable Data from User Guide

Some of the data enrichment procedures and analyses require scheduled information not obtainable from the smart card AFC system. Those include checkpoints and theirs scheduled arrival/departure time. They are necessary for estimating the spatial-temporal path of vehicles and performing schedule adherence analysis.

3.1.4 Spatial Data from External Sources

Activities within the transit network need to be related to the physical built environment. Basic spatial data such as land, water bodies, city limit, borough boundaries and street network are acquired from external sources as GIS layers. A point of interest (POI) data table is compiled in October 2007 by accessing the atlas of the Ville de Gatineau website (Figure 3.4). The list of POI includes educational establishments, shopping centres, commercial and office buildings, community centers, etc. Their exact addresses are transformed into MTM coordinates in order to assemble a point of interest data table (Table 3.6) and a GIS point layer. This information is use to study the relationship between boardings and trip generators.

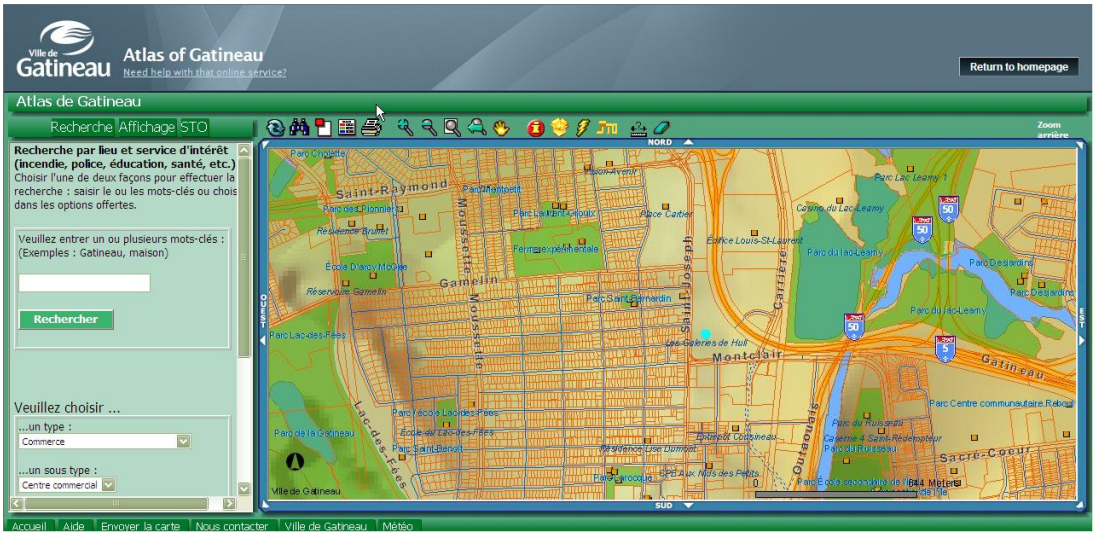


Figure 3.4 Atlas of Gatineau in the City of Gatineau website.

Table 3.6 Excerpt of point of interest data of Gatineau.

ID	Type of Trip Generator	Subtype	Name	Address	Lat	Long	X	Y
1	Récréatif	Aréna	Aréna Paul-et-Isabelle-Duchesnay	92 Rue du Patrimoine	45.39413	-75.84364	433966	5027080
2	Parc	Autre parc	Parc Frank-Robinson	Rue Frank-Robinson et Rue Principale	45.39516	-75.84095	434177	5027193
3	Communautaire	Centre communautaire	Centre Communautaire Frank-Robinson	96 Rue du Patrimoine	45.39416	-75.84309	434009	5027083
4	Éducation	École élémentaire	École des Trois-Portages	120 Rue Broad	45.40853	-75.84842	433609	5028685
5	Service Gouvernemental et Municipal	Édifice municipal	Centre de services Aylmer	115 Rue Principale	45.39489	-75.84501	433860	5027167
6	Éducation	Centre de formation	Relais de la Lièvre-Seignerie	584 Rue Maclaren Est	45.58799	-75.40393	468491	5048351
7	Culturel (objets)	Bibliothèque	Bibliothèque Lucy-Faris	115 Rue Principale	45.39489	-75.84501	433860	5027167
8	Éducation	École élémentaire	École Eardley	180 Rue North	45.40009	-75.84077	434197	5027740
9	Service Gouvernemental et Municipal	Incendie	Caserne 1 Roland-Guertin	425 Boulevard Wilfrid-Lavigne	45.39585	-75.83843	434376	5027268
10	Éducation	École secondaire	École Grande-Rivière	100 Rue Broad	45.40417	-75.8449	433879	5028198

3.1.5 Data from the OPUS Smart Card AFC System

In the last part of the research, data from the smart card AFC system of the region of Montréal are used to illustrate the limitations of the research and the potential of other types of data. The size of the multi-modal and multi-operator transit network is considerably larger and more complex than the STO network. The equipments involved are also more diverse. Three types of data, namely sales, validation and verification, are considered.

3.2 Making Sense of the Data

Passive smart card validation data solicit an understanding on their ontology and logic. The manner these data were generated influences the choice of conceptual and methodological approach.

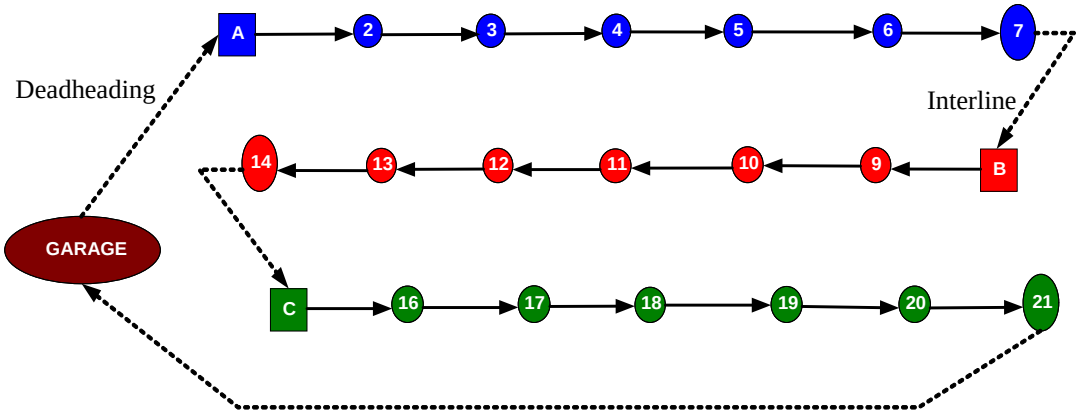
3.2.1 Understanding the Organization of Transit Service

The concept of average weekday is often used in fixed-route public transit planning in Canada and in North America in general. In the case of STO, service is repeated from Monday to Friday with some minor exceptions. With the concepts of transit planning in mind, an efficient relational database should include the following inter-related objects (here limited to fixed-route bus service, without loss of generality):

- A bus stop is the most elemental spatial object of a transit network. It represents a point of boarding, alighting or transfer for a transit user.
- A bus route is a numbered transit line with a geometry constituted of an ordered sequence of stops. Normally, two directions are defined: inbound and outbound. It can also be uni-directional or circular.
- A run implies the movement of a vehicle along a route in a specific direction with a scheduled departure time. Non-productive run, called deadheading, is an integral part of service planning although these vehicle-trips are not revealed to the public. Deadheading occurs when vehicles travel from and to the garage and during interlines. In theory, deadheading trip should not appear in smart card validation data as the vehicle is not in productive service.
- A vehicle block is defined as the vehicle operation between pull-out from the garage and the return to the garage. Each of them comprises a sequence of runs and is assigned to a specific bus and one or more drivers. From the operations planning point of view, the service of each weekday is divided into hundreds of vehicle blocks. The same vehicle blocks are repeated every weekday. Each day, different vehicles are assigned to perform the vehicle blocks.

Using validation data from 20 weekdays, the complete scheduled service of a weekday can be reconstructed by vehicle block and run. The ontology of each object in the database is examined

by simultaneously comparing the 20 days of data. An object must appear at least once in the data and should not be in logical conflict with other objects. Each validation must logically tie to a valid service run and a boarding stop. The concept of a typical vehicle block and the related information are illustrated in Figure 3.5. It explains the linkage between the spatial and temporal logics of these fundamental objects: a boarding validation represents a time-space coincidence between the vehicle and the smart card user.



Bus Route	Direction	Terminus	Departure Time	Boardings	Terminus Arrival Time
13	0	A	05:30	25	06:30
14	1	B	06:40	20	07:30
16	0	C	07:45	50	08:50

Figure 3.5 Illustration of a vehicle block composed of 3 consecutive runs (route-direction-departure time) and the related operations information (Chapleau & Chu, 2007).

3.2.2 Understanding the Sources of Error

In theory, the boardings of cardholders at different stops onto a vehicle should logically appear in the database in a correct monotonic time and stop sequence. However, a significant amount of validations turn out to be erroneous or suspect as they are not associated to a run or boarding location. It is consistent with the finding of Furth et al. (2006) regarding matching of AFC data.

The association of a vehicle block and a run with a validation record in the smart card AFC system requires driver's sign-in and initialization at the beginning of each run. According to the route, location of the vehicle is determined at the stop level in real time using X and Y coordinates from the GPS. When the smart card and the validator coincide at a certain point in time and in space, a validation record is created along with operations data, stop number and a time stamp. Inaccurate operations data, caused by erroneous sign-in and initialization, render the

run information and stop number invalid. They appear as misattributed runs and stops. Unfortunately, raw GPS coordinates are not conserved. Even if they are available, correctly matching them to the validations remains a difficult task in practice (Furth et al., 2006). Equipment failure and poor GPS signal also cause errors in validation data.

3.3 Guiding Research Principles

An understanding of the datasets leads to the following guiding research principles upon which all methodologies and analyses are based:

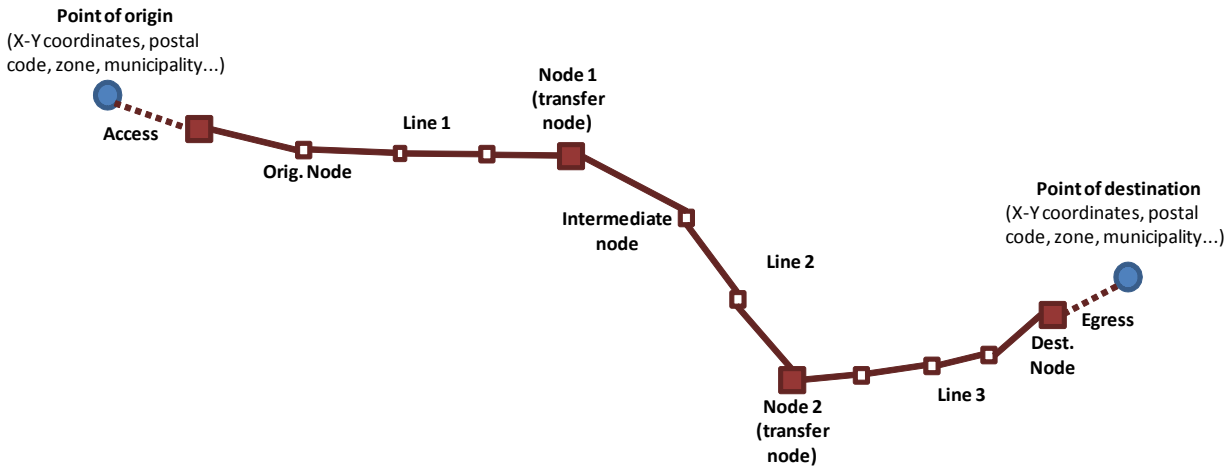
- The information (data-driven) approach;
- The totally disaggregated approach;
- The object-oriented approach;
- The concept of bootstrapping;
- The concept of entropy maximization.

3.3.1 Generalized Approaches

The information (data-driven) approach insists on the use of information based on actual data. As opposed to trip generation and synthetic data models which are based on algebraic formulations, the datasets used in the research are exclusively observed data. As a continuous passive data source, the smart card AFC system is an ideal candidate for this approach. As a result, methodologies are geared towards the extraction of information from the data.

The totally disaggregated approach emphasizes the preservation of traveler and trip components relating to each individual trip (Chapleau, 1992). Developed in the early 1980s, the approach focuses on the processing and enrichment of origin-destination trip files from large-scale regional household travel surveys. It is called “disaggregate” because every trip can be individually scrutinized with respect to all other available variables, and “totally” because the same treatment is simultaneously applied to all transportation objects related to a transit trip (Chapleau, 2003): departure time, origin and destination as X, Y coordinates, access and egress, routes and nodes (Figure 3.6). On the other hand, trip maker and household characteristics are tied to each trip, allowing in-depth multivariate analyses. Smart card data are particularly suited for this approach

as the data come in a disaggregated form. All the data processing methods proposed in the research are applied to individual validation record.



Point of origin	Point of destination	Flow	Time of day; Purpose	Origin node	Destination node	Line 1 to N	Node 1 to N	Access time	Waiting time	In-vehicle time	Age	Sex	Index
Spatialized; Accessibility; Aggregation		Sampling; Weighting; Forecasting	Trip attributs	Access and egress to the transit network		Transit itinerary		Transit assignment			Socio-demographic attributs		Various applications

Figure 3.6 An individual trip defined with the totally disaggregated approach (based on Chapleau, 1992).

The object-oriented approach allows each object to be analyzed in detail and put an emphasis on the interrelationship among them (Trépanier & Chapleau, 2001). The ability to model and relate all the objects in a transit system is essential to establish methodologies for data processing and analyses. The schematic definition of objects related to smart card validation data of the STO along with their relationships are shown in Figure 3.7. The illustration draws a parallel between the objects of the transit system and objects from a traditional transport survey (Chu, Chapleau & Trépanier, 2009). Objects with red outlines represent the interviewer of a survey in a moving vehicle. The reference unit of the interviewer is the vehicle block. On the other hand, each cardholder who boards a vehicle is interviewed at a specific time and stop. The interview is recorded as a boarding validation and is represented by objects in yellow boxes. Its reference unit is a smart card validation. Objects in green boxes are elements not captured by the system but can be inferred with data enrichment processes. For example, alighting stop and activity location are not captured directly by the system.

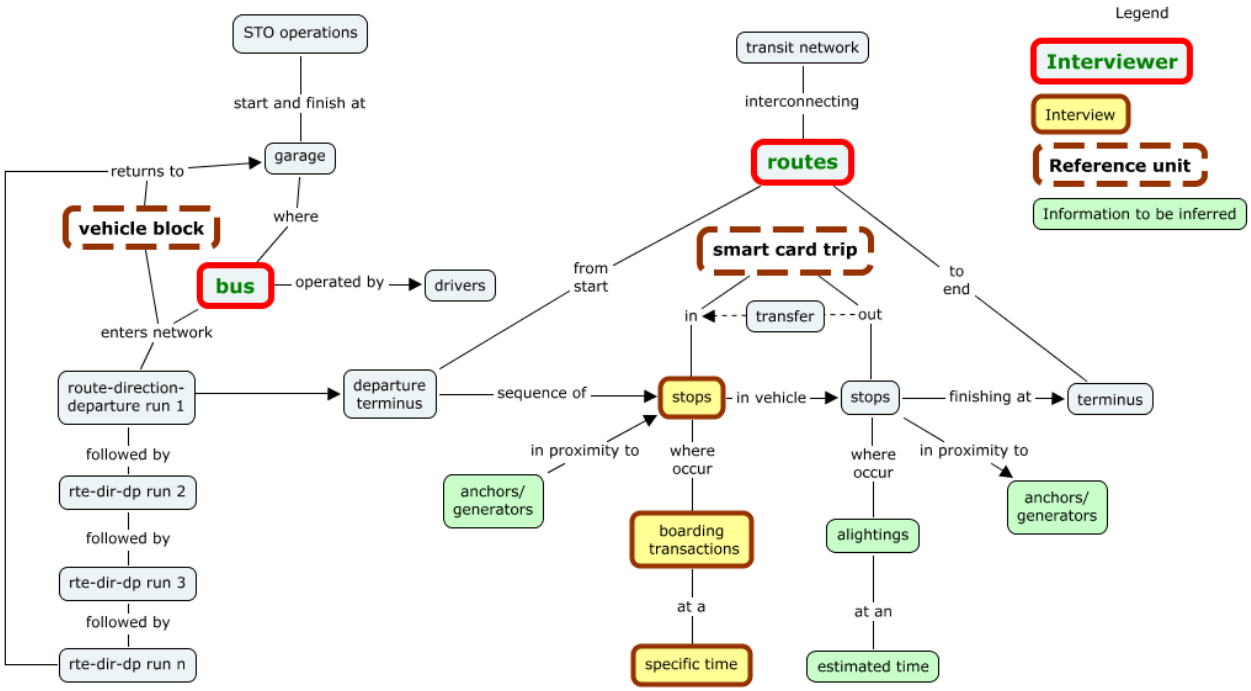


Figure 3.7 Schematic definitions and relations of various objects (Chu et al., 2009).

3.3.2 Concepts Used in Data Processing

The proposed data processing methodology is inspired by the principle of maximum entropy – assigning the most probable value to an unknown variable under the best assumptions. This concept is widely used in this research; for example: in the validation process where values are imputed to replenish data; in data enrichment processes where alighting stops for each boarding, the movement of the vehicles and the trip ends are inferred as well as in the interpretation of trip purpose.

Data processing is also inspired by the concept of bootstrapping – initial enrichment that subsequently contributes to further enrichment. For example, in order to infer trip purpose from a boarding validation, the data need to go through subsequent steps of enrichment, each contributing to new additional information.

3.3.3 Multi-day Information Approach

The guiding research principles lead to the multi-day information approach, illustrated in Figure 3.8, which is used in several processes in the proposed data processing methodology. The assumption is that each day of data represent partial information of the universe and may contain

errors. By synthesizing the correct information from each day, new objects are generated. These objects are new knowledge that is complete and clean. The completeness of the knowledge depends on the number of days in the analysis timeframe. The experiment in this research uses data from one month that includes 20 weekdays. Nothing prohibits the use of a longer timeframe. However, a short timeframe may run the risk of not containing enough information to synthesize useful knowledge. The new knowledge is in turn used as a reference to analyze and interpret the multi-day data. A cross sectional study of a multi-day dataset can be understood as looking at each day of data with respect to all the data in the analysis timeframe.

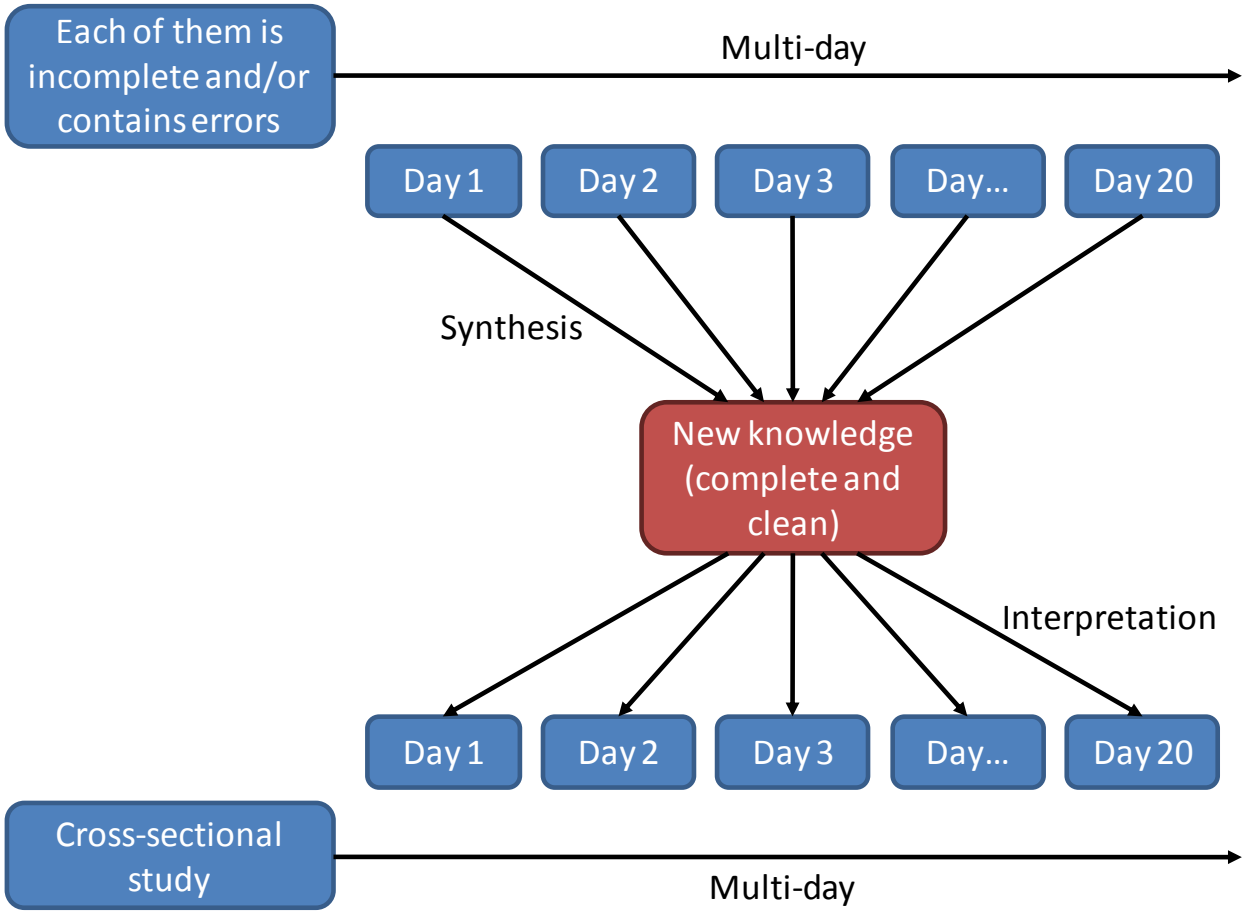


Figure 3.8 Schematic representation of the multi-day information approach.

This approach is used in two data processing steps in this research. The first one is the data validation process, which will be explained in Chapter 5. The transit service of a typical weekday is synthesized from 20 weekdays and is used as a reference dictionary for data correction. The second one is the anchor detection procedure for individual cardholder. Boarding records from 20

days are synthesized to derive a list of anchors for every cardholder. Each transaction is then analyzed with respect to the list. The methodology will be explained in more detail in Chapter 8.

CHAPTER 4 SMART CARD FARE VALIDATION DATA AS A TRAVEL SURVEY

This chapter lays the theoretical foundation of considering the smart card AFC system as a legitimate passive data collect or travel survey method. It is compared to traditional surveys in terms of data properties and quality. It also argues its potentiality as a versatile multi-purpose travel survey.

4.1 Travel Survey Type and Data Quality

Data collection represents an integral part of transit planning and travel behaviour studies. Traditional data collections are mainly done through travel surveys, which are conducted in many formats. They vary in contents and quality and can be broadly divided according to the following overlapping characteristics:

- Stated preference vs revealed preference;
- Survey method;
- Household-based vs non household-based;
- Travel mode coverage: single-mode vs multi-modal;
- Timing and duration.

4.1.1 Stated Preference vs Revealed Preference

A stated preference survey collects information on respondents' intention. They are often used in hypothetical situations, for example a fare increase, where observation on real behaviour is not possible. A revealed preference survey gathers information from activities already performed by the interviewee. It can be done through interview or observation.

4.1.2 Survey Method

Survey method describes the interface between the interviewer and interviewee. The following enumerates and briefly describes the prevalent survey methods for travel data:

- Postal survey is self-administered and generally involves sending a questionnaire to the respondent's home and asking the individual to complete and return it to the sender (Bonnel, 2003).
- "Telephone survey is conducted over the telephone, usually when the respondents are at home" (Bonnel, 2003). The interviewer administers and collects the data, often with the real time assistance of a computer software, called computer-assisted telephone interviewing (CATI).
- "Face-to-face survey is administered by an interviewer who interrogates one or more individuals in person" (Bonnel, 2003).
- Internet survey is self-administered and based on an internet questionnaire. It can be a stand-alone method or a follow-up method aiming to supplement or validate data collected by another method (Bohte & Maat, 2009).
- GPS-only (or other location-aware device, LAD) survey provides only time and position records. All other information must be determined through supplementary data collection (Stopher, 2004) or data processing and/or enrichment (Wolf, 2006).
- Survey using public transit fare validation records is a passive data collection method that requires post-processing of the data. It gathers aggregate or disaggregate data.
- On-board survey is administered by ride-checkers on board transit vehicles.
- On-site/trip generator survey is administered at specific trip generators or at locations where there is a concentration of trip makers such as an inter-modal transfer or a major boarding/alighting point.

4.1.3 Household-based vs Non Household-based

Household-based survey uses household as a basic unit of survey. It collects socio-economic characteristics of the household, captures all trip details and the travel interaction from all the members from the same household. In addition, it is compatible with the basic unit of census data. In contrast, the majority of non-household surveys present only a partial picture of travel because they relate to one particular mode or one specific generator. They are usually undertaken for a

particular purpose and are not intended to replicate the total picture of travel patterns provided by a household travel survey (Wofinden, 2003).

4.1.4 Travel Mode Coverage

Travel mode coverage is a concept closely related to survey method because there are constraints or biases on the travel modes that can be surveyed by each method. For example, private vehicle-based GPS survey is limited to trips made by automobile; wearable GPS collects details from trips made by all travel modes but requires data enrichment techniques to identify the modes used in each segment if no supplementary data are gathered (Chung & Shalaby, 2005; Tsui & Shalaby, 2006). Fare validation records and on-board surveys are only applicable on trip segments made by public transit.

4.1.5 Timing and Duration

Alternatively, surveys can be categorized according to the temporal dimensions of sampling, namely the duration (how long does it last?) and timing (when is it taken?). A multi-day survey is conducted with the same person, household or vehicle over multiple days in a continuous period (Purvis & Ruiz, 2003). It is often motivated by an attempt to capture most of the variance in travel behaviour within the survey period (Madre, 2003). A multi-period survey can be subdivided into two classes: repeated cross-sectional surveys, the same survey taken at intervals in time applied to different respondents, and panel surveys which consist of the same survey taken at intervals in time applied to the same respondents (Purvis & Ruiz, 2003).

Cross-sectional survey repeated daily is called a continuous survey by Purvis & Ruiz (2003). However, it may be more appropriate to be called an “on-going survey” because although the survey takes place everyday, it is administered to different respondents. A true continuous survey should be a panel survey repeated everyday or perpetual multi-day survey. In this setting, every respondent is continually surveyed.

4.2 Drawbacks of Traditional Survey Methods

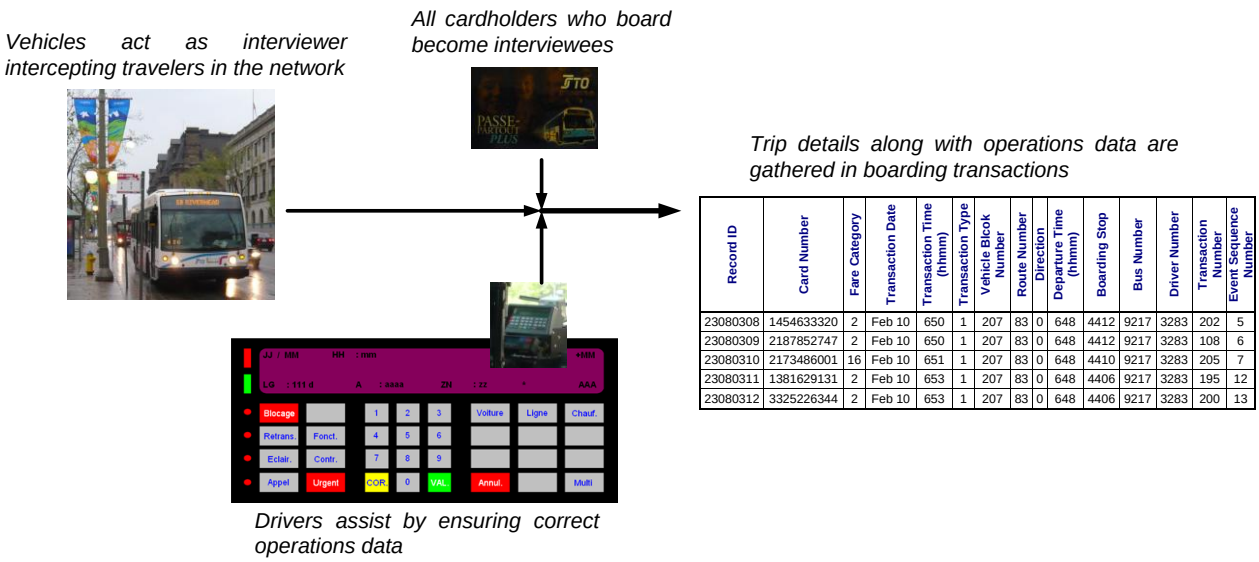
In the commonly-used household travel survey, households are sampled across an urban area and travel data are collected from trips made by each member of a household on a given day. It captures trips made by all modes of travel and is statistically expanded to represent the entire

population. The collected data are suitable for analyzing behavioural trends in a metropolitan-wide context but pose several drawbacks with respect to operations planning:

- Incompatible time-frame: household surveys are expensive to perform and are therefore carried out only once every few years. Meanwhile, a transit agency needs to adjust its service schedule several times a year.
- Insufficient sample: in places where the automobile is the prevalent mode of travel, transit trips may not be well-represented in a multi-modal travel survey. The issue is especially marked in suburban areas where the transit share is low.
- Inadequate data resolution: transit planning requires temporally and spatially detailed data that a household survey is unable to provide due to limitations on methodology and respondents. Most respondents "are not qualified geographers and do not understand or relate to the surrounding geography in a way that permits them to report it in a precise enough manner for the purposes of planning" (Stopher, Bullock & Horst, 2002).

4.3 Uniqueness of Smart Card Data

While specialized surveys address some of these concerns, they remain labour-intensive and expensive to perform. The recent adoption of smart card AFC system and AVL system by transit agencies naturally leads to a new concept in travel data collection: the Driver-Assisted Bus Interview (Figure 4.1), DABI (as opposed to CATI).



Record ID	Card Number	Fare Category	Transaction Date	Transaction Time (hhmm)	Transaction Type	Vehicle Block Number	Route Number	Direction	Departure Time (hhmm)	Boarding Stop	Bus Number	Driver Number	Transaction Number	Event Sequence Number
23080308	1454633320	2	Feb 10	650	1	207	83	0	648	4412	9217	3283	202	5
23080309	2187852747	2	Feb 10	650	1	207	83	0	648	4412	9217	3283	108	6
23080310	2173486001	16	Feb 10	651	1	207	83	0	648	4410	9217	3283	205	7
23080311	1381629131	2	Feb 10	653	1	207	83	0	648	4406	9217	3283	195	12
23080312	3325226344	2	Feb 10	653	1	207	83	0	648	4406	9217	3283	200	13

Figure 4.1 The concept of Driver-Assister Bus Interview (Chu et al., 2009).

The smart card has been acknowledged as a new and relevant technology for non-household survey (Wofinden, 2003). There are unique opportunities that a smart card AFC system can offer to transit planning:

- Continuous coverage: in theory, the smart card validations correspond to complete information over time for smart card user segment, as opposed to a small sample, allowing multi-day analysis on variability of transit use. They also cover the complete network and service.
- Spatial and temporal resolution: smart card system provides precise time stamp of the validation. With AVL, it adds reliable spatial information.
- Totally disaggregated information by cardholder: fare validations are individually recorded along with card, operations and trip details, thus offering the possibility to group validations by card and study activity pattern and travel behaviour at the card level.
- Passive data collection strategy: the data are automatically generated as opposed to self-declared or self-administered.
- Timeliness and extended timeframe: data collection is on-going and continuous. Since the data contains information of a cardholder over a long period of time, they are analogous to a panel survey.

- Integration with operations data: the systems links boarding validation with individual run, which are essential for detailed service analyses and indicators.

4.4 Smart Card Data Quality Issues

The type of survey and the data collection method undoubtedly affect the accuracy and the validity of the data. Researchers have compared the results from different practices and identified quality issues associated with each one. Several aspects of DABI relate to prominent themes on data quality and are compared to traditional survey methods.

4.4.1 On Respondents

A smart card validation record is created when a smart card is placed within the range of a card reader. The process is passive and automatically records the time and location of the validation along with other operations data. It is non-intrusive since the cardholder does not need to carry any special device other than the card. It eliminates response burden of the interviewee, which is often cited as a problem in multi-day surveys (Axhausen, Zimmersman, Schönfelder, Rindfüser & Haupt, 2007), as the cardholder is not required to perform any task other than maintaining the fare validation routine. As a result, the respondent should not experience fatigue in this continuous survey. The respondent does not need to recall or record any information for himself/herself and for other members of the same household, which would be the case in a traditional household survey. Moreover, the process does not affect the respondent's travel behaviour in any way.

4.4.2 On Response Rate

DABI survey does not encounter the issue of non response. In large metropolitan area, refusal to participate in telephone OD survey is a major problem, especially if it is time-consuming. In London, non response rates exceed 50% in several household surveys (Wofinden, 2003).

4.4.3 On Coding

Coding a transit trip in a travel survey involves the recalling and the explicit declaration of trip details including departure time, routes taken, transfers, trip origin and destination by the respondent as well as the transcription by the interviewer. In the regional travel survey carried out

in Montréal, Canada, the coding of a transit trip requires almost 2 minutes using CATI (Chapleau, 2003). Also, it is often observed that respondents round their departure time to the nearest 5 or even 15 minutes. Route taken information provided by the interviewee can be incompatible with the origin and destination that he/she declares. Although there are mechanisms in the CATI software to validate or correct the data, many records remain uncertain.

Meanwhile, the coding of a transit trip by DABI is almost completely automatic and involves no recalling. Therefore, there should be no error in recalling and transcribing. Automatic coding also avoids confusion over the definition of various transport objects such as transfer trips, which are concepts that the interviewee does not necessarily understand. The smart card AFC system thus ensures a systematic and uniform coding that is independent of the interviewer and the interviewee.

4.4.4 On Sampling

Both cross-sectional travel survey and traditional multi-day survey have drawbacks. A cross-sectional travel survey does not allow the observation of day-to-day variability in travel behaviour at the person level. It cannot take into account certain macro aspects such as economic cycles and seasonality, and micro aspects such as the weather or holidays. On the other hand, a traditional multi-day travel survey requires significantly more resources for the same sample size and inevitably causes respondent fatigue and attrition.

In order to measure transit service consumption and to monitor behavioural change in the population, a large sample with the same respondents over a continuous time period is desired. A smart card AFC system with a high market-penetration rate does have such properties. It can effectively capture transit service consumption from the majority of transit users and its evolution over time. The sample includes the whole universe of smart card users. By contrast, traditional household surveys sample a tiny portion of the population over a period of time and statistically expand the trips to represent an average weekday whereas the smart card data are reduced to an average weekday to accommodate the current practice in service planning. In the case of Montréal, a total of about 5% of all households are sampled over a three-month period in the regional travel survey. This represents a daily sampling rate of less than 0.1% repeated over three months. Taking into account that it is a multi-modal survey, the number of transit trips is very small compared to the number collected by a smart card AFC system. It is worth nothing that the

universe of smart card holders tends to have a bias towards over-representing frequent users since they are the most likely adopters of this payment method given the cost of the medium. However, it can be argued service should be geared towards those frequent users.

A personalized smart card, usually required for reduced-fare users, offers the opportunity to study historic travel patterns of a cardholder longer than any other fare payment option since the useful life of a smart card lasts several years under normal circumstances compared to days or months for magnetic fare card. In case of loss, card malfunction or card change due to the expiration of fare privilege, there is also opportunity to link the new card to a cancelled card using cardholders' registration records since some transit agencies require registration, or encourage voluntary registration by providing loss-protection service. However, it may require the creation of an additional database to keep track of those records.

4.4.5 On Survey Universe

In theory, the survey universe of a smart card AFC system includes all transit trips made by smart card holders. However, it is not necessarily the case in practice. Missing information can be caused by technical issue. It occurs when the validation equipment is out of service or when fare payment procedure is not enforced. In this case, the transaction record does not exist even if the cardholder actually boards the vehicle. This is similar to a trip that is not reported by the interviewee. Fare paid by other media such as cash or disposable ticket can be registered by the system but they do not appear in a smart card transaction history. Regular card users who forget to bring the card and use an alternative method of payment may cause biases in travel behaviour study or fare product use analysis.

In a household travel survey, the socio-economic aspect of the household, such as number of persons, level of income and number of automobiles, is surveyed. Data collected by passive source are incomplete compared to what can be gathered in a household travel survey. Trip purpose, trip ends and trips that are not trips made by public transit are not gathered. As such, the automobile segment of a bimodal trip is not known. Since the survey is based on the card (individual) level, trips made by other person in the household are not included in the survey universe.

4.5 A Multi-use Survey

Smart card validation data integrate operations data, possess desirable properties on data quality and hold many benefits over traditional surveys. Their potential uses are diverse and versatile. They can be used as a:

- Transit demand survey, which reveals the detailed spatial-temporal distribution of transit demand down to a card level.
- Continuous survey (multi-day and multi-period at the same time) which allows analyses on the variation in transit demand, performance and travel behaviour.
- Trip generator survey, which allows analyses of transit users associated with a common origin or destination.
- Resource allocation and consumption survey, which allow the calculation of indicators on transit supply, demand and level of service.
- Revealed preference survey, which records travel behaviour regarding route choice and departure time.

In the subsequent chapters, post-processing methodologies are proposed to address the shortcomings of a passive survey and to extract useful information from the data. Analytical techniques will be applied to illustrate the potential of the data on transit planning and travel behaviour study, and to demonstrate the relevance of the proposed methodologies.

CHAPTER 5 EXPLORATORY DATA ANALYSIS AND DATA VALIDATION STRATEGY

Whether the data are collected through active or passive methods, the accuracy and validity of the data are paramount in an information system. Erroneous or biased data affect the validity of analysis results. A data validation strategy in this research aims to detect erroneous or implausible values and to replace them with plausible values. A good understanding of the data collection or generation process, presented in the previous chapters, is essential to formulate an error-detection and imputation algorithm. In this chapter, an exploratory analysis is performed and a data validation strategy is proposed.

5.1 Exploratory Data Analysis

In this section, some basic statistics are compiled for various objects found in the fare validation data. Table 5.1 provides statistics for 4 analysis timeframes: a typical weekday (February 10), weekend days (Saturdays and Sundays), weekdays (Monday to Friday) and the whole month (28 days). Figure 5.1 gives a visual comparison of the number of validations by fare type. February 10, representing a typical weekday, has the highest number of validations in the month, 38,502, whereas the average number per weekday is 35,664, which is 7.4% lower than February 10. The average is affected by several weekdays, most noticeably on the 4^{th}, 7^{th} and 28^{th}, that have fewer validations than their counterparts. The number of transactions for the 4 weekends is relatively stable. There are 21,813 cards that have at least one validation during the month. All except 71 cards have been used on weekdays. In contrast, only 36% of the cards have been used on weekends.

The ratio between first boarding and transfer boarding on weekends (28.2%) is significantly higher than the ratio on weekdays (17.8%). This may be explained by the different travel patterns, such as trip purpose and destination choice, and by the level of service of the transit network. Service consumption indicators, such as the number of unlinked trips per card per day, can be calculated.

Table 5.1 Statistics on various objects. (Potentially sensitive statistics are greyed out.)

Indicators	Monthly Count	Weekday Count	Weekend Count	Thursday Feb 10
Duration (number of days)	28	20	8	1
Number of transactions	763,570	713,276	50,294	38,502
Average number or transactions per day	27,270	35,664	6,287	38,502
Number of distinct cards				
Average number of active cards per day				
Number of distinct vehicles	219	219	97	202
Number of distinct drivers	319	315	137	222
Number of distinct vehicle blocks	525	412	113	391
Number of distinct routes (direction)	175	174	54	163
Number of stops used	1,526	1,505	1,005	1,162
Number of distinct runs	34,458	30,185	4,273	1,536
Transaction type 1 (first boardings)	644,614	605,373	39,241	32,773
Transaction type 3 (transfer boardings)	118,956	107,903	11,053	5,729
Transfer to first boarding ratio	18.5%	17.8%	28.2%	17.5%
Distinct fare types	20	20	17	19
Average first boarding per card				
Average transfer boarding per card				
Average total boarding per card				
Average first boarding per active card per day				
Average transfer boarding per active card per day				
Average total boarding per active card per day				

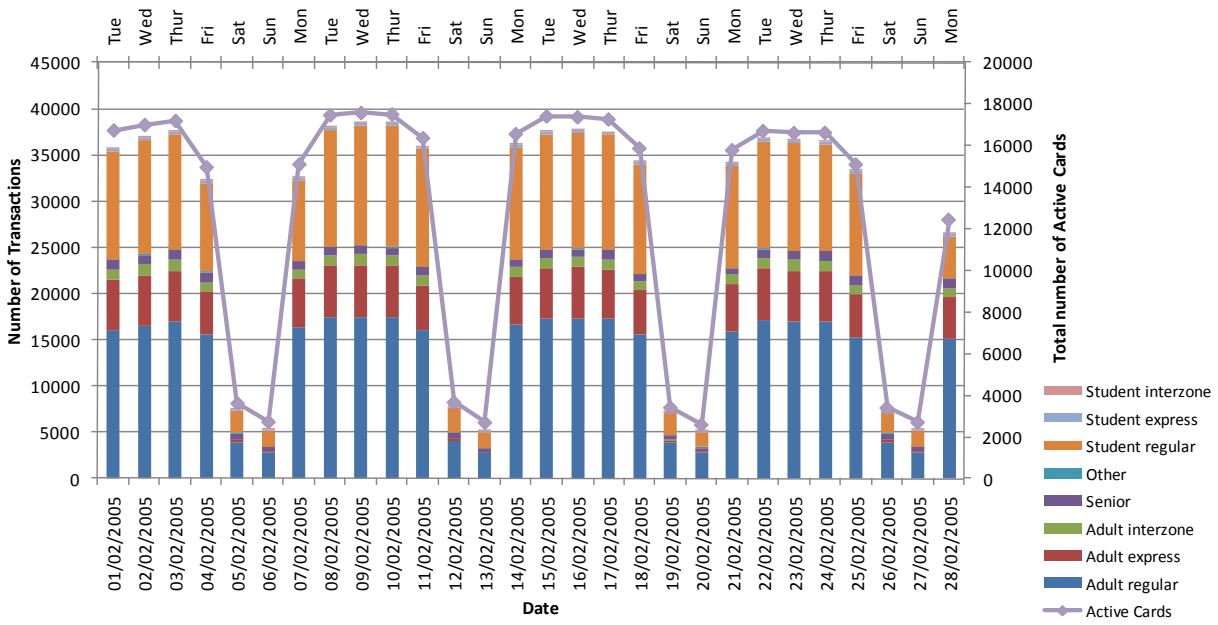


Figure 5.1 Number of transactions by fare type.

There are 20 distinct fare types used during the month and are shown in Table 5.2. Fare type 1 (Adult regular) accounts 41.7% of total transactions. Several fare types (such as Employee STO,

Étudiant intégré 2003, Étudiant régulier 2003 and Archive) have very few transactions per card. The number of active cards accounted by fare type (21,857) is slightly higher than the number of active cards (21,813) because several cards are associated with more than one fare type during the month. In subsequent analyses, those cards bear the dominant fare type according to their transaction history.

Table 5.2 Monthly statistics on various fare types. (Potentially sensitive statistics are greyed out.)

Fare Type	Fare Type Label	Aggregated Fare Type	Type 1 transactions (First boardings)	Type 3 transactions (Transfers)	Total Transactions	Active cards	First boardings per card
1	ADULTE REGULIER	Adult regular	270,424	47,935	318,359		
2	ADULTE EXPRESS	Adult express	77,321	4,187	81,508		
3	ADULTE INTERZONE	Adult interzone	16,644	1,028	17,672		
8	AINE	Senior	15,164	6,568	21,732		
10	EMPLOYE STO	Other	794	158	952		
15	CF REGULIER	Adult regular	35,716	4,220	39,936		
16	CF EXPRESS	Adult express	23,729	1,021	24,750		
17	CF INTERZONE	Adult interzone	4,322	166	4,488		
18	CF AINE	Senior	899	305	1,204		
30	ETUDIANT INTEGRE 2003	Student regular	14		14		
31	ETUDIANT REGULIER 2003	Student regular	21	1	22		
34	CAMPUS REGULIER	Student regular	21,682	4,536	26,218		
35	CAMPUS EXPRESS	Student express	1,373	239	1,612		
36	CAMPUS INTERZONE	Student interzone	124	13	137		
37	ETUDIANT INTEGRE 2004	Student regular	29,742	6,785	36,527		
38	ETUDIANT REGULIER 2004	Student regular	98,330	32,638	130,968		
39	ETUDIANT EXPRESS 2004	Student regular	3,578	1,152	4,730		
40	ETUDIANT INTERZONE 2004	Student interzone	506	236	742		
41	INTEGRE SCOLAIRE 2004	Student regular	44,118	7,759	51,877		
42	ARCHIVE	Other	113	9	122		
		Total	644,614	118,956	763,570		

In order to provide a more meaningful perspective, fare types are aggregated into 8 groups shown in Table 5.3. The number of transactions per card is different according to fare type. Cardholders with regular adult fare on average make 2 more trips than students and 9 more trips than seniors. STO employees only account for 0.1% of total first boardings. The distribution of cards according to the number of first boarding transactions made in February 2005 can be computed. The number ranges from 1 to 82 first boardings. In terms of the transit network and service, there are 525 vehicle blocks, 175 route-directions, 34,458 runs, 1,526 stops, 219 vehicles and 319 drivers that appear at least once in the transaction data table.

Table 5.3 Monthly statistics on aggregated fare types. (Potentially sensitive statistics are greyed out.)

Aggregated Fare Type	Number of first boardings	Number of transfer boarding	Number of total boardings	Number of cards	Number of first boardings per card	Transfer to first boarding ratio
Adult regular	306,140	52,155	358,295			0.17
Adult express	101,050	5,208	106,258			0.05
Adult interzone	20,966	1,194	22,160			0.06
Student regular	193,907	51,719	245,626			0.27
Student express	4,951	1,391	6,342			0.28
Student interzone	630	249	879			0.40
Senior	16,063	6,873	22,936			0.43
Other	907	167	1,074			0.18
Grand Total	644,614	118,956	763,570			0.18

The exploratory data analysis demonstrates that even at an aggregated level, unprocessed smart card validation data can provide simple indicators on trip frequency and transfer boarding rate. The benefits of a continuous source are exemplified by the monitoring of the daily variation of demand by market segment.

