Measuring transit use variability with smart-card data

Article *in* Transport Policy · May 2007

DOI: 10.1016/j.tranpol.2007.01.001

3 authors:

Catherine Morency
Polytechnique Montréal
138 PUBLICATIONS **1,297** CITATIONS
SEE PROFILE

Martin Trépanier
Polytechnique Montréal
124 PUBLICATIONS **929** CITATIONS
SEE PROFILE

Bruno Agard
Polytechnique Montréal
103 PUBLICATIONS **790** CITATIONS
SEE PROFILE

ELSEVIER

Transport Policy 14 (2007) 193–203

TRANSPORT POLICY

www.elsevier.com/locate/tranpol

Measuring transit use variability with smart-card data

Catherine Morency[b,c,*], Martin Trépanier[a,b,c], Bruno Agard[a,c]

[a]Groupe Polygistique, École Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal, Que., Canada H3C3A7
[b]Groupe MADITUC, École Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal, Que., Canada H3C3A7
[c]Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport (CIRRELT), École Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal, Que., Canada H3C3A7

Available online 12 March 2007

Abstract

The potential of smart-card data for measuring the variability of urban public transit network use is the focus of this paper. Data collected during 277 consecutive days of travel on a Canadian transit network are processed for this purpose. The organization of data using an object-oriented approach is discussed. Then, measures of spatial and temporal variability of transit use for various types of card are defined and estimated using the data sets presented. Data mining techniques are also used to identify transit use cycles and homogenous days and weeks of travel among card segments and at various times of the year.

Keywords: Smart cards; Transit system; Travel behaviour; Data mining; Variability

1. Introduction

In most urban networks, the demand for public transit constantly changes, depending on the time of travel (day of the week, season or holiday) and other factors like weather and service breakdown. Often, transit operators find it extremely difficult to adjust the service to the demand, and, clearly, better adjustment could reduce operating costs and help optimize vehicle use over the network. One of the main issues is the ability to measure the demand precisely and understand its dynamics in order to establish day-to-day predictions. Today, tools are available to planners to perform this task.

The purpose of this study is to illustrate the potential of smart-card data to measure the spatial and temporal variability of transit use. In order to do this, the object-oriented approach, data mining techniques and database management tools are used to construct systematic indicators that help evaluate the variability of travel behaviours by various population segments on a transit network.

The paper is organized as follows. First, a review of the literature in the relevant research fields is provided. Smart-card data systems and the processing of outputted data sets are discussed, as well as the potential of data mining techniques for various applications. The evaluation and measurement of the variability of travel behaviour are also discussed. The data set available for the analysis is then described, and technical details regarding its collection and processing are provided. The measurement concepts and methods are then presented. These relate to the spatial and temporal indices defined to measure the variability of travel behaviour on a transit network. The results of the analysis are subsequently presented, followed by a discussion and some future research avenues drawing on insights gained from the current research.

2. Literature review

2.1. Smart-card data

The use of smart-card automated fare collection systems in public transit is spreading throughout the world. Even

*Corresponding author. École Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal, Que., Canada H3C3A7.
Tel.: +1 514 340 4711x4502; fax: +1 514 340 5763.
E-mail addresses: cmorency@polymtl.ca (C. Morency), mtrepanier@polymtl.ca (M. Trépanier), bruno.agard@polymtl.ca (B. Agard).

0967-070X/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tranpol.2007.01.001

though the technology is relatively old (dating from the mid-seventies), the software and hardware tools required for its implementation are now more accessible (Meadowcroft, 2005). Until recently, the research related to smart-card applications in public transit was mostly technology oriented, since many technical problems had arisen in the first implementations and early daily use of such systems. Smart-card system security was examined by Attoh-Okine and Shen (1995), who reported the dual issue of protecting the card itself and at the same time preserving the privacy of the data collected (see also Clarke, 2001). Smart-card devices needed for public transport were described in detail by Blythe (1998): cards, on-board readers and a centralized information system. The challenge of integrating fare collection into other transactional activities, such as banking or shopping, is also discussed in the literature (Lambrinoudakis, 2002; Shelfer and Procaccino, 2002).

Recent work has demonstrated how much interest there is in using smart-card data for transit planning. Bagchi and White (2004, 2005) conducted three case studies in British networks to estimate turnover rates, trip rates per card and the impacts of the use of smart cards on the proportion of linked trips. They also spoke about the complementary nature of smart-card data collection in relation to other data collection methods, stating that smart cards should not replace those methods. Very few smart-card systems in the world have the capability of locating boarding points on the network (exact stalks). When this capability is available, it provides transit planners with interesting information on route load profiles, on condition that the destination location is known or can be derived (Trépanier and Chapleau, 2006).

2.2. Data mining techniques

As a result of the growing number of data generated on an everyday basis for many different reasons, new developments designed to (more or less) automatically extract knowledge from that large amount of data have appeared. The term commonly employed to unify them is "data mining", a technique which uses tools from statistics, database management and visualization, as well as new methodologies specifically developed to extract patterns from large data sets (such as machine learning). Many algorithms for data mining may be found in Fayyad et al. (1996).

Westphal and Blaxton (1998) proposed categorizing data mining functions into three groups: classification, segmentation and description (which includes visualization). This means that categories are assigned to data in comparison to historical data, grouping together sets of data that share some degree of similarity (different metrics are available) and extract patterns from the data, as well as providing the available information in a format that is understandable to the user (association rules, trees and graphical representations are common).

Many applications of data mining are already available in marketing (Berry and Linoff, 1997) as well as in product design and manufacturing (Braha, 2001). There are many more applications yet to be developed in various fields like public transit. A transit smart-card fare collection system collects a huge amount of data. In the case of the system in use in Gatineau (Canada), for example, about 600,000 entries are collected each month. Data mining techniques have been used to analyse this data with valuable results (Agard et al., 2006; Morency et al., 2006).