5.2 Data Validation Strategy

Other than providing simple indicators, exploratory analysis also sheds light on the quality of the data. As discussed in previous chapters, data accuracy is a persistent problem in AFC. A significant amount of records contain values that do not respect spatial-temporal constraints and public transit concepts. The goal of a data validation process is to improve data quality by making the data internally coherent. A comprehensive validation strategy should therefore account for the following issues:

- The procedure should quantify the extent of the errors by identifying problematic values.
- The procedure needs to be performed regularly. The complete validation process may not need to be applied continuously to all the data. A set of coherent data can be used as a reference to detect and to correct other sets of data.
- The procedure should be based on internal data or other passive data of equal quality. It is illogical to rely on external data that are less reliable and have smaller spatial and temporal coverage.

- The procedure should not discard data and should replace the problematic data with plausible values in order to conserve the continuity of the boarding history of individuals.
- The procedure should be automated to handle the large amount of data.

The strategy takes into account validation records from 20 weekdays. Erroneous values are first detected by logical rules. The validated records from each day are then used to recreate a dictionary of scheduled service in a typical weekday. Erroneous values from each day are then replaced by a plausible value according to this reference.

5.2.1 Rationale of the Data Validation Process

With the amount of data generated by the system, it may seem reasonable and harmless to remove erroneous and implausible data since data are plentiful and one may argue on the need of a data validation process. It would be much simpler to filter out unusable records given that in the long run, aggregate data can still provide adequate indicators. However, when studying the variation in demand, it is essential to assure that the variation in the number of transactions, regardless of the aggregation level, indeed comes from a real change in demand and not from errors in data or data omission. The same applies to travel behaviour analysis. Indicators on persistency or change in travel behaviour can be affected.

Several proposed methodologies and analyses take advantage of the continuity of the travel history of a card. The validation process should safeguard this distinct property. Otherwise, two potential drawbacks would arise and are illustrated below.

5.2.1.1 Spatial-temporal Discontinuity

Since errors typically occur in a sequence, a block of boarding records with mis-assigned run or stop can potentially affect many cardholders. Using or discarding an erroneous record would create a spatial-temporal discontinuity in the boarding history. In contrast, correcting the information would re-establish the spatial-temporal continuity which allows subsequent data enrichment.

5.2.1.2 Propagation and Amplification of Error

Data enrichment techniques, such as alighting stop estimation of a boarding record, require information from the subsequent record in an individual’s boarding history. Therefore, each

boarding validation has a dual function: it describes the boarding itself and at the same time holds information for enriching other boarding records. Using boarding validations with erroneous route and boarding stop not only leads to false conclusion in analyses, the error would propagate into other records in data enrichment process. If the erroneous boarding validations are discarded, information would be lost not only for the boarding itself, but also for the records which depend on it for data enrichment. In this case, the lost of information is amplified.

The following example illustrates the lost of information and the propagation of error. Figure 5.2 shows the three dimensional spatial-temporal path of a cardholder with 4 boarding transactions on a typical weekday. The bolded lines show the spatial-temporal movement of the cardholder before the validation process. The labels beside the dots indicate the known boarding stops and the alighting stops estimated with an enrichment technique. The geometry of routes 60 and 76 and the street network are shown as reference. The accompanying table provides details of each boarding validation. The route taken and the boarding stop are flagged as erroneous in trip 2 since the record is not assigned to a valid service run and boarding stop. The implication of this loss of information is not limited to the boarding of trip 2. Data enrichment processes that use information from trip 2, such as estimating the alighting stop for trip 1 and trip 2, are no longer possible. This example shows that the removal of a validation record, or an erroneous value left unchanged, not only affect the boarding itself. It also has repercussion on records that are dependent on it in an enrichment process.

Given these reasons, data correction seems to be the logical and justified step in the data validation process. Erroneous values are removed and value replenishment is done by imputation while simultaneously considering as much information as possible. Two data fields are to be corrected, the run and the boarding stop, because of their importance in demand and travel behaviour analyses. The imputation technique is based on two concepts: the repetition of transit service for an average weekday and the historic travel pattern of individual cardholders. The former is used to assign a run to a boarding validation. The latter is primarily used to assign a boarding stop to a validation, but is also used to assign a run if the first concept fails.

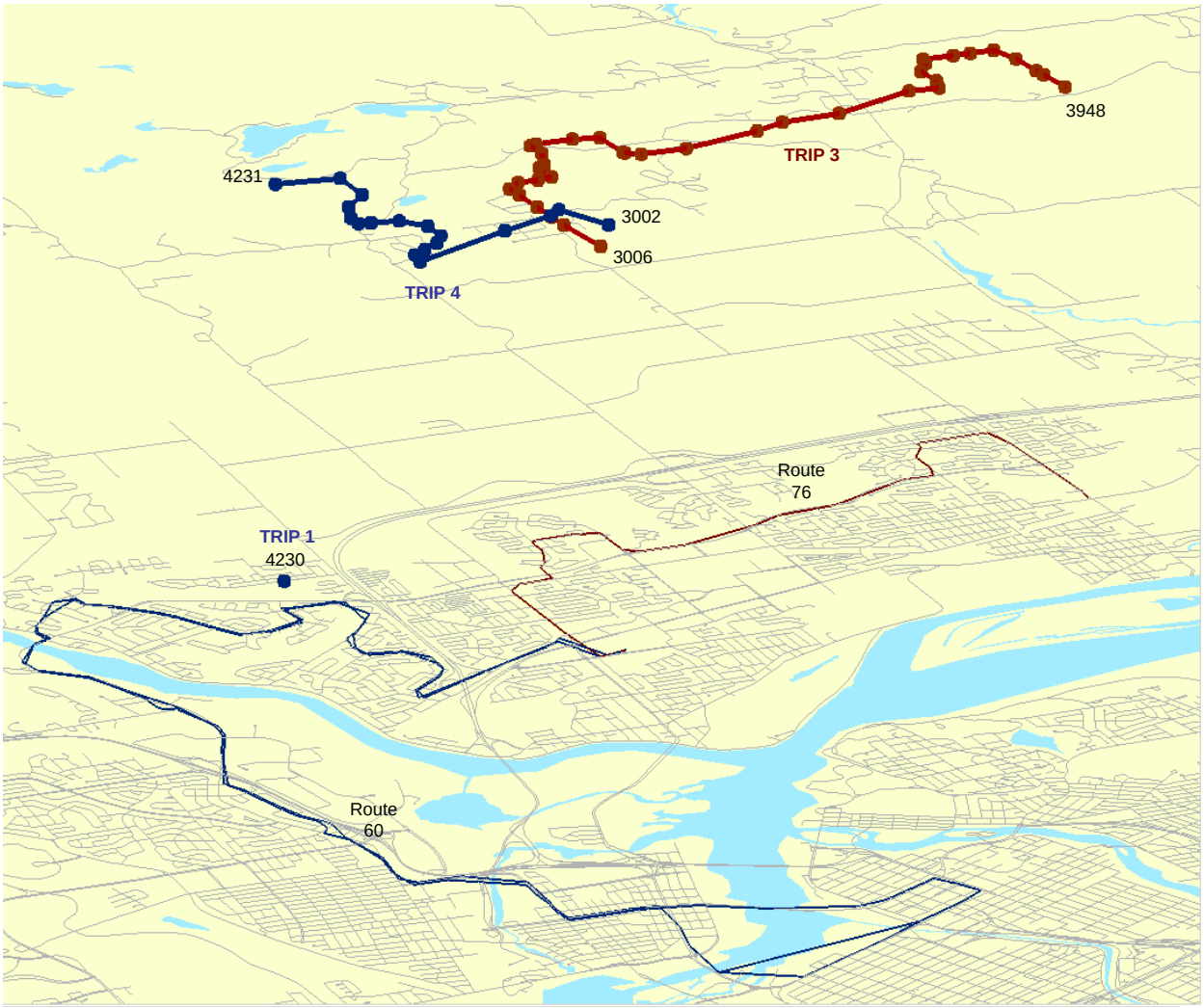
The following sections describe and demonstrate a validation strategy that first detects erroneous values by logics and produces a dictionary of scheduled service. The data imputation procedure

then replenishes data using the two aforementioned concepts. The goal is to obtain perfectly-coherent data.

Several assumptions need to be made in the data validation process. The primary assumption is that only three fields can be considered absolute, consistent and free from logical error, regardless of drivers' input. They are used as reference in the error detection and correction procedure:

- Vehicle number (or unique ID of the validation equipment associated with a vehicle or station): in bus transit, the card reader is quasi-permanently tied to a specific vehicle.
- Card number: the unique identification number of a card which is associated with a specific fare type. The card number represents one specific person for a personalized card.
- Transaction Time: system time is automatically registered when the transaction occurs.

The error detection and imputation processes both use these three objects as the reference.



Trip	Record ID	Transaction Date	Transaction Time	Route Number	Direction	Scheduled Departure Time	Boarding Stop	Derived Alighting Stop
1	23091592	10-Feb-05	06:44	60	1	06:01	4230	Unknown
2	23092894	10-Feb-05	07:02	900	0	06:38	1	Unknown
3	23119896	10-Feb-05	14:38	76	0	14:35	3948	3006
4	23123387	10-Feb-05	15:03	90	0	15:05	3002	4231

Figure 5.2 A 3-dimensional time-space diagram showing the temporal movement of the object cardholder before the data validation process (Chu & Chapleau, 2007b).

5.3 Error Detection by Logical Rules

Transit demand is usually analyzed with respect to the service. Therefore, it is important to ensure that the boarding is associated to a correct run and boarding stop. The exploratory analysis reveals that the proportion of validation records containing erroneous and suspect values reaches 15% to 20% and that most of the errors in the database can be attributed to incomplete or false sign-in information by the driver. Efficient processing requires the modelling of primary objects where disaggregate data are submitted to logics. The logics are translated into rules which can be broadly divided into two categories: public transit logics and spatial-temporal logics. The former is related to public transit planning and operations concepts, which are strictly held in a fixed-route transit service. The latter concerns the physical movement of the objects. Note that both types of logics are closely related. Values that violate the rules are flagged as erroneous or suspect.

5.3.1 Public Transit Logics

Information contained in the validations must not violate the following public transit logics:

- A boarding validation must be tied to a service run. It cannot be tied to a deadheading or interline run.
- A vehicle must follow the pre-defined order of stop during a service run.
- Every regular service run contains a route number, a direction and a scheduled departure time.
- A run, whether it is a service or non-service run, must correspond to a vehicle block.
- A valid boarding must not be made at the last stop of a route (the arrival terminus).
- Each vehicle block can only be carried out by one vehicle at a time, without any temporal overlapping.

Figure 5.3 shows all the boarding transactions linked to vehicle block 140 on 20 weekdays. Each point represents a fare validation. The points are color-coded according to the assigned run and are positioned according to the boarding time. The figure suggests that runs are confined within a specific time interval days after days. The pattern allows analysts to visually detect and even

replace the erroneous values with plausible ones. On February 10, the run 39-0-0900 (short for route 39, direction 0, scheduled departure time 9:00 AM) is missing. Instead, the number of boardings assigned to run 30-1-0810 is abnormally high and its temporal span is too long. Run 39-0-0900 represents a more realistic and logical value for the cluster between 8:50 AM to 9:40 AM. Meanwhile, several boardings are assigned to run 900-0-0655, a deadheading run. The monthly pattern suggests that they are tied to run 45-0-0706.

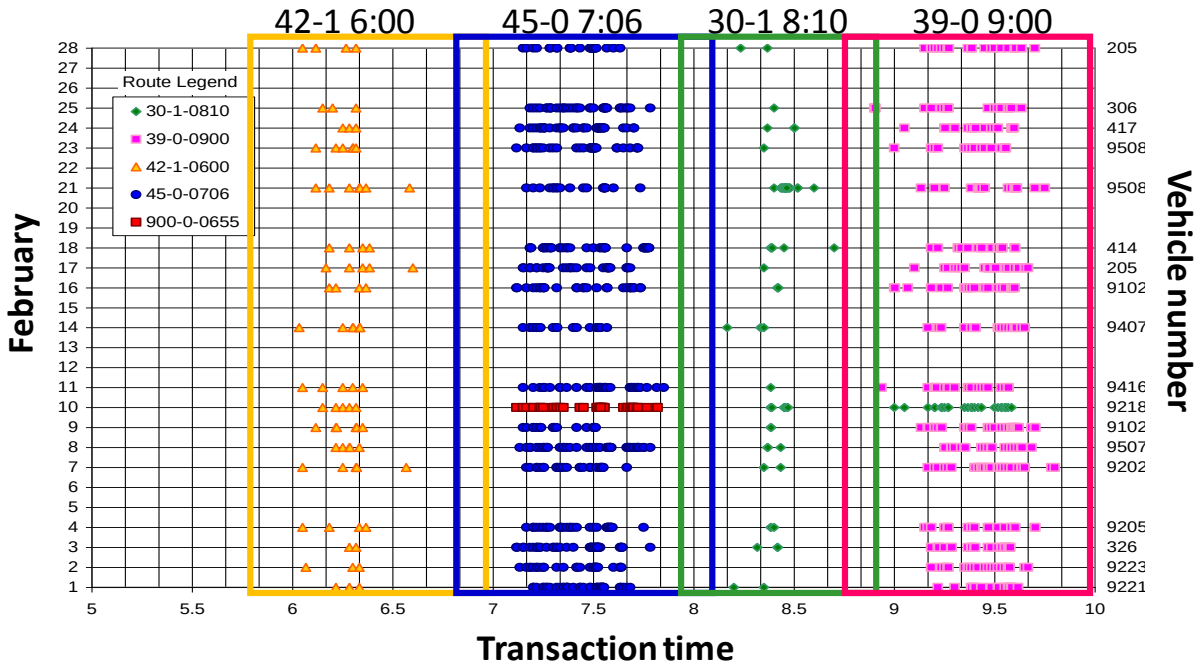


Figure 5.3 Boarding transactions associated with vehicle block 140 for 20 weekdays (Chu & Chapleau, 2007b).

5.3.2 Spatial-temporal Logics

At the same time, information contained in the validations is also scrutinized by the following spatial-temporal logics:

- A boarding must be tied to a time and a known location.
- A vehicle must move at a reasonable speed.
- A vehicle must not have an excessive dwell time at stops, except at the terminus. As a result, the temporal gap time between two boarding at the same stop cannot be excessive.

- The duration of a run must not be excessively long.

The spatial-temporal progression of a vehicle assigned to vehicle block 105, which includes eight service runs, is reconstructed according to the boarding validations in a time-space diagram (Figure 5.4). It is used to verify the spatial-temporal logic of the run and boarding stop information in the data. Cumulative linear distance of the assigned route and boarding stop are joined to each validation record. The location of the vehicle is plotted against transaction time. The size of the bubble represents the number of boardings made at a specific time (at each minute) and location (at the stop level). In theory, boardings of a run should be confined between the planned departure time of the current run at the departure terminus and that of the subsequent run on the x-axis (bounded by two red dash lines), and between the departure terminus and the last stop before the arrival terminus on the y-axis (bounded by two light-blue dash lines), although boardings before the scheduled departure time are common at departure terminus. An angled slope in the diagram indicates the vehicle moves according to the scheduled service. Erroneous run assignment can be detected when the colour-coded boarding validations appear out of bounds or have a zero slope.

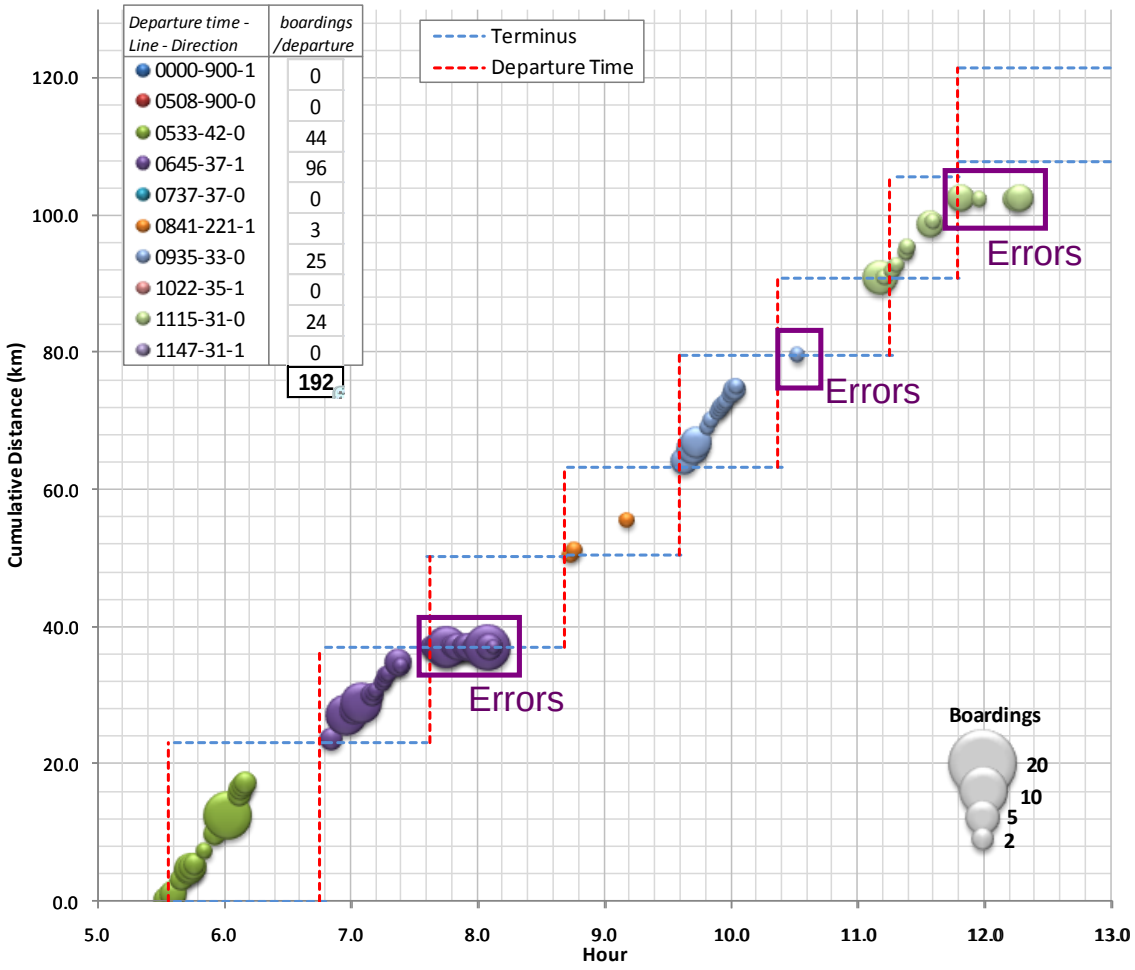


Figure 5.4 The use of a time-space diagram to detect error in run assignment (Chu et al., 2009).

5.3.3 Rule-based Detection

The previous figures visually illustrate the concept of error detection for a vehicle block. However, they are not suitable for processing large amount of data. These concepts are therefore translated into logical rules in an algorithm. Since validation records are stored on-board the bus until the vehicle transmits the data to the central server in the garage, the database is therefore organized by block of records coming from the same vehicle. When an error occurs on a vehicle, records remain erroneous until the driver redresses the problem. The error-detection strategy therefore takes advantage of the sequential pattern of erroneous records within the database. The database is parsed into smaller units and checked for erroneous or suspect values. The key steps are:

- Sorting the validations record by vehicle, followed by transaction date and time;

- Segmenting the data into smaller units according to the assigned run;
- Applying logical rules to each record;
- Detecting erroneous and implausible values;
- Flagging the whole unit as erroneous or suspect.

More specifically, the error detection defines three types of flagged records according to the level of uncertainty: irrelevant, erroneous and suspect values. An irrelevant record is not part of the scheduled service and cannot be assigned to a regular run. They include invalid vehicle number, invalid block number and runs that are not part of the regular planned service, such as transactions recorded during maintenance and special events. A record is considered erroneous when one of the values violates public transit logics and that value needs to be corrected. A record is suspect when the vehicle appears to defy the spatial-temporal logic. The threshold values incorporated in the rules can significantly alter the number of flagged records.

5.3.4 Examples of Logical Rules and Errors

The error detection procedure first filters out all irrelevant transactions, which should not be considered as regular transactions. Table 5.4 shows an excerpt of the database, with relevant fields only, to illustrate a run that is not correctly entered. The driver initialized a deadheading run 900-0-0625 but failed to initialize the service run (83-0-0648) before the first three transactions occurred. The boarding stops in the first 3 transactions are incorrect since they are dummy boarding stops (1 or 2) tied to a deadheading run (route 900). To translate this situation into an error-detection rule, all route numbers containing 900 have to be identified as errors because, by definition, there can be no boarding in a non-service or interlined run.

Table 5.4 Transactions illustrating a case where a driver failed to initialize a service run after a deadheading run.

Record ID	Card Number	Fare Type	Date	Transaction Type	Transaction Time	Vehicle Blcok Number	Route Number	Direction	Departure time	Boarding Stop	Bus Number	Driver Number	Event Sequence Number
23080308	1454633320	2	10-Feb	1	6:50	207	900	0	6:25	1	9217	3283	5
23080309	2187852747	2	10-Feb	1	6:50	207	900	0	6:25	1	9217	3283	6
23080310	2173486001	16	10-Feb	1	6:51	207	900	0	6:25	1	9217	3283	7
23080311	1381629131	2	10-Feb	1	6:53	207	83	0	6:48	4406	9217	3283	12
23080312	3325226344	2	10-Feb	1	6:53	207	83	0	6:48	4406	9217	3283	13

A run is always associated with a scheduled departure time in the smart card AFC system. Some boarding records have a “0000” schedule departure time, which indicate original values are overridden by human intervention. Therefore, all boarding validations with the departure time “00:00” are considered as erroneous (Table 5.5) and the scheduled run needs to be imputed.

Table 5.5 Transactions with departure time “00:00”.

Record ID	Card Number	Fare Type	Date	Transaction Type	Transaction Time	Vehicle Blcok Number	Route Number	Direction	Departure time	Boarding Stop	Bus Number	Driver Number	Event Sequence Number
23126308	2730621387	1	10-Feb	1	16:50	503	39	1	00:00	2610	8505	568	7
23126309	33999041	1	10-Feb	1	16:50	503	39	1	00:00	2610	8505	568	8
23126310	3068145281	37	10-Feb	1	17:03	503	39	1	00:00	2547	8505	568	26
23126311	1180472426	37	10-Feb	1	17:03	503	39	1	00:00	2547	8505	568	27
23126312	374796392	1	10-Feb	1	17:04	503	39	1	00:00	2540	8505	568	29

When a driver fails to initialize a service run following another service run, the boardings are tied to route and stop of the previous run and most likely to the last stop of one of the previous routes

in the vehicle block. Boarding stop 2008, shown in Table 5.6, is the arrival terminus of route 31 direction 1. If analysis is performed on the raw data, the number of boardings would be abnormally high for first service run while it would be abnormally low, or completely absent, for the following run. As a result, all boardings at the terminus are identified as errors. A large temporal gap between two successive boardings, usually the result of the absence of boarding towards the end of the route in addition to the layover and/or deadheading time between runs, can be used to identify a run that has not been properly initialized. On the other hand, the boarding time usually lies within the scheduled duration of the run. It is theoretically possible that it lies outside of the planned duration due to delay or other unforeseeable incidents. Such records are flagged as suspect.

Table 5.6 Transactions illustrating a case where the driver fails to initialize a service run after another service run.

Record ID	Card Number	Fare Type	Date	Transaction Type	Transaction Time	Vehicle Blcok Number	Route Number	Direction	Departure time	Boarding Stop	Bus Number	Driver Number	Event Sequence Number
23081150	383864696	38	10-Feb	3	9:21	181	31	1	8:52	2424	9122	3342	334
23081151	3602471840	38	10-Feb	3	9:21	181	31	1	8:52	2424	9122	3342	335
…	…	…	…	…	…	…	…	…	…	…	…	…	…
23081152	106782112	1	10-Feb	1	9:51	181	31	1	8:52	2008	9122	3342	358
23081153	35846603	15	10-Feb	3	9:52	181	31	1	8:52	2008	9122	3342	359
23081154	1381497547	1	10-Feb	1	9:54	181	31	1	8:52	2008	9122	3342	360

An excessively long dwell time at a stop (apart from certain locations such as the departure terminus, short-run departure terminus, and important transfer points) defies spatial-temporal logic. Therefore, consecutive boardings recorded at the same stop with a large temporal gap are flagged as suspected. Table 5.7 illustrates the implausible situation where 47 boardings are made at the same stop, the 68th stop out of 74 of route 60 direction 0, with a duration spanning of 49 minutes. A logical rule would state that a vehicle in a service run cannot spend more than a certain amount of time at the same stop. However, the amount can vary according to the type of

stop and type of route. For example, routes picking up students after class usually have a long dwell time at the school. Therefore, the rule must be flexible enough to accommodate special cases. This rule also captures a software or hardware problem. All the transactions on vehicle 9401 have boarding stops fixed at the departure terminus for the whole day regardless route information, as illustrated in Table 5.8. This may be an indication of equipment or software malfunction.

Table 5.7 Transactions illustrating a case where the dwell time at boarding stop 5006.

Record ID	Card Number	Fare Type	Date	Transaction Type	Transaction Time	Vehicle Blcok Number	Route Number	Direction	Departure time	Boarding Stop	Bus Number	Driver Number	Event Sequence Number
23102647	845316033	1	10-Feb	3	8:30	147	60	0	8:20	5006	9506	3332	176
23102648	3330703001	15	10-Feb	3	8:31	147	60	0	8:20	5006	9506	3332	177
…	…	…	…	…	…	…	…	…	…	…	…	…	…
23102693	1449361817	1	10-Feb	1	9:19	147	60	0	8:20	5006	9506	3332	222
23102694	918350952	34	10-Feb	1	9:19	147	60	0	8:20	5006	9506	3332	223

Table 5.8 Transactions illustrating a case where all boardings of a vehicle from the whole day are tied to the same stop.

Record ID	Card Number	Fare Type	Date	Transaction Type	Transaction Time	Vehicle Blcok Number	Route Number	Direction	Departure time	Boarding Stop	Bus Number	Driver Number	Event Sequence Number
23105317	2801434009	3	10-Feb	1	6:31	239	900	1	0:00	2	9401	3374	5
23105318	3533243585	15	10-Feb	1	7:01	236	48	0	6:55	1339	9401	3374	11
23105319	3792566987	1	10-Feb	1	7:01	236	48	0	6:55	1339	9401	3374	12
…	…	…	…	…	…	…	…	…	…	…	…	…	…
23105562	4139422058	38	10-Feb	1	20:27	424	439	0	20:03	2151	9401	3352	290

The last type of suspect value is characterized by an inconsistent order of boarding stops in a run as illustrated in Table 5.9. In this example, the spatial-temporal logic of a bus route is violated. Normally, the system should not allow a return to an upstream boarding stop unless there is human intervention in the system. Stop 1019 is the third stop of the route but it is repeated after the 52nd stop (2604). This type of error can be detected by checking sequence of stops within a run.

Table 5.9 The order of stops within a run is not respected in the transactions.

Record ID	Card Number	Fare Type	Date	Transaction Type	Transaction Time	Vehicle Blcok Number	Route Number	Direction	Departure time	Boarding Stop	Bus Number	Driver Number	Event Sequence Number
23094597	2251784345	1	10-Feb	1	7:30	274	46	0	7:28	1019	9408	3207	11
23094598	1991257473	1	10-Feb	1	7:31	274	46	0	7:28	1002	9408	3207	15
…	…	…	…	…	…	…	…	…	…	…	…	…	…
23094630	3070177153	1	10-Feb	1	8:07	274	46	0	7:28	2604	9408	3207	93
23094631	373658474	1	10-Feb	1	9:13	274	46	0	7:28	1019	9408	3207	129

5.3.5 Results of Error Detection

Table 5.10 summarizes the results of rule-based processing by flag category. Fewer than a thousand records (0.1% of total) are classified as irrelevant data. In total, 115,650 records are flagged as irrelevant, erroneous and suspect values, representing 15.2% of all records in the database. Many errors are captured by more than one rule. These errors are unlikely due to the system break-in period since the smart card system has been in place for several years. Figure 5.5 also shows that flagged data are approximately evenly distributed among the weekdays and are not caused by one-off incidents. Therefore, the validation process should focus on the system as a whole. The proposed error detection procedure take advantage of the information contained within the database, the data mechanism of the AFC system and public transit concepts. Further data validation would require external information not available in the database.

Table 5.10 Distribution of transactions by flag type for the month of February 2005. The same transaction can be flagged more than once by different rules (Chu & Chapleau, 2007).

Flag type	Rule	Record count
Irrelevant	Invalid vehicle number	2
	Invalid block number	376
	Special run	577
Erroneous	Boarding during deadheading	25,897
	Missing departure time	15,108
	Boarding at arrival terminus	25,212
Suspect	Excessive dwell time	84,320
	Excessive run time	24,004
	Excessive gap time	17,368
	Inconsistent sequence of boarding stop	8,224
Flagged total	All rules combined	115,650 (15.2%)
No flag	Validated	647,920 (84.9%)
Grand total		763,570 (100%)

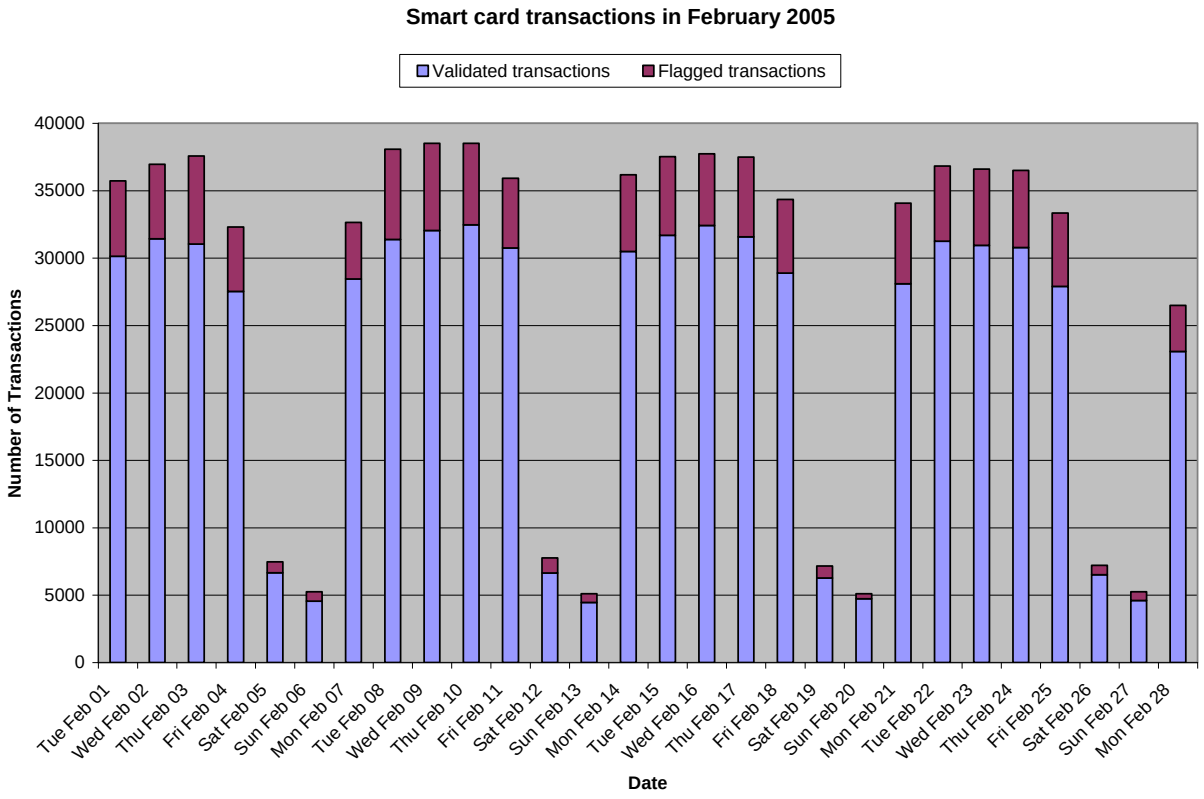


Figure 5.5 Results from the validation procedure on smart card transactions in February 2005 (Chu & Chapleau, 2007).

5.4 Data Replenishment by Imputation

The erroneous and implausible data represent a significant loss of information. There is a need to recover the lost information in order to re-establish the spatial-temporal continuity of each cardholder, and to avoid the propagation and amplification of error and loss of information. The aim is to replenish as much data as possible in a perfectly coherent way. The fact that a transaction record is created means that a boarding occurred. The fare validation record must therefore be tied to a run and a boarding stop. With an informational approach, imputation is done for flagged records according to the repetition of transit service and by mining through cardholders' historic boarding patterns. The following paragraphs give an overview of the data correction procedure.

5.4.1 Concept 1: Repetition of Transit Service

As mentioned earlier, public transit service revolves around an average weekday and is divided into vehicle-blocks which are carried out by vehicles. Since the vehicle number in the validation records is considered infallible, the imputation technique seeks to establish a relationship between the vehicle number and the vehicles trips on each day, through the concept of vehicle blocks. The procedure involves the following steps:

- Building a correct scheduled service dictionary of an average day based on vehicle-blocks;
- Linking vehicles to vehicle-blocks on each day;
- Assigning a run to each transaction.

5.4.1.1 Building an Operations Dictionary with an Informational Approach

The first step is to build an operations dictionary of an average weekday from the validation records based on the information approach. Although not all the scheduled service runs appear in the boarding records everyday (which is exactly the reason why imputation is needed), the complete service can be derived by integrating knowledge accumulated from a number of days. This assumes that a longer period yields a more complete and reliable knowledge of the operations because the probability of missing a run diminishes as the number of days increases.

Using validation records from February 2005, all the runs and their associated vehicle-blocks are enumerated.

By sorting the list by vehicle block and planned departure time, not only the sequence of runs with planned departure time within a vehicle blocks can be revealed, it also allows the derivation of additional information, such as planned run duration. Paradoxically, flagged data, which are initially regarded as failure to initialize the correct service run, can contribute to refine information on scheduled run durations since they provide the planned departure time of the deadheading trips between two service runs.

Table 5.11 provides an excerpt of the complete service that is reconstructed from 20 weekdays of February 2005 during which the timetable remained largely unchanged, with only minor exceptions. At the same time, an indicator shows the number of days that the run is present. For example, vehicle block 140 contains the following 5 consecutive runs:

- Route 42, direction 1 departing at 6:00 AM (42-1-0600) is present 18 days out of 20;
- 900-0-0655 is a deadheading run which appears once;
- 45-0-0706 is present 17 days out of 20;
- 30-1-0810 is present 17 days out of 20;
- 39-0-0900 is present 17 days out of 20.

Logically, the four service runs should be present in the database in all 20 days. This means some boardings from the missing days are incorrectly assigned to another run. The dictionary needs to be enriched in order to serve as a reference for imputation. Although the planned departure times are present, but in practice, they cannot be use as the temporal boundary for run assignment because early boardings at departure termini are very common. At the departure terminus, drivers often allow cardholders to board the vehicle several minutes before the scheduled departure time. This is especially true for runs preceded by a layover time or deadheading run. This situation would confuse the algorithm by assigning the boarding to the preceding run. Therefore, the temporal boundaries need to be redefined. The redefined beginning-of-line and end-of-line takes into account this observation. If a deadheading run is present, it is incorporated into the following run. If not, the last minutes of the run duration is transferred from the preceding run to the following run. The amount can be determined by analyzing the latest boarding time from the

validated records. However, this remains a delicate operation as there is a need to balance between early boarding and delay from the previous run.

Table 5.11 A typical vehicle block for a weekday derived from transactions taken from the enriched operations dictionary.

Vehicle Block Number	Route Number	Direction	Departure Time	Type of Run	Days Present (Out of 20)	Number of Transactions	Number of Validated Transactions	Latest Transaction Time	Redefined Beginning-of-Line	Redefined End-of-Line
140	42	1	6:00	Service	18	80	80	6:36	5:46	6:55
140	900	0	6:55	Deadheading	1	52	0	-	-	-
140	45	0	7:06	Service	17	567	567	7:51	6:56	7:55
140	30	1	8:10	Service	17	99	72	8:36	7:56	8:45
140	39	0	9:00	Service	17	380	380	9:48	8:46	10:03

This operations dictionary forms the basis for imputing run information. If the operations vary according to the day of week, imputation would require more than one dictionaries. For the STO, the timetable is based on an average weekday. The vehicle blocks are practically the same for all weekdays although there are a few additional trips in Tuesdays, Thursdays and Fridays. Imputation for Saturdays and Sundays would require two additional dictionaries.

5.4.1.2 Linking a Vehicle to a Vehicle-block

Although vehicle-blocks are repeated daily, it is not the same vehicle that is assigned to carry out the same vehicle-block everyday. Therefore, a link needs to be established between the vehicle and the vehicle-blocks for each day. A direct relationship can be established in the majority of cases since each validation record contains the vehicle block number and the vehicle number at the same time. Problems arise when there is conflict among the data such as a vehicle-block linked to more than one vehicles with overlapping time. To deduce the correct vehicle-block,

those records are matched against vehicle-blocks that have not appeared on the day according to the range of boarding times. Another way is to use the concept of cardholders' boarding histories.

5.4.1.3 Assigning a Run to a Boarding

In order to assign a run to a boarding, the transaction time is simply compared against the redefined beginning-of-line and end-of-line times in the enriched operations dictionary.

5.4.2 Concept 2: Boarding History of Individual Cardholders

The second concept uses the boarding history of individual cardholders to obtain runs and boarding stops in the imputation process. The underlying assumption is that there exists a pattern or some regularities in individual cardholders travel behaviour which are revealed in the multi-day boarding history. This is supported by Chapin (1974) who proposes that there is recurrent activity pattern of an individual in space and time. The smart card AFC system stores all boarding transactions of travelers who use a smart card as their mode of payment. Since each card represents a unique cardholder, a cardholder's historic boarding records – 20 days in this case – are captured by the system. According to the theory of activity-based travel demand modeling, travel is considered a derived demand arising from the need or desire to participate in an activity, which is in turn characterized by its location in space and time. Space-wise, the majority of travelers have a pre-defined set of locations where they visit, such as their home, workplace or a shopping center. Also, activities take place at a certain point in time. The periodicity and persistence of these two elements in a traveler's activity schedule, whether daily, weekly or for an even longer period, constitute a travel pattern which is imprinted in their boarding records. Since most of the cardholders are frequent travelers, the historic travel patterns of these cardholders can reveal regularities in run and boarding stop. The concept can even be extended to identify boarding locations from validation data without location stamp.

5.4.2.1 Imputation of Run

As previously described, run imputation can be done more easily with the concept of the repetition of transit service. In some circumstances, such as duplicated vehicle-blocks, it may be relevant to use the boarding history of individual cardholders. The underlying assumption is that cardholders who boarded a vehicle run in a particular day are not governed by randomness.

Rather, it is their aggregate travel pattern which links them together on the same run. It is important to stress that only validated records are used under this concept. The procedure of run imputation for a sequential unit of flagged records involves the following steps:

- Isolating the boarding history of each individual on-board the vehicle;
- Determining the most-probable run for each individual using a Bayesian statistical approach;
- Validating and assigning the run to all the on-board individuals.

The first step is to perform queries to isolate the boarding history of each individual on-board the vehicle.

Figure 5.6 The transaction history of a card in the month of February 2005 (Chu & Chapleau, 2007).

illustrates all transactions made by a typical cardholder in February 2005. Each transaction is represented by a point indicating its date and time. Different symbols denote the routes as assigned by the system. A filled point represents a valid transaction; a hollow point represents a flagged transaction for which a run is imputed. The number above or below the point shows the stop order of the boarding stop. 0 represents the departure terminus.

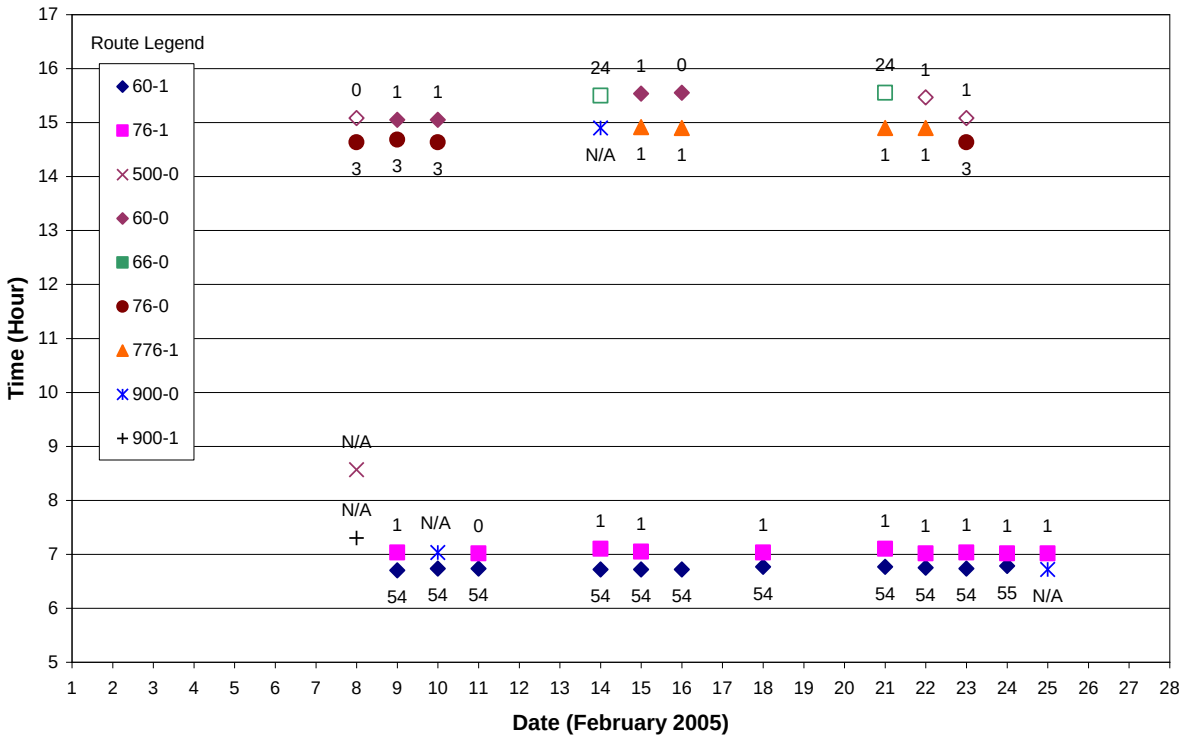


Figure 5.6 The transaction history of a card in the month of February 2005 (Chu & Chapleau, 2007).

In the field of statistics, there are two approaches to probability, namely the frequentist (classical approach) and the Bayesian approach. Given a distribution, the frequentist approach assumes that parameters are fixed and the randomness lies in the data, while the Bayesian approach considers the parameters as random variables and uses the observed data to provide information on the parameters. According to data mining literature, the imputation of run or boarding stop information can be formulated as a classification problem, since both the run and boarding stop values constitute categorical or discrete classes. The goal of the Bayesian classification is to learn the distribution of instances from validated records and to predict the most likely class for each of the flagged records. For each record, the conditional probability for each class is calculated based on explanatory variables from the boarding records of a card. The class with the highest probability is retained. The simple implementation and the strong interpretability of the Bayesian classifier merit the choice.

The procedure is first implemented and tested in WEKA, with the class being the run. The Naïve Bayes Classifier in WEKA takes into account the boarding time, previous and following routes

taken on the same day as well as their transaction times as attributes. The algorithm uses all validated data as the training set as the goal of the procedure is not the prediction of trip details of future trip. Instead, it is viewed as a data cleaning process. The procedure is repeated for each individual.

Another implementation of run imputation draws on attributes from users' historic travel patterns using day of the week (weekdays, Saturday or Sunday) and transaction time as explanatory variables. An algorithm looks for the most frequent run that a user takes at about the same time within the analysis timeframe. Boardings made within plus or minus 5-minute window are considered. The imputed run is filtered by the subset of runs associated with the vehicle block according to the operations dictionary. This means that predictions are filtered out if they do not belong to the subset of runs in the vehicle block.

The public transit logic dictates that a vehicle can only carry out one run at a time. Transactions made by people on-board the same vehicle at a particular time must therefore have the same vehicle-block and run. Since run imputation does not guarantee that condition, the values need to be validated by examining the aggregate travel behaviour. The run, and as a consequence the vehicle-block, for that sequence of transactions is most likely to be the most frequent value among the cardholders. The results based on the regularity in public transit operations and on the regularity in users' historic travel pattern can be cross-validated to improve accuracy.

5.4.2.2 Imputation of Boarding Stop

Similar to run imputation, boarding stop can be estimated using users' historic travel patterns. The procedure involves the following steps:

- Isolating the boarding history of each individual on-board the vehicle;
- Determine the most probable boarding stop for each individual using a Bayesian approach;
- Validating the stop for each individual;
- Assign stops to the remaining transactions using interpolating or extrapolation.

It is possible to impute boarding stops since activity not only recurs in time, but also in space. Upon the imputation of run, boarding stop probabilities can be computed by following a similar

Bayesian approach, using boarding stop as the class and adding run or route as explanatory variable. An algorithm determines the most frequent boarding stop when the user takes a specific run within the analysis timeframe. The temporal aspect is implicitly considered in the run. In cases where the run is taken only once, the condition is relaxed to include boarding on the same route (same route and direction but not the same scheduled departure time).

Similar to run imputation, the imputed boarding stops can be in conflict with the other transactions. The imputed boarding stops need to be validated to ensure that the bus object respects the public transit and spatial-temporal constraints: the order of boardings must be coherent with stop order and the travel speed must be reasonable. The validation filters prioritize stops that are imputed from run information as opposed to route information only. They identify imputed stops that are incompatible with the preceding and following boarding records. Several iterations of filtering are required.

For records without an imputation or with a rejected value, linear interpolation and extrapolation of known positions are used to estimate boarding stop. Since the temporal resolution of the transaction time is one minute and the distance between bus stops is generally short, there is uncertainty on the estimated boarding stop. Although more sophisticated methods can be used, linear interpolation and extrapolation are often sufficient to improve the quality of the data for demand and travel behaviour analyses. A schematic summary of the data validation strategy is shown in Figure 5.7.