2.3. Travel behaviour variability

Even though most transportation models are based on a typical weekday, it is widely recognized that travel behaviour is subject to temporal variability. Actually, the study of the day-to-day variability of travel behaviour began more than 30 years ago. However, analyses remain difficult to conduct because of the cost and burden related to the collection of continuous data on individual travel, although advances in travel survey methods and related technologies have facilitated the collection of these types of data with less of a burden on the respondents. This helps promote the gathering of rich data sets to examine the variability of behaviours. Multi-day travel surveys, in which 2–7 days of travel data are collected, have recently become more common (for instance, the CHASE survey: Doherty and Miller, 2000), and significant travel diaries such as Mobidrive (Axhausen et al., 2002), a 6-week diary, have shown the potential to explain the underlying rhythms of daily life. These surveys collect significant data, but sample sizes are generally small due to the classical quality (depth of questions)/quantity (sample size) trade-off faced by all data collection processes. GPS-assisted travel surveys have also become popular, since they reduce the responsibility of the respondent by automatically collecting spatio-temporal variables (Murakami and Wagner, 1999; Chung and Shalaby, 2005; Draijer et al., 2000; Wolf et al., 2001).

Finally, measurement issues related to the variability of travel behaviours have been the focus of much research. Some authors have used cluster analysis to classify travellers with similar daily activity patterns in groups (Pas, 1983, 1988; Pas and Koppelman, 1986; Jun and Goulias, 1997). Using the Mobidrive survey, Schlich and Axhausen (2003) have measured the similarity between days of travel. Kitamura et al. (2006) examine the spatio-temporal variability (time–space prism) of day-to-day behaviours with the same data set. Gärling and Axhausen (2003) have also discussed the habitual nature of travel.

Smart-card systems continuously monitor the use of the transit system by all the card holders. The relevant processing of these continuous data sets can contribute to our understanding of travel behaviour rhythms by measuring the similarity between days and weeks of travel and by estimating the variability of what could be called a "typical" day or week of travel.

3. Data set

3.1. Source

The data set is provided by the Société de Transport de l'Outatouais (STO), a transit authority serving the 240,000 inhabitants of Gatineau, Quebec. The STO authority is a Canadian leader in transit smart-card fare collection. This system has been in use since 2001, and a large proportion of STO users have a smart card. The smart card carries the photo of the user, which ensures that the card is used by one person only. The data set provided is related only to cards and is completely anonymous. This ensures the full privacy of the data, since the user's details cannot be determined.

The STO smart-card collection system is composed of four subsystems: smart-card readers aboard the buses, along with recharging and maintenance equipment; an integrated information system, which separately stores user information and boarding information (logs) because of privacy concerns; the service operation information system, which provides operational data to the integrated information system (bus routes, vehicle assignments, employee list, etc.) and the accounting system, which takes care of the financial transactions (the main purpose of the smart-card system being fare collection).

3.2. Object model

An object model has been developed so that a better understanding can be gained of all the elements related to the smart-card system within the transit network. The method used for this task is transportation object-oriented modelling (TOOM, see Trépanier and Chapleau, 2001). TOOM classifies data into four metaclasses of objects: static (supporting transportation), kinetic (describing movements), dynamic (transportation actors) and systemic (networks, systems). Fig. 1 illustrates the objects of the smart-card system at the STO. The number of instances is also shown for the January to October 2005 data set. To simplify the object model, we describe four major groups of objects as follows:

(A) Network objects, which are the public elements of the STO transit network (drivers, buses, routes, route stops).
(B) Operations objects, which are the internal elements of the STO bus operations (drivers, buses, workpieces, garage).
(C) Administrative objects, which are the smart-card system data elements.
(D) Trips, trip chains and all-month trip chains (trip habits), which are related to the transit user's behaviour.

The object model helps to explain the relationship between the elements of data available. For example, a typical STO user has made trip chains during the month. A trip chain is composed of a sequence of trips. Each trip is related to one or more route sections and associated with

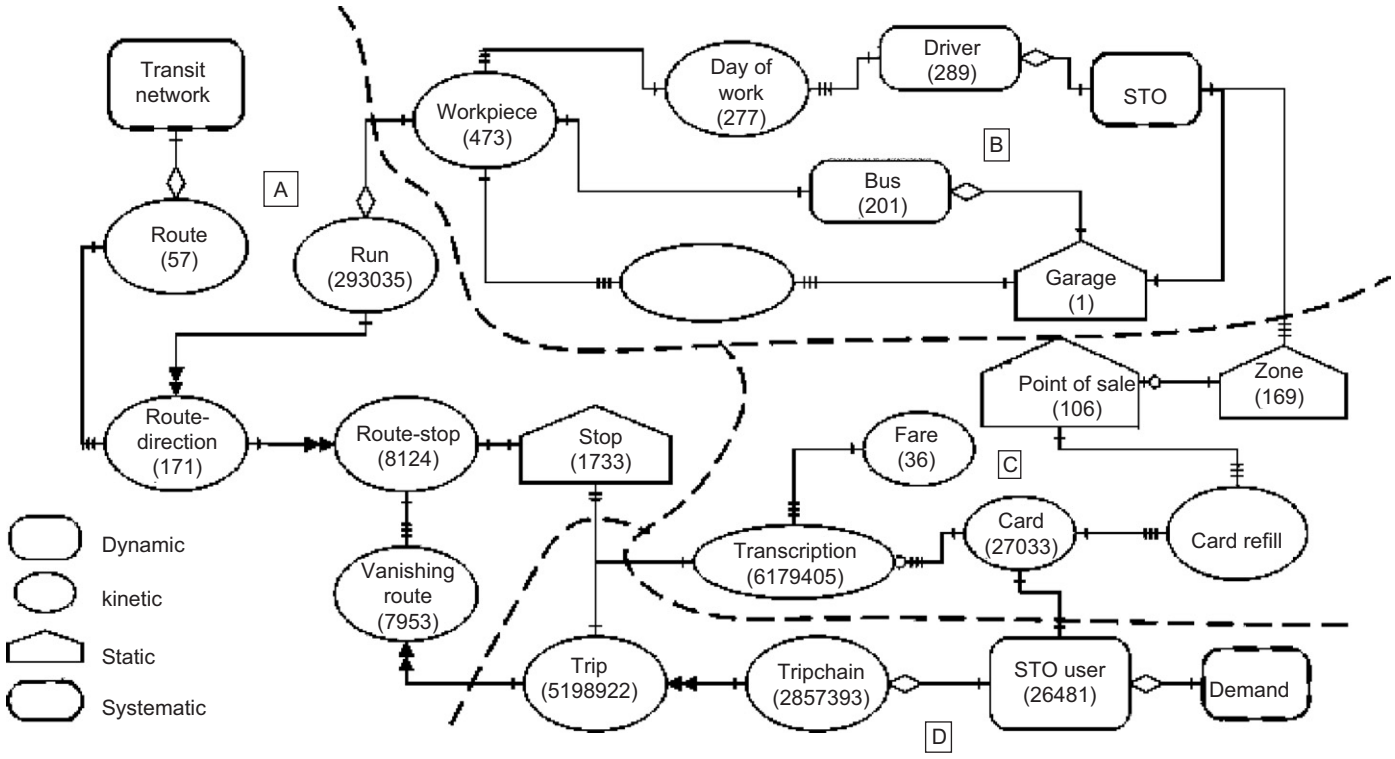

Fig. 1. The STO smart-card system object model.

operational bus runs. Runs are related to workpieces that have been completed by a specific driver on a specific bus. The figure shows us that relationships between a user and a bus, or other relationships, can be retrieved, if needed. The TOOM approach can be easily integrated into existing information systems to help in data processing. A recent example is the linking of GPS data to planned routes in the case of road network monitoring (Marzolf et al., 2006). TOOM has also been used to analyse trip calculator log files collected on transit authority websites (Trépanier et al., 2005).

3.3. Data set structure

The data set was extracted from the integrated information system and corresponds mainly to the "Transaction" object shown in Fig. 1. The table contains a record for each of the 6.2 million boardings made by a card aboard a bus during the day during the period from January 1 to October 10, 2005. Records may include a valid boarding, a transfer or a refusal owing to the lack of the right to board the bus. Table 1 presents the contents of the data set.

When additional information is needed, like stop locations and the sequence of stops on a route, it is extracted from other tables in the database. It can be observed that alighting locations are not reported in the system. This is because there is no validation when users leave the buses. The estimation of alighting locations is currently being addressed in another research project.

3.4. Selected sample

The selected sample for this article consists of information related to more than 2.2 million boardings recorded by 7118 different smart cards between January and October 2005. Only cards which were validated both before January 10 and after October 1 were included in the sample in order to maximize the observation period for each of them. The card type (adult, senior or student) is the only information available for classification, in addition to an indication of the privilege of using specific parts of the network without penalty (express, interzone or regular routes). For analysis purposes, these cards were aggregated in five classes. Table 2 summarizes the selected sample: number of cards and boardings.

Fig. 2 describes the sample over the 10-month period. It shows the total number of boardings per week as well as the proportion of boardings occurring each week per class of card type.

From this distribution, we observe:

- *Senior cards*: Fewer boardings for seniors during the winter (January and February), but a fairly stable number of boardings from April to September.
- *Student cards*: A drop in the number of boardings during the spring school break (the week of March 7), a decline with the end of the school year (in universities, from the beginning of May) and a minimum number of boardings in July and August.
- *Adult cards*: Fewer boardings for all adults during the summer, with a drop during the week of August 8, and

Table 1
Contents of the data set for the period January 1–October 10, 2005

Card ID	The card number is not related to an individual, and is used only for cross-relating records
Rare (card) type	The fare type related to the card (regular, student, express, etc.)
Date time	The date and time the card is read aboard the bus.
Route ID	Route number of the STO network on which the bus is operating
Direction	Direction of the route
Stop ID	Stop number of the boarding, obtained with the help of a GPS device aboard the vehicle
Validation result	Indicates whether the boarding was valid, was a transfer or was refused by the reader
Operational info	Information on run number, vehicle number and bus driver is also available

Table 2
Classification of card types according to type of card and privileges on the network

Class of smart card	ID	Nb. cards	Nb. boardings	Boardings per card
Adult-interzone[a]	A-I	288	81,880	284
Adult-express[b]	A-E	1657	452,090	273
Adult-regular[c]	A-R	4379	1,407,040	321
Elderly	E	443	158,900	359
Student	S	351	133,380	380
Total cards	Tot	7118	2,233,300	314

[a]Unlimited access to regular, express and interzone network.
[b]Unlimited access to regular and express network, supplement ($) on interzone network.
[c]Unlimited access to regular network, supplement ($) on express and interzone network.