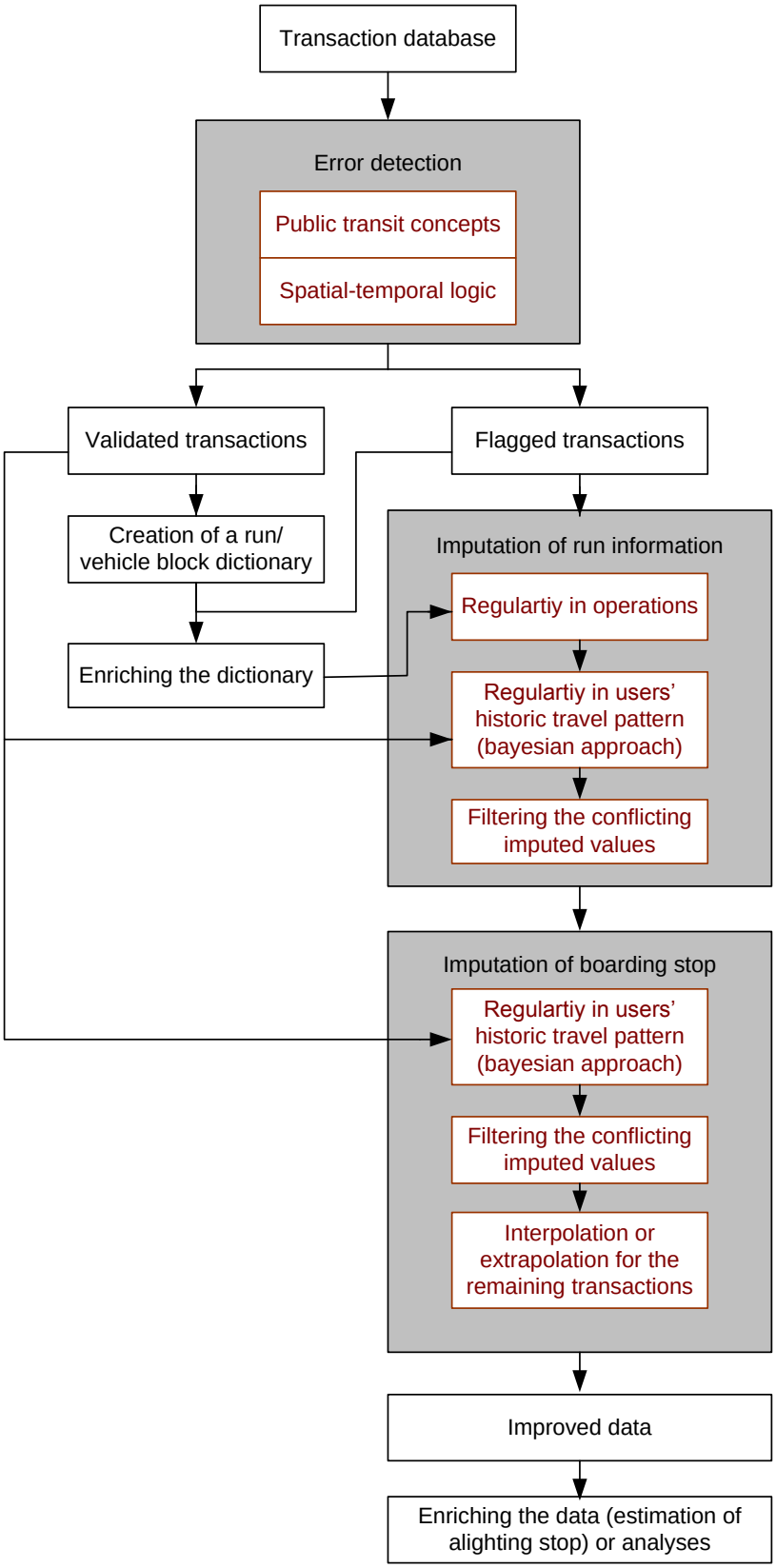


Figure 5.7 Summary of the data validation strategy (Chu & Chapleau, 2007).

5.4.3 Analysis of Results from Data Validation

The relevance and efficacy of the validation process are evaluated. A typical weekday containing about 35,000 transactions can be processed automatically in less than 10 minutes by a desktop computer. There are 38,502 transactions in total on February 10, 2005. 6,032 (15.7%) are flagged as irrelevant, erroneous or suspect. There are 4 irrelevant transactions that are not included in the imputation process. Run and stop for 5,311 out of 6,028 transactions are successfully imputed, which translates to a success rate of 88.1%. Transactions without imputed values are removed from subsequent analyses. After the procedure, 37,781 of 38,502 (98.1%) transactions are considered as coherent, against 84.3% in the raw data. Subsequent refinement to the algorithm further improves the results.

Figure 5.8 shows three condensed time-space diagrams of all the active vehicle objects on February 10. They are used as a visual tool to validate the results of the imputation procedure. Boarding validation records are segmented into individual runs. The vehicle locations in linear distance, revealed by the validation records, are joined together to form a spatial-temporal path. The horizontal axis represents the linear distance of a run and the vertical axis represents time of day. The vertical gap between the runs is adjusted so that each of them is visible. Since a vehicle in service follows stops in a specific order, a monotonic progression is expected. The raw data (red) contain many inconsistencies: a vertical line indicates that the vehicle is stationary and a line with a negative slope means that the vehicle is moving in the wrong direction. This indicates that the boarding locations from the raw data are not reliable. Data excluding all flagged records (blue) and after the imputation process (green) are significantly cleaner and more coherent, confirming the utility of the validation process.

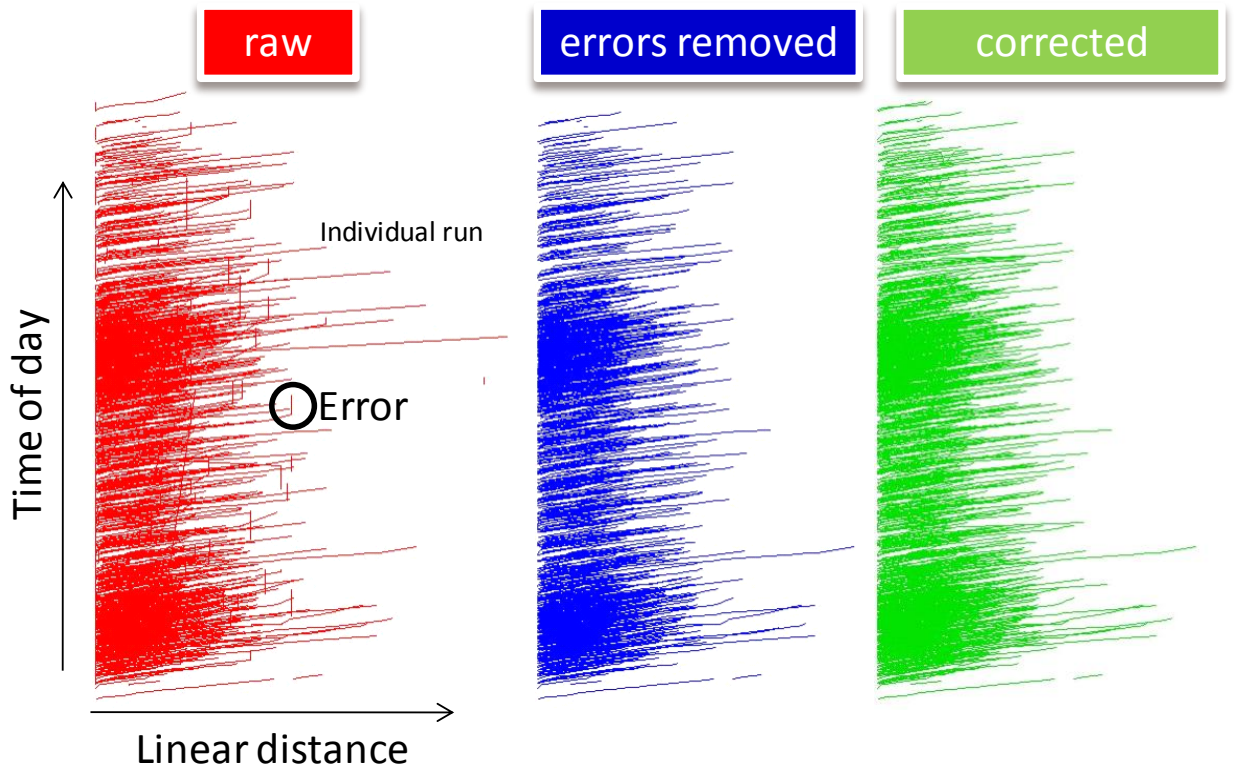
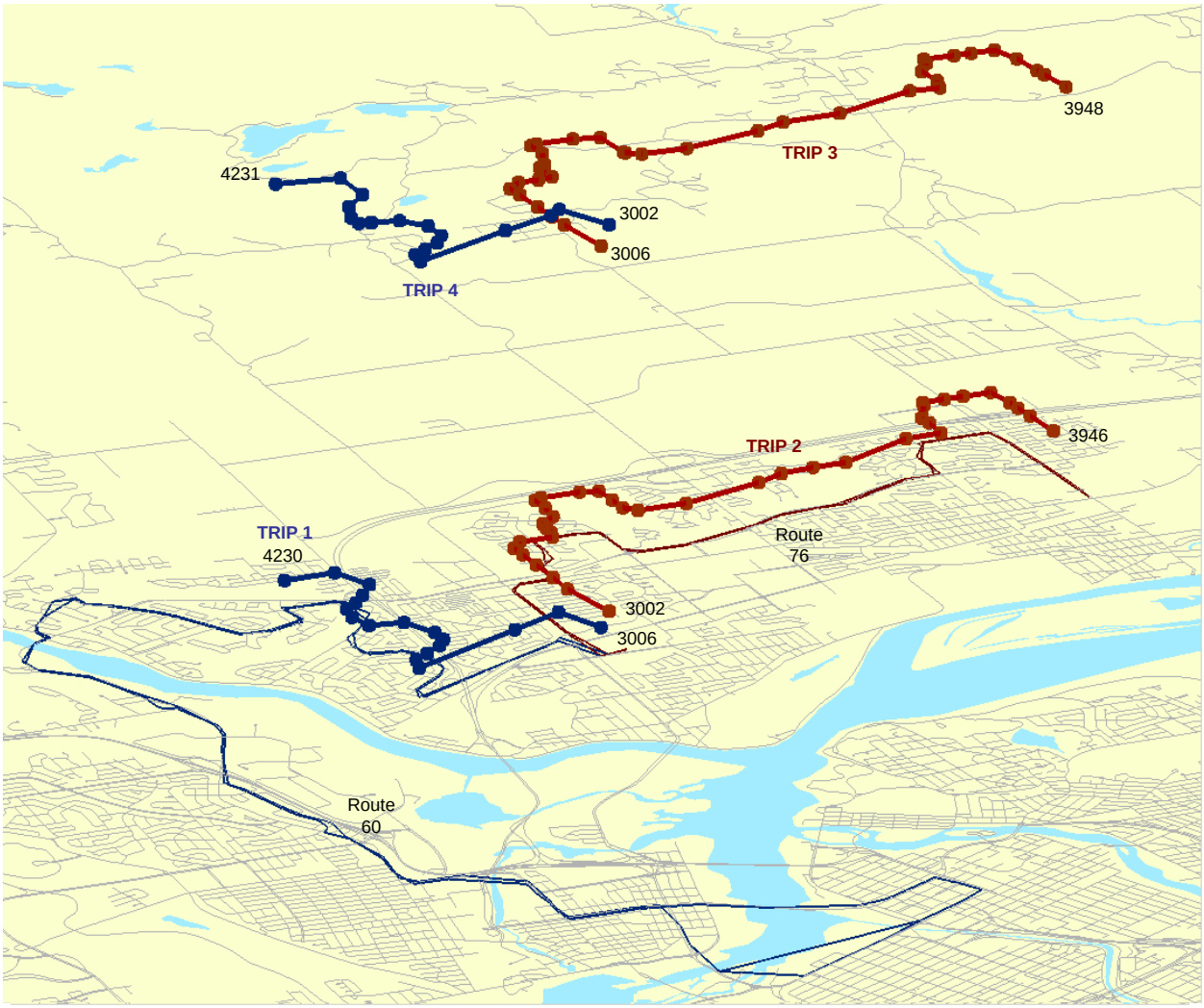


Figure 5.8 Time-space diagrams of bus objects used to evaluate the imputation procedure (Chu & Chapleau, 2007).

Although the procedure does not guarantee that all the imputations are correct and reflect the reality, the procedure recovers most unknown values and re-establishes the spatial-temporal continuity of bus and card objects. In addition, the filtering of the imputed values assures that the retained values satisfy public transit and spatial-temporal constraints, and are coherent as a whole. In order to measure the relative effectiveness and the contribution of the two concepts used in data replenishment, statistics that compare the retained imputed values and the imputed values by each concept is computed. The procedure based on the regularity in planned service imputes 5,732 (95.1%) out of the 6,028 retained run values. 101 (1.7%) imputed values are different from the retained values. 195 (3.2%) transactions have no imputed value. On the other hand, 3,160 out of 6,028 (52.4%) of run values imputed by regularity in cardholders' historic travel patterns are retained. 46 (0.8%) imputed values are different from the retained values. 2,822 of transactions (46.8%) has no imputed values or the values have been filtered out. The statistics confirm both concepts can contribute to the imputation process but suggests that the regularity in operations is significantly more reliable than the regularity in users' travel patterns

in run imputation. However, by combining both methods, the number of incorrect and unsuccessful imputations can be reduced. It must be noted that certain types of cardholders, as revealed by the fare type, can display a more regular travel pattern. Therefore, this concept may be more reliable for routes that have a more regular clientele. Meanwhile, the regularity in users' historic travel patterns represents a reliable concept for the imputation of boarding stop. This procedure provides 4,036 imputed values out of 5,311 retained boarding stops, which constitutes a success rate of 76.0%.

Figure 5.9 illustrates the result of the validation process of the same 4 boardings described in the previous section. Route 900-0-0638 in the second transaction is replaced by the imputed value 76-1-0705. The transaction time of 07:02 constitutes an early boarding, which is common at departure terminus. According to the historic travel patterns of the card (43 transactions in the month), all 10 transactions on route 76-1 are made at stop 3002, and all of them made between 07:01 to 07:06. Therefore, there is a strong likelihood that this particular boarding is made at stop 3002 on route 76-1. With this additional information, subsequent imputation of alighting stop of the previous boarding is achievable. Without the correct route and boarding stop in trip 2, the alighting stop of trip 1 cannot be derived. In this example, not only is the apparently lost information recovered, but the spatial-temporal continuity of the objects is also re-established. The example illustrates that, without proper processing, a database with 15% of records containing errors can potentially affect 30% of the records. The goal of minimizing information loss by improving data accuracy is achieved through the validation process described above.



Trip	Record ID	Transaction Date	Transaction Time	Route Number	Direction	Scheduled Departure Time	Boarding Stop	Derived Alighting Stop
1	23091592	10-Feb-05	06:44	60	1	06:01	4230	3006
2	23092894	10-Feb-05	07:02	76	1	07:05	3002	3946
3	23119896	10-Feb-05	14:38	76	0	14:35	3948	3006
4	23123387	10-Feb-05	15:03	90	0	15:05	3002	4231

Figure 5.9 A 3-dimensional time-space diagram showing the re-established temporal movement of the cardholder object after the data validation process (Chu & Chapleau, 2007).

5.5 Assessing the Integrity of Databases: a Case study

The validation process and the subsequent enrichment processes depend heavily on the validity of the data used as reference. Although fix-route transit service is in general well structured and organized, in reality some services do not necessarily follow public transit logics. Examples include planned or unforeseen temporary detour, short-run trips, extras, vehicles dispatched to cope with unforeseen incidents and service organized for accommodating the need of specific clientele such as school routes. These details may not be accurately coded and reflected in the fare collection system. Therefore, it is necessary to validate information in databases and to keep them coherent, making incidental changes if necessary. The inference of the complete service dictionary from boarding transactions assumes that the service database in the AFC system is correct and up-to-date. It also assumes that the planned service did not change during the whole analysis timeframe. The following case study shows how errors in a data table can affect the validation procedure and subsequent analyses as well as how they can be detected.

5.5.1 Case Study: School Routes

Boarding validations with student fare represent one third of the total number of transactions (252,847 out of 763,570) in February 2005. The STO provides special school routes for this clientele in the morning and afternoon to facilitate access to schools. Those routes are numbered in the 600s and 700s in the route data table. There are 5 runs of route 760 direction 0 with the same departure time. They belong to different vehicle blocks: 147, 371, 375, 463 and 468. The following observations from raw data reveal that the coding of these routes is problematic and does not reflect the reality. Further analysis suggests that they do not represent the same routes:

- Implausible vehicle speed: according to the scheduled trip time and route geometry, the required speed is too high to complete the trip.
- Non-symmetric ridership: route 760 has 2 directions, one direction in the AM and another direction in the PM. The monthly total number of validations is 206 in the morning and 2,735 in the afternoon. The latter is more than 10 times more important. Although the route dictionary indicates that the route is symmetric in both directions, the empirical data suggests otherwise.

- Distinctive clientele: one would expect for completely interchangeable routes, their clientele would not be distinct. Two analyses suggest that the clientele is not similar. With the exception of blocks 147 and 375 where there are numerous interchanging cardholders, only a small proportion of cardholders uses another block. A spatial analysis of the origin of the cardholders (using their first validation in the AM) as well as the spatial-temporal constraints of cardholders who subsequently transfer to other routes indicates the vehicles are heading to different directions and does not follow the stop order in the route dictionary.

5.5.1.1 Recoding the Routes

The empirical findings suggest that the vehicle follow parts of the geometry of route 760-0 and continue on as a regular route as indicated in the service dictionary. Cardholders already on-board would not need another fare validation and the "transfer boarding" is therefore not revealed.

The immediate impact of this coding inconsistency would be the poor estimation of the most probable alighting stop in the data enrichment process. According the public transit logics, cardholders can only alight at stops served by route 760-0. In reality, they may alight at stops of the subsequent route. The problem necessitates a recoding of the routes. Although this example highlights a particular case, it illustrates the type and scale of impact that erroneous data in the reference can cause. The analyses successfully pinpoint and rectify the issue. The coding of route geometry may differ from one transit operator to another and issues may arise when analyzing data from a multi-operator AFC system.

CHAPTER 6 METHODS AND ANALYSES FOR OPERATIONS PLANNING

6.1 Spatial-temporal Description of Boardings

Smart card AFC data can act as an alternative data collection method to traditional surveys by offering high-resolution spatial and temporal information. Many tasks in transit planning may be served directly by applying standard GIS procedures and spatial statistics to validated AFC data. The boardings can be described in an aggregate manner to generate demand indicators and reveal general trends. They can also be described in a disaggregate manner in order to generate trip itinerary for transit assignment, and for studying trip details and travel behaviour. Three levels of aggregation on transit network objects are examined: stop, route, and link and node at the network level.

6.1.1 Levels of Aggregation

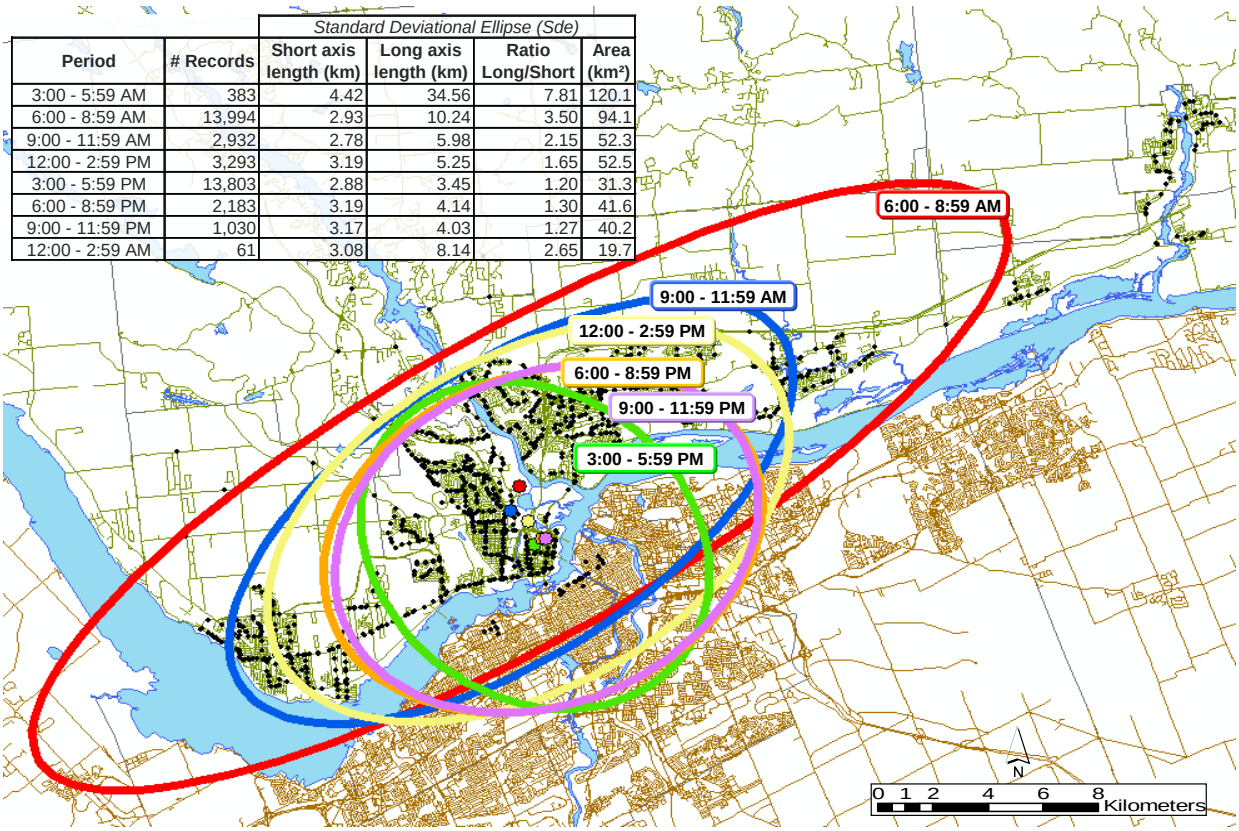
The most basic spatial unit of analysis is at the stop level since the data contain boarding stop location. At the stop level, it is possible to associate the trip generators in proximity to each stop, thus linking user activities to trip generators. Boardings can be analyzed by route. Each route contains many runs which are defined by a direction and a scheduled departure time. Load profile of individual runs can be summarized and compared alongside. Service consumption indicators such as passenger-kilometres and occupancy rate can be estimated. Planners are also able to examine the characteristics of each route segment.

Analyses by stop and run allow the fine-tuning of schedule on individual route and at the stop level, but lack the ability to synthesize and to create a reference demand for the network. For service planning at the network level, such as network geometry and transfer coordination, route-level data are incompatible because of overlapping segments. In the STO network, similar routes are coded as different route numbers and many routes share major corridors, giving the users the possibility of choosing among several routes. Therefore, in transit demand modeling, the ability to aggregate separate data into route-neutral network information while conserving individual trip details is indispensable.

6.1.2 Spatial-temporal Distribution of Boardings

Aggregate boarding spatial pattern can be described using spatial statistics. Standard deviation ellipses measure the concentration or the spread of a spatial distribution of boardings. The length of the long and short axes indicates the level of dispersion in two directions. The ratio between them shows whether the distribution is skewed towards a particular direction. The more compact the ellipse, the more concentrated the boardings are. Meanwhile, the barycentre reveals the spatial mean of all boardings. In order to characterize the temporal evolution of boardings within a day, transaction time is separated into 8 3-hour periods. For each period, a standard deviation ellipse and a barycentre are calculated from all the transactions using Crime Stat III (Levine, 2004). The generated shapefiles (.shp) are put into a standard GIS environment for visualization.

Figure 6.1 shows standard deviation ellipses for six time periods along with the corresponding barycentres. Statistics are also included in the table within the figure. Two time periods are not included because there are few observations. The evolution of the ellipses shows that during the AM peak (6:00 – 8:59 AM), the boardings are more spread out. During the course of the day, the ellipses progressively decrease in size and in the ratio between the long and short axes until the PM peak (3:00 – 5:59 PM). At the same time, the barycentres move southward. These indicate that the boardings are becoming progressively more concentrated and are moving towards the south. It is interested to note that the orientation of the long and short axes are reversed meaning that the distribution are more spread out in the north-south then in the east-west direction in the PM peak. This observation can be explained by the locations of several high schools. After the PM peak, ellipses increase in area with the barycentres staying at about the same location.

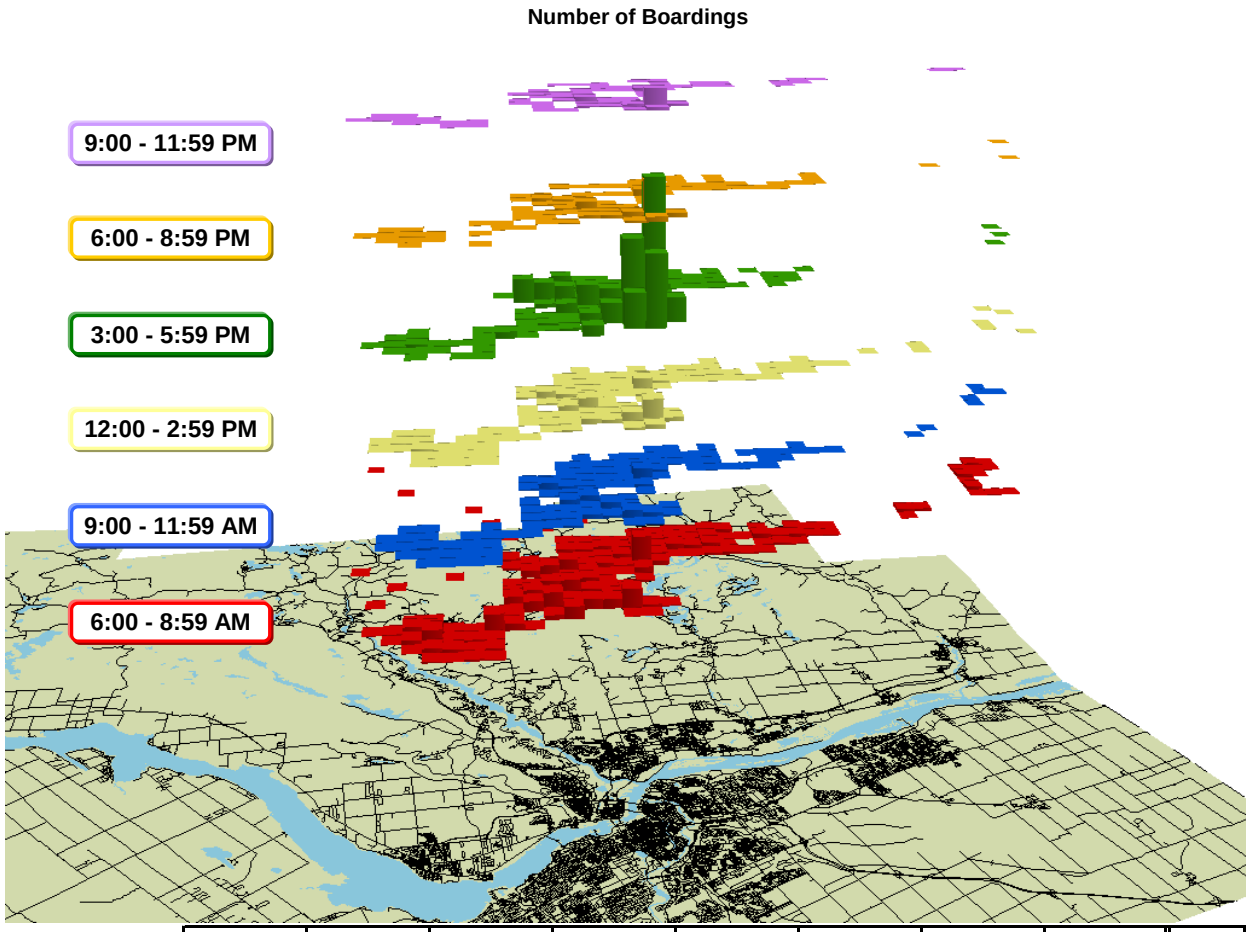


Period	# Records	Standard Deviational Ellipse (Sde)			
		Short axis length (km)	Long axis length (km)	Ratio Long/Short	Area (km²)
3:00 - 5:59 AM	383	4.42	34.56	7.81	120.1
6:00 - 8:59 AM	13,994	2.93	10.24	3.50	94.1
9:00 - 11:59 AM	2,932	2.78	5.98	2.15	52.3
12:00 - 2:59 PM	3,293	3.19	5.25	1.65	52.5
3:00 - 5:59 PM	13,803	2.88	3.45	1.20	31.3
6:00 - 8:59 PM	2,183	3.19	4.14	1.30	41.6
9:00 - 11:59 PM	1,030	3.17	4.03	1.27	40.2
12:00 - 2:59 AM	61	3.08	8.14	2.65	19.7

Figure 6.1 Standard deviation ellipses of transactions during the course of a typical day (Chapleau & Chu, 2007).

Another technique to describe the boarding count is to perform a grid analysis. The territory is partitioned into cells measuring 1 km by 1 km. Boardings within each cell are summed and represented by three dimensional columns. By aggregating boarding into cells, the density of boardings in each cell can be analyzed by time of day. Similar to the previous analysis, transaction time is divided into eight 3-hour periods in order to reveal the temporal evolution of the boardings. Traditional statistics, such as count, maximum, minimum, density and standard deviation, are calculated.

Figure 6.2 shows the result. The two peak periods account for three-quarter of the boardings of the day. The AM peak has the highest number of cells with at least one observation (155 out of 161 active cells). The PM peak contains cells with extremely high concentrations, with almost 3,000 transactions in a one square kilometre cell. The average density of a non-empty cell is 47% higher during the PM peak than during the AM peak.



	3:00 - 5:59 AM	6:00 - 8:59 AM	9:00 - 11:59 AM	12:00 - 2:59 PM	3:00 - 5:59 PM	6:00 - 8:59 PM	9:00 - 11:59 PM	12:00 - 2:59 AM	Whole Day
# obs.	383	13,994	2,932	3,293	13,803	2,183	1,030	61	37,679
% obs.	1.0%	37.1%	7.8%	8.7%	36.6%	5.8%	2.7%	0.2%	100%
# cells with obs.	73	155	115	114	104	86	69	14	161
Density (obs./km²)	5.2	90.3	25.5	28.9	132.7	25.4	14.9	4.4	234.03
Min/Cell	1	1	1	1	1	1	1	1	1
Max/Cell	21	741	277	509	2963	727	356	32	2963
Standard deviation	4.6	112.3	45.1	66.8	384.8	82.9	44.7	8.1	165.8

Figure 6.2 Grid analysis of boardings by time of day (Chapleau & Chu, 2007).

The interpretation of the phenomenon is that during the AM peak, many users board at their residential locations which are more spread out in the region. Since most of the employments or education institutions are concentrated in the CBDs of Gatineau and Ottawa, during the PM peak, the users return home from those areas. The spatial description helps understand the within-day migration of the public transit population.

A detailed temporal distribution of boarding is shown in Figure 6.3. The number of observations per 15-minute interval follows a curve with two distinct intense periods corresponding to the typical AM and PM peak periods. The AM peak is smooth while the PM peak has several crests

which can be explained by boardings made by students leaving school. The jagged curve during mid-day and evening can be attributed to service at each half or whole hour. The ratios between the long and short axes of the standard deviation ellipse for each period are shown along with the graph.

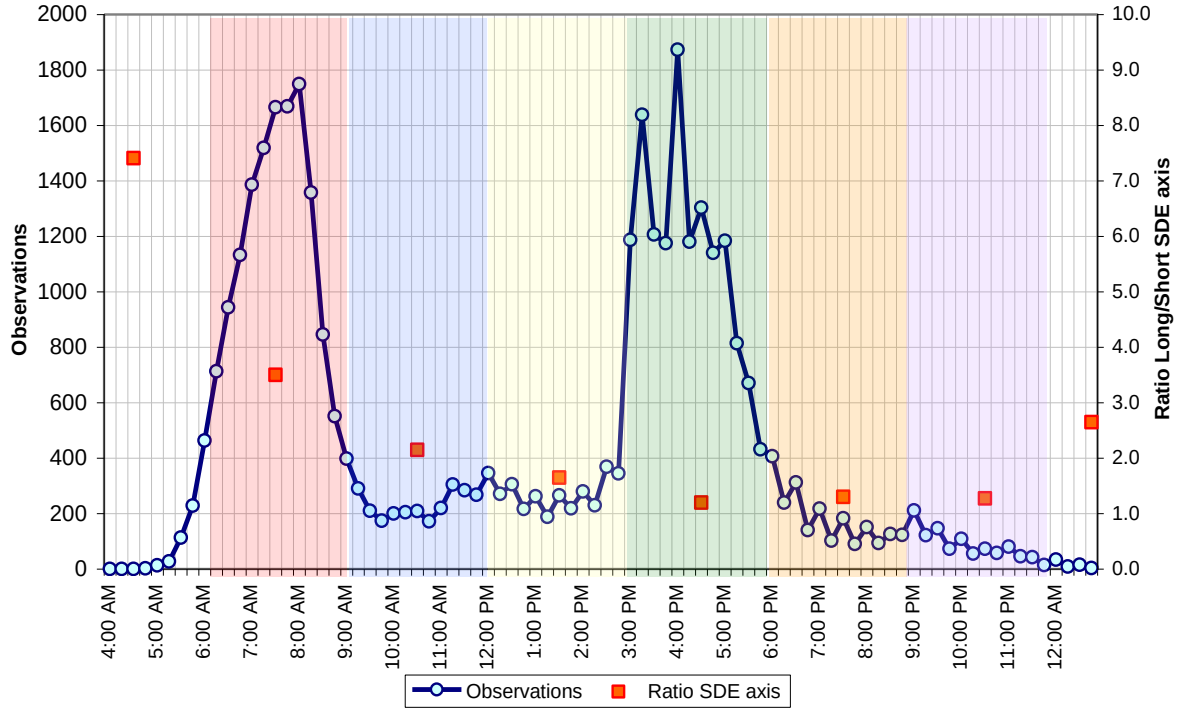


Figure 6.3 Temporal boarding intensity by 15-minute interval (Chapleau & Chu, 2007).

6.1.3 Stop-level Boarding

Stop-level aggregate boardings and derived alightings during the AM peak in a typical weekday are shown in Figure 6.4. The estimation of the most probable alighting stop is discussed in Chapter 7. If the travel pattern is symmetric, similar result of the figure on the right can also be obtained by aggregating boarding during the PM peak. The colour within each pie represents fare type and its size is proportional to the number of boardings. Several patterns can be observed by comparing and contrasting the two maps:

- Adults and students under 21 are the dominant fare types among the cardholders.
- The majority of student over 21, mainly university students, alight in a few stops in Ottawa, nearby educational institutions such as the Ottawa University.

- Seniors and STO employees have few transactions.
- Boardings are distributed among a greater number of stops when compared to alightings.
- The number of boardings is also more evenly distributed among the stops than the number of alightings.
- The number of boardings in the CBD of Ottawa is very low compared to what it receives in alightings. This shows a passenger flow migrating towards Ottawa.
- The destinations of the cardholders form several clusters.
- The destinations of students under 21 remain largely in Gatineau whereas those of adults are concentrated in the CBD of Ottawa and Gatineau.

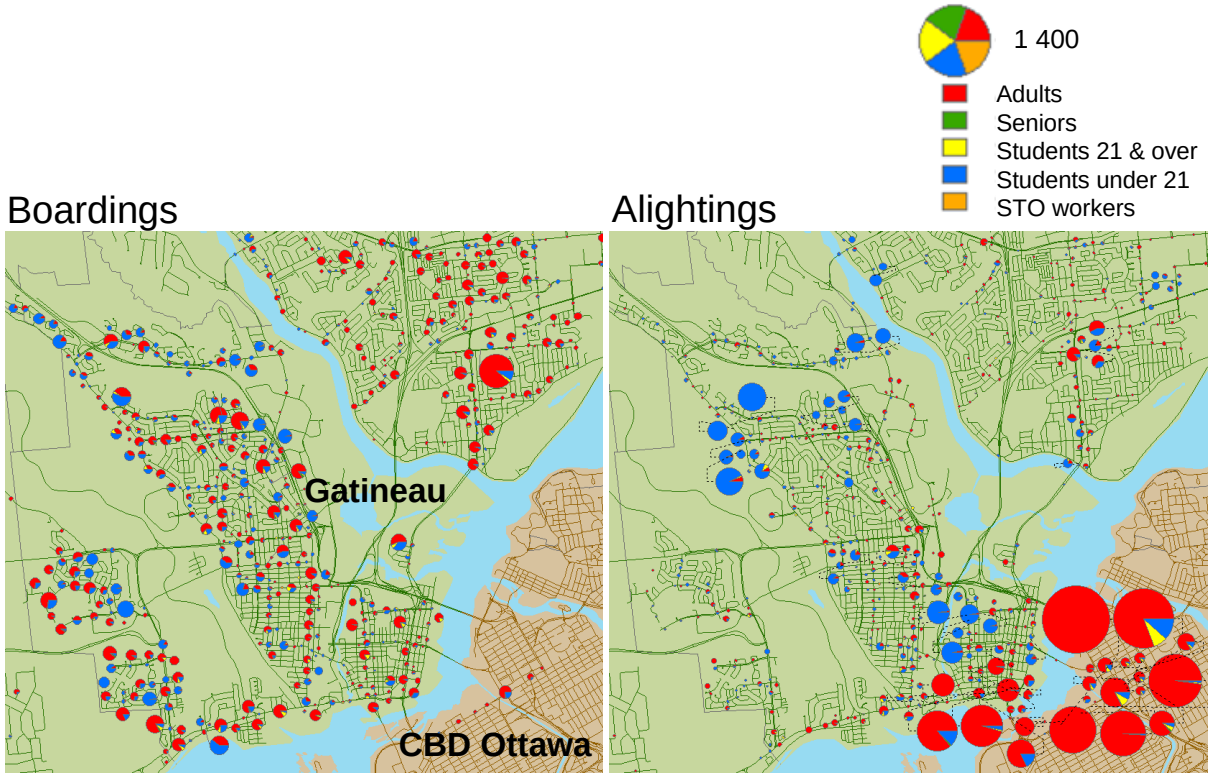


Figure 6.4 Stop-level boardings and derived alightings for a typical AM peak (Chu et al., 2009).

6.1.4 Route-level Boarding

Boardings can also be aggregated into route-level to characterize the function and the users of a route. Figure 6.5 illustrates the temporal distribution of the transit users on a specific bus route on a typical weekday. Of the total of 1,604 bus runs in that day, route 37 and its variants have 108

departures (54 inbound and 54 outbound) and the highest boarding count. The figure underlines the importance of scrutinizing the spatial and temporal aspects of the travel demand. Each route, depending on the locations served and service frequency, has a distinctive boarding signature. Route 37 features two important peak periods and a small peak in midday. The AM peak is slightly higher and narrower than the PM peak. Both the inbound and outbound directions of the route are frequented by the population during the peaks, meaning that the function of the route is not limited to transporting users to and from the CBD of Gatineau and Ottawa. “Reverse commute” on this route is observed.

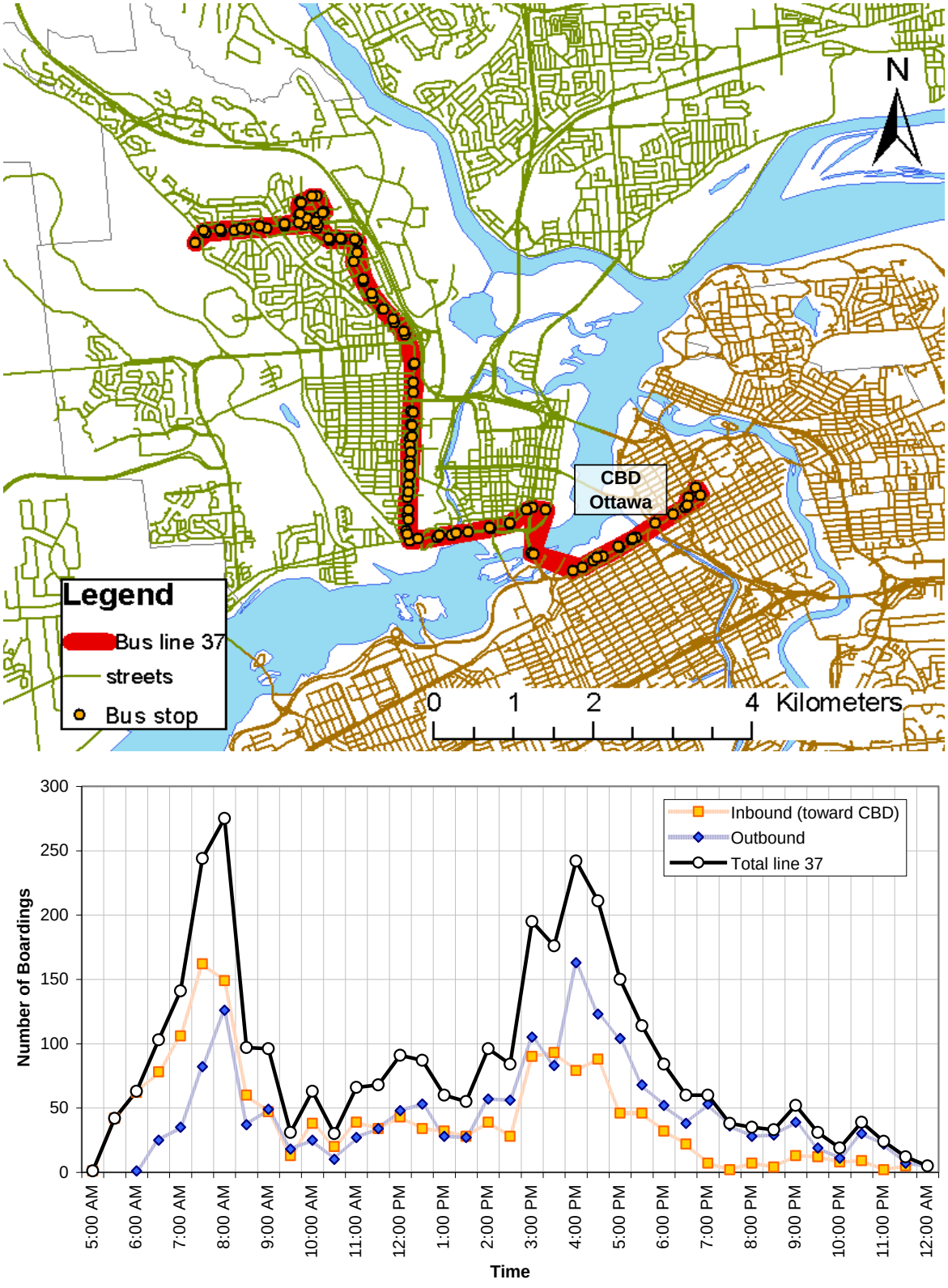


Figure 6.5 Stop locations and hourly route-level boardings of route 37 (Chapleau & Chu, 2007).

6.1.5 Run-level Boarding

Planners also benefit from the analysis of highly-detailed space-time diagrams, such as the comparison between boardings in AM and PM peak periods for both directions of route 37 (Figure 6.6). The size of the bubbles is proportional to the number of boardings. The time and location of boarding activities of each run can be clearly seen. Moreover, the concept of average speed by segment can be observed from the slope. The presence of bus bunching is also revealed in the diagrams.

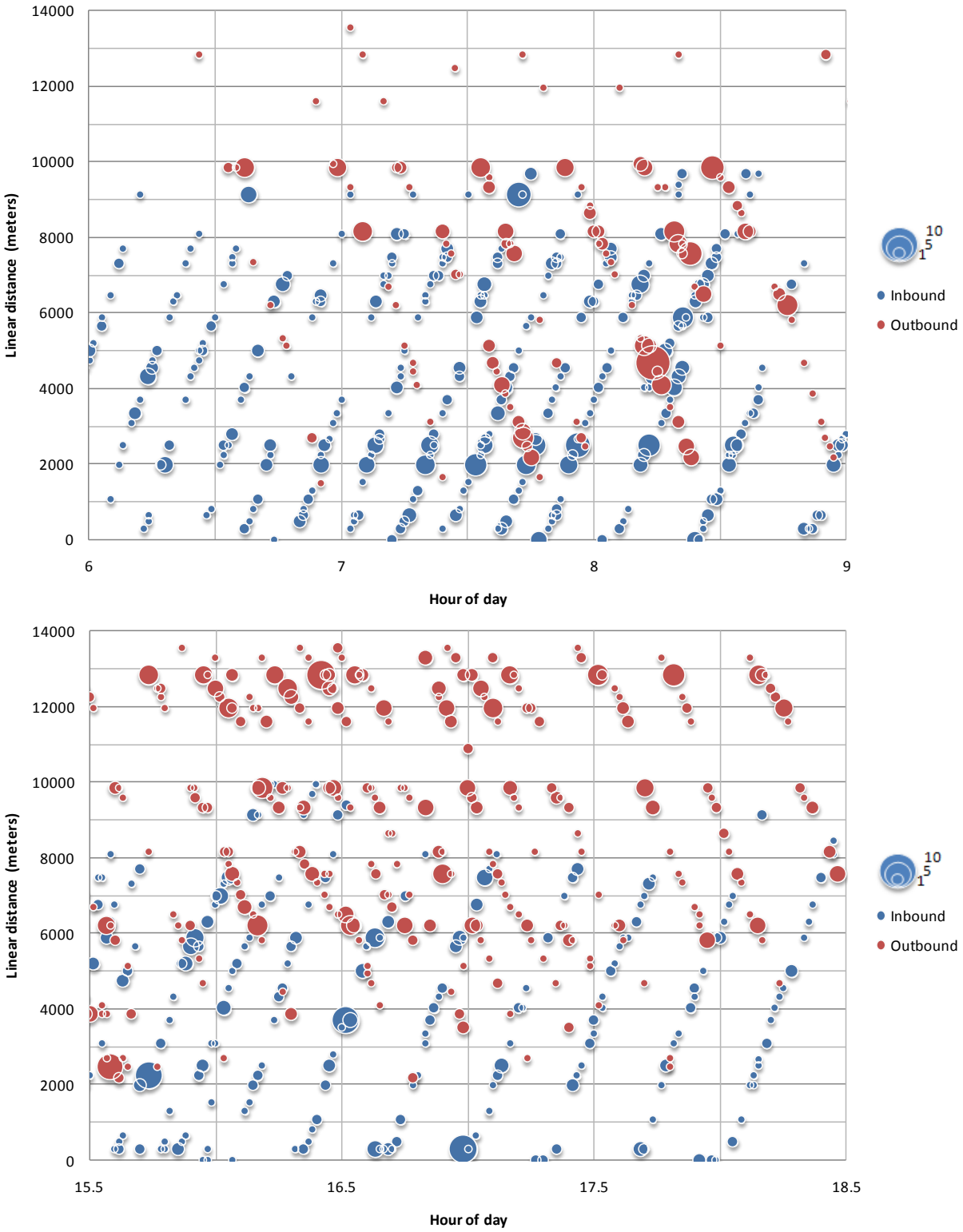


Figure 6.6 Space-time diagrams of bus route 37 illustrating vehicle location and boarding intensity during the AM and PM peaks (based on Chapleau & Chu, 2007).