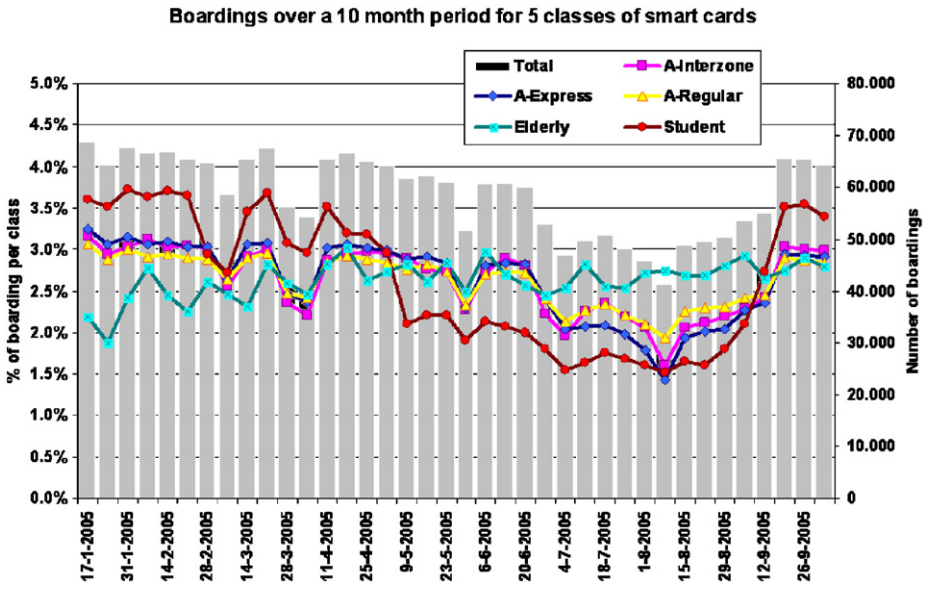

Fig. 2. Total number of boardings per week and the proportion of the boardings for the five classes of smart cards.

other in the last week of March and the first week of April (also observed for the student cards because of the long Easter weekend).

- *Overall*: A decline in the number of boardings starting around the end of April, leading to fewer boardings during the summer months, as well as occasional drops (school break at the end of March).

4. Concepts and method

4.1. Indicators of spatial variability

The spatial variability of transit use is examined through the enumeration of all the bus stops used for boarding. It is worth remembering that the observed transaction is boarding, and that this transaction can represent the start of a trip or be the consequence of a transfer between two bus routes. First, the overall number of bus stops used for boarding is examined, as is the temporal structure of their acquisition (cumulative structure of the first use of these bus stops). Then, the frequency of use of the bus stops is studied, in order to express a level of regularity. It allows the number of bus stops that cover the main proportion of transit paths observed via the smart-card data to be measured. These measures are performed globally as well as by card type. From this perspective, the importance, in terms of cumulative boardings, of the bus stops most often used is examined.

At this point, the study does not take into account the spatial proximity between stops. Hence, further analyses are needed to find equivalent bus stops (nearby stops on the same route or on a set of parallel routes) in order to refine this measurement.

4.2. Indicator of temporal variability and clustering methods

The temporal variability of transit use is evaluated using data mining techniques. For this purpose, a data set containing boardings per hour per day is constructed for every card (see example in Table 3).

Table 3 shows that day 1 of the observation period (January 3, 2005) was a Saturday (7). Card number 2988642241 was validated somewhere on the network between 6:00 p.m. and 6:59 p.m. (H18). Using this information and clustering algorithms, typical temporal patterns of boardings, for cards of similar classes, are identified. Details regarding the clustering methods used are given below.

For the current experiment, a *k*-mean algorithm is used to partition the data set into a predefined number of clusters. The *k*-mean algorithm minimizes the sum, over all clusters, of the distance to the centroid of each cluster. The Hamming distance, a distance which represents the percentage of data between two elements that differ, is used to measure the closeness of those two elements. Each centroid is the component-wise median of points in that cluster.

The algorithm used with Matlab 7.0 (Seber, 1984; Spath, 1985) has two phases:

The *first phase* uses what the literature often describes as "batch" updates: each step consists of reassigning points to their nearest cluster centroid, all at once, followed by recalculation of the cluster centroids.

The *second phase* uses what the literature often describes as "online" updates, where points are individually reassigned if doing so reduces the sum of distances, and the cluster centroids are recomputed after each reassignment. During this second phase, each step consists of one pass through all the points.

Table 3
Structure of a typical record of the data set constructed to analyse the temporal variability of boardings

ID card	DAY	DTYPE	H00	H01	H02	H03	H04	H05	…	H16	H17	H18	H19	H20	H21	H22	H23
2988642241	1	7	0	0	0	0	0	0		0	0	1	0	0	0	0	0
2795002016	1	7	0	0	0	0	0	0		0	0	0	0	0	0	0	0
3059308960	1	7	0	0	0	0	0	0		0	0	0	0	0	0	0	0
3053531265	1	7	0	0	0	0	0	0		0	0	1	0	0	0	0	0
2250525560	1	7	0	0	0	0	0	0		0	0	0	0	0	0	0	0
2532072608	1	7	0	0	0	0	0	0		0	0	0	0	0	0	0	0
2531995521	1	7	0	0	0	0	0	0		0	0	0	0	0	0	0	0
2525842296	1	7	0	0	0	0	0	0		0	0	0	1	0	0	0	0
3865739417	1	7	0	0	0	0	0	0		0	1	0	0	0	0	0	0