CHAPTER 7 METHODS AND ANALYSES FOR TRANSIT DEMAND MODELING

As mentioned in previous chapters, data from the smart card AFC system need additional processing to derive information not captured by the system. Spatial-temporal description of boardings is presented in Chapter 6. This chapter proposes data enrichment methods to derive additional trip details for transit demand analyses and modeling.

7.1 Reconstructing an Itinerary

7.1.1 Definition of an Itinerary

The disaggregate property of the smart card validation data enables planners to reconstruct individual itineraries from separate trip segments. The concept of itinerary has been used in travel survey and public transit operations planning model, such as MADITUC (Chapleau, Allard & Canova, 1982), and is illustrated in Figure 7.1. Itineraries are considered superior to an OD matrix because they not only store the origin and the destination of a trip, but also the departing time, route choice, transfer activity and cardholder's attributes in a disaggregate form. In addition, itineraries can readily be aggregated into one or several OD matrices. This enrichment makes the analysis of linked trips, trip chains, activity space and activity duration possible.

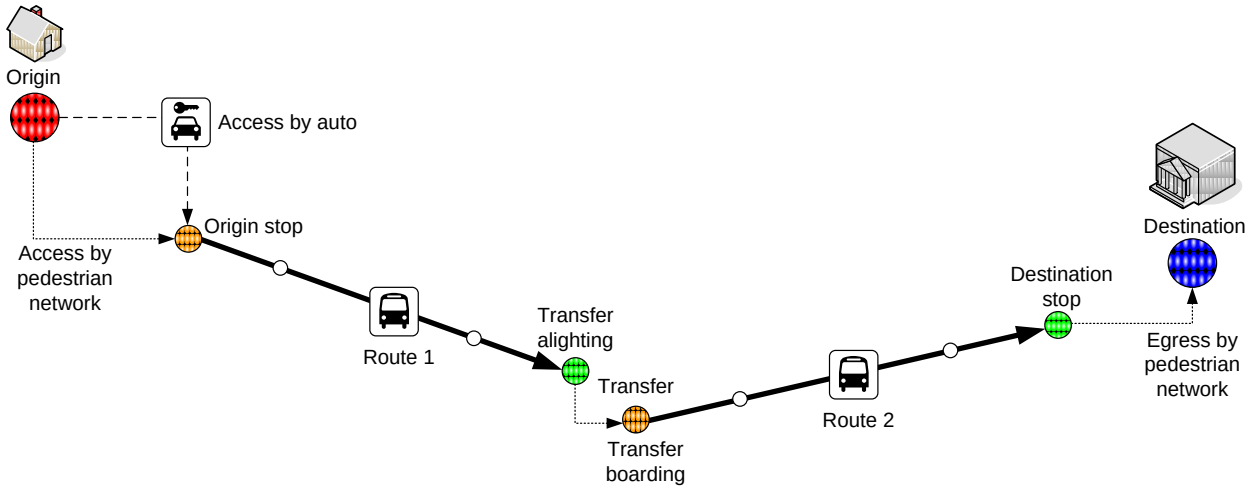


Figure 7.1 The schematic definition of an itinerary of a linked trip (Chu & Chapleau, 2008).

Each itinerary constitutes a person-trip, which is defined as a one-way movement of a person between two points, namely the origin and the destination, for a specific purpose (Kittleson &

Associates, Inc. et al., 2003). Since the actual origin and destination are not known, this chapter focuses only on the origin and the estimated destination stops. In the Chapter 8, methodology is proposed to estimate actual trip origin and destination under certain assumptions.

7.1.2 Determining the Most Probable Alighting Stop

Trépanier et al. (2007) propose a methodology to estimate the most probable alighting stop for a given boarding. The methodology considers the route geometry by only allowing alighting on the "vanishing route", which is defined as the set of stops to be visited by the vehicle after the boarding stop. This procedure requires a correct route dictionary which lists the stop sequence of each particular route-direction. The results show that more than 92% of the alighting locations lie within 1,000 metres of the next boarding stop and more than 81% of them within 300 metres. However, there are some minor drawbacks in the algorithm:

- The most-probable alighting stop is validated by comparing the distance between the alighting stop and the next boarding stop. In cases where there are missing boarding transaction due to equipment malfunction or the trip is made using another transportation mode, the algorithm can provide result that is acceptable in distance but not in logic. For example, if the next boarding stop is the same stop as the current boarding stop, which happens when the returning trip is not made by transit, the estimated alighting stop would be the next stop in the vanishing route. The individual thus travels only one stop on the vehicle.
- By the same token, estimated alighting stop that are outside the distance threshold are not re-estimated.

A modified version of the algorithm is used in the later stage of the research. It is based on the multi-day boarding records of an individual. The algorithm breaks the transactions into pairs of two boardings. The unit is formed with 2 successive boardings within the same day regardless of time of transaction. The unit is invalid when the distance between two successive boarding stop is less than 500 meters. Meanwhile, a valid alighting stop must be within 2,000 meters of the next boarding stop in the pair. When a pair cannot be formed, the following rules are used:

- If the boarding is the last transaction of a day, the chain will be form using the first transaction of the same day as the next boarding. If it is not present or not valid, it will use

the first transaction of the next day as the next boarding. If the latter is not present or not valid, it will use the most frequent first boarding location of the day within the analysis period. No pair can be defined if there is no other transaction. Estimation of the most probable alighting stop is done based on each pair.

- If no valid alighting stop can be found, the algorithm will use the individual's most frequent boarding stop as the next boarding stop of the pair. If the result is not valid, it will assign the most frequently-used alighting stop (by all users) in the vanishing route as the alighting stop for the boarding. Assuming that boarding and alighting are symmetric on both directions of the route, the most frequently-used alighting stop is the stop closest to the most frequently-used boarding stop of the opposite direction. An alighting stop cannot be estimated if there are not enough boarding transactions within the analysis period.

Chapter 8 shows the reason why it would be essential to integrate the multi-day travel pattern into the alighting estimation algorithm.

7.2 Estimating Vehicle Spatial-temporal Path with Timetable Data

7.2.1 Running Time Estimation

With an objective to more accurately identify transfer activities and to construct spatial-temporal load profiles for bus runs, the spatial-temporal path of the vehicles needs to be estimated from partial information. The procedure aims to take into account all available temporal constraints in order to produce a likely estimate. With the knowledge of scheduled departure and arrival times for each run, the stop sequence and the linear distance between stops, several assumptions are made in the estimation process. First, the vehicle departs from the terminus according to the scheduled departure time unless information contained in the validation data dictate otherwise. Second, cardholders' boarding times on their next routes serve as the upper bounds at the estimated alighting stops. Third, given the constraints, linear interpolation is used to estimate the arrival time at stops where there is no boarding. It takes into account the first and the last transaction times at each stop and the distance between stops. Fourth, planned running time is used to extrapolate vehicle path beyond the last boarding of a run. Most transit agencies dispose planned departure time at check points. Using this information, the running time between each

stop can be interpolated. The arrival time beyond the last boarding would be equal to the last boarding time plus the planned running time. This procedure assumes that the vehicle moves toward the terminus at the planned speed, regardless if it is late or in advance with respect to the schedule. The more boarding validations there are along the route, the more accurate the estimation would be. The complexity of this problem can be greatly reduced if detailed GPS logs of vehicles are available and fused with the validation records.

7.2.2 Spatial-temporal paths of vehicles

Table 7.1 shows the resulting arrival time at the first 11 stops (out of 70) of the inbound route 44. The values in bold indicate actual boardings at the stops and only the first transaction time at each stop is shown. Other arrival times are estimated based on the assumptions described above. This enrichment process is applied to all runs on a typical weekday. Plausible spatial-temporal paths of all vehicles are therefore reconstructed at the stop level.

Table 7.1 Results of running time estimation. Bold values denote times from actual boarding validations (Chu & Chapleau, 2008).

Route	Direction	Planned Departure Time	Stop 0 Terminus	Stop 1	Stop 2	Stop 3	Stop 4	Stop 5	Stop 6	Stop 7	Stop 8	Stop 9	Stop 10
44	0	5:50	**5:49**	5:51	**5:51**	5:52	5:52	**5:53**	**5:54**	5:54	5:55	**5:55**	5:56
44	0	6:20	**6:17**	6:21	**6:22**	6:23	**6:23**	**6:24**	6:25	**6:25**	6:26	**6:26**	6:26
44	0	6:38	6:38	6:39	6:40	6:41	**6:42**	**6:43**	6:44	6:44	6:45	6:45	**6:45**
44	0	6:52	6:52	6:54	**6:55**	6:56	**6:56**	**6:57**	**6:58**	6:58	**6:59**	**7:00**	**7:01**
44	0	7:10	**7:10**	7:11	**7:12**	7:13	**7:13**	**7:15**	7:16	7:16	7:17	7:17	7:18
44	0	7:24	**7:24**	7:27	**7:28**	7:29	**7:29**	**7:30**	**7:31**	7:31	7:32	7:32	7:33
44	0	7:36	**7:33**	7:36	**7:36**	7:38	**7:38**	**7:40**	**7:40**	**7:41**	7:42	7:42	7:43
44	0	7:49	**7:48**	7:50	7:51	7:51	**7:52**	**7:53**	**7:53**	**7:54**	**7:55**	7:55	**7:56**
44	0	8:00	8:00	8:01	**8:02**	8:03	**8:03**	8:03	**8:04**	8:04	**8:05**	**8:06**	8:06
44	0	8:33	**8:33**	8:34	**8:34**	8:35	**8:35**	**8:36**	8:37	8:37	8:37	8:38	8:38
44	0	9:00	**9:00**	9:04	9:05	9:05	**9:06**	9:07	**9:08**	9:08	9:09	9:09	9:10

7.3 Identifying Transfer Activities

A linked trip in public transit refers to a person-trip that is carried out with more than one route on a given day. To reconstruct the itineraries and to identify the origin and the destination stops, one must correctly identify linked-trip, which involves an alighting stop from the first route and a boarding stop from the following route. These two stops usually have to be physically close and accessible on foot. Also, there should also be a temporal relationship between the time of alighting and the time of next boarding. Ideally, cardholders are rational and have the complete knowledge of the network and timetable. An ideal transfer activity can be revealed by a spatial-temporal coincidence where the cardholder boards the first run of the desired route after his/her arrival at the boarding stop. It is assumed that there is no capacity constraint, meaning that the first vehicle has enough capacity for all the transferring cardholders.

7.3.1 Algorithm to Detect Transfer Coincidence

The complete spatial-temporal paths of the vehicles in the network allow planners to detect transfer coincidence. Since each pair of boarding and alighting is associated with a run, the estimated alighting stop is assigned an alighting time according to the running time information. This time is then compared with the next boarding time of the same cardholder. In order to take into account the transfer access distance between the alighting and the boarding stop, the transfer access time is estimated using an average walk speed of 1.2 m/s and the straight-line distance between the stops. An algorithm is set up to check whether the run boarded by the transferring cardholder represents the first run upon his/her arrival at the stop. Figure 7.2 illustrates the concept.

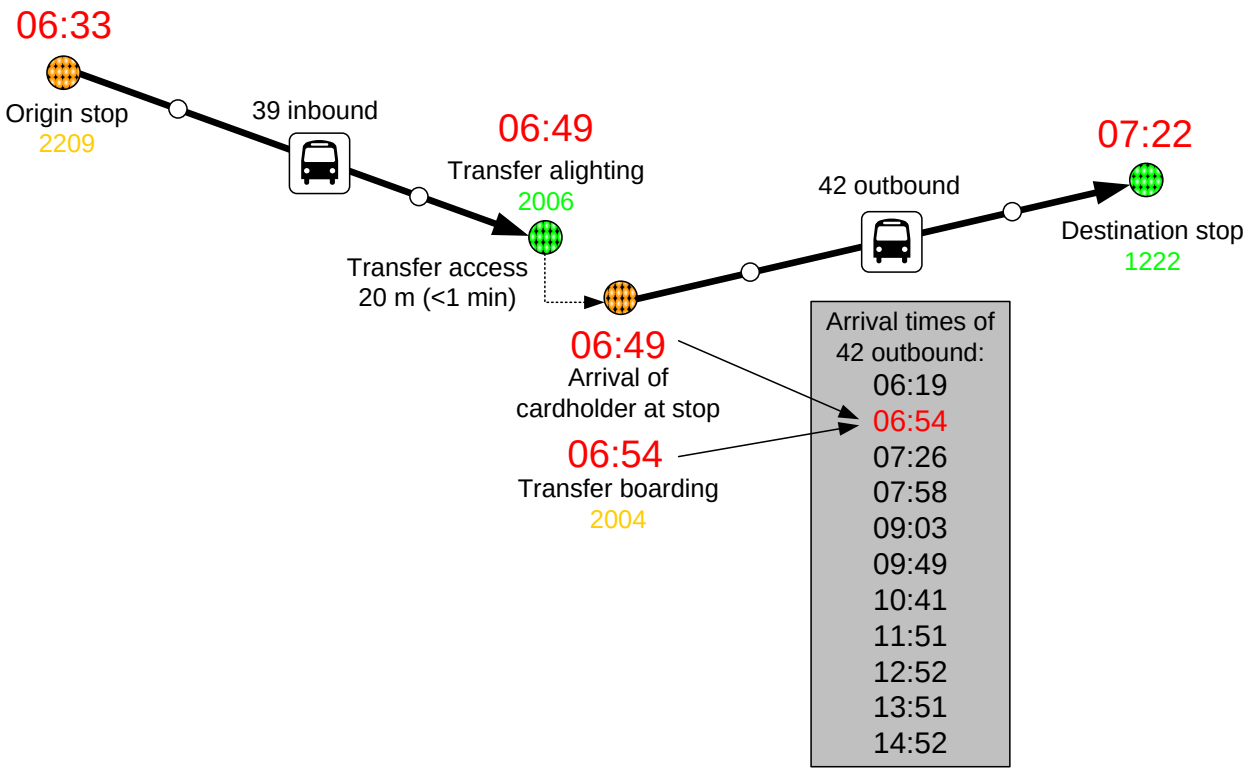


Figure 7.2 An example of a transfer coincidence (Chu & Chapleau, 2008).

To compensate for the fact that some cardholders do not have perfect knowledge of the network and timetable, the variation in walking speed as well as the capacity constraint for routes with a high frequency, a temporal leeway of 5 minutes is added to the cardholder's arrival time at the stop. This allows the algorithm to take into account those uncertainties. Routes with long headway are assumed to have a low ridership and enough capacity for transferring passengers.

The algorithm also looks at the routes taken by the cardholder. Even when the alighting-boarding pair satisfies the conditions of a spatial-temporal coincidence, the route taken must not be same in both boardings, regardless of route direction. Two successive boardings onto the same route and the same direction may indicate a stopover with its own trip purpose while a different direction suggests that the second boarding is a return trip. Variants or deviations of routes are also considered. Future development of the algorithm can take into account not only the route information, but also the proximity of the departure and arrival stops in order to detect return trips.

7.4 Cardholder Itinerary with Trip Details

7.4.1 Reconstruction of Cardholder Itinerary

Once the transfer has been identified, the public transit segments of the cardholder itinerary can be reconstructed. An itinerary of a person-trip contains a boarding stop and time, an estimated alighting stop and time, and in linked trips, transfer alighting and boarding stops and their associated times. These data allow the derivation of additional information and indicators for objects such as a person-trip, a vehicle-trip and activity.

7.4.2 Activity Duration

Activity duration can be more precisely defined as the period between alighting time at the destination stop and the boarding time at the origin stop of the next trip. Access and egress times cannot be considered as the trip ends cannot be identified.

7.4.3 Distance traveled and Trip Duration

Distance traveled on-board transit vehicles, measured in passenger-kilometres, is used in network usage estimation and revenue sharing mechanism. Circuity, which is a ratio of the actual distance traveled by the user against the shortest path, can be a relevant indicator of effectiveness of the network geometry with respect to the demand.

Trip duration or travel time is an indicator similar to distance traveled, but measured in time instead of distance. It comprises wait time and in-vehicle time. It is a measure which can be compared with other modes of transportation in order to model the competitiveness of public transit. With trip distance and duration, average in-vehicle speed or trip speed can be calculated.

7.4.4 Example of Itineraries from a Cardholder

The boarding validations are classified into linked and non-linked trips. In total, the algorithm recognizes 33,775 person-trips on February 10, 2005. Figure 7.3 provides a three-dimensional GIS representation of the trips that a cardholder made during the day. The paths of the vehicles are shown in pink and the actual movement of the cardholder is shown in blue. Derived activities, including transfer movements, are shown in orange. Table 7.2 displays the detail of each

analytical component. The cardholder made six transactions in total of which three are considered as transfers. There are three person-trips with different trip purposes. In-vehicle time, transfer time as well as activity duration are derived according to the running time information. The cardholder spent 111 minutes on-board and traveled about 48 kilometres, giving an in-vehicle speed of 25.9 km/h. When transfer time is considered, the overall speed decreases to 22.1 km/h. The reconstruction would not be possible if data quality was poor or if the data had not been enriched. All this information contributes to the disaggregate analysis of travel and activity patterns.

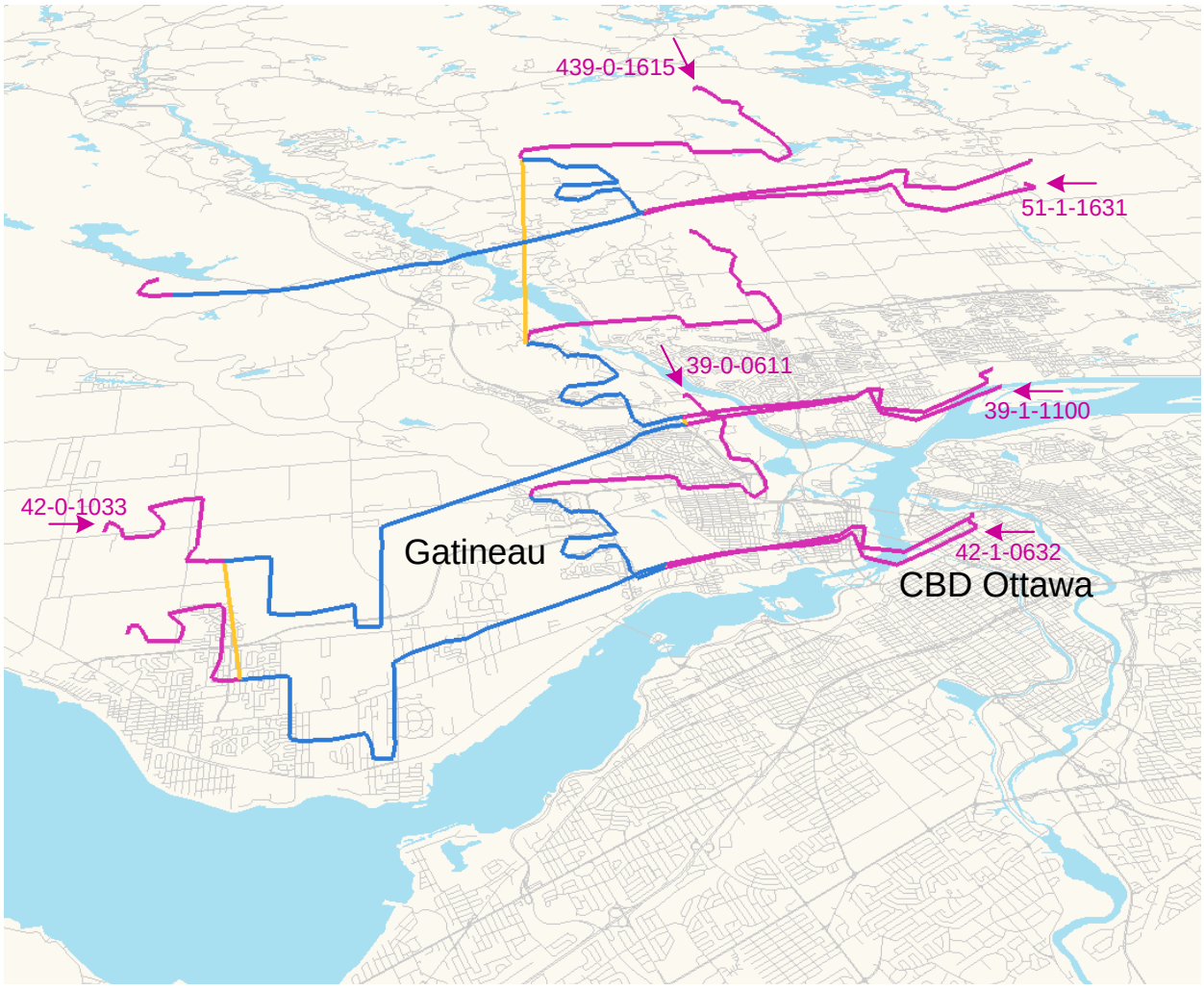


Figure 7.3 A three-dimensional representation of the itineraries of a cardholder (Chu & Chapleau, 2008).

Table 7.2 Derived trip and activity details (Chu & Chapleau, 2008).

Itineraries	Start	End	Duration (minute)	Derived Activity	Bus Run	Stop Number	Associated Objects	In-vehicle Distance (m)	Access Distance (m)	Travel Speed (km/h)	Trip In-vehicle Distance (m)	Trip In-vehicle Time (minute)	Trip Travel Speed (km/h)	Trip Time (minute)	Trip Travel Speed (km/h)
Trip 1	06:33	06:49	16	In-vehicle	39-0-0611	2209	Bus 9129	6651		24.9	17708	44	24.1	49	21.7
	06:49	06:54	5	Transfer		2006			20						
	06:54	07:22	28	In-vehicle	42-1-0632	2004	Bus 9421	11057		23.7					
	07:22	10:46	204	Major Activity		1222			68						
Trip 2	10:46	11:06	20	In-vehicle	42-0-1033	1220	Bus 309	10829		32.5	17503	34	30.9	47	22.3
	11:06	11:19	13	Transfer		2025			45						
	11:19	11:33	14	In-vehicle	39-1-1100	2023	Bus 9410	6674		28.6					
	11:33	16:39	306	Major Activity		2211			40						
Trip 3	16:39	16:55	16	In-vehicle	439-0-1615	2213	Bus 8905	5909		22.2	12731	33	23.1	34	22.5
	16:55	16:56	1	Transfer		2002			93						
	16:56	17:13	17	In-vehicle	51-1-1631	2000	Bus 9417	6822		24.1					
	17:13					1222			68						
Total											47942	111	25.9	130	22.1

7.5 Applications for Transit Planning

7.5.1 Transfer Analysis

Transfer is an integral component of a modern transit network, which is often multi-modal. Although transfer allows users to reach a greater number of origins and destinations, research reveals that transfer, characterized by additional wait and access times, reduces the attractiveness of transit. Therefore, it is imperative to get a comprehensive understanding of the transfer pattern in the network. Indicators such as the number of transfers, transfer access distance and transfer wait time need to be studied.

In the STO network, a boarding is automatically labeled as a transfer if a validation is made within 120 minutes of the first fare validation of the cardholder's boarding chain. Comparison between the transfer boardings identified by the smart card system and by transfer coincidence is essential to understand the level of overestimation of transfer activities reported by the system. Of the 37,781 transactions, 5,516 (14.6%) were originally labeled as transfer boardings by the

system. Following the definition of transfer coincidence, the algorithm identifies 4,002 transactions (10.6%) as transfer boardings. The experimental result suggests that the definition used by the AFC system overestimates transfer trips by nearly 40%. The implication is that detecting transfer with a simple temporal threshold may not be sufficient to accurately detect transfer activities. Each transfer activity varies depending on the headway of the route and the transfer access distance between stops. A fixed temporal threshold cannot capture the subtlety of those interactions. An alternative to a fixed temporal threshold is to use specific thresholds for each type of transfer in a multi-modal transit network, as proposed by Seaborn et al. (2009).

Analysis of derived transfer wait time allows researchers to study user behaviours on transfer and transfer efficiency. Figure 7.4 shows the temporal distribution of the derived transfer time, which is defined as the time difference between the cardholder's estimated alighting time on the previous route and actual transaction time at the boarding stop. The general trend shows that as transfer time increases, the number of observations decreases. The cumulative percentage curve also reveals that about 50 percent of transfer trips have a transfer time of 7 minutes or less while over 80 percent have a transfer time of 18 minutes or less. The number of observations with transfer time between 0 and 1 minute can be explained by the fact that some transaction times are used as upper bounds for vehicle running times. Observations with a long derived transfer wait time (for example: 60 minutes or greater) can be explained by routes with long headway. The latter can be the result of a route operating only part of the day or a route following slightly different paths but coded with different route numbers. The transfer identification algorithm can be improved if these special cases are taken into account.

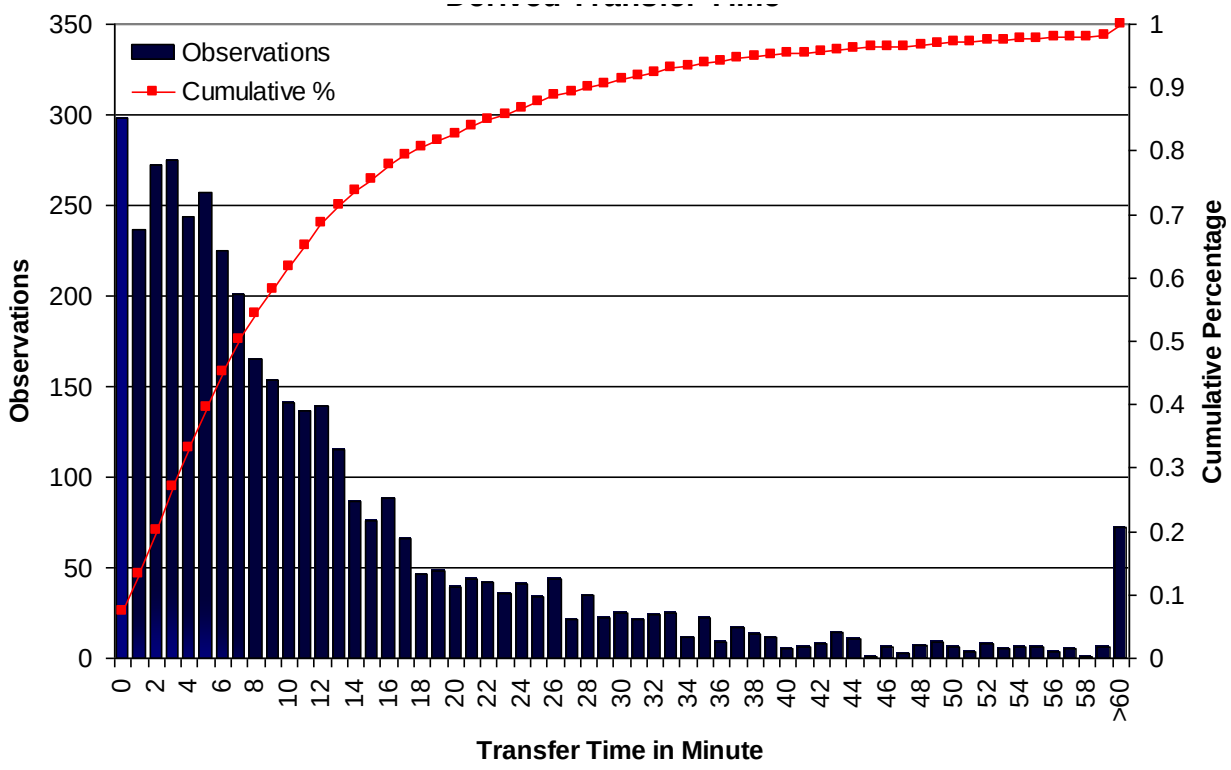


Figure 7.4 The temporal distribution and the cumulative percentage of the derived transfer times (Chu & Chapleau, 2008).

While transfer time reveals the whole duration of the transfer activity, transfer wait time takes into account the transfer access distance between the alighting stop and the boarding stop. Wait time is defined as the time difference between the cardholder's estimated arrival time at the boarding stop and the actual boarding transaction time. Figure 7.5 shows the ratio between the derived wait time and the planned headway of the connecting route. This is possible because of the disaggregate nature of the data and the analysis takes into account 3,405 observations. Undefined ratios are mainly due to the absence of a previous run or excessively large headways from routes which operate only part of the day. Observations for headway class under 5 minutes is not shown because the number of observations is low and the uncertainties involved with walk speed and estimated alighting times. Theoretically, the expected wait time for users with a uniform arrival distribution is half the headway. The figure reveals that the average ratio for each headway class is lower than 0.5. Furthermore, as the planned headway increases, the ratio tends to decrease. This suggests that users have a shorter than expected wait time. In fact, the average ratio of 0.31 means that users wait about 40% less than the expected wait time. Possible

explanations include users' familiarity of the network and timetable as well as coordinated transfers integrated in the timetable.

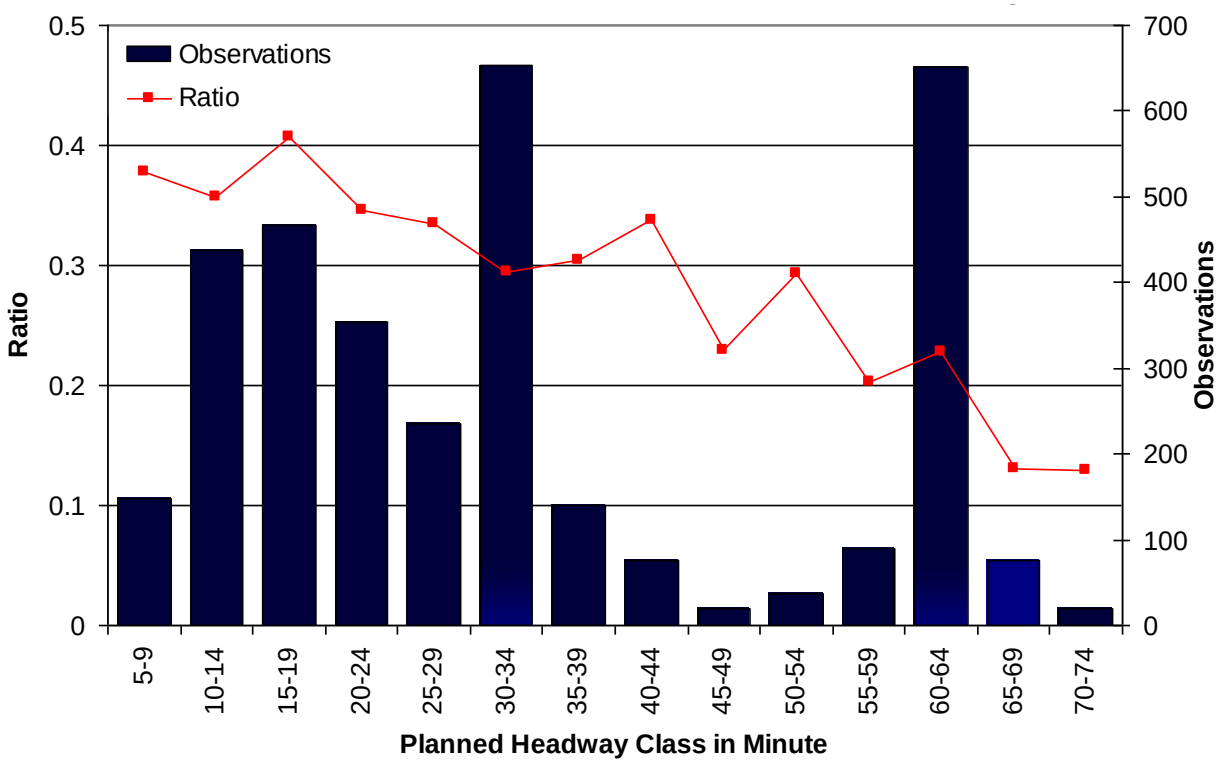


Figure 7.5 The ratio between derived transfer wait time and planned headway (Chu & Chapleau, 2008).

Figure 7.6 illustrates the spatial distribution of all boardings. Due to the network size and level of detail, only part of the network is shown. The height of the column indicates the number of boardings at each stop. The part in green shows the proportion of first (non-transfer) boardings and the red colour shows the proportion of transfer boardings. The cluster of transfer boardings are consistent with major transfer points published in the user guide. Meanwhile, route by route transfer patterns computed according to time of the day and stops would be useful to planners in order to prioritize timetable coordination.

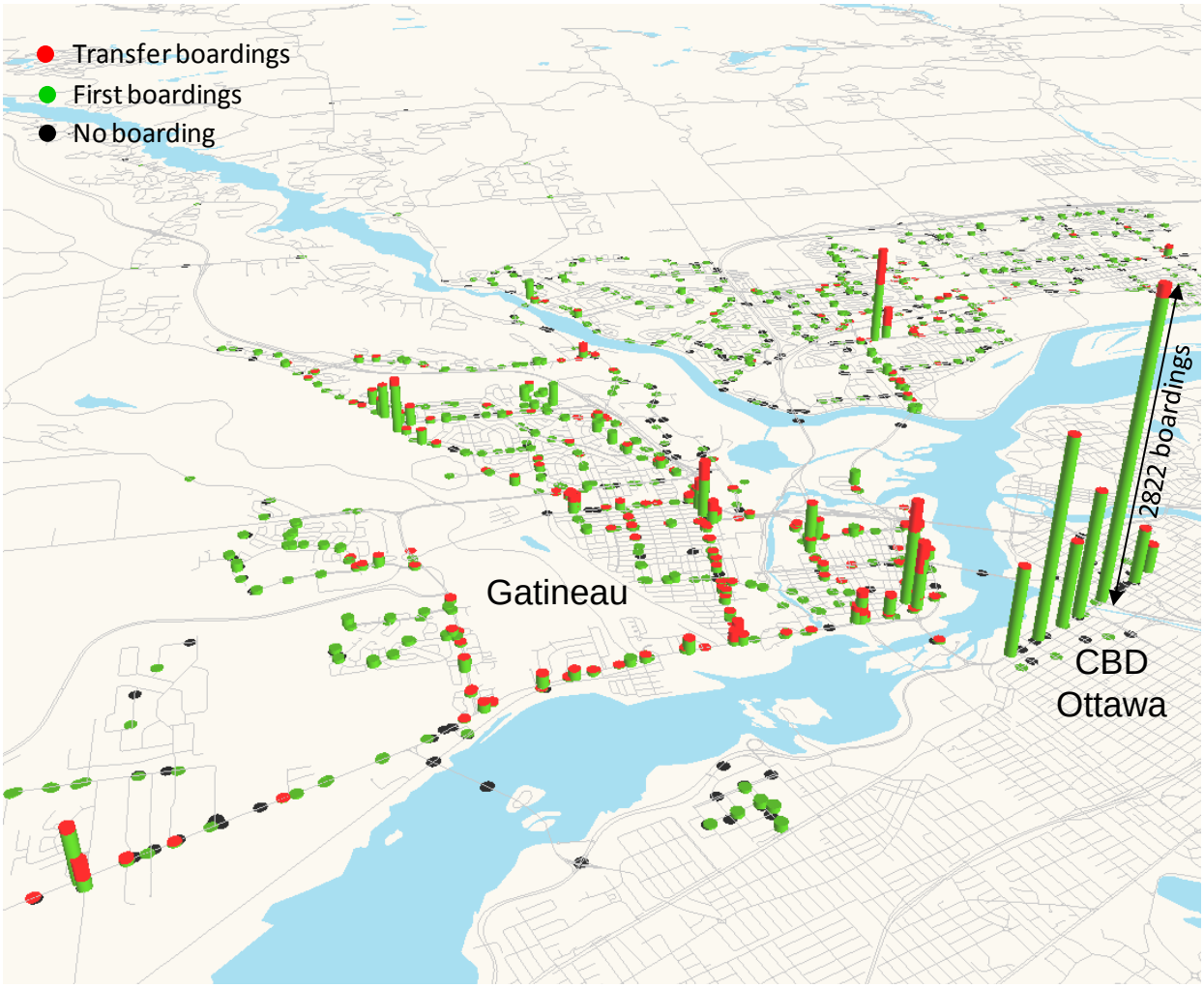


Figure 7.6 The spatial distribution of the first and transfer boardings in a typical day (Chu & Chapleau, 2008).

7.5.2 Load profiles

Load profile remains one of the most fundamental tools in transit planning. Figure 7.7 illustrates a stop-level load profile of a typical run using enriched smart card validation data. The top-half of the figure shows ridership at an aggregated level: the cumulative number of boardings, the cumulative number of alightings and the resulting load. The bottom-half displays the OD information in a disaggregate manner illustrating the boarding and the alighting locations of each cardholder. The last transaction times at each boarding stop are shown. There are 21 boardings in this run and the maximum load is 12. Passenger-kilometres of this run can easily be obtained by calculating the area under the load profile or by adding the length of each individual OD pair.

The main advantage of this load profile compared to a traditional one is the availability of disaggregate data on run, OD, time and distance.

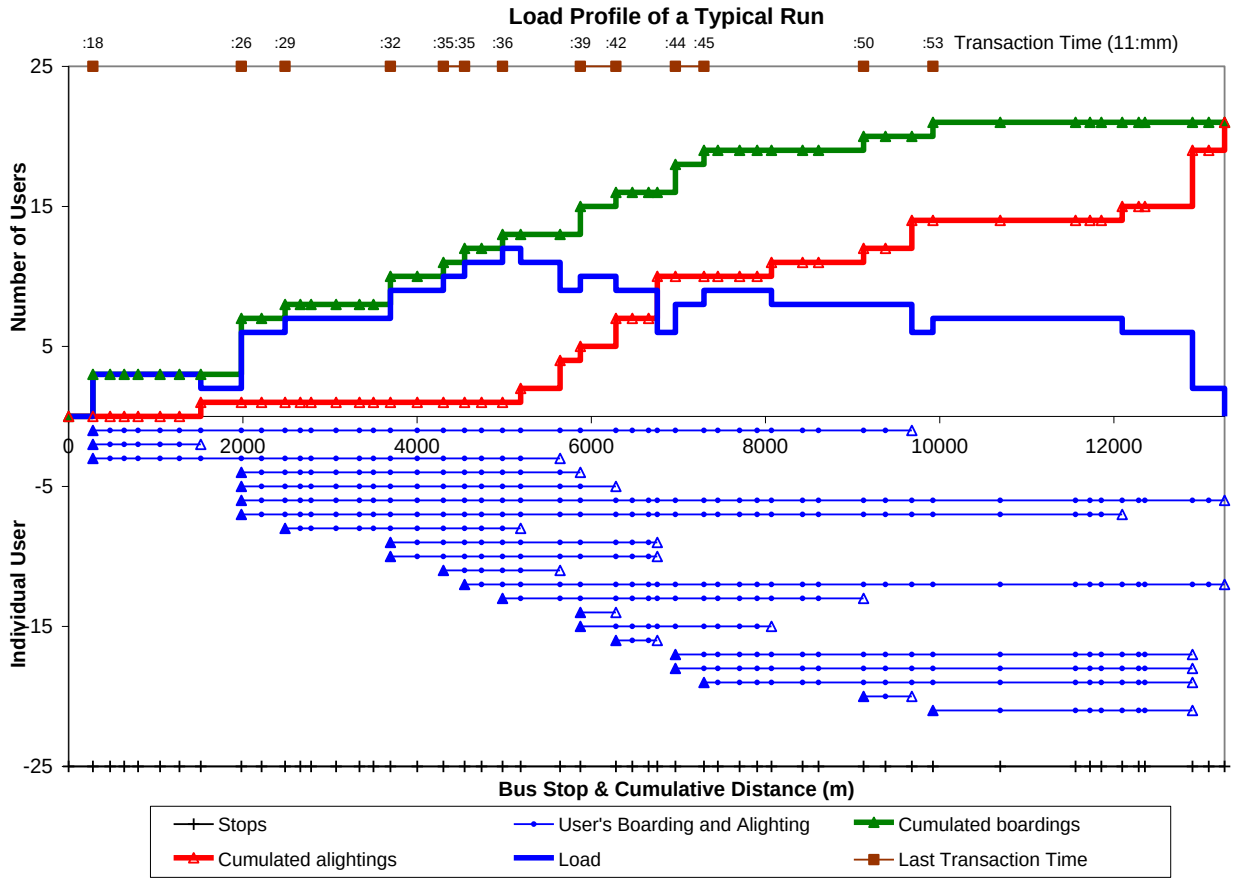


Figure 7.7 Load profile of a run showing ridership at both aggregate and disaggregate levels (Chu & Chapleau, 2008).

Planners can benefit from further integrating information from time and successive runs of the same route. When combining the load profiles of several runs, one can create the complete three-dimensional spatial-temporal load profiles of a route at the stop level during the course of a day. The required labour and resource would be prohibitive if done with manual data collection methods. The information allows planners to examine the within-day variation of the demand, changes in location of the maximum load point (or more precisely the maximum load inter-stop segment) as well as to perform schedule adherence analysis. Figure 7.8 shows the load profiles of all inbound runs of route 44 on a typical weekday. This inbound route travels from the Aylmer sector of Gatineau to the CBD of Ottawa and operates only during the morning peak. There are 11 runs in total from 5:50 AM to 9:49 AM. The x-axis shows the linearized distance of the route

along with the check points. Planned departure times and the planned arrival times at both termini are shown besides the blue dots on the y-axis. The loads of the vehicles are illustrated in various shades of colour along the z-axis, with the maximum load point shown in the darkest colour. The dash lines in blue represent the interpolated scheduled trajectory of the runs. The dash lines in red and orange delimit the zone where the vehicle is less than 2 minutes in advance and less than 5 minutes behind schedule. As mentioned earlier, actual transaction times are only available up to the last boarding of the run. The remainders come from estimated vehicle paths with the assumption that the vehicle arrives at the terminus at the planned arrival time if the resulting speed of the vehicles is reasonable.

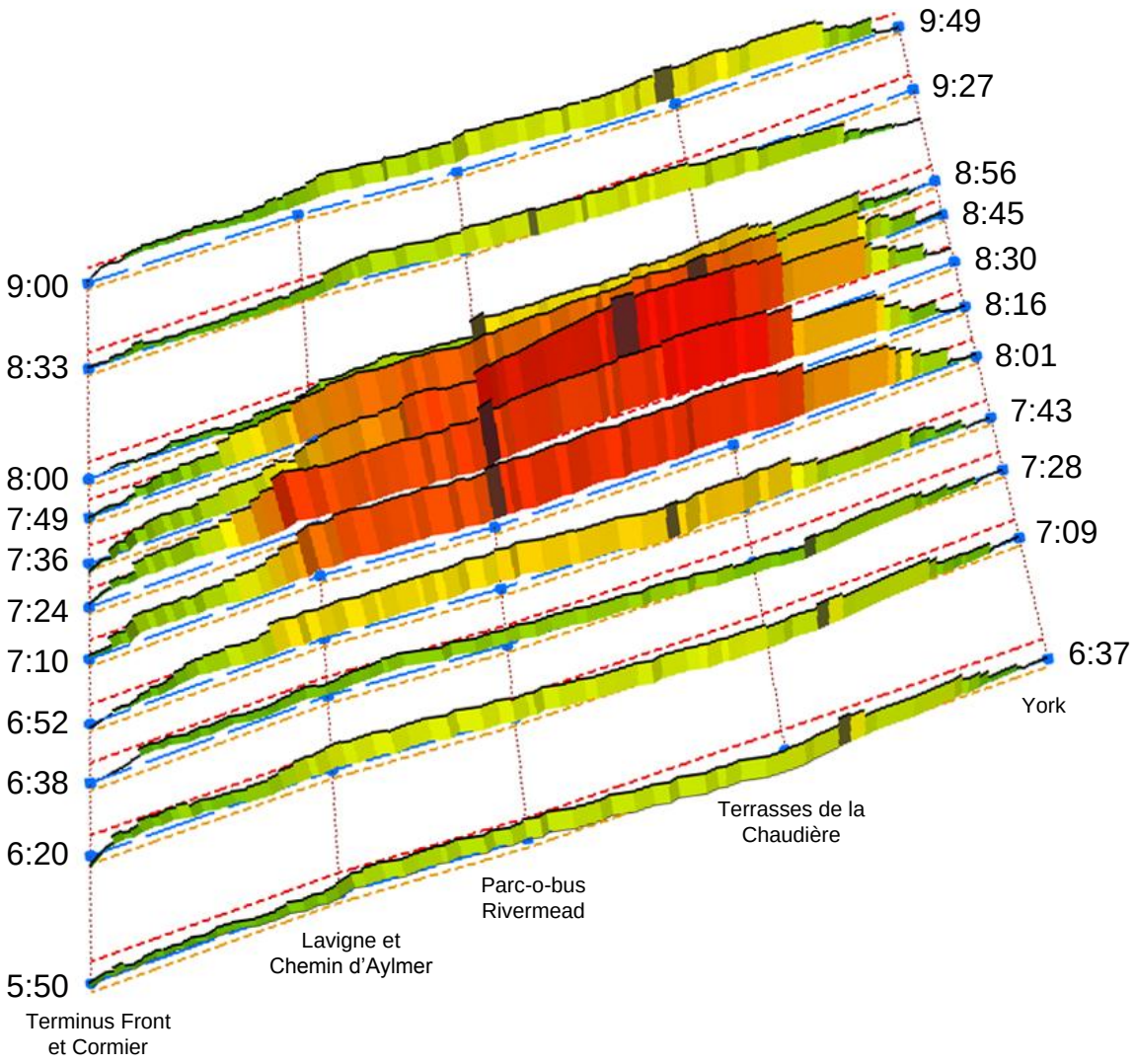


Figure 7.8 Three-dimensional stop-level spatial-temporal load profiles of inbound route 44 (Chu & Chapleau, 2008).

Figure 7.9 depicts a similar analysis on route 64 direction Gabrielle-Roy during the AM peak on a typical weekday. It differs from the previous analysis in two aspects. First, after the last boarding of the run, the vehicle movement follows the running time of the timetable. Second, an extra dimension on fare type allows analysts to distinguish the travel behaviour of different groups of clientele. Route 64 is 17.7-km in length and transports users within the City of Gatineau. The terminus Gabrielle-Roy is located amid several educational institutions and this characteristic is revealed by the predominant presence of a student population (in darker blue). 158 out of 186 boardings (85%) are cardholders with student fare. The runs encounter several

delays with respect to the schedule during the morning peak. The paths suggest a recurrent problem between the checkpoints Saint-Louis et Avenue Gatineau and Freeman et Saint-Joseph as several runs require more time than planned. The Alonzo Wright Bridge located between the checkpoints may explain the delay. The maximum load sections (in black) vary in space and are usually located towards the end of the route.

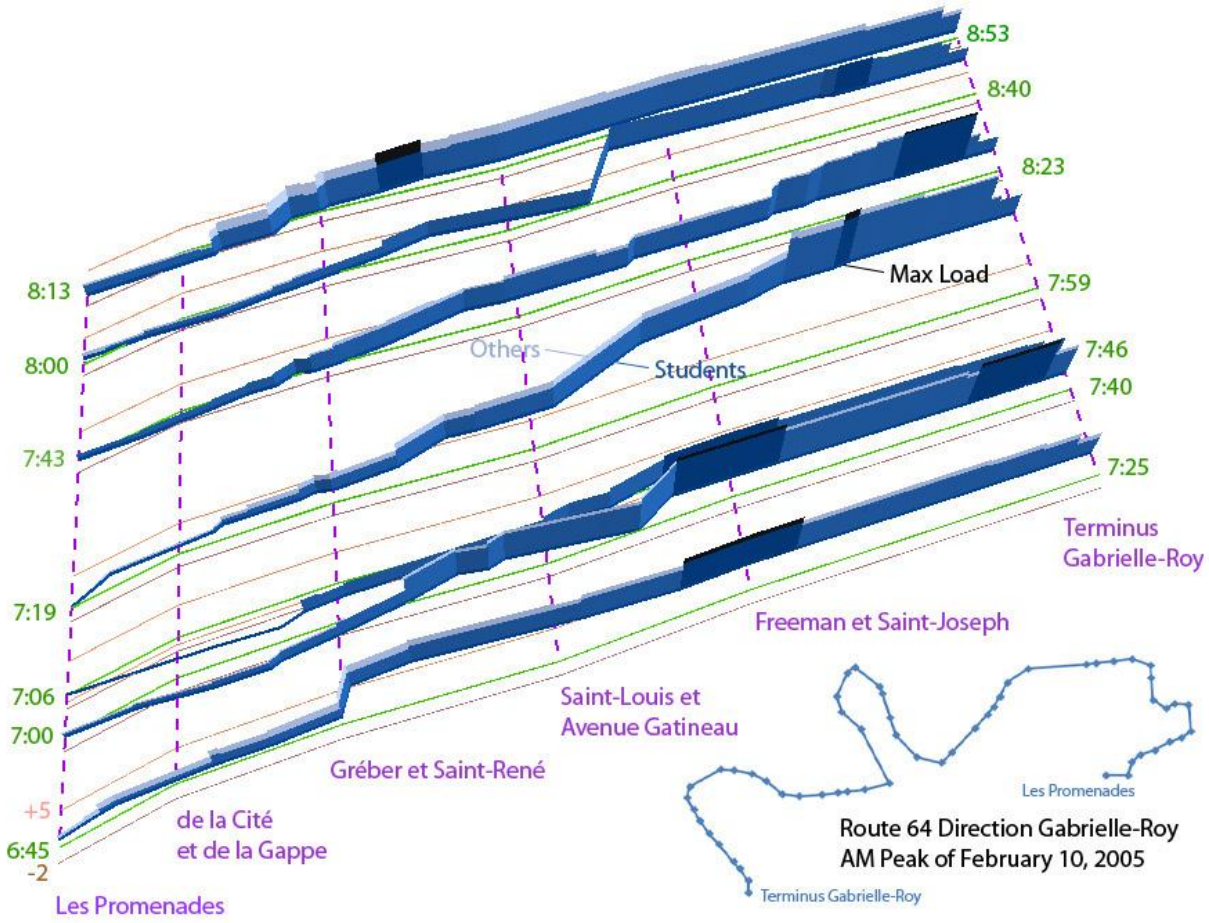


Figure 7.9 Three-dimensional stop-level spatial-temporal load profiles of route 64 direction Gabrielle-Roy.

7.5.3 Detailed Analysis of a Route

The enrichment techniques allow the calculation of indicators on users, on-time performance and trip distance by run. Table 7.3 shows some of the performance indicators for all the 11 runs of inbound route 44. They result from synthesizing validation records and the following features:

- Estimated alighting stop for each individual boarding;

- Planned arrival time at each stop by interpolating time table of check points;
- Estimated arrival time at each stop of the vehicle;
- Distance traveled for each trip;
- Aggregation of stops into route segments.

Table 7.3 Performance indicators for inbound route 44.

Route 44 Direction Ottawa Run	Boarding	% First Boarding	% Transfer Boarding	Adult	Senior	Students over 21	Students 21 or under
5:50	28	82%	18%	28			
6:20	28	89%	11%	26		1	1
6:38	20	70%	30%	17	1		2
6:52	28	100%	0%	23			5
7:10	41	100%	0%	32		3	6
7:24	54	100%	0%	47	1	1	5
7:36	61	100%	0%	53		2	6
7:49	55	98%	2%	39	2	1	13
8:00	37	100%	0%	34			3
8:33	27	100%	0%	22			5
9:00	27	89%	11%	18		3	6

Route 44 Direction Ottawa Run	Boarding at Segment 1	Boarding at Segment 2	Boarding at Segment 3	Boarding at Segment 4	Average delay at boarding (mins)	Average delay at aligting (mins)	Boarding delay outside -2 to 5	Sum of Distance Traveled (km)	Average Distance Traveled (km)	Users using the reverse direction
5:50	15	7		6	-0.6	-1.7		250.8	9.0	4.0
6:20	22	2	1	3	0.0	-2.0		292.5	10.4	7.0
6:38	9	4		7	0.5	-2.2		182.9	9.1	4.0
6:52	23	3	2		1.2	-0.1		354.1	12.6	7.0
7:10	36	5			1.2	0.6		561.9	13.7	12.0
7:24	41	8	5		2.9	5.3	4	570.7	10.6	11.0
7:36	30	21	5	5	2.3	6.4	10	547.7	9.0	19.0
7:49	33	17	3	2	0.9	1.9		499.7	9.1	6.0
8:00	13	19	5		-0.1	-1.3		273.9	7.4	6.0
8:33	14	10	2	1	-1.1	-3.4		234.7	8.7	2.0
9:00	12	8	6	1	3.5	3.3	1	265.1	9.8	2.0

Indicators on boarding, first boarding or transfer boarding come from results of queries on corrected data. The numbers shows that the run of 7:36 has the highest ridership even though it has the shortest headway at only 12 minutes. The route receives few transfer boardings which are not evenly distributed among the runs. Some are present at the beginning and at the end of the peak period but are almost absent between 6:52 and 8:33. It may be explained by a higher level of service or connectivity during the middle of the peak period.

Indicators regarding user category and boarding by route segment require the aggregation of fare types and stops into route segments respectively. Adults represent the majority of the users of this route while students account up to one third of ridership in some runs. More than 61% of the boardings are located within the first route segment and more than 86% are located within the first two.

Smart card validation data and scheduled arrival time at the stop level allow the calculation of on-time performance indicators. It can be weighted by the number of boardings at each stop. Delay, or advance (denoted by a negative sign), at boarding is calculated by subtracting the planned arrival time of the vehicle at the stop from the actual transaction time of the boarding. With the estimated vehicle paths, on-time performance can be assessed at locations where there is no boarding but may have important alighting activities, such as check points and at the termini. Schedule adherence for alighting uses planned arrival time of the vehicle and the estimated alighting time of the user. The two measures are computed for each transaction and the average values are shown by run. Three runs have negative value for boarding meaning that on average, the vehicles arrive earlier than planned. Three runs have average delay at boarding greater than 2 minutes and the same number of runs has average delay at alighting greater than 2 minutes. If the criteria on-time performance at boarding has 2-minute advance and 5-minutes delay thresholds, 10 out of 61 boardings from the 7:36 run are considered inacceptable. The on-time performance for this run is 84%. It coincides with the run with the highest ridership.

Distance traveled for a user is the on-board distance between the boarding stop and the estimated alighting stop. The total distance traveled, measured in passenger-kilometres, is the sum of distance for all individuals. It is interesting to notice that the average distance traveled varies significantly on some runs. On the other hand, only 80 out of 406 users (20%) take both inbound and outbound route 44 on this day, meaning most of them take another route for the return trip or travel on another mode altogether.

The planned trip duration can also be compared against actual trip duration for individual itinerary. The planned on-board duration comes from the timetable whereas the actual on-board duration can be derived from estimated vehicle paths.

7.6 Transit Network Analysis

7.6.1 Transit Assignment with a Totally Disaggregated Approach

A transit network deals with flows of passengers. It is usually modeled by a graph composed of links and nodes. When coded at the stop level in automobile network standard, links and nodes permitting transfers become extremely complex. As an illustration, an intersection with 4 route-directions and one stop at each corner requires a total of 12 links. It is not practical to model a complete public transit network with this standard. A solution is to use a MADITUC node to aggregate stops in proximity and allow free movement within the group of stops (Chapleau et al., 1982). Developed in the 1980s, the concept underlying a MADITUC node is that stops within walking distance are aggregated to form a new node within which free transfers are allowed.

Figure 7.10 shows a specific location where, according to the derived information, a large number of boardings and alightings occur. The points with the arrow pointer symbolize the revealed boarding stops whereas the other ends show the derived alighting stops. The numbers beside the arrows indicate the movement between pairs of bus routes. The short access distance among the routes suggests transfer movements. For a typical route-based transit trip assignment model, these seven bus stops can be coded as a single node (a radius of around 100 metres).

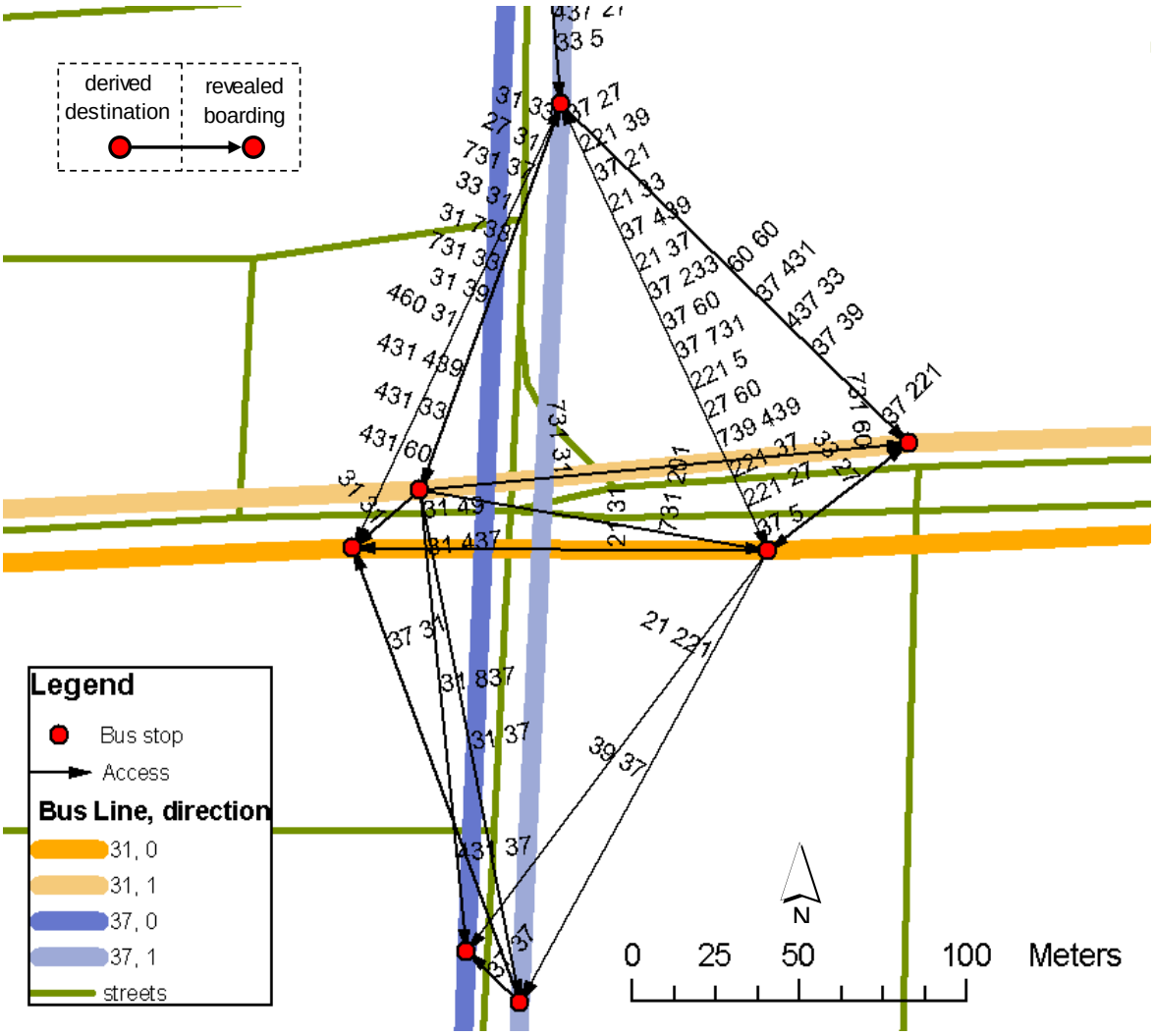


Figure 7.10 An intersection with several transit routes providing multiple transfer possibilities (Chapleau & Chu, 2007).

Based on this concept, instead of using the original route data based on stops, routes are recoded according to the new nodes. Itineraries derived from smart card data can either be loaded onto the network according to the declared routes or be used to calibrate a model for transit assignment simulation. Example of this aggregation from the STO network is shown in Figure 7.11. The highlighted node has 8 individual stops and 69 route-stops. Transfer movement revealed by smart card transaction data are shown in red. Only two-dimensional distance is used in the STO model. In real life, there can be barriers, such as an escarpment, that impede movement of cardholders between stops that appear close together in a two-dimensional representation. Field knowledge and revealed transfers can help refining the aggregation process.

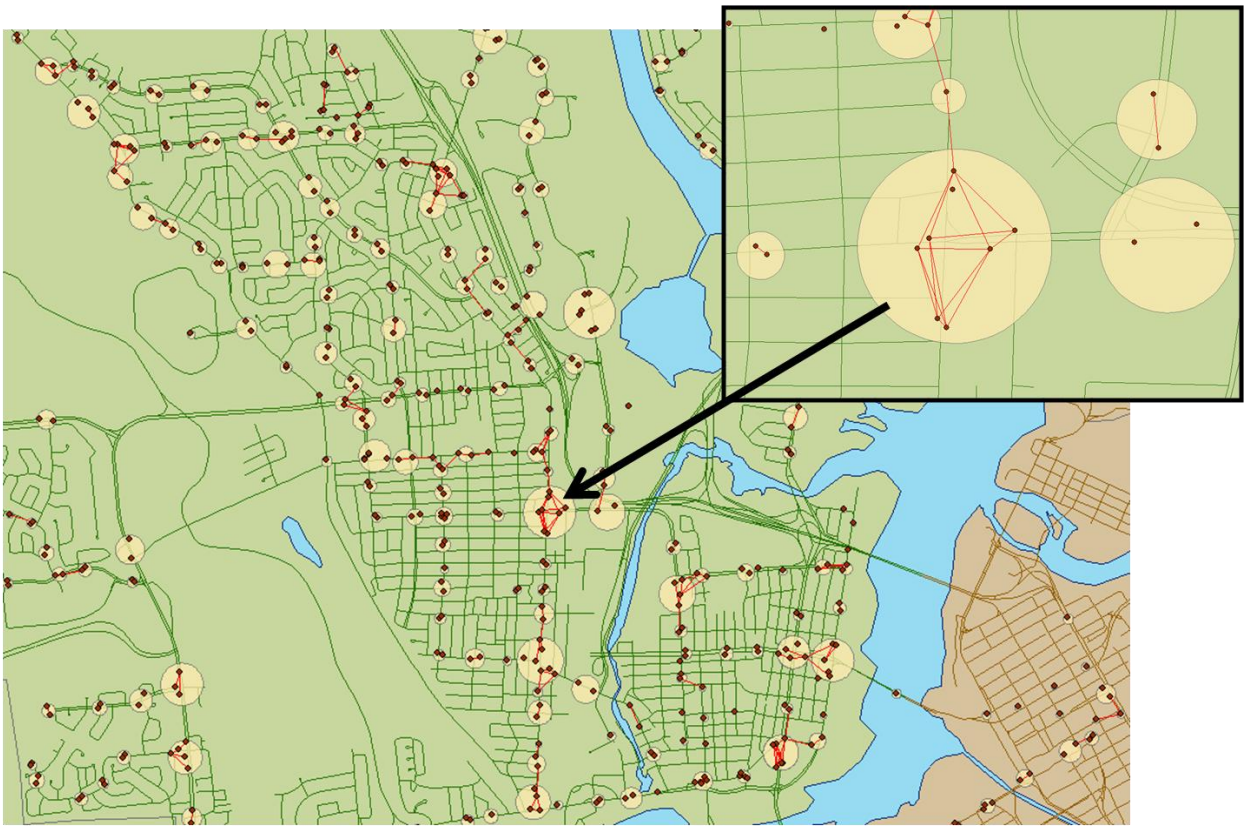


Figure 7.11 Example of MADITUC nodes.

Since the model uses a totally disaggregated approach, itineraries loaded onto the network keep their attributes and can be analyzed individually and simultaneously. The OD trip file acquired from the validation records can be loaded into the network to generate a number of performance indicators. For example, load profiles can be divided into distinct portions to illustrate the fare type. They can also be used to show the sector where the cardholder originates. They are presented in the following sections.

7.6.2 Examples of Transit Network Assignments

With a complete set of enriched smart card validation data in hand, this section focuses on the estimation of OD trip files for several time periods, which are common in regular modeling and planning setup. According to the level of resolution required for the analysis, the supply part of the model would usually use a route-based approach. The route-based definition of a transit network is essentially derived from the geometry attributes of the bus routes. While the GIS network model is useful to visualize the services both for planners and users, the planning model

should consist of a simplified geometric route-based network, with a level of service derived from the operations data (commercial speed and average headway) and OD trip files corresponding to chosen planning time periods. The latter can serve as a reference demand to evaluate different network geometries and service plans.

7.6.2.1 Transit Network Assignments with Revealed Routes

Figure 7.12 shows the three-dimensional result of a transit assignment onto a stop-level network using trips itineraries derived from smart card data as input. The flows on common lines are summed together and the colour illustrates flow intensity. The total load represents about 93% of all trips. Trips with alighting location over one kilometre from the subsequent boarding location are not included. The transit assignment allows the calculation of most travel demand statistics, including passengers-kilometres, passengers-hours, load factor, transfer wait time, etc.

Figure 7.12 Three-dimensional representation of the result of a transit assignment (Chapleau & Chu, 2007).

Figure 7.13 illustrates the resulting load profiles of the transit assignment with the itineraries from a typical AM peak. Flow direction and common lines are considered. The flow during the AM peak is mostly directed towards the CBD of Gatineau and Ottawa. The colours put an emphasis on cardholders' trip origin. The region served by the STO is divided into eight areas: Plateau, Hull, Hull Island, East Gatineau, West Gatineau, Chelsea/Masson/Buckingham and part of Ottawa which are colour-coded with lighter shades. The darker shades in the load profile correspond to the areas where the cardholders made their first boarding. For example, West Gatineau the highest share of cardholders entering the CBD of Ottawa, followed by East Gatineau. This type of illustration is useful for analysis of mobility, network usage, equity and revenue/cost sharing.

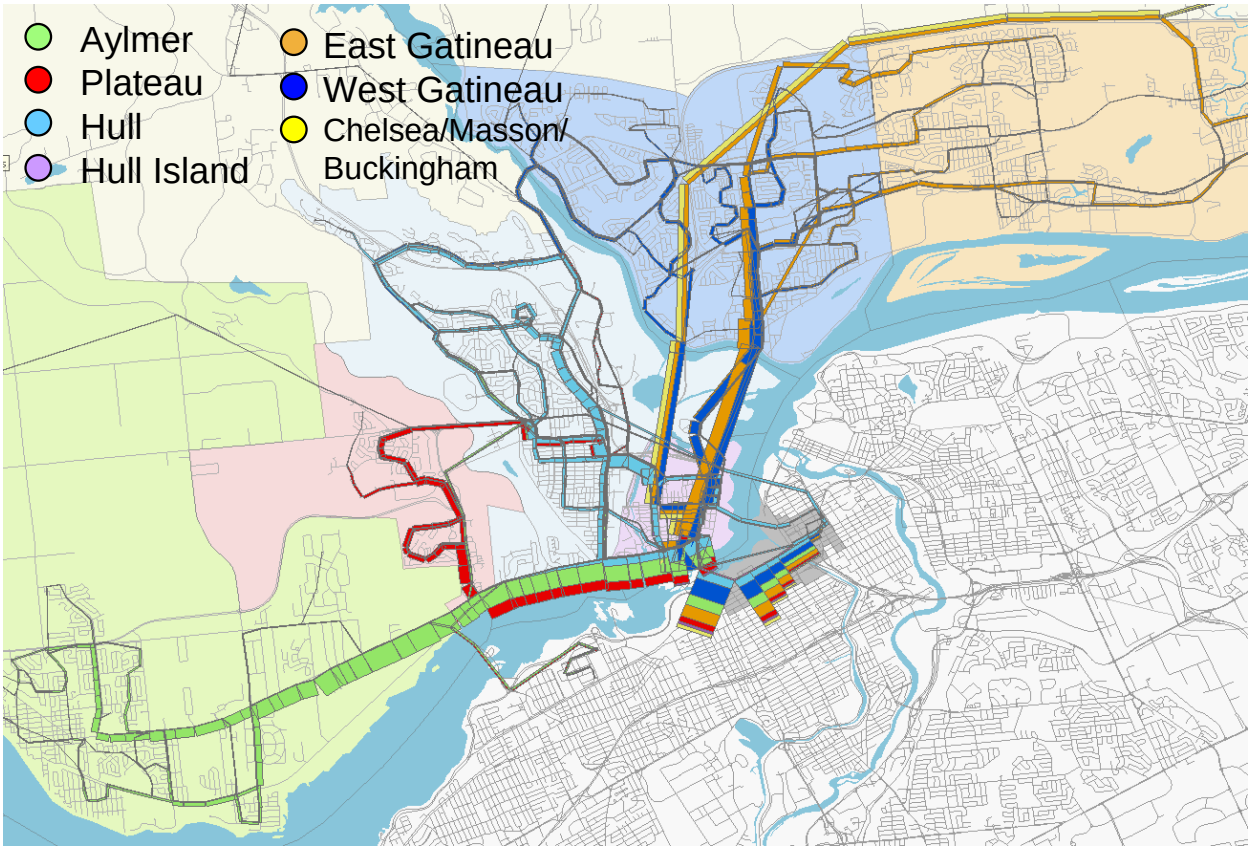


Figure 7.13 Assigning smart card itineraries onto a MADITUC transit network (Chu et al., 2008).

Figure 7.14 shows the result of a similar transit assignment procedure. The focus of the analysis is placed on fare type. Since almost 97% (739,560 out of 763,570) of the boarding in the month are made by either adults or students, the results are dominated by those fare types. The spatial distribution of adults and students are not equal. The majority of the cardholders who enter the CBDs of Gatineau and Ottawa are adults and the flow direction is mostly toward Ottawa. The student population has a particular niche. It is concentrated in the northwest part of Hull where several educational establishments and student residences on the Boulevard de la Cité-des-Jeunes are located. Interestingly, the flow intensity on some links in that area is comparable in both directions during the AM peak period.

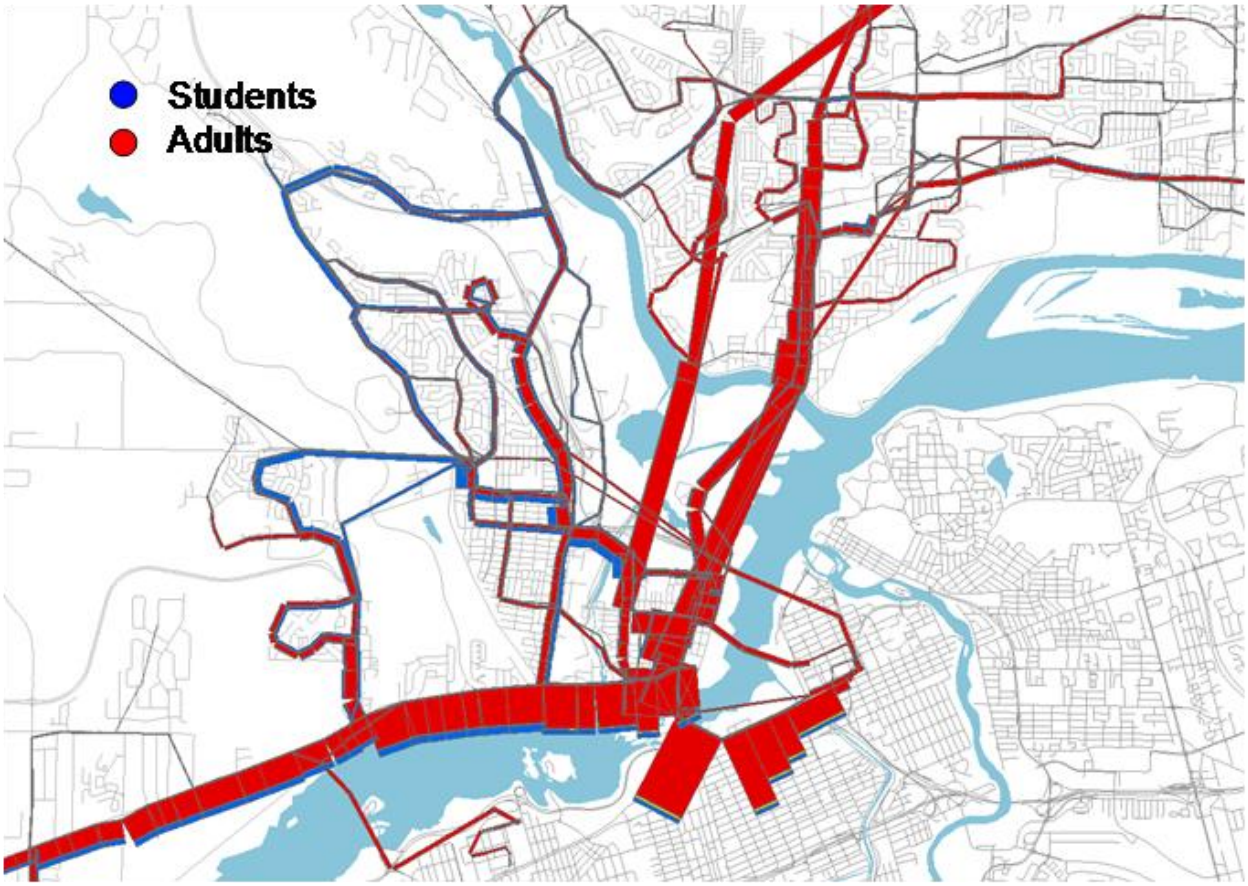


Figure 7.14 Load profile from transit assignment according to fare type.

Figure 7.15 presents the load profile of the network in conjunction with passenger movements during a typical morning peak. The boardings are spatially dispersed except where the park-and-ride facilities are located. This reflects the fact that at the start of a day, the population is sparsely distributed across the region. Transfers represent about 10% of total boardings and are limited to a few nodes at major intersections. Alighting activities are highly concentrated in the CBDs of Ottawa and Gatineau as well as near major trip generators. The migration of population during the AM peak reveals that activity location during the day is significantly more concentrated than the residential location of the cardholders.

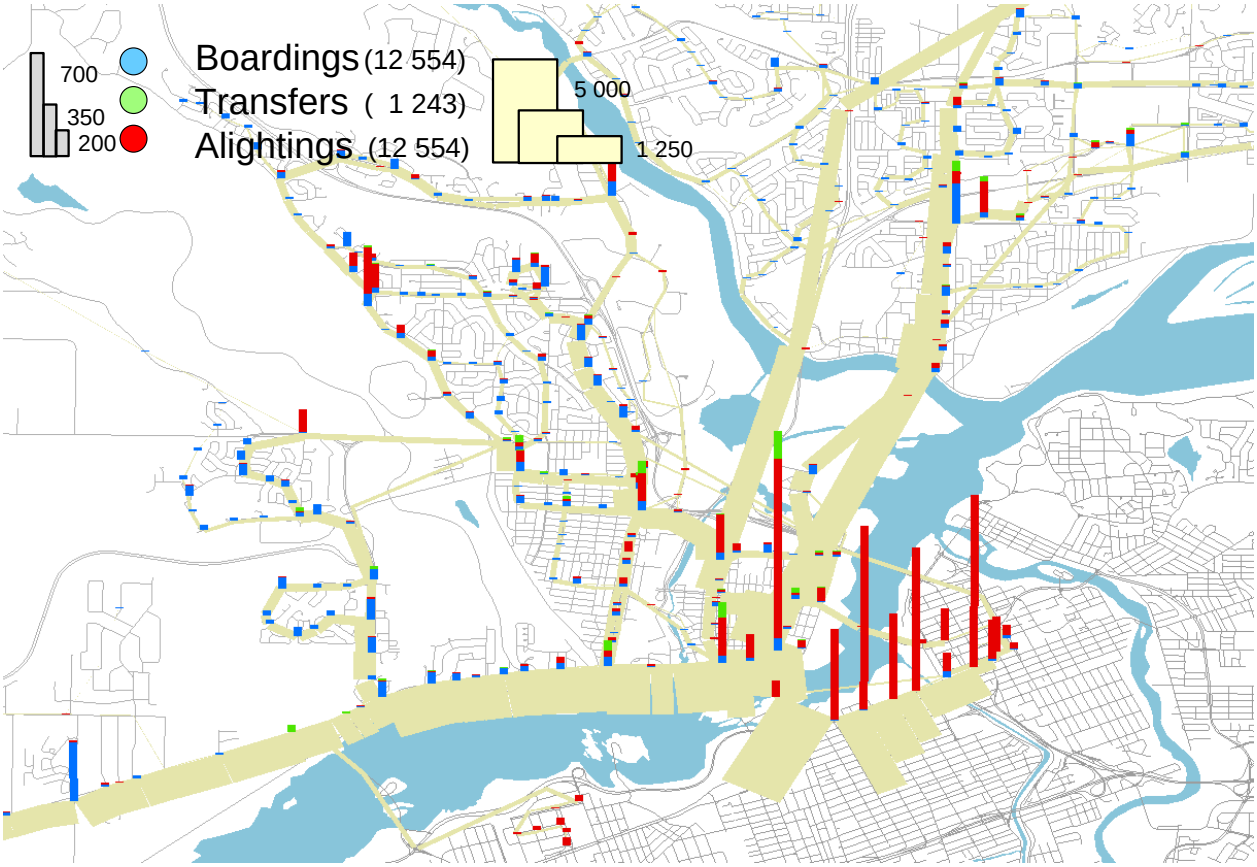


Figure 7.15 Load profiles and passenger movements from transit assignment.

7.6.2.2 Transit Network Assignments with Simulated Routes

Instead of using declared route information, route can be simulated for a trip from an origin node to a destination node. This, however, would require complete operation details of the transit network including the frequency of each route and the commercial speed. It would also require a calibration of the transit assignment model to find the parameters that best describe the behaviour of the cardholders.

7.7 Deriving Activity-space Profile

7.7.1 Modeling Activity Space

Passenger transport need is tightly linked to the locations where activities take place. As an OD survey can reveal activity locations of a person for the duration of the day, the transit itineraries derived from smart card can similarly reveal activity locations of cardholders. A grid analysis

allows one to recreate the spatial-temporal activity pattern, or the land occupation pattern, of transit users at various moments of a day. Figure 7.16 provides a schematic representation of the assumptions behind the technique and most of them are similar to previous analyses:

- The smart card system captures all the transit trips activity of cardholders within a day. Cardholders only travel by public transit within the day.
- The location of the cardholder is defined as the cell in which the estimated alighting occurs. The occupation starts at the time boarding and ends at the time of next boarding.
- The original location of the cardholder is defined as the location of first boarding of the day while the last location is defined as the location of the last alighting of the day, estimated using the first boarding of the day.
- Cardholders en-route are considered as transitional occupation and are not modeled.

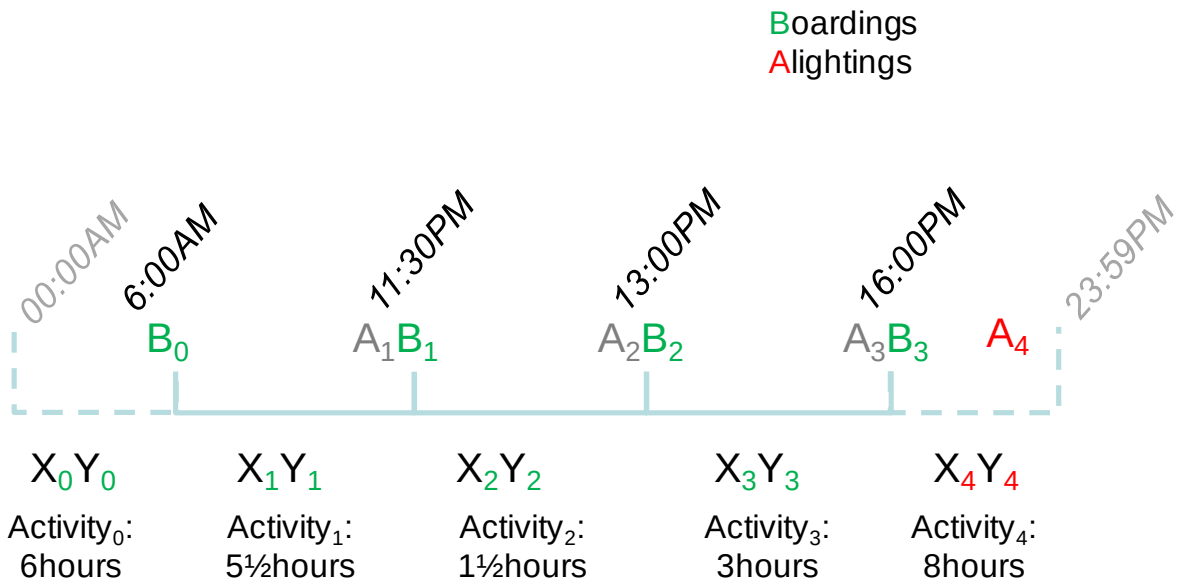


Figure 7.16 Assumptions used to derive land occupation profile.

A layer of cell 250 by 250 metres is created on top of the network. The size of the cell is chosen in order to obtain a detailed enough profile and at the same time smooth out the surface shaped by stop locations. When images of the occupation profile of successive time intervals (every 30 minutes) are placed in sequence, they illustrate the spatial-temporal movements of the cardholders within a day. Figure 7.17 contains 6 snapshots from a typical day and is originally an animation that chronologically captures the location of cardholders in the Gatineau region. It

shows that at 6 AM, cardholders are more evenly distributed across the territory with the exceptions of park-and-ride installations. During the AM peak, a large proportion of the population progressively migrates towards the CBD of Ottawa and Gatineau and towards educational institutions. The situation remains steady during the day. The reverse happens during the PM peak – the concentration of population dissipates back to their original residence. This type of travel information illustrates the evolution and the dynamic of this population and allows activity modelers to create a link between night-time location (residential location revealed by census data) and daytime location (for work, study and leisure). The cardholder concentration at park-and-ride locations before 6 AM likely indicates their entry point to the network and not their actual residence. The activity location of cards with a lone boarding during the day cannot be estimated. They remain at the boarding location for the whole day and are shown by the red portion of the column.

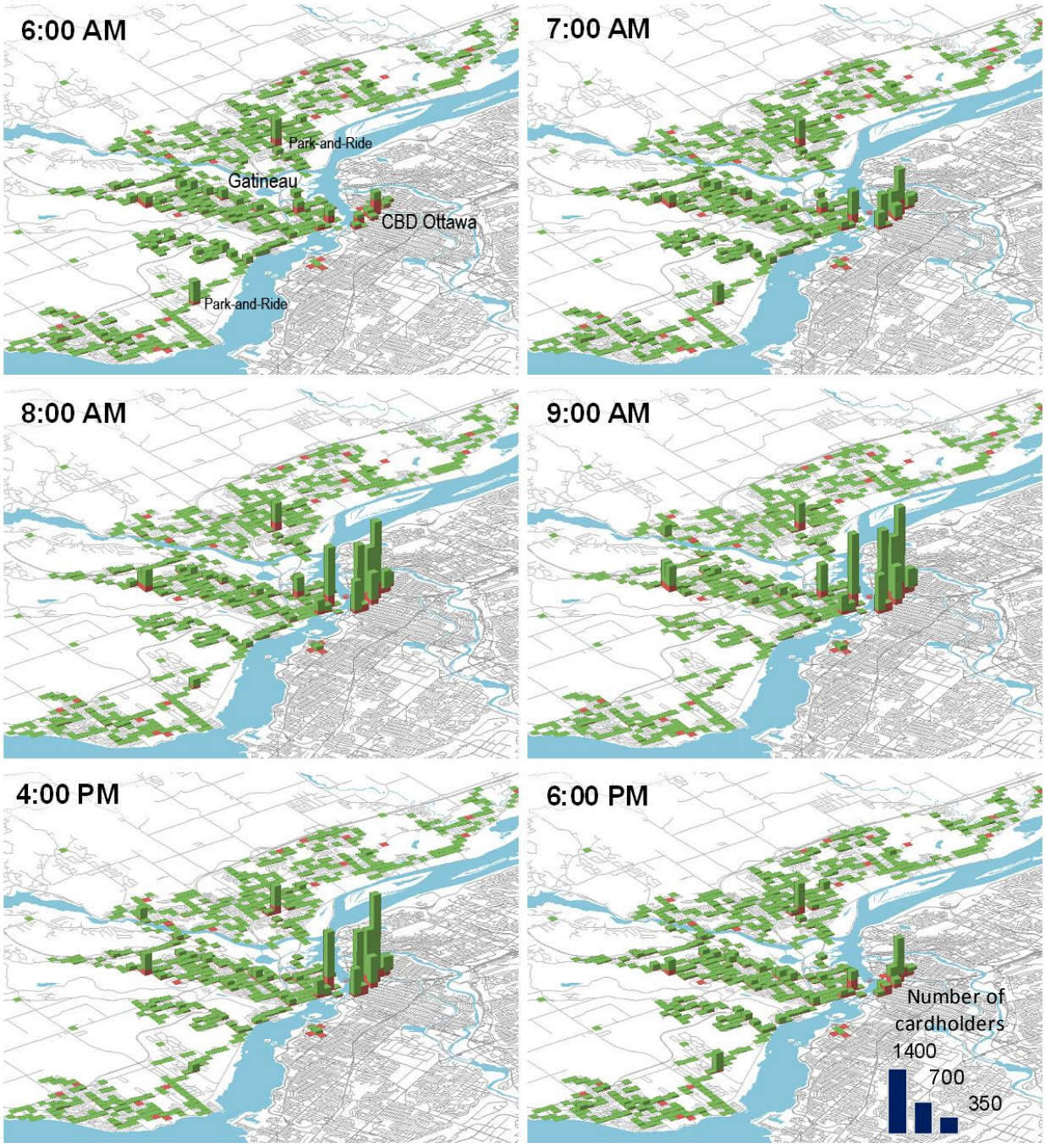


Figure 7.17 Derived land occupation profile of smart cardholders on a typical weekday (Chu et al., 2009).

7.7.2 Activity Space with Trip Generators

Trip ends and trip purpose are closely linked. Similar to GPS travel survey, both elements are not explicitly collected by passive methods and require inference in post-processing. Since travel is a

derived demand from the need to participate in activities, the assumption is that the boarding and alighting stops are in proximity to the cardholder's actual activity location. Chapleau, Trépanier & Chu (2008) explore the possibility of inferring activity location of cardholders with the use of external trip generator data. A georeferenced trip generator database, which includes educational institutions, hospital, shopping centres and park-and-ride facilities, etc., is associated with the transit network using GIS. It is assumed that a user's point of entry into the transit network (first boarding of an itinerary) is strongly linked to the location of the previous activity.

Figure 7.18 illustrates the possible linkage between three objects: boarding time, boarding location and trip generator. Major trip generators are revealed by the number of boardings in the nearby stops. In addition, some trip generators, such as educational establishments, have a distinctive temporal signature which is characterized by the activity schedule of their occupants.

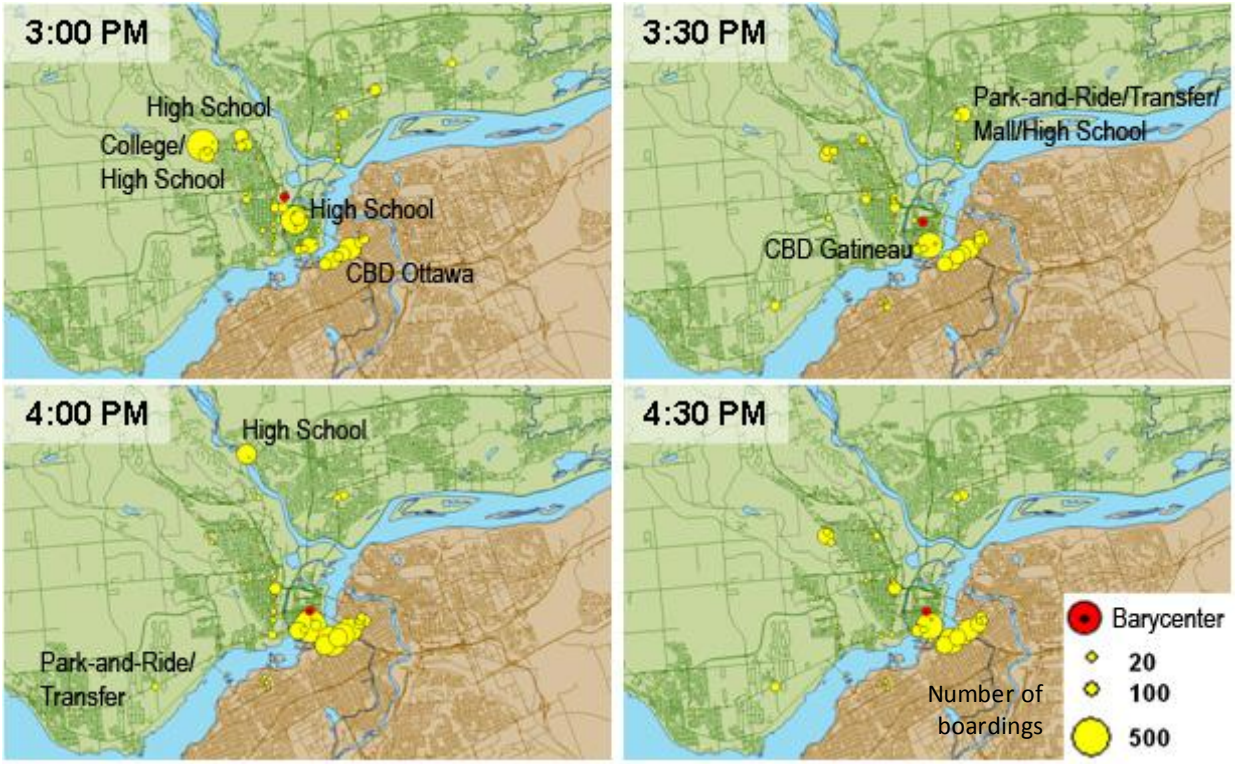


Figure 7.18 Trip generators are revealed by activities in the transit network (Chu et al., 2009).

Meanwhile, fare type can be effectively served as a proxy of the cardholder's status. The spatial coincidence across multiple days indicates a location of importance, or an anchor point, while the temporal coincidence would indicate a recurring trip. The notion of identifying anchor points therefore corresponds to the discovery of primary travel and activity pattern within the boarding

history of a card. For students, as illustrated in the next section, the identification is simpler and more precise because it can safely be assumed that one of the anchor points is an educational institution and one of the main activities is study.

7.7.3 Analysis by Trip Generator

Transit demand can be studied in relation to trip generators. Figure 7.19 shows a portion of inbound route 44. The stops are indicated by brown circles. All the points of interest in the database are drawn in squares. The colour of the squares represents the type of point of interest. Given that both the stops and POIs are georeferenced, it is possible to associate stops with POIs that are within a distance of 300 metres. It is chosen to represent the maximum access or egress distance on foot to and from the stops. Points of interest that lie within the perimeters can be visualized. The demand of a stop without a major trip generator can come from the surrounding residential area, minor trip generators or transfers from other routes.

Figure 7.19 Various types of trip generators located along the inbound route 44.

Planners can gain insight from the number and type of major trip generators in proximity to transit stops to explain the ridership. Figure 7.20 depicts all boarding and alighting activities of route 44 in both directions on a typical day along with the number of generators within 300 metres. It shows distance-based association between the origin and destination stops and the trip generators. The boarding and alighting activities are partitioned into first or transfer boardings. The geometry of the route is reduced to one dimension and each stop is identified by the linearized distance from the departure terminus. The CBD of Ottawa is treated as a single entity because of its complex land use pattern and the lack of complete points of interest data. Points of interest are counted twice if they are located inside the perimeter of two stops.