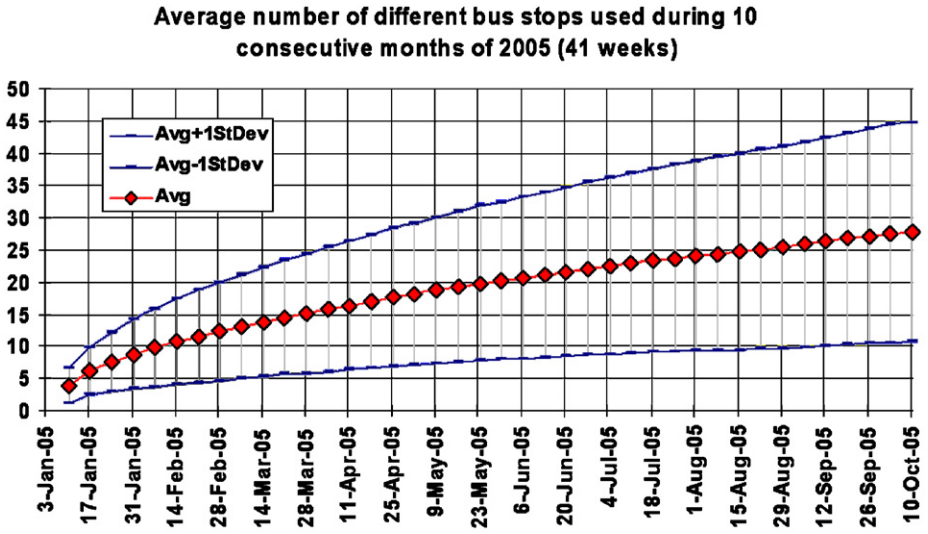

Fig. 3. Average number of different bus stops used during 10 consecutive months of observation (first 41 weeks of 2005).

k-means can converge to a local optimum, in this case a partition of points in which moving any single point to a different cluster increases the total sum of distances. This problem can only be solved by a clever (or lucky or exhaustive) choice of starting points. The number of clusters is an input of the method that depends on the level of granularity expected for the analysis. Four clusters appear to constitute a good experimental value for the data set considered here.

5. Results

5.1. Indicators of spatial variability

5.1.1. Enumeration of all the bus stops used for boarding

Fig. 3 presents the average cumulative number of different bus stops used by the card holders during 41 consecutive weeks of 2005, as well as the confidence interval ($\pm$one standard deviation). The variability is surprisingly quite stable over the entire period (variation coefficient $\approx 61.4\%$). On average, approximately 0.7 of a new stop is acquired per week, for an average total of 27.7 bus stops used for boarding during the whole period of observation.

When these statistics are compiled for the main classes of smart card, the figures are quite different. Table 4 presents the total number of different stops used during the observation period for every class of card, as well as the number used during the first week. An average acquisition rate (number of new stops acquired per week) is, moreover, estimated.

The temporal structure for the accumulation of new bus stops for these classes is also presented in Fig. 4. Different spatial patterns are outlined using this classification, which seems relevant since the variation within the classes is less than that for the entire data set. Actually, the student card type represents the more diversified use of bus stops, followed by the senior card type. The two classes board at more different bus stops during a typical week and keep on adding new stops to their background at higher rates than the adult card type (0.92 and 0.83 new stops per week for students and seniors, respectively, compared to 0.33 new stops per week for interzone adults). The interzone and express classes of adult card type, both of which are commuter-type cards, are more likely to use transit

Table 4
Number of different stops used for boardings during the observation period and during the first week, as well as the average acquisition rate (number of new stops per week) for the main classes of card

Class of smart card	Nb. diff. stops over 10 months	Variation coefficient (%)	Nb. diff. stops first week	Avg. acq. rate (per week)
Adult-interzone	15.7	41.6	2.75	0.33
Adult-express	17.5	44.2	2.95	0.37
Adult-regular	30.4	57.3	4.26	0.67
Elderly	37.3	46.1	4.88	0.83
Student	40.1	51.4	4.27	0.92
Total cards	27.7	61.4	3.93	0.61

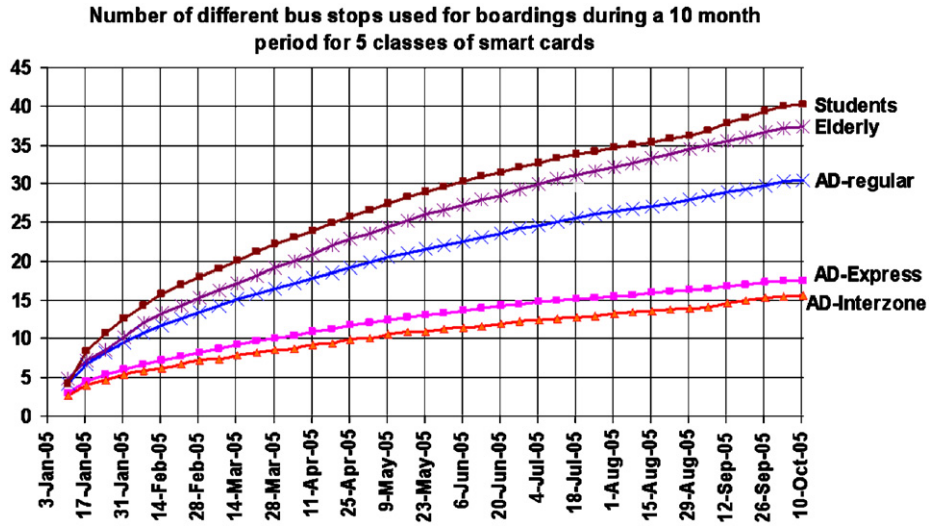

Fig. 4. Average number of different bus stops used during 10 consecutive months of observation (first 41 weeks of 2005).

exclusively for their journey to work, and consequently have a smaller range of bus stops, as well as lower acquisition rates.

5.1.2. Frequency of use of all the bus stops used for boarding

In addition to measuring the number of different bus stops, it is relevant to measure the frequency of use of these bus stops as well, in order to better understand the spatial regularity of the boarding patterns in time and uncover cycles of irregular travel patterns. In order to do this, the proportion of boardings occurring at the most frequently used bus stops is examined. On average, the most frequently used bus stop accounts for almost 37% of all the boardings observed during the 10-month period, while the next two most frequently used ones account for 64% of the boardings. Actually, among all the different bus stops used for boarding by means of a smart card, almost 43% were used only once during the observation period; this directly relates to irregular and punctual activities. Moreover, almost 70% of these bus stops were used four times or less during the observation period.