The route 44 operates only during the peak periods. Inbound route 44 has 11 runs during the AM peak period and outbound route 44 has 12 runs during the PM peak period. Fare validation data from a typical weekday have been used. On the inbound route, all boardings occur in Gatineau (distance 0 to 15,718) and most travellers alight towards the end of the line. This indicates a movement towards the CBD of Gatineau and Ottawa. The outbound route has a reversed but symmetric pattern, although the ridership is about 13% lower than the inbound route. It is remarkable to see the importance of boarding and alighting activities that occurred at distance 8,617 (inbound) and 11,060 (outbound) where only one point of interest associated with the stop – the Rivermead park-and-ride facility. It has a capacity of more than 500 spaces and there are other routes that converge at the junction.

Such a tool allows planners to examine the functional characteristics of each stop as origin, destination or transfer point with respect to the trip generators in proximity. Transfer activities at specific stops can be identified such that major transfer points can be used as time points to increase reliability for transferring patronage.

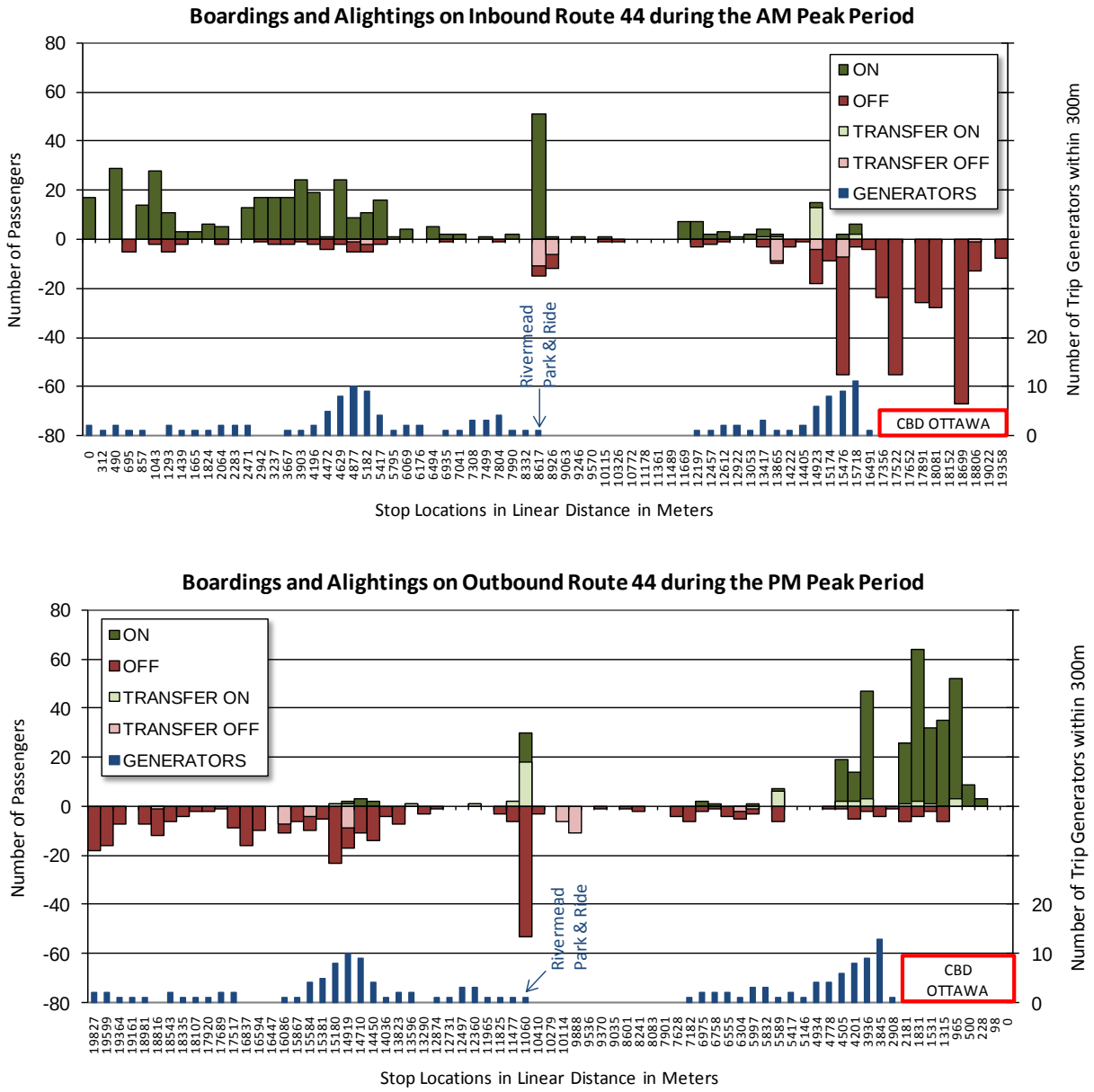


Figure 7.20 Functional analysis of transit stop with ridership and trip generator data (Chu et al., 2009).

7.8 Deriving Trip Ends for Individual Itinerary

In previous sections, trip generators are linked to boarding and alighting to explain aggregate travel pattern. A logical step to follow is to link trip generators to trip ends for each individual itinerary. This additional information would contribute to trip purpose inference and in-depth

travel behaviour analysis. Travel pattern can be revealed when boarding record data are consolidated and can be done by aggregating specific objects.

7.8.1 Travel Pattern Consolidation by Aggregations

In order to reveal travel pattern at trip ends, it is proposed to aggregate the spatial and temporal attributes into larger and logical units:

- Aggregation of boarding records into itinerary: as explained in previous sections, each person-trip is defined as a one-way movement of a person between two trip ends for a specific purpose. The inclusion of the transfer boardings would cloud the analysis. Person-trip with one or more transfer boardings is aggregated into one itinerary.
- Temporal aggregation: many activities in the society are collective and the time of activity is determined by communal convention. For example, high school usually ends in the afternoon and a student would make a trip after school in the afternoon. Under this assumption, transaction times are aggregated into several periods of day: AM peak, midday, PM peak, evening and night.
- Spatial aggregation: boarding stops are recorded in validation records as individual stop. Given a trip generator can give access to more than one route or stop, the use of nodes consolidates spatial pattern.

7.8.2 Example: Linking Student Boardings to Schools

An obvious illustration is to assign an educational establishment to every card with student fare. The assumptions are that boardings with student fares are made by students who are associated with a specific educational establishment. Given the number of boarding records in the card's history is sufficient, it is possible to identify the establishment by examining the historic boarding pattern. The longer the timeframe of the data, the more accurate the assignment would be. Starting with the georeferenced database of all educational establishments in the region, recurrent boardings in proximity and in the same time period indicate an association with the institution.

As will be discussed in the next chapter, the methods on consolidating travel pattern and linking trip generators to individual itinerary allow the in-depth study of travel behaviour of an individual

or people sharing similar attributes. The identification of trip ends will also be extended to an individual level to include cardholders' residence.

CHAPTER 8 METHODS AND ANALYSES FOR THE STUDY OF TRAVEL BEHAVIOUR

Trips, which are what a transport system handles, have been characterized by various levels of abstraction. Passenger trips, which concern the movement of a person from an origin to a destination, are primarily classified according to trip purpose, time of day, day of week, mode, person type, frequency, activity duration and route choice (Meyer & Miller, 2001). The rationale behind such classification is that planners and modelers recognize that the demand of transportation is highly differentiated (Ortúzar & Willumsen, 2001). Trips with different characteristics behave differently in a transport system. It is hoped that those characteristics would provide a more complete portrait of the demand and an improved understanding towards the underlying travel behaviour.

One major factor that dictates the level of detail in trip characterization is the availability of data. The level of detail is a direct result of the attributes describing the trip and the level of resolution used to measure each attribute. In traditional trip-based models, often in the absence of more precise information, trip ends and purposes are classified into generalized categories such as home-based work, home-based study, home-based others and non home-based, and are organized by period of day. Meanwhile, in household travel surveys or diaries, respondents are often asked to provide details on household, trip makers and trip itineraries. The concepts of tour, trip chain, daily schedule, go beyond the notion of individual trip and consider the interdependency among trips made by the same individual.

Although there is much progress in traditional travel data collection, it is costly and difficult to gather multi-day data of the same person in a large-scale basis due to issues like respondent burden, accuracy in reporting and decreasing response rate. The advent of passive data collection technologies, namely portable/wearable GPS, AFC systems and computerized sensors in vehicles, offers a new avenue to continuously gather large amount of high-resolution data. As mentioned in previous chapters, fare validation records from a smart card AFC system with AVL capability continuously capture disaggregate spatial-temporal information on network activities. The proposed data processing and enrichment procedures resolve problems inherent to passive data collection methods, namely data quality and missing trip attributes due to the absence of user input. The resulting multi-day data provide enormous potential for travel behaviour study.

The purpose of this chapter is twofold: it proposes methodologies to derive attributes which add a multi-day dimension to public transit trip characterization and to analyze multi-day travel behaviour of subgroups and individual cardholders. Based on a month of processed public transit smart card validation records, procedures which detect anchor points for each cardholder, associate boardings and alightings with anchors and derive additional trip attributes are elaborated. Techniques, such as spatial statistics, spatial analyses by GIS, visualizations and data mining, are applied to scrutinize travel behaviour of a subgroup or an individual transit user.

8.1 Framework for Analyzing Multi-day Data

The use of cross-sectional travel data accompanies the assumption of an average day whereas multi-day data follow the activity rhythm and cycle of individuals over time. The methodological framework on analyzing multi-day data, illustrated in Figure 8.1, revolves around a central theme – "with respect to" (WRT). The ultimate goals are to characterize a trip "with respect to" all trips made by the same individual within the analysis timeframe as well as to compare travel behaviour of an individual "with respect to" other individuals. The basic unit of analysis is a card. Boarding records enriched with alighting stops are grouped and analyzed by card. A longer timeframe validates travel patterns, reveals less frequent events and extended activity cycles.

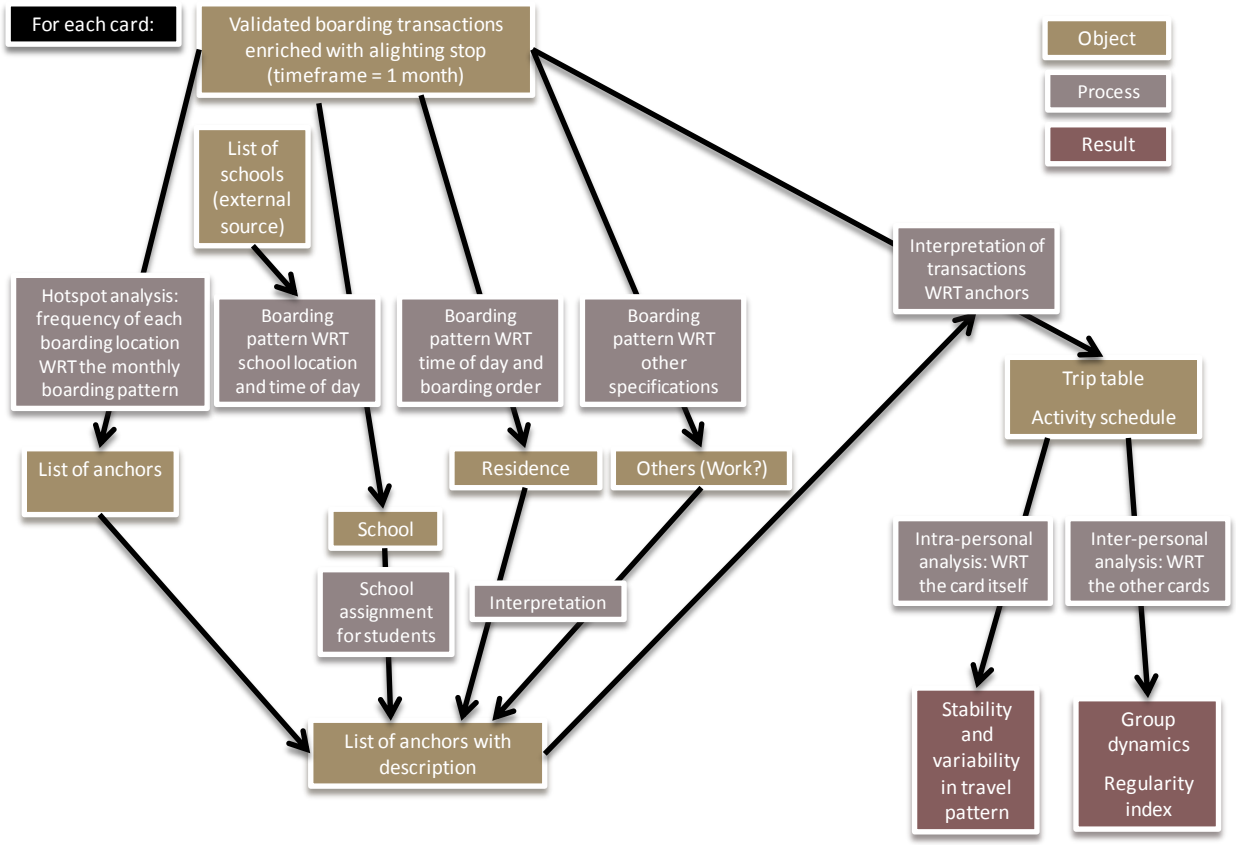


Figure 8.1 Schematic representation of the multi-day travel behaviour analysis framework.

The procedure involves the identification of anchors and their association with trip ends. Anchor points are defined as places where a person repeatedly visits. On a daily or short-term basis, they usually include the residence, work or study locations. With a longer timeframe, places of worship, shopping centers and friends' or relatives' homes may also be revealed as anchors. A hotspot analysis of boarding location can be defined as the frequency of each boarding location with respect to the monthly boarding pattern. The school assignment process, linking each card with student fare to an educational establishment, can be accomplished by looking at boarding pattern with respect to school location and time of day. The approximate residential location of a cardholder can be estimated by studying the boarding pattern with respect to time of day and order of boardings in a day. The execution of these processes generates a list of anchors for each cardholder. The itineraries are in turn interpreted with respect to the list of anchors.

Travel behaviour can be analyzed at two levels: the intra-personal level and the inter-personal level. The former compares partial information of the cardholder's derived trip table and activity schedule with respect to the cardholder's complete information within the analysis timeframe in

order to gain knowledge on the stability and variability in travel pattern. The later studies the travel pattern of one cardholder with respect to another cardholder or subgroup of cardholders in order to segment transit users by travel attributes. Details of each process will be explained in the following sections.

8.2 Characterizing Trips with Multi-day Data

The first step to characterizing a cardholder's trips from multi-day data is to create a cardholder's profile which will be used as a reference. This procedure synthesizes data from the analysis timeframe into information that would help interpret each transit trip. The information can include a status, an affiliation, places that a cardholder frequently visits such as home, school and workplace. Each trip would be systematically processed with respect to the profile, allowing boarding and alighting stops to be tied to trip ends, the interpretation of trip purpose and the calculation of mobility indicators.

8.2.1 Identifying Anchor Points for Each Cardholder

Since the use of public transit is derived from one's need to participate in an activity, travel pattern should be tightly related to the person's activity schedule.

8.2.2 Discovering a Card's Anchor Points

Recurring boardings at a specific location over a period of days suggest a place with significant importance to the cardholder. Discovering those concentrations of events, known as hotspot analysis, lays the foundation to associate those special places with actual points of interest. In addition to the spatial component of the events, anchor point identification needs to take into account the temporal aspect, land use and transportation concepts.

Since boarding events are discrete points in space and time, they need to be aggregated into larger and logical units. Spatial aggregation consolidates boarding and alighting patterns. Stops that lie within 50 metres of each other are aggregated to form a new node. A temporal aggregation groups discrete transaction times into time periods or classifies them by order. Trip aggregation combines multiple boardings from a linked trip into a single itinerary. Failing to do so would inadvertently create artificial transfer anchors.

The boarding intensity at each node during the analysis timeframe determines whether a node is considered as an anchor of a cardholder. An intensity threshold is required to filter out occasional stops. A minimum of 5 boardings at the node and a minimum share of 20% over total boardings are sufficient to distinguish an anchor from an occasional stop. The procedure generates a list of anchors for each cardholder. The results reveal that there are 8 cards with four anchors, 730 with three, 15,157 with two, 3,953 with one and 1,851 with none.

One way to visualize anchors is to use GIS (Figure 8.2). The anchors from two cards are illustrated. Card A uses a regular adult fare and has 39 first boardings and 1 transfer boarding. The kernel density of the boarding events is rendered into a 3-D surface. The properties of kernel density means there is no need to aggregate boarding stops but the surface must be divided into cells. The algorithm searches events within a distance of 500 metres which is chosen to represent the maximum walking distance to and from a bus stop. Two distinct anchors can be visually identified.

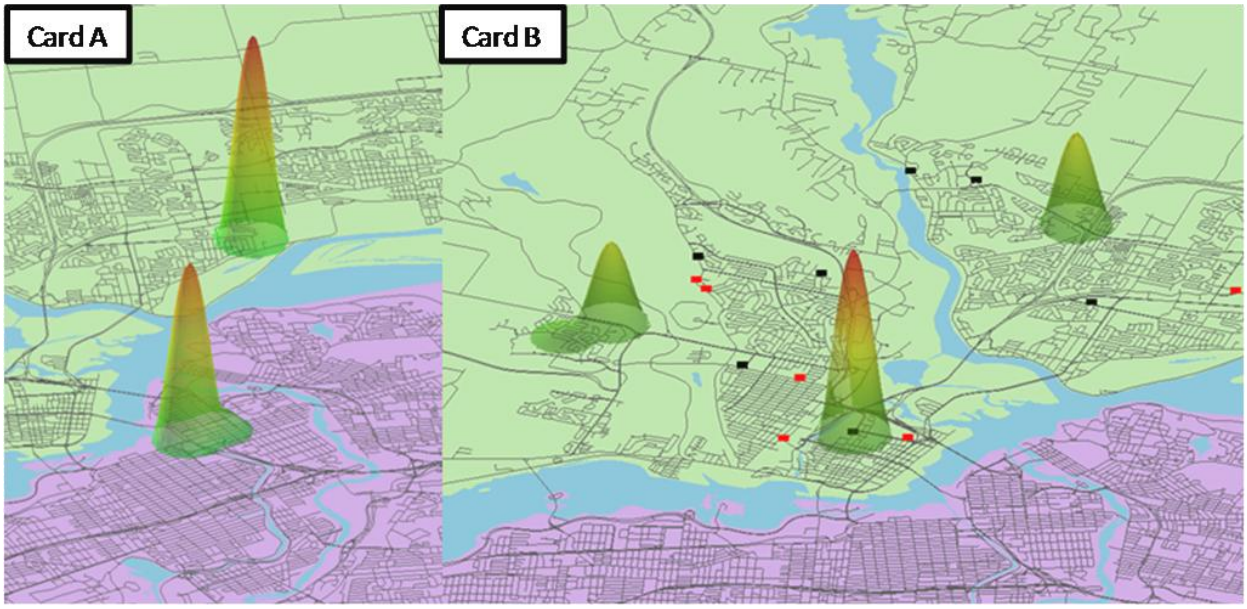


Figure 8.2 Two anchor points for card A and three for card B (Chu & Chapleau, 2010).

The same is done to card B which uses student fare. There are 36 first boardings and 1 transfer boarding. The fare type indicates that the cardholder belongs to an educational establishment. High schools (black dots) and colleges (red dots) in Gatineau are shown. One of them, École de l'Île, lies within one of the three hotspots, strongly suggesting an association between the cardholder and the school.

8.2.2.1 Associating Anchors with Specific Locations

A stop is a predefined position that is not exclusively tied to a particular activity location. It is assumed that cardholders minimize the access distance to a bus stop. Therefore, the only known and useful relationship is that the activity location is situated in the vicinity of the boarding stop. Three types of anchors can be distinguished:

- Anchor point: a finite set of locations with known spatial coordinates. Only in special cases can a boarding or alighting stop be associated to a specific anchor point. Example: educational establishments for students
- Anchor area: a predefined area which can be an artificial grid or geopolitical region. Example: CBD
- Fuzzy area: in an ideal situation, every trip end should be identified by exact spatial coordinates. However, it is impossible in most cases. Instead, an area around the stop or node is used to represent the anchor. Example: cardholder's residence

8.2.2.2 Assigning Anchor Points to a Set of Locations with Known Spatial Coordinates: the Case for Students

When assigning anchor points to a set of locations with known spatial coordinates, two additional pieces of information prove to be valuable:

- Fare type indicates whether the cardholder is an adult, a student under 21, a student 21 and over or a senior. Most transit agencies require a proof of affiliation to an educational establishment for purchasing a student fare product. It is therefore assumed that student fare cards must be associated with an educational establishment.
- By convention, classes in high school end between 2 to 5 in the afternoon. The first boarding records within the period are therefore used in the school assignment process. Since class time in college is more variable, all records between noon and 6 PM are used.

The procedure assigns cards to either a high school or a college. Since the fare type does not distinguish whether a student attends a high school or a college, separate assignments are performed. Starting with a georeferenced database of all educational establishments in Gatineau, the boarding locations from each card are assigned one by one to the nearest school. A school is

chosen for each card based on the frequency and on the average access distance. A score, which is the product of accuracy and count, is calculated to measure the confidence of each assignment. A minimum score of 5 prevents assignments that are too uncertain. For example, card B is assigned to École de l'Île and has a score of 18.

Assignment results are listed in Table 8.1. Out of the 6,030 cards, 2,565 cards are assigned to a high school or a college. Cards are not assigned due to the following reasons:

- Students who study outside of Gatineau (for example, in Ottawa)
- Students under 21 who study at universities and adult learning centres
- Not enough boardings in the month to create a profile or absence of clear travel pattern

Table 8.1 Results of the school assignment (Chu & Chapleau, 2010).

Assigned Educational Establishment	Number of Cards	Average Score	Avg Number of 1st Boardings
École Mont-Bleu	771	10.9	33.3
École de l'Île	541	11.5	34.1
Cégep de l'Outaouais Campus Gabrielle-Roy	481	10.1	30.6
École Saint-Alexandre	267	11.3	28.1
École Philémon-Wright	129	11.1	30.6
Collège Nouvelles-Frontières (High school and College)	105	10.1	29.5
Cégep de l'Outaouais Campus Felix-Leclerc	94	8.3	32.9
Cégep Heritage College	85	8.4	32.5
Collège Multicollège de l'Ouest du Québec	50	9.1	33.1
Establishments with fewer than 50 cards	42	9.1	30.8
Not Assigned	3465	0.9	23.4
Grand Total	6030	5.0	27.1

8.2.2.3 Assigning Anchors to Locations without Known Spatial Coordinates

Unlike educational institutions where their spatial coordinates are known, anchors such as residence and work locations are less clearly defined because:

- An exhaustive and up-to-date database of residence and workplace are rarely available.
- The spatial resolution at the bus-stop level does not allow analyst to pinpoint the exact location.

Therefore, a probabilistic approach is envisioned. Similar to school assignment, this approach takes into account the time of day, distance, frequency of travel of cardholders' multi-day travel

pattern. Instead of linking this information to a specific location, a kernel density analysis associates the range of activity location of the cardholder with a probability. The exact density function needs to be refined as the true density will depend on multiple factors including population distribution around the node and access distance on foot.

For each cardholder, a residence anchor is chosen among the list of anchors. The procedure considers boarding locations of the first transaction of the day. The interpretation is that most cardholders make the first boarding of the day from their residence. For the student population, no residence is assigned if the chosen anchor coincides with the school anchor. This suggests that the student travels only one-way from the school by public transit.

8.2.3 Linking Trip Ends to Anchors and Nodes

In order to characterize trip origin and destination at the cardholder's level, boarding and alighting stops need to be linked to anchors. A trip end is tied to one of the anchors in the list if it is located within 500 metres from the anchor. Otherwise, it is tied to the node that the stop belongs to and is considered as an occasional location. The procedure is performed on all trips for every cardholder.

8.3 Multi-day Travel Behaviour Analysis

The enriched smart card trip data provide 20 days of longitudinal observations on cardholders. Analyses performed on cardholders sharing a common anchor and on a single card presented in the next sections demonstrate the potential of the data in understanding aggregate and individual travel behaviour.

8.3.1 Analyzing Travel Behaviour of Cards Tied to a Specific Anchor

By combining the results from school assignment with the derived residence, one can study the travel behaviour of cardholders tied to the same school, answering questions on students' travel needs of a particular school. A kernel density map shows the residences of students attending École de l'Île (Figure 8.3). The surface does not pinpoint exact locations but identify areas where the residences are most likely to be. The absence of residence immediately around the school suggests that students who reside close to the school choose other modes of travel.

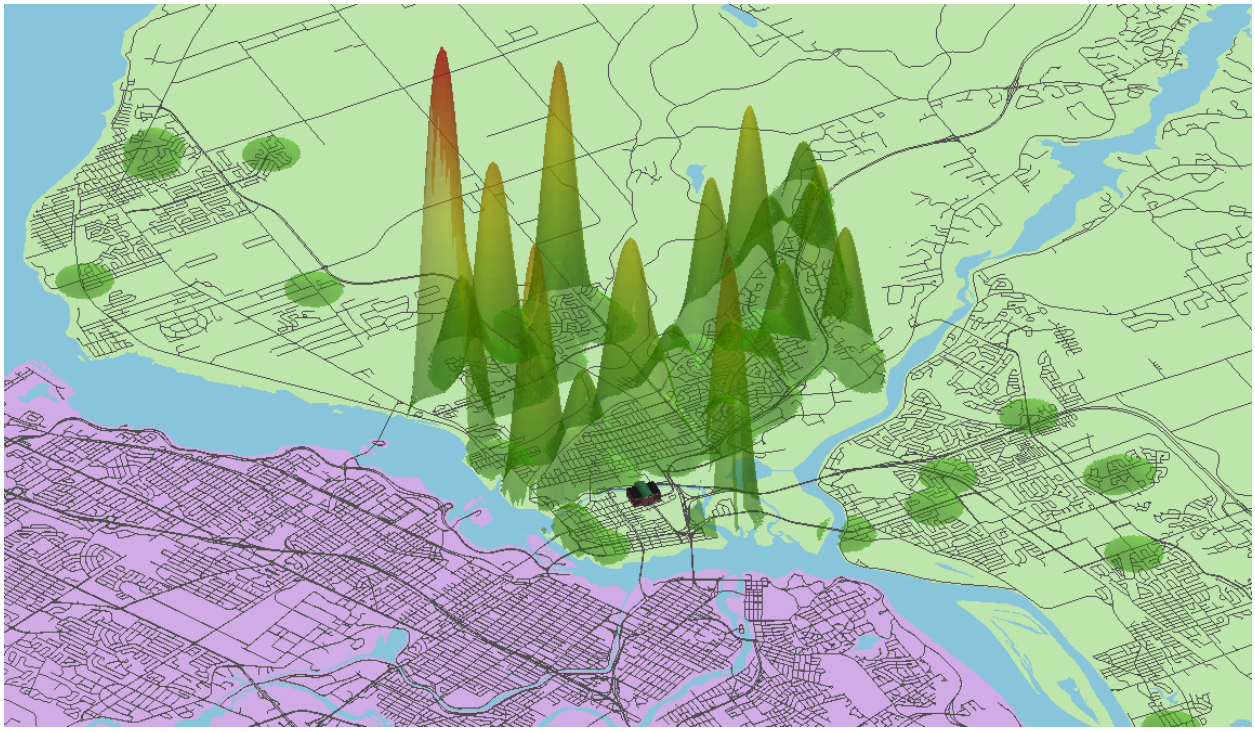


Figure 8.3 Kernel density showing the derived residence of cardholders attending École de l'Île High School (Chu & Chapleau, 2010).

Figure 8.4 exhibits alternative techniques, such as spatial statistics, to describe the residences. Orange dots represent spatial counts at the node level. The mean straight-line from the school is 3.55 km, with the shortest and longest being 0.87 and 10.97 kilometres respectively. The mean centre, representing the center point of the group of residences, is identified by a yellow cross at the northwest of the school. The standard deviational ellipse, in red, measures the dispersion of the residences. It is drawn one standard deviation around the mean centre. It is slightly elongated in the north-south axis meaning the spread along that direction is more important. A convex hull, in blue, is created by linking the exterior points of the residences. It defines the area within which all students using transit with smart cards reside.

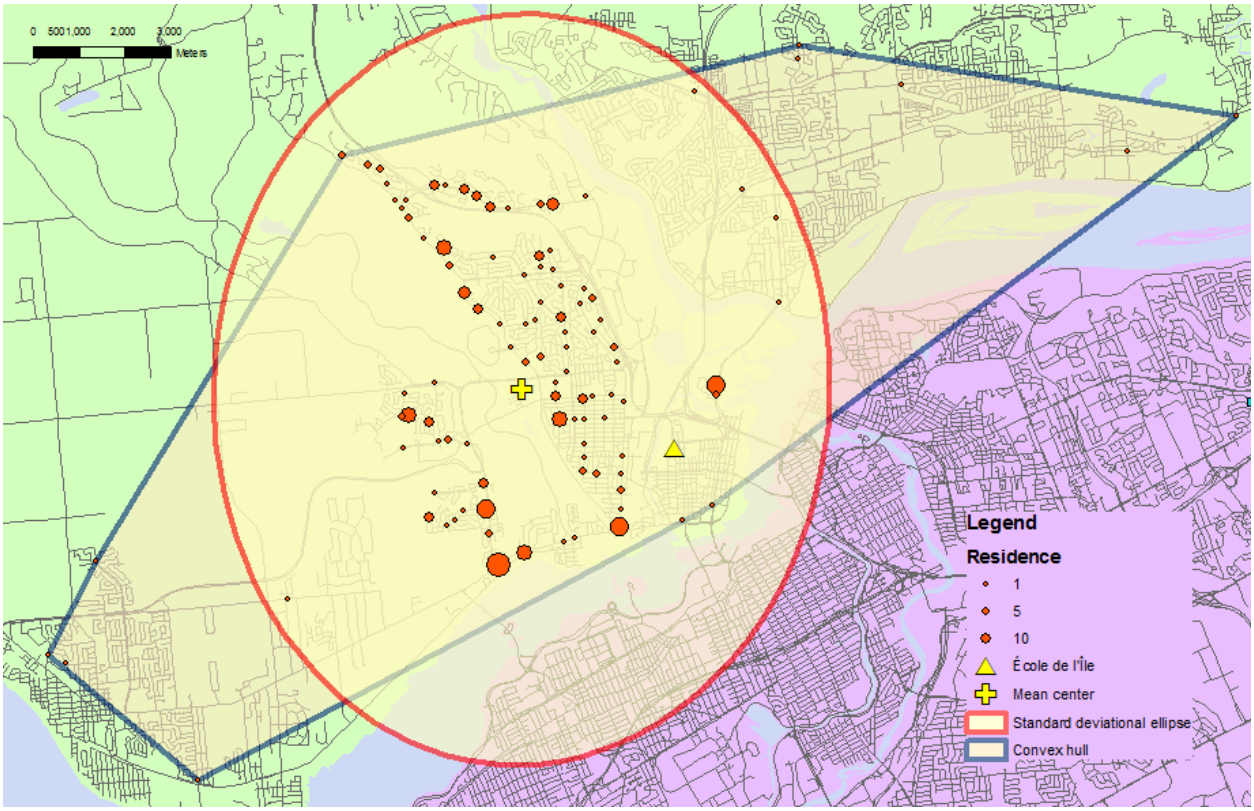


Figure 8.4 Various spatial measures describing the derived residence of cardholders at node level (Chu & Chapleau, 2010).

An interesting way to study the trip-making dynamics of students is to examine the relationship between their departure time (loosely defined here as the transaction time) and trip origins. Figure 8.5 contains nine snapshots pinpointing the origins of trips heading to École de l'Île at 15-minute interval. Common knowledge implies that students aim to arrive on time. Although students have the freedom to arrive early, the images, with concentric rings drawn at every two kilometres, reveal patterns in the evolution of trip intensity and trip origin over time.

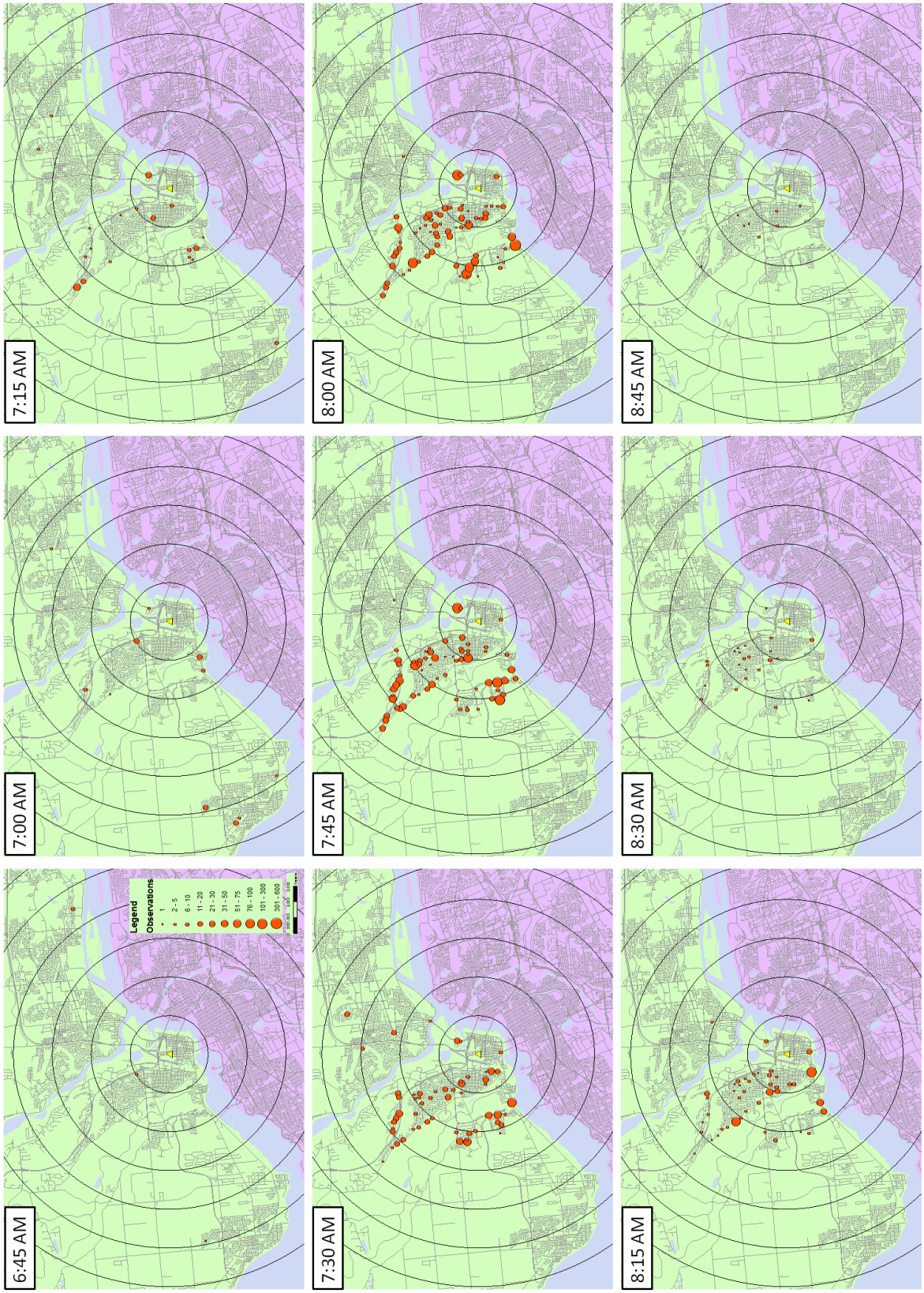


Figure 8.5 Origins of school-bounded trips at 15-minute interval (Chu & Chapleau, 2010).

Complementing Figure 8.5, Figure 8.6 shows the numeric distribution of departure time per 15-minute interval for the same trips. Almost 95% of them start between 7:30 and 8:29 AM. The most-intense 15-minute interval (from 8:00 to 8:14 AM) accounts for over 38% of the trips. The two curves outline the mean and median on-board distance per trip. At the aggregate level, longer distance traveled tends to associate with earlier departure time. Few survey data have the required spatial-temporal resolution and precision to demonstrate this subtlety. The longest on-board distance is 29.0 km and the shortest, 0.7 km. The mean distance is 6.2 km and is 1.7 time longer than mean straight-line distance from the residences. The combined distance traveled of all these trips in the analysis timeframe accounts to 35,665 kilometres. These values can be used to calculate indicators from the network consumption, environmental or active transport perspectives. For example, the volume of greenhouse gas emitted from these trips, whether by public transit or by automobile, can be estimated. Alternatively, knowledge of home locations and travel habits of students can help promote the use of active transport.

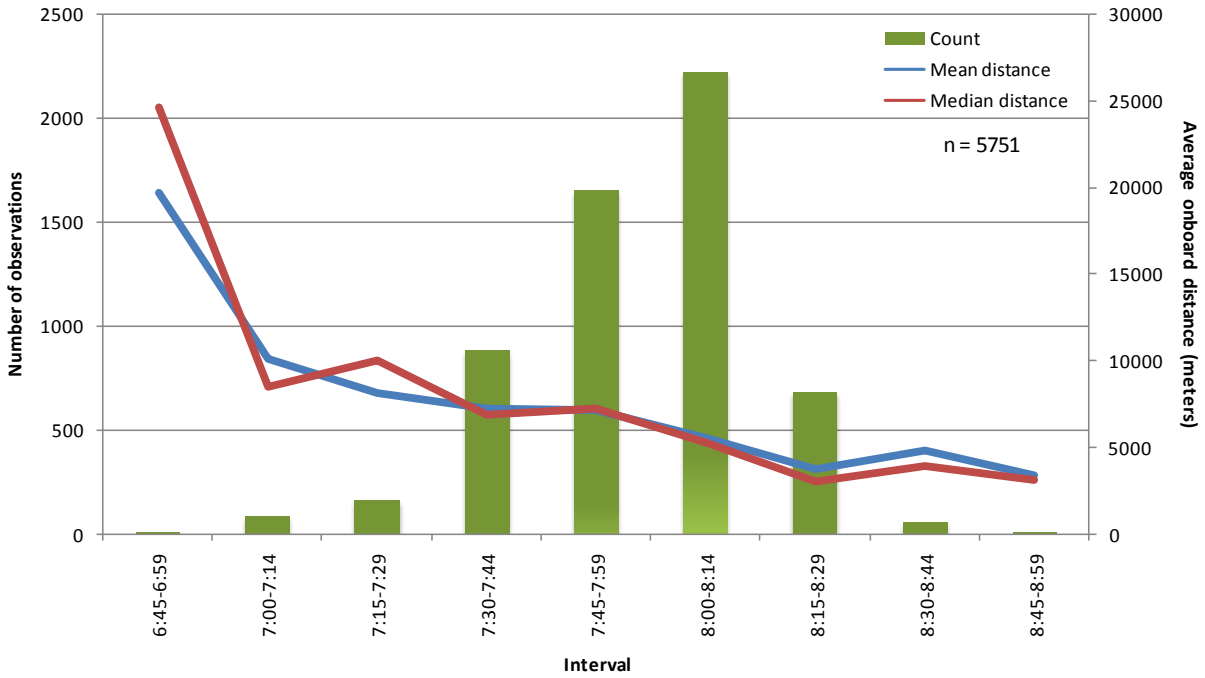


Figure 8.6 Distribution of departure time and on-board distance for trips heading to École de l'Île High School (Chu & Chapleau, 2010).

When the school is analyzed as an origin, the distribution of departure time needs to be examined in a finer temporal resolution. Figure 8.7 shows the number of boardings between 15:00 and 15:30 at one-minute interval. The outpouring of students indicates the end of class and the

presence of dedicated vehicles waiting for the students. It is interesting to note that the number of departing trips is 23% higher than arriving trips in this subgroup, suggesting that more students are using public transit when they leave school.

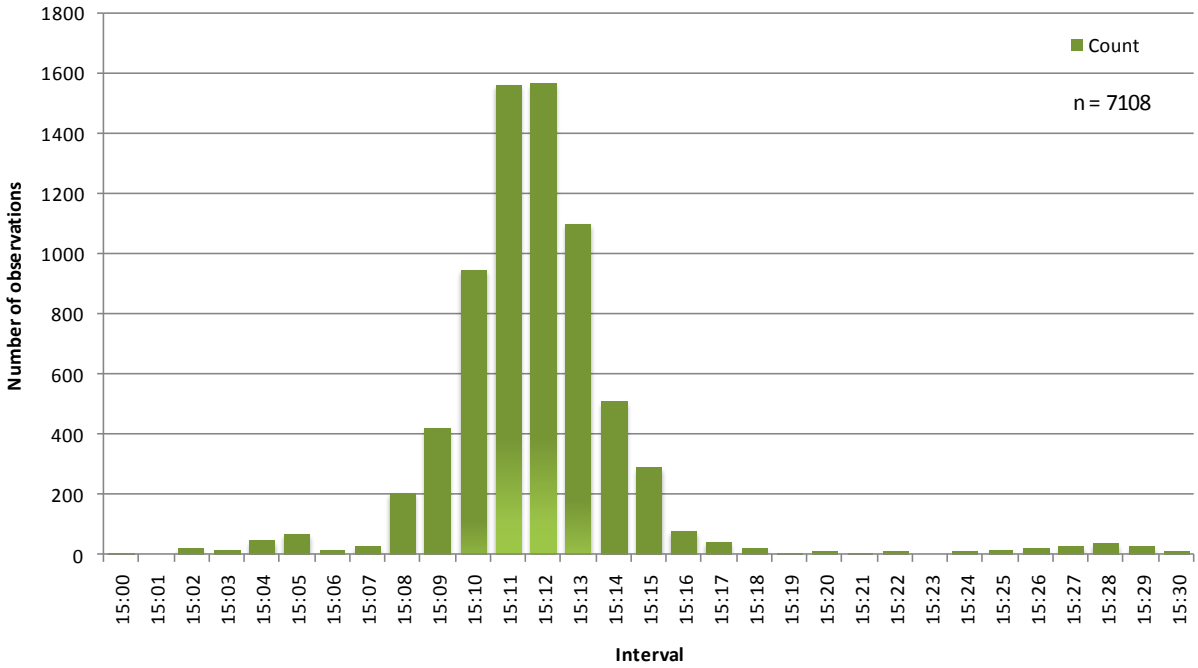


Figure 8.7 Distribution of departure time from École de l'Île High School by the minute (Chu & Chapleau, 2010).

8.3.2 Analyzing Travel Behaviour of a Specific Card

Figure 8.8 contains a wealth of information on the trips made by card B. The pies indicate the number of boardings associated with each node and the destinations of those trips. The cardholder's three anchor points are identified. Trip origins, destinations and frequency are drawn with desire lines. A convex hull, with a buffer of 500 metres to represent access and egress distance on foot, is generated from the boarding nodes to delimit the observed activity space of the cardholder.

Table 8.2 serves as an accompanying trip table to Figure 8.8. The level of trip detail is remarkable. It incorporates a node-level origin-destination matrix in tabular form and attributes that describe the multi-day dimension of the trips: activity duration (defined as the temporal gap between two boardings), boarding time, on-board distance and their respective variability. In total, the cardholder traveled an estimated 280,983 metres by public transit in twenty weekdays.

The monthly activity amplitude, defined as the span between earliest boarding time and the latest boarding time, is a temporal equivalent of activity space. Daily and monthly activity amplitude as well as the day-by-day variation can be computed. The monthly activity amplitude of this particular card is 7 hours 48 minutes. The trips express a highly regular pattern in boarding time. This is the combined result of the regularity of the cardholder and the transit vehicles.

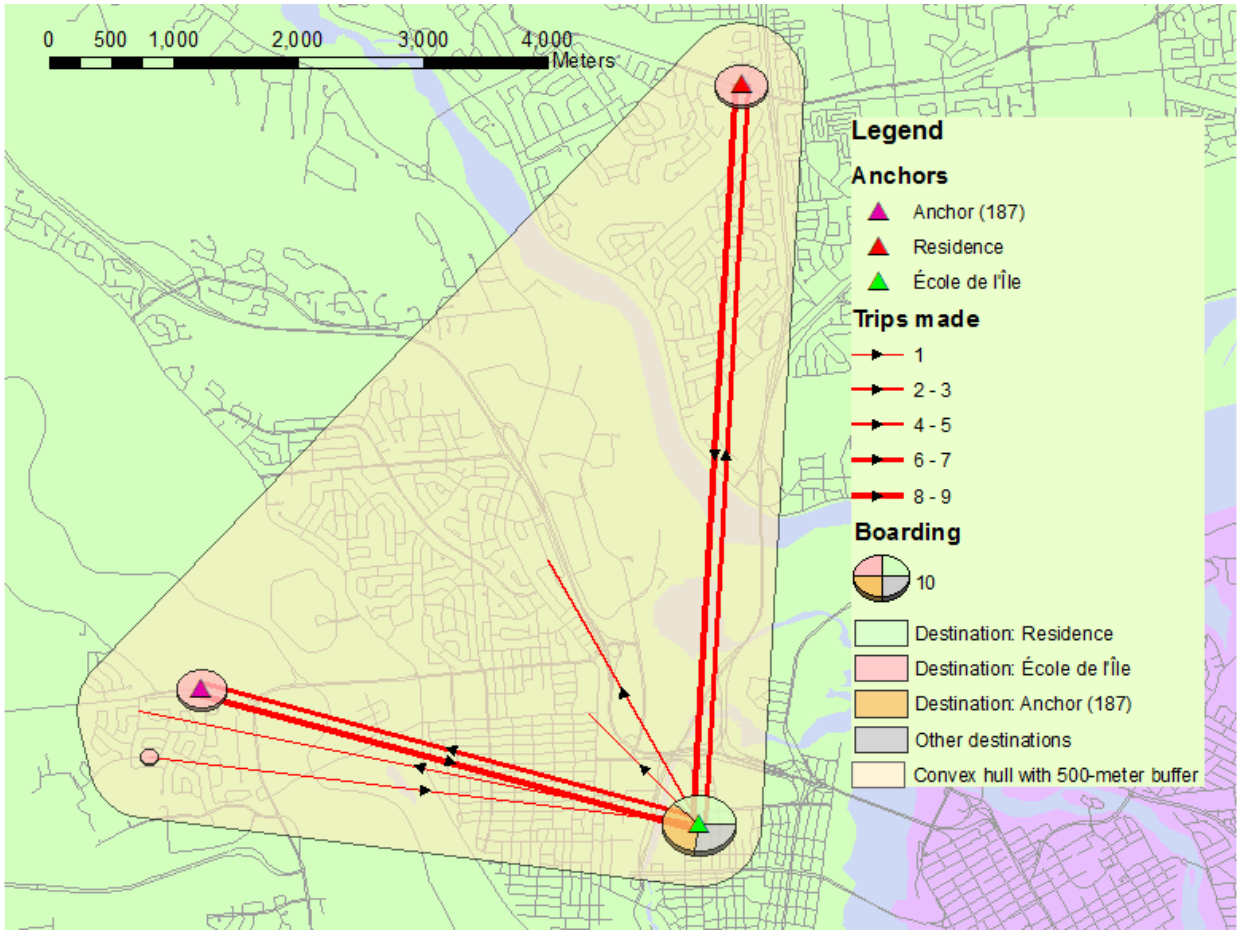


Figure 8.8 Trips details of card B (Chu & Chapleau, 2010).