Fig. 5 presents some of these results segmented per card type. It shows the cumulative proportion of boardings owing to the number of different bus stops (in descending order of frequency). It shows, for instance, that the two most frequently used bus stops account for 77.9%, 75.2%, 61.4%, 48.6% and 47.8% of all the boardings for adult-interzone (A-I), adult-express (A-E), adult-regular (A-R), senior (E) and student (S) cards, respectively. Hence, commuter-type cards show more regular behaviours, since a smaller number of stops accounts for a higher proportion of the boardings. This is quite compatible with the regular behaviours between home and the workplace. Moreover, student and senior card types have a wider spectrum of boarding stops, probably revealing a more dispersed and diverse use of the transit system.

5.2. Indicators of temporal variability

5.2.1. Primary results: comparison of the clusters per card type

Data mining techniques were similarly applied to the temporal boarding behaviours of cards according to the classes previously defined. The results of the clustering process are, on the one hand, cluster centroids expressing the main behaviours helping to discriminate between classes, and, on the other hand, the membership of every day of observation in one of these clusters.

Fig. 6 shows the temporal centroid of the four clusters of each card type. The temporal centroids are the common

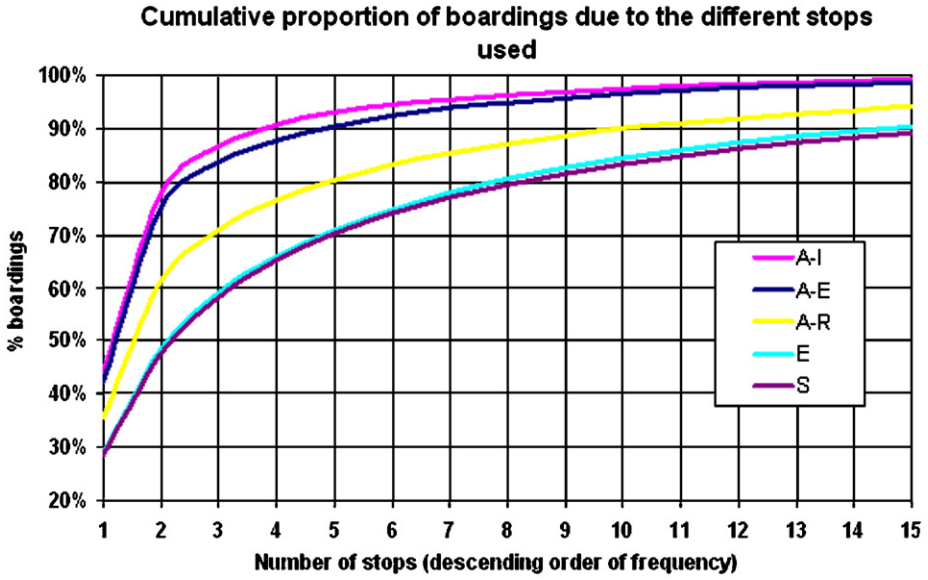

Fig. 5. Cumulative proportion of boardings owing to the different bus stops used (in descending order of frequency of use) according to card type.

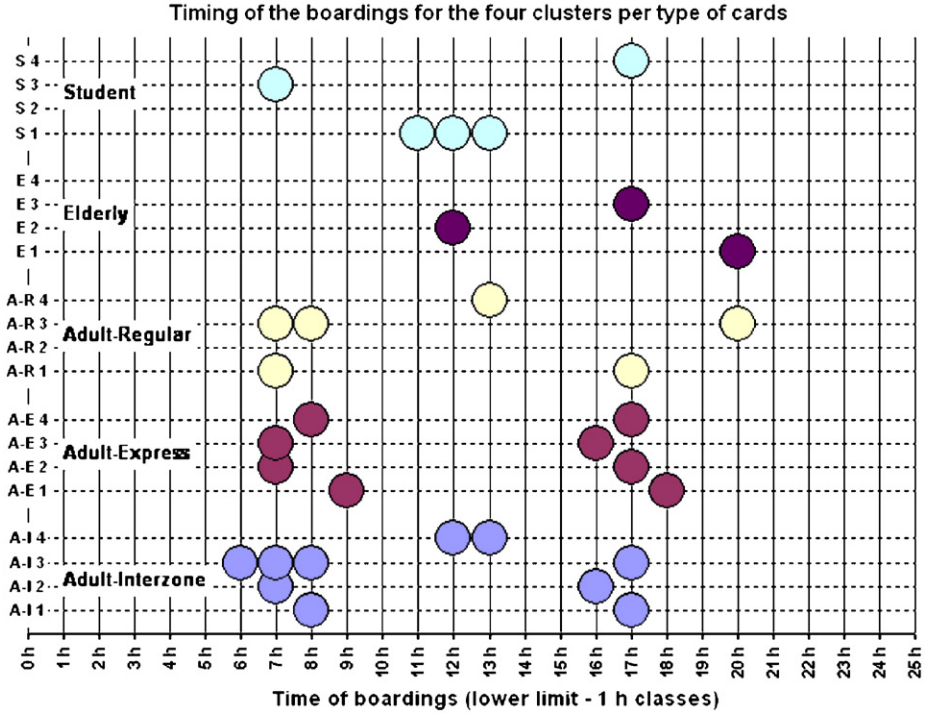

Fig. 6. Temporal centroids of the clusters for the five classes of cards.

boarding times for clusters. This means that most boardings occur at the cluster time centroid for each card within this cluster.