Table 8.2 Origin-destination trip table with multi-day attributes (Chu & Chapleau, 2010).

Origin (Node)	Destination (Node)	Number of observations (whole day, 20 weekdays)	Average distance between origin stop and nearest anchor (meters)	Average distance between destination stop and nearest anchor (meters)	Average activity duration (minutes)	Average boarding time	Earliest boarding time	Latest boarding time	Standard deviation of boarding time (minutes)	Average distance traveled (meters)
Anchor (187)	École de l'île (429)	8	45	307	441	8:11	08:09	8:18	2.73	5155
École de l'île (429)	Anchor (187)	6	23	45	422	15:12	15:11	15:13	0.58	5295
	Residence (452)	7	23	57	465	15:24	15:21	15:26	2.10	11700
	Node 342	2	23	2476		15:24	15:22	15:26	2.00	3148
	Node 174	1	23	528		15:12	15:12	15:12	0.00	5847
	Node 373	1	23	1268		15:13	15:13	15:13	0.00	1762
	Unknown	1	23			15:10	15:10	15:10	0.00	
Residence (452)	École de l'île (429)	9	52	93	446	7:40	7:38	7:44	2.02	11758
Node 177	École de l'île (429)	1	684	62	412	8:20	8:20	8:20	0.00	6346
		36	54	3077	442	11:37	7:38	15:26	221.3	7805

Going even further, the complete activity schedule related to the use of public transit can be derived (Figure 8.9). Trip purpose and activity duration, interpreted according to cardholder's profile, destination anchors, derived activity duration, are indicated by colour bars. Boarding times and locations are shown as coloured triangles. The transit route number and direction are also labeled.

The monthly schedule exposes the agenda of a stereotypical student who goes to school in the morning and leaves school in the afternoon. At the same time, the student is atypical because of a travel rhythm that alternates every other week. During the first and third weeks, the student starts the day around 07:40 from his residence to go to school. During the other weeks, the student starts the day about 30 minutes later from anchor 187. One can interpret that the anchor 187 is a second residence where the student resides every other week. This illustrates an excellent example of what cross-sectional travel data cannot provide. Even with multi-day data which last less than one week, this bi-weekly pattern cannot be easily captured and interpreted. Although in some cases, the exact trip purposes and the identity of anchors may never be known, it raises the fundamental question on whether those are really necessary, as the multi-day trip attributes may be more precise in characterizing those trips.

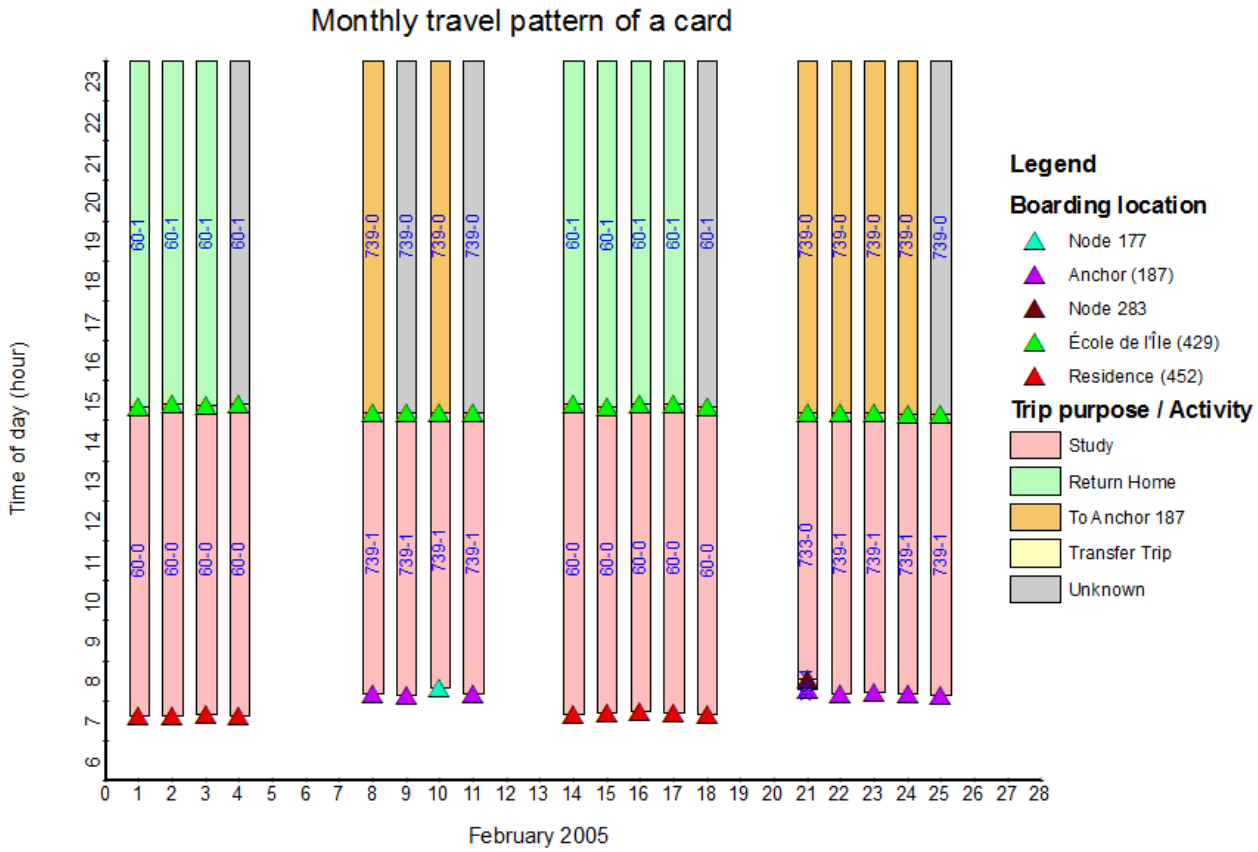


Figure 8.9 Derived activity schedule of card B (Chu & Chapleau, 2010).

8.3.3 Data Mining as a Tool for Travel Behaviour Analysis

So far, travel patterns are detected by statistical and spatial analyses. They require a significant amount of involvement from the analysts to go through the data. Another approach is to use data mining to automatically and efficiently find patterns in datasets. Two data mining techniques are applied to the enhanced trip data. Even though the following examples use a small dataset, they nevertheless demonstrate the value of this analytical tool and the methodology.

8.3.3.1 Applying an Association Rule Algorithm to Analyze Travel Behaviour

Association learning aims to find attribute values that frequently co-occur in the same transaction in a database. Rules are usually built and evaluated primarily on two measures: support and confidence. Support is expressed by the proportion of transactions that the rule holds. Confidence measures the accuracy of the rule given the antecedent. The procedure aims to uncover hidden travel pattern and provide insights to the understanding of travel behaviour.

The dataset contains 37 instances, 43 distinct values from 8 attributes which described trip origin, destination, trip start time, activity duration and end time, route taken, distance traveled and day of the week. Since association learning does not handle continuous values, time and distance are discretized into nominal values, to the nearest half hour and half kilometre respectively. With a minimal support of 15% and a minimal confidence of 40%, 5,874 rules are generated. Since the number of rules increases exponentially with the number of attribute values, a strategy to filter out less interesting rules is needed. Given the same support and confidence, rules with short antecedents and long consequents are more interesting because they contain all the simpler rules. Figure 8.10 shows a subset of rules generated with by the dataset grouped by antecedent.

Rules	Supp	Conf	Lift	Lev	Strg	Cov
▷ Route=739_1						
◢ OrigNode=Node452						
-> Route=60_0	0.243	1.000	4.111	0.184	1.000	0.243
-> Distmeter=12000m	0.243	1.000	2.313	0.138	1.778	0.243
-> StartTime=450min	0.243	1.000	4.111	0.184	1.000	0.243
-> EndTime=930min	0.243	1.000	4.111	0.184	1.000	0.243
-> DestNode=Node429	0.243	1.000	2.056	0.125	2.000	0.243
-> Route=60_0 Distmeter=12000m	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 StartTime=450min	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 EndTime=930min	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> Distmeter=12000m StartTime=450min	0.243	1.000	4.111	0.184	1.000	0.243
-> Distmeter=12000m EndTime=930min	0.243	1.000	4.111	0.184	1.000	0.243
-> Distmeter=12000m DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> StartTime=450min EndTime=930min	0.243	1.000	4.111	0.184	1.000	0.243
-> StartTime=450min DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> EndTime=930min DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 Distmeter=12000m StartTime=450min	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 Distmeter=12000m EndTime=930min	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 Distmeter=12000m DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 StartTime=450min EndTime=930min	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 StartTime=450min DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 EndTime=930min DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> Distmeter=12000m StartTime=450min EndTime=930min	0.243	1.000	4.111	0.184	1.000	0.243
-> Distmeter=12000m StartTime=450min DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> Distmeter=12000m EndTime=930min DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> StartTime=450min EndTime=930min DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 Distmeter=12000m StartTime=450min EndTime=930min	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 Distmeter=12000m StartTime=450min DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 Distmeter=12000m EndTime=930min DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 StartTime=450min EndTime=930min DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> Distmeter=12000m StartTime=450min EndTime=930min DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
-> Route=60_0 Distmeter=12000m StartTime=450min EndTime=930min DestNode=Node429	0.243	1.000	4.111	0.184	1.000	0.243
▷ Distmeter=12000m						
▷ Route=60_0						
▷ DestNode=Node429						
▷ StartTime=450min						
▷ EndTime=930min						
▷ Distmeter=5000m						
▷ Route=739_0						

Figure 8.10 A subset of rules describing card B's travel pattern (Chu & Chapleau, 2010).

As an example, the following rule is put into the cardholder's context:

OrigNode=Node452 -> Route=60_0 Distmeter=12000m StartTime=450min EndTime=930min DestNode=Node429

The interpretation would be:

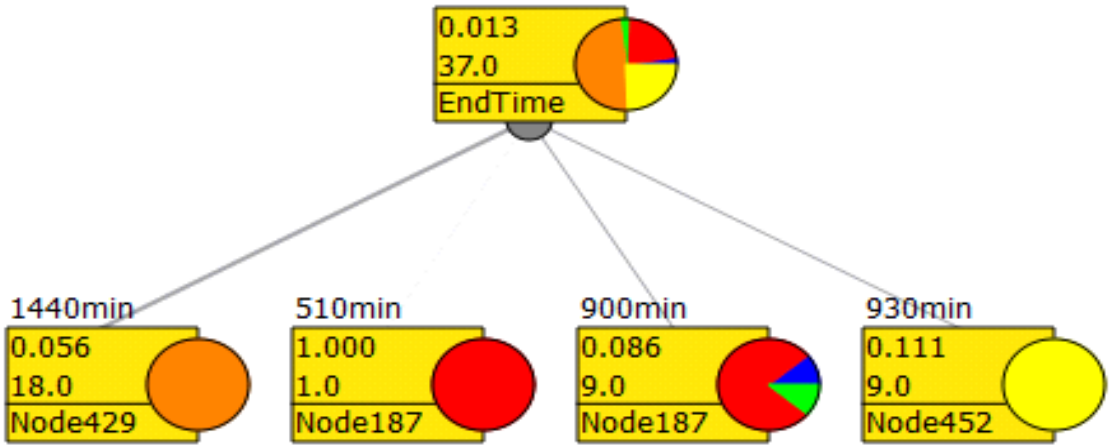
> If the boarding location is the residence anchor, it is observed in 9 times out of 9 (support of 24.3% times 37 observations at 100% confidence), the cardholder boards route 60 direction 0 at 07:30, travels 12 km to École de l'Île and leaves at 15:30.

8.3.3.2 Applying an Classification Algorithm to Measure Regularity in Travel Pattern

Another type of learning is classification. Based on a set of known examples, a classification algorithm constructs a decision tree that best reflects the underlying structure of the data with respect to the predefined class attribute. Most often, the resulting model is served to predict the class of unclassified examples. The performance of the model is evaluated by the proportion of correctly classified examples. A confusion matrix shows the difference between the predicted and the actual classes.

Since a classification algorithm is designed to find pattern in a dataset and can simultaneously handle numerous attributes, it is more adept to describe travel pattern than any algebraically formulated indicator that only considers a limited number of observations and attributes. By applying a classification algorithm to the trips made by a cardholder, one can obtain a travel behaviour model, albeit a simple one, along with its performance measure, with respect to the chosen class attribute.

A model yields a poor performance because little pattern exists between the class attribute and the other variables. In contrast, a strong pattern is revealed by a model with good predictive power. The comparison of model performances across cardholders can be interpreted as the relative regularity in trip pattern among the cardholders. Figure 8.11 shows the decision tree and confusion matrix, generated by the C4.5 algorithm (a supervised statistical classifier using the concept of information entropy), with the class being the boarding location.



		Predicted					
		Node177	**Node187**	**Node283**	**Node429**	**Node452**	**Total**
Actual	**Node177**		1				1
	Node187		8				8
	Node283		1				1
	Node429				18		18
	Node452					9	9
	Total		**10**		**18**	**9**	**37**

Figure 8.11 A decision tree generated by the C4.5 algorithm along with the confusion matrix (Chu & Chapleau, 2010).

Since the classification tree has only two levels, it suggests that among all attributes in the dataset, activity end time (expressed as minutes from midnight) alone is the best predictor of boarding location. The classification accuracy is 94.6% for the training set, which means the cardholder's boarding location can be correctly predicted by activity end time in 36 out of 37 times. This suggests a highly regular travel pattern. It is imperative to avoid overfitting, which describes a phenomenon of a highly branched tree that is too specific for a dataset, by selecting appropriate pruning parameters.

8.4 Analyzing Travel Behaviour of Cards Tied to a Specific Transit Service

Another possibility of multi-day data is to examine travel behaviour with respect to the transit service. The resolution of the data allows researchers to analyze travel behaviour down to the run.

Figure 8.12 examines cardholder's loyalty towards bus runs on two separate occasions. In this context, two loyalties are measured. Cardholders can be perfectly loyal if they have taken the same run of route 37 on both Thursdays (February 10 and 17, 2005). They can also be considered "somewhat loyal" if they have taken two runs that are within an hour of each other on both Thursdays. The gap between the two loyalties is the greatest during the AM and PM peaks when route 37 has a high frequency. The flexibility of service reduces the likelihood of taking exactly the same run. The opposite is true when services are less frequent. The gap between two loyalties is small. Also, the loyalties are higher during the AM and PM peaks because they tend to be commute trips whereas during other time periods, they tend to be occasional trips.

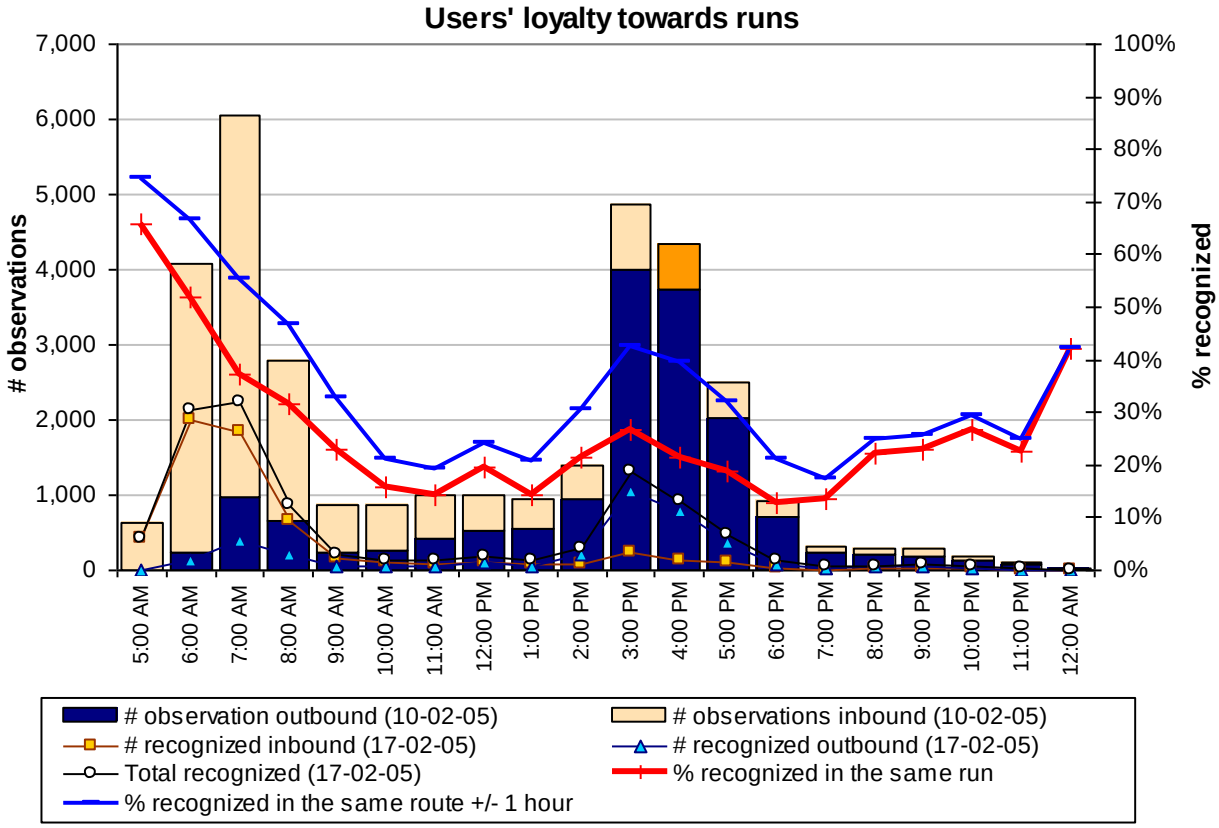


Figure 8.12 Cardholders' loyalty towards a route and a run (Chapleau & Chu, 2007).

8.5 Observations from Travel Behaviour Analysis

Spatial statistics, spatial analysis with GIS, visualizations and data mining are among the tools used to scrutinize the enhanced trip and activity data in order to better understand aggregate and individual travel behaviour. The results, which include multi-day trip attributes, detailed trip

table, activity schedule, demonstrate the possibility and value of smart card data in transit planning and travel behaviour analysis.

The analysis leads to several interesting observations:

- The one-month analysis timeframe allows analysts to uncover extended activity cycle. A bi-weekly cycle can be observed in the derived activity schedule (Figure 8.9).
- The richness of the multi-day trip attributes raises the fundamental question on whether certain traditional trip attribute, for example trip purpose, is necessary for some applications such as transit planning (Madre, Massot & Armoogum, 2000). Trip purpose is used to characterize trip frequency and spatial-temporal regularity in a cross-sectional travel survey. A good illustration is that work trips are assumed to be stable on all weekdays. Passively collected multi-day data do not have trip purpose but provide instead the frequency, spatial-temporal regularity and activity duration. These data seem more appropriate to characterize a trip than the trip purpose variable. Whether to label the trip seems to be a matter of interpretation rather than a necessity.
- There is potential for an augmented transit demand origin-destination matrix that incorporates multi-day attributes such as trip time flexibility and route preference.
- An implication from the results of travel behaviour analysis is that the use of multi-day pattern is fundamental in refining algorithms for imputation and alighting stop estimation. The derived activity schedule (Figure 8.9) suggests that the unknown values may be caused by wrong assumptions in the alighting stop estimation algorithm and that the use of multi-day trip attributes may fill in the unknown.

CHAPTER 9 LIMITATIONS AND GENERALIZATION OF FINDINGS

Methodological and analytical approaches have been proposed and demonstrated in the previous chapters. Although smart card AFC systems have become more prominent, not all transit agencies have disaggregated located-stamped validation data at their disposal due to various reasons such as the unavailability of AVL systems or data access right issues. On the one hand, it is important to recognize that each smart card AFC has a unique setup and each transit network has specific details in its planning and operations. They must be carefully studied and considered when devising methodological and analytical procedures. On the other hand, concepts proposed in this research, such as the multi-day informational approach, the use of spatial-temporal coincidence to detect transfer trips and boarding stop imputation based the boarding history of individuals, can easily be applied or transposed to other sets of data with similar structure for processing and enrichment.

9.1 Limitations

Some of the limitations on data and methodology of this research are discussed in the following paragraphs.

- The smart card data from the AFC system of the STO are relatively simple and uniform:
    - The system is managed by one medium-size transit agency which operates only one mode;
    - There are only monthly fare products available to smart card holders;
    - There is only one type of fare validation equipment;
    - All vehicles are equipped with GPS system which is integrated with the AFC system.
    - The amount of data involved is relatively small, with less than one million validation records a month.
- The proposed validation strategy handles errors in validations data by reducing the number of transactions which would otherwise be discarded or unaltered. There is no perfect solution to recover all data, nor guarantee that the imputed values are correct

because of the difficulties to obtain the "ground truth". The strategy however guarantees the coherence in data. The same logic applies to the proposed data enrichment procedures.

- There is opportunity to improve the proposed data enrichment algorithms. The multi-day informational approach can be integrated into the alighting stop estimation by taking into account the historic boarding pattern of cardholders. Furthermore, the distance between the estimated alighting stop and subsequent boarding stop cannot be the only indicator to measure the "goodness-of-fit" of the estimated values. Transfer trips can be more accurately identified using spatial-temporal coincidence that takes into account routes that serve similar destinations.

9.2 A Multi-modal and Multi-operator AFC System

Part of the main attractiveness of a smart card AFC system is its ability to interoperate among different transit agencies and across various travel modes (Acumen Building Enterprise, Inc. et al., 2006). The data from the STO, which involves only one transit agency and the bus mode, do not necessarily reflect the complexities that can occur in a large-size multi-modal and multi-operator AFC system, such as the smart card AFC system from the Greater Montréal Area, OPUS. In the following sections, partially available data from OPUS will be used to illustrate the complexity and the potential of other type of data from a multi-modal and multi-operator AFC system.

The territory of the regional transport authority in the Greater Montréal Area, Agence métropolitaine de transport (AMT), encompasses 83 municipalities and one Indian reserve, and has a total population of about 3.6 million. In general, each municipality is responsible to organize and operate transit service within its territory. Each transit operator establishes the local fare structure, fare products and fare control strategy. 17 public transit operators provide transit service with multiple modes in the area:

- the métro, a completely underground heavy rail network with 68 stations on four lines;
- the commuter rail network with 5 lines and 51 stations;
- the express and local bus networks;
- the paratransit service.

Those networks are interconnected to form a regional network and transfers among modes are possible at regional termini, train and métro stations. Transit operators use different types of fare validation equipments, and some have access to AVL and APC systems. The nature of transit services is diverse among the operators: the central city, inner-suburb and the outer-suburb operators have very different clienteles, level of service and route geometry, resulting in very different planning and operations approaches. There are hundreds of fare products within the system and the number of validation records can reach more than 1.5 million a day. This presents technical challenges and requires additional resources in order to make systemic use of the data.

As mentioned in chapter 3, smart card AFC systems generate not only validation records, which the previous chapters have focused on. They also generate at least two other types of data, namely sales data and verification data, which can complement validation data and contribute to the planning and operations of transit systems (Figure 3.1). The following sections illustrate some potential applications of those data:

- spatial-temporal measure of the sales of local and integrated fare products with sales data;
- spatial-temporal usage and purchasing pattern of fare products with validation and sales data;
- boarding and transfer activities of a mode or a network with validation data;
- fare evasion measure with verification data.

9.3 Potential Applications of Sales Data

Other than for fare collection purposes, sales data provide information in a diverse context and can be used for applications such as:

- the management and marketing of fare products by studying sales volume and spatial distribution;
- the study of purchase behaviour of transit users from a spatial-temporal perspective;
- the usage pattern of various types of sale equipments.

9.3.1 Sales Volume and Spatial Distribution

Figure 9.1 maps the spatial distribution of the sales of integrated fare monthly pass products by zone (denoted by TRAM 1 to 8). The black dots represent points of sale that sell integrated fare products with OPUS. They include regional termini, commuter rail stations and retail stores. The relative sales intensity for each product group is spatially shown by kernel density and distinctive sales patterns are revealed:

- The CBD, where the termini of five commuter rail lines and regional buses are located, is a major point of sales for all products.
- Sales are important at major intermodal junctions: commuter rail-métro, métro-regional bus, commuter rail-métro-regional bus.
- Otherwise, the sales location of products is highly spatial according to the fare zones.

This information can be used to study spatial sales pattern and purchase behaviour as well as to target specific clientele in communication campaign.

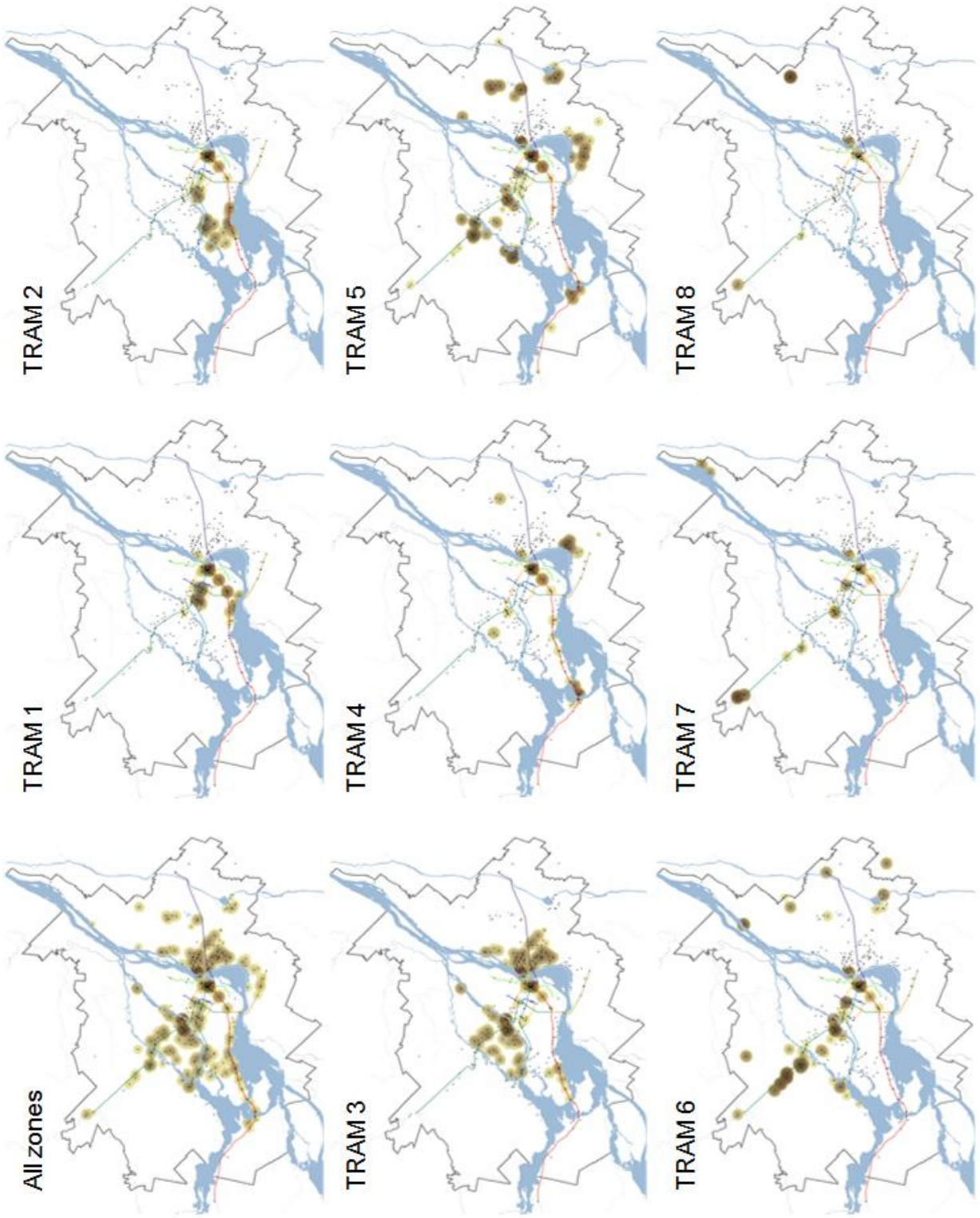


Figure 9.1 Spatial distribution of sales of monthly integrated fare products (Chu & Bergeron, 2010).

9.3.2 Purchase Behaviour of Transit Users

Figure 9.2 presents the sales of all integrated fare monthly passes by OPUS as cumulative percentage. The pass is valid for unlimited travel within a calendar month and is available for sale five months in advance. With “1” representing the first day of a month and the first day of use for a monthly pass, the figure shows the temporal distribution of sales by day, shedding light on traveler’s aggregate purchase behaviour. Each of the monthly passes has more than 70,000 units sold. About 6% of the sales are completed more than a week before the first day of the month and 98% are completed within the first week of the month. Sales pattern varies slightly from month to month. Only 48% of all November passes are sold one day before the start of the month, compared to 66% and 69% for the October and December passes. It is probably due to the fact that November 1st is a Sunday. As a consequent, about 20% of the purchases are made on the first weekday of the month, which is November 2nd. The data and analysis suggest that very few people purchase the monthly pass in advance even though it is put on sale five month in advance. In addition, the day of week plays an important role on the purchase pattern of monthly pass. The information can be used by the marketing and public relations departments to encourage advance purchase and to prevent long queue.

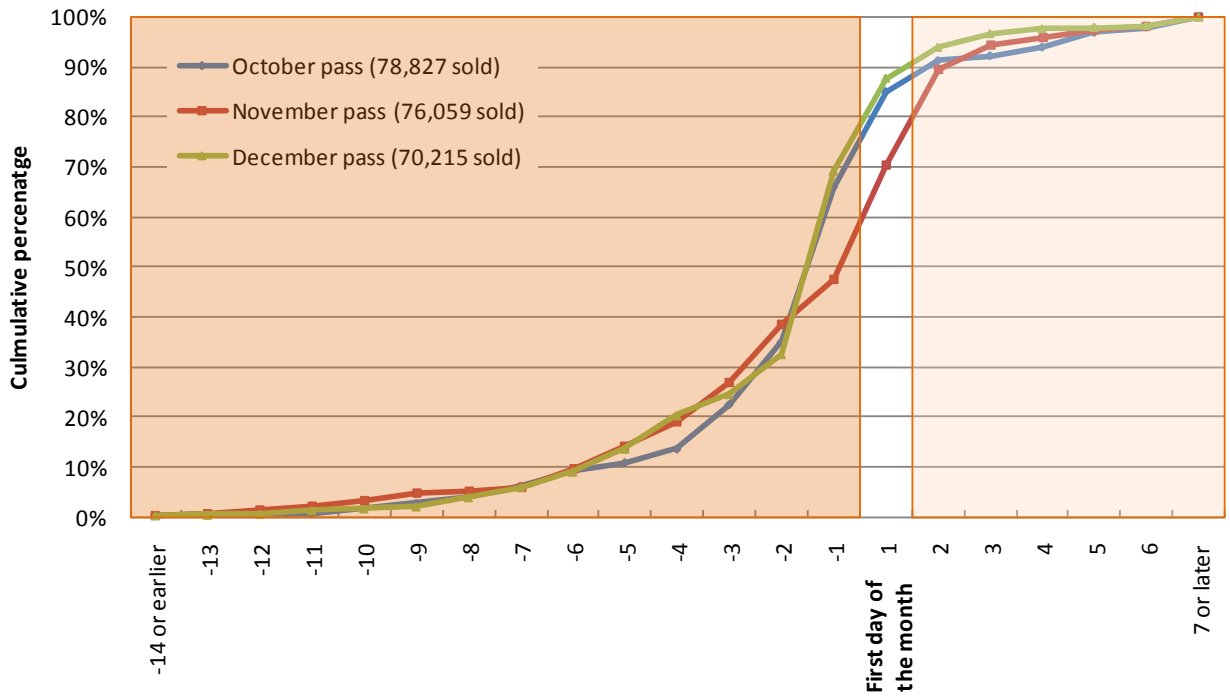


Figure 9.2 Monthly pass sales by day of month for purchase behaviour study (Chu & Bergeron, 2010).

9.3.3 Usage Pattern of Sales Equipments

Using disaggregate sales data with transaction time precise to the second, Figure 9.3 illustrates the combined temporal sales pattern on November 30, 2009 from four automatic vending machines at the Longueuil Terminus. The important sales of December monthly passes provide an interesting setting to study the capacity and level-of-service of the sale equipments. First, the duration of a sale is calculated from the gap between two successive transaction times, using both successful and unsuccessful sale attempts. Second, the average duration for successful sales for each 15-minute interval is derived. The observed minimum average duration during the day can then be identified. On the one hand, the duration is bounded by the user delay in the lower limit. On the other hand, the duration is determined by a successive transaction. Therefore, by assuming that the capacity is reached at some point during that day (no gap between two users), the observed operational capacity can be estimated.

The observed minimum duration per 15-minute interval is 62 seconds, which gives a maximum capacity of 58 sales per 15 minute for 4 machines combined. This would represent the best

scenario in the observation. Considering variance in user delay and time lost due to unsuccessful sale attempts, during intervals with an average duration close to the minimum, namely from 7:00 to 9:29 and 15:15 to 19:14, the automatic vending machines are operated at or close to capacity. A more elaborate statistical model incorporating variance in user delay can be incorporated. As shown in the purchase behaviour study, special pattern can only be detected with the analysis of several months of data. Therefore, analysis should be performed continuously in order to verify the hypotheses.

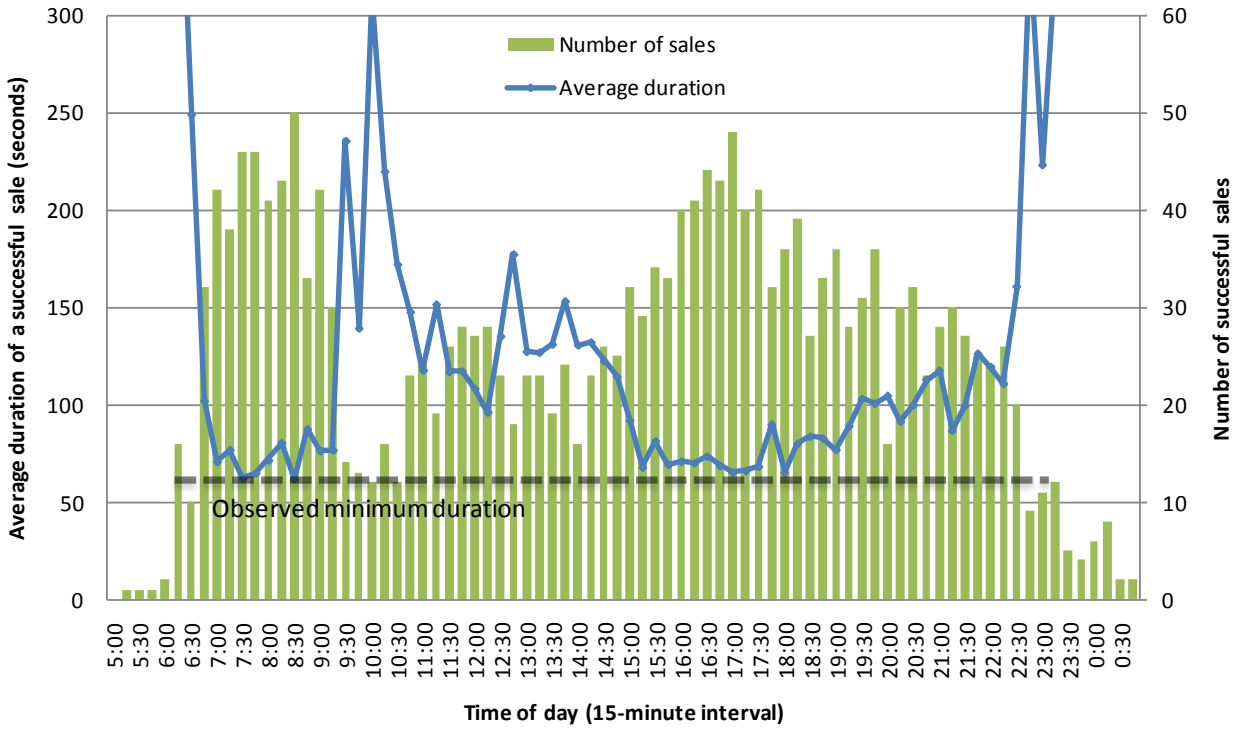


Figure 9.3 Fare product sales pattern of automatic vending machines for level of service study (Chu & Bergeron, 2010).

9.4 Applications of Validation Data

9.4.1 Macroscopic Analysis of Validation Data

Validation transaction records are generated when a client enters the transit network or transfers between modes. As such, the data provide an accurate measure of network usage, including variations due to seasons, incidents and special events. Figure 9.4 shows the total number of validations in the métro network for the month of November, 2009. They include both first

validations (entry into the local transit network) and transfer validations (within the same or into other transit networks) as well as local and integrated fare products. The size of the pie indicates the cumulative entries in the month and the ridership of each station. This information is up-to-date and can be used for the distribution of government subsidies.

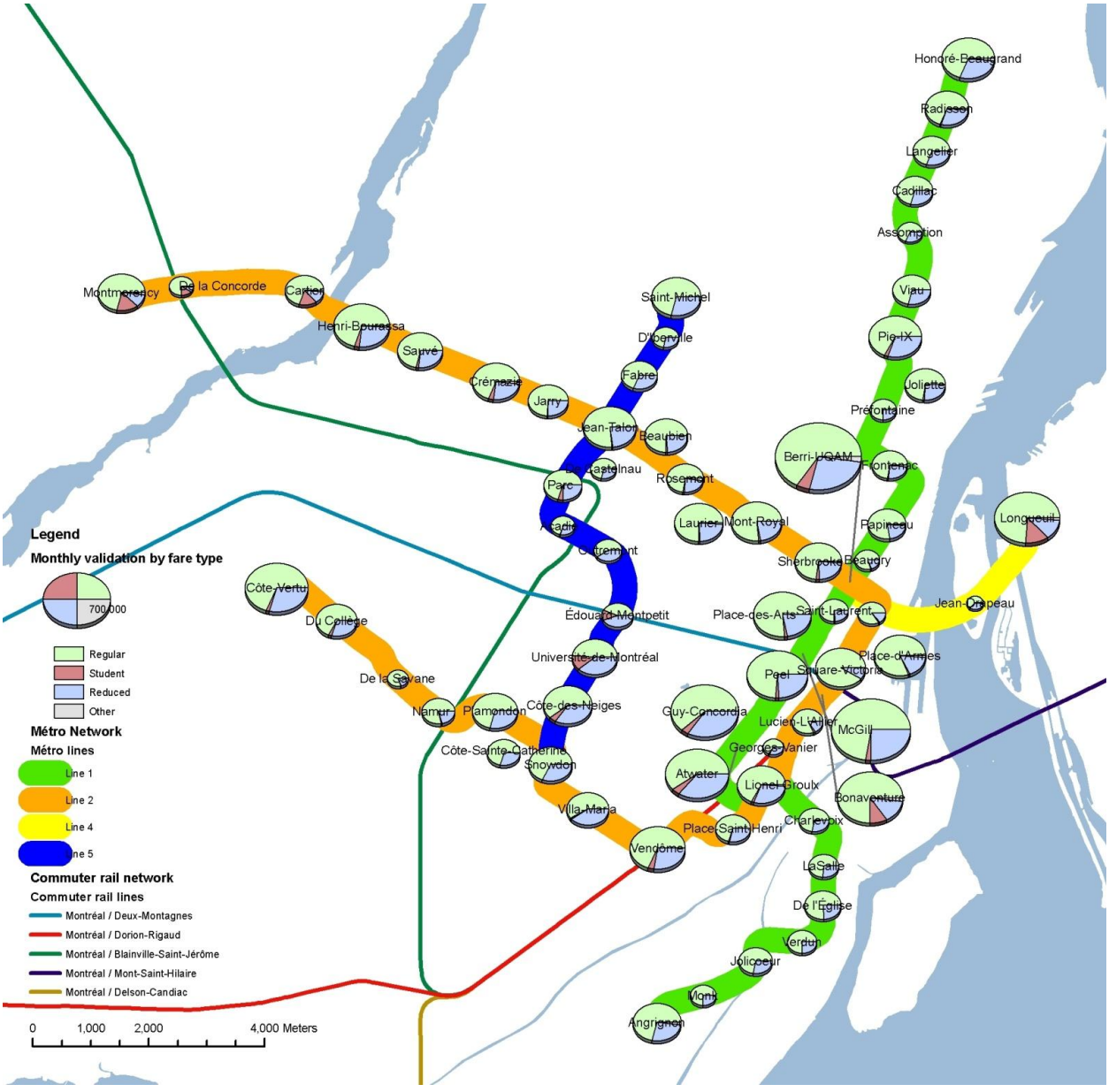


Figure 9.4 Monthly validation (November, 2009) in the métro network by fare type (Chu & Bergeron, 2010).

Moreover, each entry point of the transit network can be characterized by the type of clientele that it serves by using the fare product as an attribute. Figure 9.4 displays the share of each fare type at the entry station of the métro network. Fare products are grouped into four types:

- regular – local and integrated fare products used by adults
- student – integrated fare products for students between the age of 18 and 25
- reduced – local fare products for students up to the age of 25, integrated products for children up to the age of 18 and senior aged 65 and over
- other – employee, free and promotional fare products

Université-de-Montréal and Édouard-Montpetit stations on Line 5 have over half of their clientele using student or reduced fare. They are mainly college and university students as there are several educational institutions located nearby. The same dominance cannot be observed at other stations close to universities, namely Guy-Concordia, McGill and Berri-UQAM, due to a more diverse range of destinations in the CBD. The use of integrated fare products is more prominent at stations outside the Island of Montréal and at Bonaventure where the regional bus terminus is located.

9.4.2 Microscopic Analysis of Validation Data

In a microscopic scale, the within-day validation pattern can be examined. The total number of entries by 15-minute interval in a weekday (Thursday) of three métro stations, namely Longueuil, McGill and Montmorency, is illustrated in Figure 9.5. Although the number of entries at Longueuil station doubles that of Montmorency, both stations exhibit very similar usage pattern. They have a strong AM peak (6:00 to 8:59), accounting for almost half of the daily ridership; a noticeable PM peak (15:30 to 18:29) with about 16% of the daily ridership; a stable mid-day ridership and a low off-peak. Their similarities may be explained by the fact that they both are regional transit terminus located in the suburb with many transfer activities. In contrast, McGill station, serving important destinations in the CBD, displays another usage pattern. It is characterized by an intense PM peak, accounting for half of the daily ridership. There are also important off-peak activities. The small peak from 21:00 to 21:14 coincides with the store closing hour on Thursdays. The temporal dynamic allows operators and planners to evaluate the level-of-

service of transit services and of validation equipments, to estimate expansion factors in demand forecasting, and to plan alternative transit service in case of an emergency.

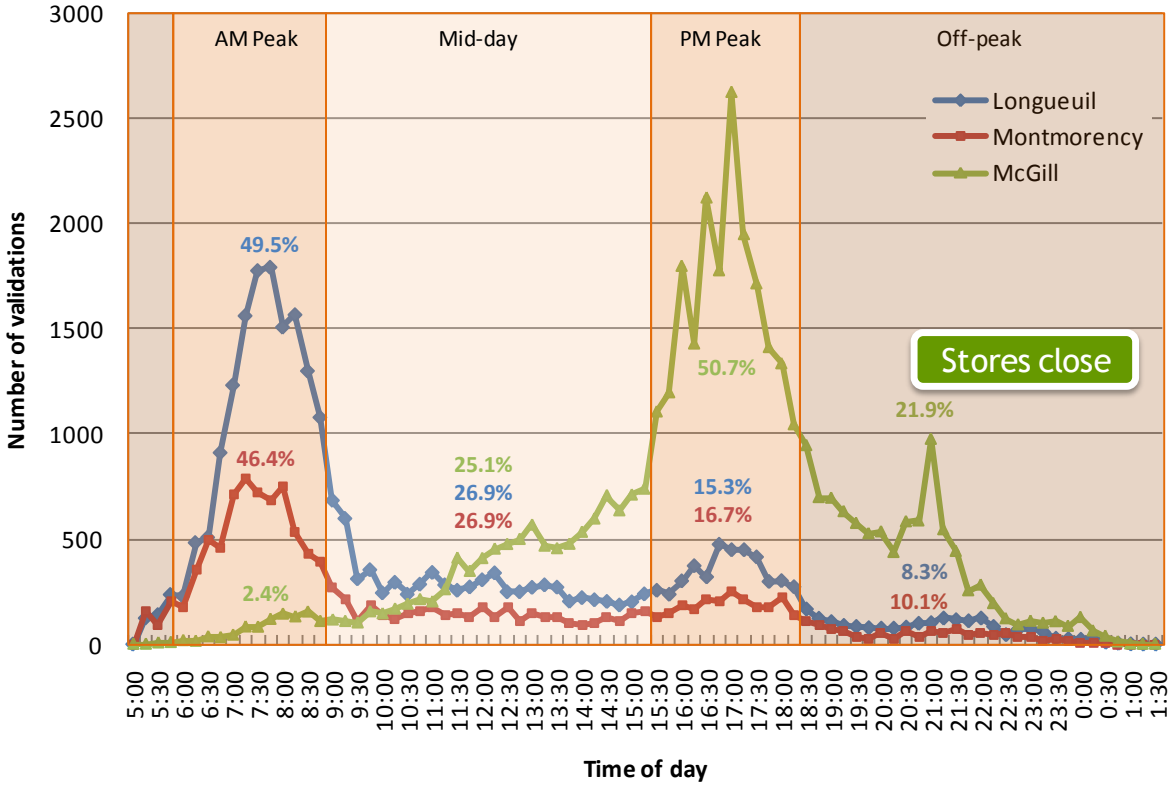


Figure 9.5 Temporal distribution of entry validations in three métro stations (Chu & Bergeron, 2010).