The size of each cluster is presented in Table 5 in terms of card-days. The card-day is used because a card may change cluster with the day of observation, depending on its related travel behaviour. We can see that some clusters are very important within a card type. Cluster 1 of adult-regular (A-R1), with departure times at 7:00 a.m. and 5 :00 p.m., accounts for about 37% of the boardings of this card type. Student and senior card types are distributed more equally within clusters.

5.2.2. *Proportion of zero-boarding days*

The temporal rhythms of activity on the transit network can be appreciated by a study of the zero-boarding days for each card type and day. These are the days when there was no trip for a card. Table 6 summarizes these statistics. It shows a hierarchy of behaviours between card types. At one end of the scale, the A-I cards are more closely

Table 5
Size of each cluster in card-days, for the five classes of cards

Student		Elderly		Adult-regular		Adult-express		Adult-interzone	
Cluster	Size	Cluster	Size	Cluster	Size	Cluster	Size	Cluster	Size
S1	2117	E1	2071	A-R1	247,628	A-E1	30,605	A-I1	14,337
S2	34,697	E2	14,549	A-R2	377,458	A-E2	47,716	A-I2	26,460
S3	2700	E3	7540	A-R3	10,937	A-E3	75,238	A-I3	186
S4	10,484	E4	40,085	A-R4	41,843	A-E4	81,718	A-I4	281

Table 6
Average proportion of zero-boarding days per card and day type

Average proportion of zero-boarding days per card and day type	Day of observation (%)						
	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Adult-interzone	34.0	22.5	21.6	22.5	36.9	90.4	92.7
Adult-express	34.6	24.4	23.7	24.5	36.4	85.4	87.9
Adult-regular	33.1	25.1	24.5	24.6	33.0	65.2	69.8
Elderly	44.8	40.0	39.8	38.9	39.3	55.1	65.3
Student	41.5	37.1	35.5	36.1	41.1	65.3	71.6

associated with weekday travel, obviously between home and the workplace. They have the highest proportion of zero-boarding days on weekends (more than 90% of all Saturdays and Sundays were zero-boarding days for this card type). At the other end of the scale, senior cards are associated with more fluctuations in behaviour throughout the week, with the highest proportion of zero-boarding days on weekdays and the lowest proportion on weekends.

5.2.3. Overall cluster's membership

The regularity of temporal behaviour can also be evaluated by the days of travel belonging to similar clusters of behaviours. The idea is to measure how often a particular card is associated with one particular cluster. The study of the most popular cluster for each card gives this type of information. Fig. 7 shows the regularity of behaviours for all classes of cards: the percentage of boardings that belong to the most popular cluster (for the class), as are functions of the percentage of users (within the class). It reveals that about 50% of the senior card type has an overall regularity of 64%, while 50% of the A-I card type has an overall regularity of 93%. Also, all cards, whatever their type, have an overall regularity higher than 30% (49% for the less regular A-I card observed).

5.2.4. Weekday cluster's membership

The same analysis has been conducted for weekdays. Table 7 presents the percentage of observed days belonging to the most frequently occurring cluster for each card type, per day of the week. It shows that membership is quite stable for weekdays.

6. Conclusion

The paper shows that smart-card data have the potential to give a continuous profile of transit use by various types of cards. To arrive at this conclusion, the data were well formatted with the help of an object-oriented approach, which identifies all the objects of the system. From the experiments, we were able to conclude that it is possible to observe regularity indicators by using raw information on boardings only, even though little individual information was available. One of the next tasks in this project will be to implement the method on a larger scale, with day-to-day usage. More work will be required on information system design in order to build a friendlier tool that will be suitable for use by transit operators.

There are limits to the model. The use of clustering methods on pooled behaviours by smart-card classes is relevant for homogenous classes such as A-E and A-I, but not for classes with few constrained movements. A further analysis will be needed of the behaviour of these ''non-regular'' users, possibly by using more clusters or by formatting the data in a different manner. The lack of individual user data makes it difficult to drill down the result within card types.

The potential of using transit smart cards is endless. Smart-card data could, for example, help in examining the impacts of weather on transit demand. They could also be used to analyse transit network performance by comparing observed data (from boardings) with planned data (from a schedule). There is also the possibility of analysing spatial travel behaviour with a more precise level of resolution, like bus stops or small zones. The increasing number of

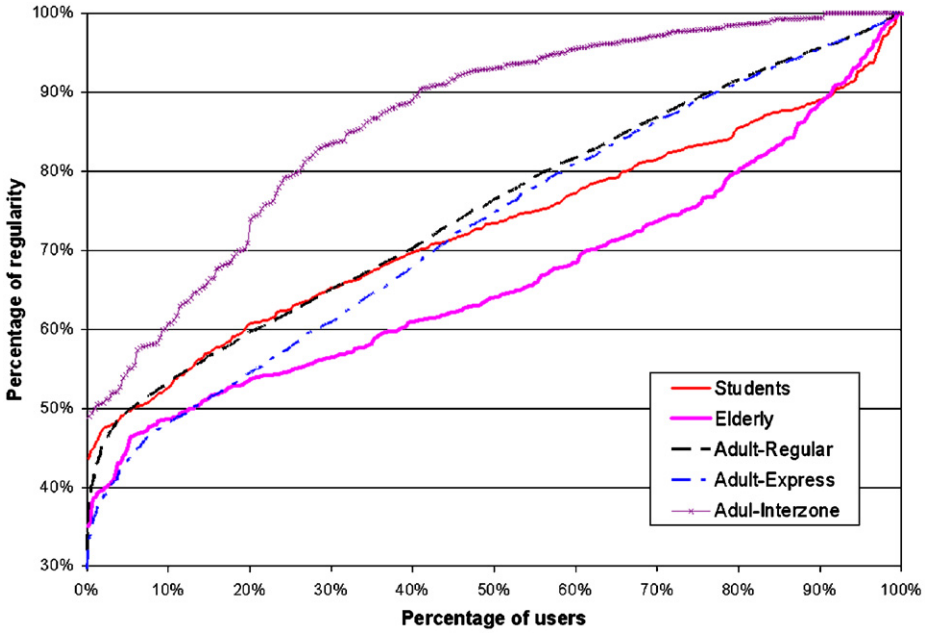

Fig. 7. Overall clusters' membership.

Table 7
Weekday cluster's membership (percentage of the most frequently occurring cluster)

Day	Student (% m.f.c.)	Elderly (% m.f.c.)	Adult-regular (% m.f.c.)	Adult-express (% m.f.c.)	Adult-interzone (% m.f.c.)
Sunday	80	61	70	59	73
Monday	67	62	54	35	65
Tuesday	68	63	54	36	65
Wednesday	68	63	54	35	65
Thursday	67	62	54	35	64
Friday	69	62	56	35	61
Saturday	79	62	69	67	69

transit networks equipped with a smart-card fare collection system will probably help to open up this new research field in the coming years.

Acknowledgements

The authors wish to acknowledge the STO (Gatineau's transit authority) which graciously provided data for this study. The research project is also supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Fonds Québécois de Recherche sur la Nature et les Technologies (FQRNT) and the Agence Métropolitaine de Transport of Montreal (AMT).

References

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

Attoh-Okine, N.O., Shen, L.D., 1995. Security issues of emerging smart card fare collection application in mass transit. In: IEEE Vehicle Navigation and Information Systems Conference. Proceedings. In conjunction with the Pacific Rim TransTech Conference. Sixth International VNIS. 'A Ride into the Future', pp. 523–526.

Axhausen, K.W., Zimmermann, A., Schönfelder, S., Rindsfüser, G., Haupt, T., 2002. Observing the rhythms of daily life: a six-week travel diary. Transportation 29 (2), 95–124.

Bagchi, M., White, P.R., 2004. What role for smart-card data from a bus system? Municipal Engineer 157, 39–46.

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transport Policy 12, 464–474.

Berry, M., Linoff, G., 1997. Data Mining Techniques: for Marketing, Sales, and Customer Support. Wiley, New York.

Blythe, P., 1998. Integrated ticketing smart cards in transport. In: IEE Colloquium: Using ITS in Public Transport and in Emergency Services, pp. 1–21.

Braha, D., 2001. Data Mining for Design and Manufacturing. Kluwer Academic Publishers, Boston, MA.

Chung, E., Shalaby, A., 2005. A trip reconstruction tool for GPS-based personal travel surveys. Journal of Transportation Planning and Technology 28 (5), 381–401.

Clarke, R., 2001. Person location and person tracking: technologies, risks and policy implications. Information Technology and People 14 (2), 206–231.

Doherty, S.T., Miller, E.J., 2000. A computerized household activity scheduling survey. Transportation 27 (1), 75–97.

Draijer, G., Kalfs, N., Perdok, J., 2000. Global positioning system as a data collection method for travel research. Transportation Research Record 1719, 147–153.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., 1996. Advances in Knowledge Discovery and Data Mining. AAAI Press, The MIT Press, Cambridge, MA.

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

Jun, M., Goulias, K., 1997. A dynamic analysis of person and household activity and travel patterns using data from the first two waves in the Puget sound transportation panel. Transportation (24), 309–331.

Kitamura, R., Yamamoto, T., Susilo, Y.O., Axhausen, K.W., 2006. How routine is a routine? An analysis of the day-to-day variability in prism vertex location. Transportation Research Part A (40), 259–279.

Lambrinoudakis, C., 2002. Smart card technology for deploying a secure information management framework. Information Management and Computer Security 8 (4), 173–183.

Marzolf, F., Trépanier, M., Langevin, A., 2006. Road network monitoring: algorithms and a case study. Computers and Operational Research Journal 33 (12), 3494–3507.

Meadowcroft, P., 2005. Hong Kong raises the bar in smart card innovation. Card Technology Today 17 (1), 12–13.

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

Murakami, E., Wagner, D.P., 1999. Can using a global positioning system (GPS) improve trip reporting? Transportation Research—C 7 (C), 149–165.

Pas, E.I., 1983. A flexible and integrated methodology for analytical classification of daily travel-activity behaviour. Transportation Science 17 (4), 405–429.

Pas, E.I., 1988. Weekly travel-activity behaviour. Transportation (15), 89–109.

Pas, E.I., Koppelman, F.S., 1986. An examination of the determinants of day-to-day variability in individuals' urban travel behaviour. Transportation (13), 183–200.

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

Seber, G.A.F., 1984. Multivariate Observations. Wiley, New York.

Shelfer, K.M., Procaccino, J.D., 2002. Smart card evolution. Communications of the ACM 45 (7), 83–88.

Spath, H., 1985. Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples (Translated by J. Goldschmidt). Halsted Press, New York, 226pp.

Trépanier, M., Chapleau, R., 2001. Analyse orientée-objet et totalement désagrégée des données d'enquêtes ménages origine-destination. Revue Canadienne de Génie Civil, Ottawa 28 (1), 48–58.

Trépanier, M., Chapleau, R., 2006. Destination estimation from public transport smartcard data. In: The 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM), Saint-Étienne, France.

Trépanier, M., Chapleau, R., Allard, B., 2005. Can trip planner log file analysis help in transit service planning? Journal of Public Transportation, Miami 8 (2), 79–103.

Westphal, C., Blaxton, T., 1998. Data Mining Solutions. Wiley, New York.

Wolf, J., Guensler, R., Bachman, W., 2001. Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. Presented at the 80th Annual Meeting of the Transportation Research Board, Washington, DC.