For finance and marketing purposes, detailed information on how often and when a fare product is used is captured. This is especially important for unlimited travel passes, since this information is required for equitably distributing revenue from integrated fare products and government subsidy on fare integration.

The advantage of disaggregate fare validation can be highlighted with the following illustration: instead of computing an average number of transactions for a fare product as a ratio of total number of transactions and the total number of product sold, disaggregate data provide the number of transactions for the product on each card. As a result, a distribution of usage rate across all holders of the same product can be known. At the same time, it is possible to merge information on the usage rate of other fare products loaded on the same card.

Validation data can be used to validate or replace APC or turnstile data, monitor validations made by various fare categories and products, calibrate transit network assignment model, measure network usage for revenue or deficit sharing among transit agencies and complement other surveys such as the large-scale origin-destination survey where public transit trips may not be sufficiently sampled in areas with low transit share.

9.5 Applications of Verification Data

The commuter rail network in Montréal has a self-service and barrier-free fare system. There is no turnstile at the entrance and users are responsible to validate their ticket when entering the controlled area and to keep a proof of payment during the journey. Using portable card readers, a team of inspectors regularly perform three types of inspection to deter fare evasion:

- on-board;
- at the entrance station;
- at the exit station.

Data generated by the AFC system have the potential to address some concerns regarding a self-service and barrier-free system (Multisystems et al., 2002), particularly in:

- measuring more accurately the fare evasion rate;
- monitoring the productivity of the inspectors;
- improving the inspection strategy.

9.5.1 Estimating Fare Evasion Rate

Verification data are generated by a handheld verification unit which reads a rechargeable or a disposable smart card. According to the parameters entered by the inspector, which include the inspector identification number, the type and the location of inspection, the unit determines whether the fare product is valid. In addition, the unit records the exact time and the reason of refusal. In contrast, methods that rely on inspectors' field report only provide approximate data, which affects the reliability of fare evasion measure.

Fare evasion can be analysed by date. Figure 9.6 shows the daily number of verifications for November 2009 in the commuter rail network. It also shows the evasion rate which is the ratio between the number of infractions and the number of verifications. Fare evasion rate can also be analyzed by day of week (Table 9.1). The preliminary result suggests that the level of evasion may vary across days. It is also possible that the fluctuation can be explained by the fact that evasion can be more easily captured on some days. Data on the type and location of inspection allows analyses by train line and fare zones, as illustrated in Table 9.2.

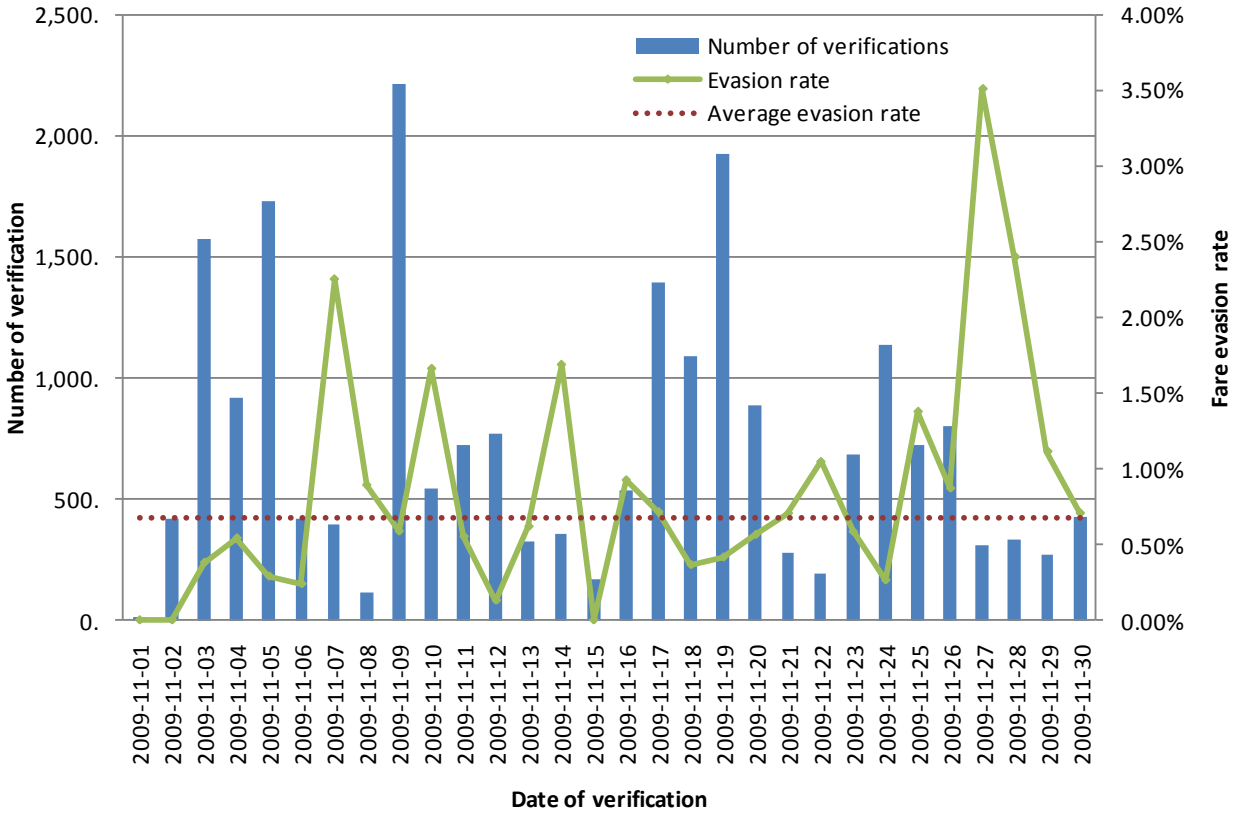


Figure 9.6 Verification and fare evasion rate in November 2009 (Chu & Bergeron, 2010).

Table 9.1 Evasion rate by day of week (Chu & Bergeron, 2010).

Day of week	Number of infractions	Number of verifications	Evasion Rate
Monday	25	4,278	0.58%
Tuesday	28	4,648	0.60%
Wednesday	23	3,464	0.66%
Thursday	21	5,221	0.40%
Friday	19	1,937	0.98%
Saturday	25	1,370	1.82%
Sunday	6	751	0.80%
Total	147	21,669	0.68%

Table 9.2 Evasion rate by fare zone and station on the Montréal / Deux-Montagnes train line (Chu & Bergeron, 2010).

Line	Fare zone	Station	Number of verifications	Number of infractions	Evasion rate
Deux-Montagnes	1	Centrale	1407	8	0.57%
		Canora	592	15	2.53%
		Mont-Royal	158	1	0.63%
		Montpellier	1545	8	0.52%
	2	Du Ruisseau	871	7	0.80%
		Bois-Franc	248	8	3.23%
		Sunnybrooke	264	1	0.38%
		Roxboro	1062	4	0.38%
	3	Île-Bigras	69	0	0.00%
		Ste-Dorothée	263	3	1.14%
		Grand-Moulin	42	0	0.00%
	5	Deux-Montagnes	186	0	0.00%
		St-Eustache	2	0	0.00%
	Unknown		72	0	0.00%
	Total		6781	55	0.81%

The burden of manually recording the number of verifications, date and time can be lifted from the inspectors with smart card AFC systems. At the same time, the productivity of inspectors can be measured as they are required to log onto the handheld verifier. However, analysis results are largely dependent on the accuracy of the parameters entered into the unit by the inspector. The unit merely acts as a device which facilitates the inspection and the recording of data. For example, analysis of productivity implies that inspectors log on with the correct ID number and fare evasion analyses assume the type and location of inspection are corrected input.

Since inspections are performed on a sample of transit users, the evasion rate can be biased if there is no sound inspection strategy. Analysing verification data helps to identify weakness in strategy and to improve it. For on-board inspection, there is a possibility of associating inspection records with train departure using disaggregate data. By integrating with the time stamp of verification data with timetable or, preferably GPS data, evasion rate by train departure can be studied.

CHAPTER 10 CONTRIBUTIONS AND PERSPECTIVES

10.1 Summary of Key Findings

The research proposes conceptual, theoretical, methodological and analytical frameworks to use of data from smart card AFC systems with the purpose of improving the understanding of public transit. This understanding is multi-faceted: it includes the service perspective, the demand perspective, the travel behaviour perspective and the management perspectives. Based on actual data from smart card AFC systems and information technologies such as relational database, GPS, GIS, data mining, spatial statistics and visualization, the research spawns new findings and confirms previous results by other researchers. Findings can be divided into two groups: contributions in methodology and experimental results from the specific datasets.

10.1.1 Contributions to Methodology

10.1.1.1 Conceptual Framework

The research inherits from the MADITUC group the tradition of the totally disaggregated approach and the object-oriented approach. They become the guiding principles in the research. Objects in a smart card validation record are defined and examined with respect to components of a public transit system. From that gains an understanding of the smart card data and establishes methodological procedures for subsequent data processing and analyses.

10.1.1.2 Theoretical Framework

The research provides a critical analysis of current knowledge on a wide range of subjects, mostly importantly on transit planning, processing and applications of passively collected data, travel behaviour studies and information technologies, through a comprehensive literature review. Benefiting from the extensive experience on CATI regional household origin-destination travel survey in the Montréal region, survey data quality issues are discussed. It is argued that with their unique properties and benefits over traditional survey methods, smart card validation data can be used as a multi-use travel survey instrument.

10.1.1.3 Methodological Framework

Methodology on data processing and enrichment are proposed and demonstrated to address the issues inherent to all passive data as well as those specific to smart card validation data.

To assure consistency in data, a validation strategy is put forward. An error-detection procedure uses logical rules to flag irrelevant, erroneous and suspect values. An imputation procedure replaces them with probable values based on repetition in planned transit service and historic boarding pattern of cardholders.

Enrichment procedures aim to infer information not captured by smart card AFC data. Several algorithms are applied to the data with a view to reconstruct individual itineraries from individual boarding validations and stop-level spatial-temporal load profiles: the alighting location estimation makes use of the card-level boarding chain and transit network geometry to determine the most probable alighting stop; the estimation of spatial-temporal paths for vehicles provides stop-level vehicle running time and is based on boarding times and scheduled arrival times; the transfer activity identification uses the concept of spatial-temporal coincidence. The resulting itineraries serve as input in a transit assignment model, namely MADITUC, for assignment based on the observed routes or for simulation. The set of itineraries can be used as a reference demand to test different route geometries and service plans. By associating stop locations with an external data base of points of interest, trip ends can be linked to specific trip generators. The procedure is especially promising for certain sub-group of cardholders, such as students, who have additional constraints on time and location choice.

The procedure is transposed to a multi-day context for the study of travel behaviour, both for an individual and individuals belonging to a sub-group. Multi-day boarding data, which can be aggregated into larger units to reveal travel pattern, allows the discovery of anchor points of individual cardholder with hotspot analysis. The hotspots and trip ends are tied to actual or symbolic locations. Trip details, such as trip purpose, multi-day characterization and activity schedule related to transit trips can therefore be inferred and interpreted.

10.1.1.4 Analytical Framework

Smart card data provide opportunities for analyses that could not be completed with data gathered by traditional methods. Data that have undergone processing and enrichment procedures are

analyzed. The techniques are chosen with the aim of highlighting the distinctive properties and analytical potential of the data: the disaggregate property, continuous and complete coverage as well as detailed spatial and temporal resolutions. In terms of transit planning, the research provides an exploratory analysis on key objects; studies the spatial-temporal distribution of boardings and alightings at various levels of aggregation with GIS and spatial statistics; analyzes transfer activities; performs transit network analyses; generates indicators on service supply and consumption from the operator's perspective and trip details from the user's perspective; models movements in activity space; investigates multi-day travel pattern and behaviour with tools such as data mining.

Applications for transit management involve the use of sales data, validation data and verification data. They provide information on sales volume and spatial distribution of fare productions; on the purchase behaviour of transit users; on the usage pattern of sales equipments; on the spatial-temporal distribution of validations and on fare evasion rate.

The analyses take advantage of the multi-dimensional aspect of the data by simultaneously considering multiple attributes. The use of visualizations is emphasized to reduce dimensionality of the data and facilitate the understanding of results.

10.1.2 Experimental Results from the Specific Datasets

The methodology and analyses regarding data processing and enrichment are demonstrated with a month of location-stamped validation data from the smart card AFC system of the STO. Sales, validation and verification data from the OPUS system in Montréal are also used to illustrated analyses for transit management purposes.

10.1.2.1 Results from the STO Dataset

The error-detection procedure finds that more than 15% of fare validation records contain erroneous or suspect data for the month of February 2005. On an average weekday, the data recovery process imputes route and stop information for 88% of flagged records, improving the consistency of the dataset from 84% to 98%. The concepts of repetition in planned transit service and historic boarding pattern of cardholders both contribute to the process while the former provide more reliable run values. Subsequent refinement to the algorithm further improves the results.

Enrichment procedures are applied to the corrected data with the aim of reconstructing itineraries and synthesize spatial-temporal load profiles. In total, 33,775 person-trips are identified. Originally, 5,516 out of 37,781 (14.6%) boarding records were originally labelled as transfer by the fare system. Based on the definition of spatial-temporal coincidence, the algorithm identifies 4,002 (10.6%) records as transfer boardings. The result suggests that the fare system overestimates the proportion of transfer trips by nearly 40%. From a traveler's perspective, detailed itineraries of a cardholder are illustrated. Indicators such as travel time, on-boarding distance and activity are calculated. From the transit operator's perspective, spatial-temporal load profiles provides information on stop-level load factor, maximum load point, passenger-kilometres, on-time performance and travelers by fare type for each run or a group of runs. Ridership can be interpreted when stops are spatially associated with points-of-interests.

Totally disaggregated transit assignment model is used to load itineraries derived from smart card boarding data. About 93% of all trips are used because some trips do not have a valid alighting location. The results allow the calculation of most travel demand statistics, including passengers-kilometres, passengers-hours, load factor, transfer wait time, etc., for the network during an analysis timeframe. Spatial analyses and statistics on boardings and alighting, along with the activity space profile, show daily migrations patterns in and out the CBD of Ottawa. Distinctive travel patterns in space and time are also detected for specific sub-group, such as cardholders with student fare.

In terms of travel behaviour, a hotspot analysis with multi-day data detects zero to four anchors for 21,691 out of 21,742 cardholders. For the student populations, 43% (2,565 / 6,030) of the cards are associated with specific high schools and pre-university colleges by an assignment algorithm. Each trip of a cardholder is analyzed with respect to the multi-day travel profile to augment the trip characterization and trip interpretation. A demonstration of association rule algorithm on the 37 boardings of one cardholder yields interesting rules on travel pattern. The result of a classification algorithm can be used as a proxy for measuring regularity in travel pattern.

10.1.2.2 Results from the OPUS Dataset

From the analysis for sales data, it is shown that the sale location of products is highly spatial according to the fare zones, except in major hubs and termini. Daily sales pattern of calendar

month products varies slightly from month to month, depending on whether a month ends with non-work day. Sales data from automatic ticketing machines can reveal whether the latter are working at capacity. Aggregated fare validation data reveal distinctive temporal signature and user composition of different types of stations. Fare verification data improve the accuracy of estimating fare evasions rate and help to enhance fare verification strategy.

10.1.3 Implications of the Research

10.1.3.1 Implication for Data Collection and Processing

Data gathered from passive sources, specifically from a smart card AFC with GPS, are an excellent source of information for transit planning and travel behaviour study. However, they contain a significant amount of errors. Data need to be validated before any detailed analysis. Data enrichment procedures compensate for the shortcomings inherent to passive data.

10.1.3.2 Implication for Transit Planning and Travel Behaviour Study

The proposed methods and analyses demonstrate the potential of located-stamped smart card AFC data for transit planning. A variety of data needs for planning can be satisfied. The data themselves can reveal transit trip pattern and at the same time, provide a source for other travel behaviour studies.

10.1.3.3 Other Perspective

Another important subject related to the use of smart card AFC data is privacy issue (Dempsey, 2008) and data ownership issue. They are however out of the scope of this research.

10.1.3.4 Contributions to the Transportation Literature

The findings from this research have led participations in conferences of academic or professional nature:

- “Décortication de données de cartes à puce en vue d'un modèle de planification d'un réseau T.C.” presented in the 2007 Congrès annuel de l’Association québécoise du transport et des routes (Chu & Chapleau, 2007a);

- "Imputation techniques for missing fields and implausible values in public transit smart card data" (Chu & Chapleau, 2007b) and "Modeling transit travel patterns from location-stamped smart card data using a disaggregate approach" (Chapleau & Chu, 2007) presented in the 2007 World Conference on Transport Research. "Leveraging data of a multi-modal and multi-operator smart card automatic fare collection system from the perspective of a regional transit authority" (Chu & Bergeron, 2010) presented in the 2010 World Conference on Transport Research.
- "Enriching archived smart card transaction data for transit demand modeling" (Chu & Chapleau, 2008), Driver-assisted bus interview: passive transit travel survey with smart card automatic fare collection system and applications" (Chu, Chapleau & Trépanier, 2009) and "Augmenting transit trip characterization and travel behavior comprehension with multi-day location-stamped smart card transactions" (Chu & Chapleau, 2010) are presented in the 2008, 2009 and 2010 Annual Meeting of the Transportation Research Board respectively. These papers are subsequently published in the following peer-reviewed journal:
    - Transportation Research Record 2063: Transit: Management, Technology, and Planning 2008 (Chu & Chapleau, 2008);
    - Transportation Research Record 2105: Information Systems, Geographic Information Systems, and Advanced Computing 2009 (Chu, Chapleau & Trépanier, 2009);
    - Transportation Research Record (pending publication, Chu & Chapleau, 2010).
- "The ultimate survey for transit planning: complete information with smart card data and GIS" presented in the 2008 International Conference on Survey Methods in Transport (Chapleau, Trépanier & Chu, 2008);
- "Analysing multi-day travel behaviours with public transit smart card transaction data" presented in the 2009 International Association on Travel Behaviour Research (Chu & Chapleau, 2009).

10.2 Future Research Directions

This research paves the way for further research on smart card AFC data. The following list comprises relevant subjects that address the limitation of current research:

- Large-scale applications of data from a multi-operator and a multi-modal smart card AFC system;
- Synthesis of passive data from different sources for public transit applications;
- Studies of transit route-choice behaviour;
- Development of a transit assignment model which handles multi-day itineraries.

10.2.1 Large-scale Applications of Data from a Multi-operator and Multi-modal Smart Card Automatic Fare Collection System

An urban area often involves multiple jurisdictions with each of them organizing their own public transit service and fare products. One of the appeals of a smart card AFC system is the possibility to inter-operate among transit agencies, thus offering the flexibility of providing integrated fare products for transit users. In previous chapters, methods and analyses are proposed and demonstrated with data from the STO, which has a single mode and standardized equipments and fare structures. However, in an urban area involving multiple transit agencies, the processing and application of smart card data will encounter numerous complexities, among which:

- Multiple transit modes: a large urban area usually offers more than one mode of transit services such as a subway, commuter rail, express and local bus networks and paratransit. The data requirement and the algorithms, such as for alighting location estimation, need to be adapted to specific mode. The transferring opportunity and the route choice multiply as the number of modes and routes increases.
- Multiple fare structures: even though an integrated fare structure is possible, local fare structures may still remain. Single fare, inclusive zonal fare and inter-zonal fare structures can be adopted by different transit agencies. Each card is also allowed to hold several fare products. These may provide additional information for estimating boarding and alighting locations but their complexities can also prove to be an obstacle.

- Multiple fare validation strategies: fare validation is not standardized among modes. Strategies such as entry-exit validation, on-board validation, at-station validation, honour-system validation or combination of these can all be present in a multi-mode system.
- Non-standardized equipments: the adoption and choice of data collection technologies by various transit agencies are different. Some vehicle fleets may have AVL system integrated to the smart card AFC system while some do not. The spatial aspect of the fare validations may be lacking.
- Non-standardized operational procedures: data processing techniques and analyses are dependent on the schedule of planned service and the association of vehicle-block with a vehicle. The algorithm needs to be adapted to the operational procedures and cultures of various transit agencies.
- Changes in the transit network: a multi-modal and multi-agency transit network is in constant evolution. Changes from each of the sub-network need to be systematically incorporated for data processing and analysis.
- The amount of data: data processing needs to cope with the amount of data generated in a large multi-modal system.

Hence, deriving itineraries in a multi-modal network from boarding-only smart card AFC data for a transit assignment model would be a challenging problem, while at the same time, would extend the methodologies demonstrated in this research.

10.2.2 Synthesis of Passive Data from Different Systems

Passive data collection technologies are increasingly common in the public transit industry. However, due to the nature of measurement and the precision of the instruments, passive data from different sources cannot easily be reconciled. For example, data from APC, AFC, AVL and odometer all provide partial truth of the reality due to measurement errors and operational procedure. They need to be reconciled and synthesized in order to generate reliable performance and travel demand indicators for public transit. Chapleau & Allard (2010) study this problem with data from the Société de Transport de Montréal.

10.2.3 Transit route-choice behaviour

Smart card AFC systems provide an unprecedented amount of observed transit route-choice data. The route-choice behaviour can be analyzed by comparing the optimal itineraries to the observed itineraries. Insights can be gained for parameters in route-choice or transit assignment models, which are difficult to obtain under current practice. For example, transfer penalty by combination of modes or by specific location can be estimated. Before and after study of travel behaviour after a service change can also performed since data are gathered continuously.

Since smart card data contain detailed time information, other phenomena that vary in time such as transit service interruption, traffic conditions, weather and special events can be integrated and analyzed.

10.2.4 Transit Assignment with Multi-day Smart Card Itineraries

The current model of transit network assignment and road network assignment uses an average demand as input. This is mainly due to the unavailability of multi-day data. Multi-day data not only provide demand in multiple days. They also provide information on transit trip rates for each cardholder, along with trip time flexibility and route preference. A transit assignment model that incorporates these pieces of information would provide more truthful outputs, resulting in better allocation of resource and service that would be more adapted to the needs of the users.

CONCLUSION

The interaction between services supplied by transit operators and the travel demand of the public is in constant evolution and difficult to measure. There is a continuous quest for information and methodology that can help reveal and facilitate the understanding of this dynamic relationship. Due to operators' focus on performance and customer as well as the implementation of transit technologies, recent paradigm shift in data needs and data availability provides an opportunity to leverage new data sources. This research is based on a set of validations data from a smart card automatic fare collection (AFC) system. Using information technologies, including relational database, geographic information system (GIS), spatial statistics, data mining and visualization as the main tools, it proposes new methods in data processing, enrichment and analysis in order to better quantify transit demand, enhance operations planning, improve system management and understand travel behaviour. Operators can therefore use these spatial-temporal data to monitor service consumption in a more timely matter, estimate more precisely the demand for various level of resolution as well as perform analyses for planning, operations and management by generating relevant indicators. Three overall principles guide the research: the information (data-driven) approach, the totally disaggregated approach and the object-oriented approach. These principles lead to a multi-day information approach, an important concept used in the proposed data processing and enrichment procedures.

Within the context of transit planning and operations, the research demonstrates that information provided by smart card data is more timely, has a better coverage and a higher spatial-temporal resolution than a regional origin-destination survey. A data validation strategy, which contains an error-detection procedure and a data correction procedure, is proposed. Many tasks in transit planning can be served directly by applying standard GIS procedures and spatial statistics to the validated data. Since a disaggregate transit assignment implies a trip in the form of an itinerary, several data enrichment procedures are proposed and undertaken: alighting stop estimation for each boarding with the concept of "boarding chain"; interpolation of stop arrival time for each vehicle run according to temporal information embedded in the transactions and transfer identification with the concept of "spatial-temporal coincidence". These enrichments allow the reconstruction of complete itinerary from boarding validation data and the derivation of related objects. The proposed trip characterization with the multi-day approach allows in-depth study of

travel behaviour of an individual as well as comparison between a subgroup of cardholders who share a common anchor. The multi-day trip characterization leads researchers to rethink some fundamental aspects of trip description.

Given that the validation data from the STO are relatively simple, data from a multi-operator and multi-modal AFC system in the Greater Montréal Area, OPUS, are used to illustrate the complexity and technical challenges. They are also used to introduce the potential of other types of data, namely sales and verification data, suitable for transit planning, operations and management. Since the setup of each smart card AFC system, the transit network and its fare structure is unique, the needs on data processing and enrichment vary and are specific to each system. However, the principles, the methodological approaches and the analyses proposed in this research can be adapted and transferred to similar datasets. Perspective research topics are also proposed: the large-scale applications of smart card data from a multi-operator and a multi-modal the synthesis of passive data from different sources for public transit applications; the studies of transit route-choice behaviour and the development of a transit assignment model which handles multi-day itineraries.

REFERENCES

Acumen Building Enterprise, Inc. & Booz Allen Hamilton, Inc., (2006). *TCRP Report 115: Smartcard Interoperability Issues for the Transit Industry*. Washington, D.C.: Transportation Research Board of the National Academies.

Asakura, Y., Iryo T., Nakajima Y., & Kusakabe T. (2009). Estimation of behavioural change of railway passengers using smart card data. Paper presented at CASPT 2009, Hong Kong.

Attoh-Okine, N.O., & Shen L.D. (1995). Security issues of emerging smart cards fare collection application in mass transit. *Proceedings of the Conference on Vehicle Navigation and Information Systems (VNIS)*, pp. 523-526.

Axhausen, K.W., Zimmersman A., Schönfelder S., Rindfüser G., & Haupt T. (2002). Observing the rhythms of daily life: a six-week travel diary. *Transportation, 29*, 95-124.

Bagchi, M., & White, P.R. (2004). What role for smart-card data from bus systems? *Proceedings of the Institution of Civil Engineers: Municipal Engineer, 157*(1), 39-46.

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

Barry, J.J., Newhouser, R., Rahbee A., & Sayeda S. (2002). Origin and destination estimation in New York City with automated fare system data. *Transportation Research Record: Journal of the Transportation Research Board, No. 1817*, 183-187.

Bayarma, A., Kitamura, R., & Susilo, Y.O. (2007). On the recurrence of daily travel patterns: a stochastic-process approach to multi-day travel behavior. *Transportation Research Record: Journal of the Transportation Research Board, No. 2021*, 55-63.

Berkow, M., El-Geneidy, A., Bertini, R., & Crout, D. (2009). Beyond generating transit performance measures: visualizations and statistical analysis with historical data. *Transportation Research Record: Journal of the Transportation Research Board, No. 2111*, 158-168.

Bertini, R.L., & El-Geneidy, A. (2003). Generating transit performance measures with archived data. *Transportation Research Record: Journal of the Transportation Research Board, No. 1841*, 109-119.

Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, *17*(3), 285-297.

Bonnel, P. (2003). Postal, telephone, and face-to-face surveys: how comparable are they? In P. Stopher, & P. Jones (eds.), *Transport Survey Quality and Innovation* (pp. 215-237). Oxford: Pergamon.

Boyle, D.K. (1998). *TCRP Synthesis of Transit Practice 29: Passenger Counting Technologies and Procedures*. Washington D.C.: Transportation Research Board of the National Academies.

Bryan, H., & Blythe, P. (2007). Understanding behaviour through smartcard data analysis. *Proceedings of the Institution of Civil Engineers: Transport*, *160*(4), 173-177.

Bullock, P., Jiang, Q., & Stopher, P.R. (2005). Using GPS technology to measure on-time running of scheduled bus services. *Journal of Public Transportation*, *8*(1), 21-40.

Ceder, A. (2001). Public Transport Scheduling. In K.J. Button, & D.A. Hensher (eds.), *Handbook of Transport Systems and Traffic Control* (pp. 539-558). Oxford: Elsevier Science.

Ceder, A. (2007). *Public Transit Planning and Operation: Theory, Modeling and Practice*, Oxford: Elsevier.

Chapin, F.S. (1974). *Human Activity Patterns in the City: Things People Do in Time and in Space*. New York: Wiley.

Chapleau, R., Allard, B., & Canova, M. (1982). *MADITUC, un modèle de planification opérationnelle adapté aux entreprises de transport en commun de taille moyenne* (Publication #265). Montréal: Centre de recherche sur les transports, Université de Montréal.

Chapleau, R. (1992). La modélisation de la demande de transport urbain avec une approche totalement désagrégée. *Selected Proceedings of the 6th World Conference on Transport Research, Lyon*, (Volume II, pp. 937-948). Lyon: World Conference on Transport Research Society.

Chapleau, R., Allard, B., & Trépanier, M. (1996). Transit path calculation supported by special geographic information system - transit information system. *Transportation Research Record: Journal of the Transportation Research Board, No. 1521*, 98-107.

Chapleau, R. (2003). Measuring the internal quality of the Montreal CATI household travel survey. In P. Stopher, & P. Jones (eds.), *Transport Survey Quality and Innovation* (pp. 69-88). Oxford: Pergamon Press.

Chapleau, R., & Allard B. (2007). Spatio-temporal analysis of paratransit trips. Paper presented at TRANSED 2007, Montréal, Canada.

Chapleau, R., & Chu, K.K.A. (2007). Modeling transit travel patterns from location-stamped smart card data using a disaggregate approach. Paper presented at the 11th World Conference on Transport Research, Berkeley, CA.

Chapleau, R., Trépanier, M., & Chu, K.K.A. (2008). The ultimate survey for transit planning: complete information with smart card data and GIS. Paper presented at the 8th International Conference on Survey in Transport, Lac D'Annecy, France.

Chapleau, R., & Allard, B. (2010). Merging AFC, APC, GPS and GIS-T data to generate productivity indicators and travel demand models in public transit. Paper presented at the 12th Conference of Transport Research, Lisbon, Portugal.

Chira-Chavala, T., & Coifman, B. (1996). Effects of smart cards on transit operators. *Transportation Research Record: Journal of the Transportation Research Board*, *No. 1521*, 84-90.

Chu, K.K.A., & Chapleau, R. (2007a). Décortication de données de cartes à puce en vue d'un modèle de planification d'un réseau T.C. Paper presented at the 42nd Annual Conference of the Association québécoise du transport et des routes, Montréal, QC.

Chu, K.K.A., & Chapleau, R. (2007b). Imputation techniques for missing fields and implausible values in public transit smart card data. Paper presented at the 11th World Conference on Transport Research, Berkeley, CA.

Chu, K.K.A., & Chapleau, R (2008). Enriching archived smart card transaction data for transit demand modeling. *Transportation Research Record: Journal of the Transportation Research Board*, *No. 2063*, 63-72.

Chu, K.K.A., & Chapleau, R. (2009). Analysing multi-day travel behaviours with public transit smart card transaction data. Paper presented at the 12th International Conference on Travel Behaviour Research, Jaipur, India.

Chu, K.K.A., Chapleau, R., & Trépanier, M. (2009). 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, No. 2105*, 1-10.

Chu, K.K.A., & Chapleau, R. (2010). Augmenting transit trip characterization and travel behavior comprehension with multi-day location-stamped smart card transactions. Paper presented at the 89th Annual Meeting of the Transportation Research Board, Washington, D.C. (Publication forthcoming in the TRR)

Chu, K.K.A., & Bergeron, D. (2010). Leveraging data of a multi-modal and multi-operator smart card automatic fare collection system from the perspective of a regional transit authority. Paper presented at the 12th Conference on Transport Research, Lisbon, Portugal. (Publication forthcoming in the Selected Proceedings)

Chung, E.-H., & Shalaby A. (2005). A trip bases reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology, 28*(5), 381-401.

CrimeStats (2005). *CrimeStats Manual Version 3.0*. Last accessed on December 20, 2010. From http://www.icpsr.umich.edu/CrimeStat/

Cui, A. (2006). *Bus passenger origin-destination matrix estimation using automatic data collection systems*. M.S. Thesis, Massachusetts Institute of Technology, M.A., United States.

De Cea, J., & Fernández, E. (2000). Transit-assignment models. In K.J. Button, & D.A. Hensher, (eds.), *Handbook of Transport Modelling* (pp. 497-508). Oxford: Elsevier Science.

Dempsey, P.S. (2008). *TCRP Legal Research Digest 25: Privacy Issues with the Use of Smart Cards*. Washington, D.C.: Transportation Research Board of the National Academies.

Desharnais, M-C., & Chapleau, R. (2010). A disaggregate investigation of demand patterns for paratransit. Paper presented at the 89th Annual Meeting of the Transportation Research Board, Washington, D.C.

Dionne, V. (2006). *Diagnostic et performance d'une ligne de train de banlieue à l'aide de données GPS*, Mémoire M.Sc.A., École Polytechnique de Montréal, Q.C., Canada.

Elango, V., Guensler R., & Ogle J. (2007). Day-to-day travel variability in the commute Atlanta, Georgia, study. *Transportation Research Record: Journal of the Transportation Research Board, No. 2014*, 39-49.

Farhan, A., Shalaby, A., & Sayed, T. (2002). Bus travel time prediction using GPS and APC. *Proceedings of the 7th International Conference on Applications of Advanced Technologies in Transportation*, (August 2002, pp. 613-623). Cambridge: ASCE.

Farzin, J. (2008). Constructing an automated bus origin-destination matrix using farecard and GPS data in São Paulo, Brazil. Paper presented at the 87th Annual Conference of the Transportation Research Board, Washington, D.C.

Frawley W., Piatetsky-Shapiro G., & Matheus C. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine, Fall*, 213–228.

Furth, P.G., Hemily, B., Muller, T.H.J., & Strathman, J.G. (2006). *TCRP Report 113: Using Archived AVL-APC Data to Improve Transit Performance and Management*. Washington, D.C.: Transportation Research Board of the National Academies.

Golani, H. (2007). Use of archived bus location, dispatch, and ridership data for transit analysis. *Transportation Research Record: Journal of the Transportation Research Board, No. 1992*, 101-112.

Hammerle, M., Hayes, M., & McNeil, S. (2004). Use of automatic vehicle location and passenger count data to evaluate bus operations. *Transportation Research Record: Journal of the Transportation Research Board, No. 1903*, 27-34.

Hand, D., Mannila H., & Smyth P. (2001). *Principles of Data Mining*. Cambridge: MIT Press.

Hanson, S., & Huff J.O. (1988). Repetition and day-to-day variability in individual travel patterns: implications for classification. In R. Golledge, & H. Timmermans (eds.), *Behavioral Modelling in Geography and Planning* (pp. 368-399). New York: Croom Helm Ltd.

Hofmann M., & O'Mahony, M. (2005a). The impact of adverse weather conditions on urban bus performance measures. *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria* (pp. 84-89).

Hofmann M., & O'Mahony, M. (2005b). Transfer journey identification and analyses from electronic fare collection data. *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria* (pp. 34-39).

Huff, J.O., & S. Hanson (1990). Measurement of habitual behavior: examining systematic variability in repetitive travel. In P. Jones (ed.), *Developments in Dynamic and Activity-Based Approaches to Travel Analysis* (pp. 229-249). Aldershot: Gower Publishing Co.

Iowa Department of Transportation (2005). Glossary of Terms. *Transit Manager's Handbook*. Last accessed on December 1, 2006. From http://www.iatransit.com/links/handbook/glossary.asp

Jang, W. (2010). Travel time and transfer analysis using transit smart card data. Paper presented at the 89th Annual Meeting of the Transportation Research Board, Washington, D.C.

Kittleson & Associates, Inc., KFH Group, Inc., Parsons Brinckerhoff Quade & Douglass, Inc., & Hunter-Zaworski, K. (2003). *TCRP Report 100: Transit Capacity and Quality of Service Manual* (2nd ed.). Washington, D.C.: Transportation Research Board of the National Academies.

Kuhnimhof T., & Wassmuth, V. (2002). Do you go to the movies during your lunch break? Trip-context data-based modeling of activities. *Transportation Research Record: Journal Of The Transportation Research Board, No. 1807*, 34-42.

Lehtonen et al. (2002). Use of Smart Card Payment System Data. *Nordic Road and Transport Research, 3*, 4-5.

Levine, N. (2004). *CrimeStat III: a Spatial Statistics Program for the Analysis of Crime Incident Locations (Version 3.0)*. Washington, D.C.: Ned Levine & Associates, Houston, T.X. / National Institute Of Justice.

Madre, J.-L., Massot M-H., & Armoogum J. (2000). Monthly frequency versus previous day description of trips: what information is needed on urban mobility? Paper presented at the 9th International Conference on Travel Behaviour Research, Gold Coast, Australia.

Madre, J-L. (2003). Multi-Day and Multi-Period Data. In P. Stopher, & P. Jones (eds.), *Transport Survey Quality and Innovation* (pp. 253-270).Oxford: Pergamon Press.

Meyer, M.D., & Miller E.J. (2001). *Urban Transportation Planning* (2nd ed.). New York: McGraw-Hill.

Mojica, C. (2008). *Examining changes in transit passenger travel behavior through a smart card activity analysis*. M.S. Thesis, Massachusetts Institute of Technology, M.A, United States.

Morency, C. (2004). *Contributions à la modélisation totalement désagrégée des interactions entre mobilité urbaine et dynamiques spatiales*. Thèse de Doctorat, École Polytechnique de Montréal, Q.C., Canada.

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

Morency, C., Trépanier M., & Martin, B. (2008). Object-oriented analysis of carsharing system. *Transportation Research Record: Journal Of The Transportation Research Board, No. 2063*, 105-112.

Morency, C., & Trépanier M. (2010). Assessing transit loyalty with smart card data. Paper presented at the 12th Conference on Transport Research, Lisbon, Portugal.

Munizaga, M., Palma, C., & Mora, P. (2010). Public transport OD matrix estimation from smart card payment system data. Paper presented at the 12th Conference on Transport Research., Lisbon, Portugal.

Multisystems, Inc., Mundle & Associates,, Inc. & Parsons Transportation Group, Inc. (2002). *TCRP Report 80: A Toolkit for Self-Service, Barrier-Free Fare Collection*. Washington D.C.: Transportation Research Board of the National Academies.

Multisystems, Inc., Mundle & Associates, Inc., & Simon & Simon Research and Associates, Inc. (2003). *TCRP Report 94: Fare Policies, Structures and Technologies: Update*. Washington, D.C.: Transportation Research Board of the National Academies.

Navick, D.S., & Furth, P.G. (2002). Estimating passenger miles, origin-destination patterns, and loads with location-stamped farebox data. *Transportation Research Record: Journal Of The Transportation Research Board, No. 1799*, 107-113.

Nielsen, O.A. (2000). Stochastic transit assignment model considering differences in passengers utility functions. *Transportation Research Part B, 34*(5), 377-402.

Okamura, T., Zhang J., & Akimasa F. (2004). Finding behavioral rules of urban public transport passengers by using boarding records of integrated stored fare card system. Proceedings of the 10th World Conference on Transport Research. Istanbul, Turkey.

Ortúzar J., & Willumsen L. (2001). *Modelling Transport* (3rd ed.). Chichester: John Wiley.

Park, J.Y., Kim D.J., & Lim Y. (2008). Use of the smart card data to define public transit in Seoul, South Korea. *Transportation Research Record: Journal Of The Transportation Research Board, No. 2063*, 3-9.

Pelletier, M-P., Trépanier, M., & Morency, C. (2009). *Smart Card Data in Public Transit Planning: A Review* (CIRRELT-2009-46). Montréal: Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport.

Purvis, C.L., & Ruiz, T. (2003). Standards and practice for multi-day and multi-period survey. In P. Stopher, & P. Jones (eds.), *Transport Survey Quality and Innovation* (pp. 271-282). Oxford: Pergamon Press.

Rajbhandari, R., Chien S.I., & Daniel J.R. (2003). Estimation of bus dwell times with automatic passenger counter information. *Transportation Research Record: Journal of the Transportation Research Board, No. 1841*, 120-127.

Richardson, A.J. (2003). Estimating average distance travelled from bus boarding counts. Paper presented at the 82nd Annual Meeting of the Transportation Research Board, Washington, D.C.

Sammer, G. (2009). Identifying and reconciling the data needs of public transit planning, marketing and performance measurement. In P. Bonnel, M. Lee-Gosselin, J. Zmud, & J-L. Madre (eds.), *Transport Survey Methods: Keeping Up with a Changing World* (pp. 321-348). Bingley: Emerald Group Publishing Ltd.

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

Schönfelder, S. (2006). *Urban Rhythms - Modelling the Rhythms of Individual Travel Behavior*. Ph.D. Thesis, Swiss Federal Institute of Technology, Zurich, Switzerland.

Schuessler, N., & Axhausen K.W. (2009). Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board, No. 2105*, 28-36.

Seaborn, C., Attanucci J., & Wilson N. (2009). Analyzing multimodal public transport journeys in London with smart card fare payment data. *Transportation Research Record: Journal of the Transportation Research Board, No. 2121*, 55–62.

Schaller, B. (2002). *TCRP Synthesis 43: Effective Use of Transit Websites: a Synthesis of Transit Practice*. Washington D.C.: Transportation Research Board of the National Academies.

Stanford University Libraries and Academic Information Resources (2010). *What is GIS?* Last accessed on April 18, 2010. From http://www-sul.stanford.edu/depts/gis/whatgis.html

Stopher, P., Clifford, E., & Montes, M. (2008). Variability of travel over multiple days: analysis of three panel waves. *Transportation Research Record: Journal of the Transportation Research Board, No. 2054*, 56–63.

Stopher, P.R., Bullock P., & Horst F.N. (2002). Exploring the use of passive GPS devices to measure travel. In K.C.P. Wang, S. Medanat, S. Nambisan, & G. Spring (eds.), *Proceedings of the 7th International Conference on Applications of Advanced Technologies to Transportation* (pp. 959-967). Reston: ASCE.

Stopher, P.R. (2004). GPS, location, and household Travel. In A. Hensher, K. Button, K. Haynes, & P. Stopher (eds), *Handbook of Transport Geography and Spatial Systems* (pp. 443-449). Oxford: Elsevier.

Strathman, J.G., Kimpel, T.J., Braoch, J., Wachana, P., Coffel, K., Callas, S., Elliot, B., & Elmore-Yalch, R. (2008). *TCRP Report 126: Leveraging ITS Data for Transit Market Research: A Practitioner's Guidebook*. Washington, D.C.: Transportation Research Board of the National Academies.

Susilo, Y.O., & Axhausen K.W. (2007). How firm are you? A study of the stability of individual activity-travel-location patterns using the Herfindahl-Hirschman index. Paper presented at the 11th World Conference of Transport Research, Berkeley, CA.

Trépanier, M. & Chapleau, R. (2001). Analyse orientée-objet et totalement désagrégée des données d'enquêtes ménages origine-destination. *Revue canadienne de génie civil*, *28*(1), 48-58.

Trépanier, M., Barj, S., Dufour, C., & Poipré, R. (2004). Examen des potentialités d'analyse des données d'un système de paiement par carte à puce en transport urbain. Paper presented at the Transportation Association of Canada Annual Conference, Québec City, Canada.

Trépanier, M., Chapleau, R., & Allard, B. (2005). Can trip planner log files analysis help in transit service planning? *Journal of Public Transportation*, *8*(2), 79-103.

Trépanier, M., Chapleau, R., & Tranchant, N. (2007). Individual trip destination estimation in transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, *11*(1), 1-15.

Trépanier, M., & Vassivière, F. (2008). Democratized smartcard data for transit operators. Paper presented at the 15th World Congress on Intelligent Transport Systems, New York, NY.

Trépanier, M., Morency, C. & Blanchette, C. (2009). Enhancing Household Travel Surveys Using Smart Card Data? Paper presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C.

Trépanier, M., Morency, C., & Agard, B. (2009). Calculation of transit performance measures using smartcard data. *Journal of Public Transportation*, *12*(1), 76-96.

Trépanier, M. (2010). L'exploitation des données de cartes à puce à des fins de planification des réseaux de transport collectif urbains. Paper presented at the 12th Conference on Transport Research, Lisbon, Portugal.

Tseytin, G., Hofmann, M., O'Mahony, M., & Lyons, D. (2006). Tracing individual public transport customers from an anonymous transaction database. *Journal of Public Transportation*, *9*(4), 47-60.

Tsui, S.Y.A., & Shalaby, A. (2006). An enhanced system for link and mode identification for GPS-Based personal travel surveys. *Transportation Research Record: Journal of the Transportation Research Board, No. 1972*, 38–45.

Uniman, D., Attanucci, J., Mishalani, R., & Wilson, N. (2010). Service reliability measurement using automated fare card data: application to the London underground. Paper presented at the 89th Transportation Research Board Annual Meeting, Washington D.C.

Utsunomiya, M., Attanucci, J., & Wilson, N. (2006). Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation Research Record: Journal of the Transportation Research Board, No. 1971*, 119-126.

Vuchic, V.R. (2005). *Urban Transit Operations, Planning and Economics*. Hoboken: John Wiley & Sons.

Wilson, N. H. M., Zhao, J. & Rahbee A. (2008) The potential impact of automated data collection systems on urban public transport planning. In N. H. M. Wilson, & A. Nuzzolo (eds.), *Schedule-Based Modeling of Transportation Networks* (pp.75-99). New York: Springer.

Witten, I.H., & Frank E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations*. San Francisco: Morgan Kaufmann Publishers.

Wofinden, D. (2003). Non-Household Surveys. In P. Stopher, & P. Jones (eds.), *Transport Survey Quality and Innovation* (pp. 377-402).Oxford: Pergamon Press.

Wolf, J., Guensler, R., & Bachman, W. (2001). Elimination of the travel diary: experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record: Journal of the Transportation Research Board, No. 1768*, 125-134.

Wolf, J., Schönfelder, S., Samaga U., Oliveira, M., & Axhausen, K.W. (2004). Eighty weeks of global positioning system traces: approaches to enriching trip information. *Transportation Research Record: Journal of the Transportation Research Board, No. 1870*, 46-54.

Wolf, J. (2006). Applications of new technologies in travel survey. In P. Stopher, & C. Stecher (eds.), *Travel Survey Methods: Quality and Futures Directions* (pp. 531-544). Oxford: Elsevier.

Zhang L., Zhao S., Zhu Y., & Zhu Z. (2007). Study on the method of constructing bus stops OD matrix based on IC card data. Paper presented at the 3rd International Conference on Wireless Communications, Networking and Mobile Computing, Shanghai, China.

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

Zhao, J. (2004). *The Planning and Analysis Implications of Automated Data Collection Systems: Rail Transit OD Matrix Inference and Path Choice Modeling Examples*. Master's Thesis, Massachusetts Institute of Technology, Cambridge, M.A., United States.