Proceedings of the IEEE ITSC 2006
2006 IEEE Intelligent Transportation Systems Conference
Toronto, Canada, September 17-20, 2006

Analysing the variability of transit users behaviour with smart card data

Catherine Morency, Martin Trépanier, Bruno Agard

Abstract - **This paper proposes various measures regarding the variability of travel behaviours of transit users. The analyses are performed with smart card data collected over a ten months period. The variability in terms of boarding per day and new stops frequented with the days of travel on the transit network is examined. Data mining techniques are then used to classify days of travel according to the similarity of the boarding time periods. In this view, the use of two specific smart cards is examined in more details. These experiments first show that the behaviours of regular transit users evolve with time both in terms of transit stops frequented and time of boarding. Hence, variability of behaviours also changes for various user types.**

I. INTRODUCTION

INCREASING popularity of Smart Card Automated Fare Collection (SCAFC) Systems in urban transit take advantage of new ITS technologies. It makes it possible in the same time to improve the customer's satisfaction (unique card, various tickets, more flexibility in the possible uses, secured uses, and so on), and to facilitate revenue collection for public transport authorities.

Such systems generate every day large amount of data about the utilization of the public transport system. Each transaction, in addition to revenue collection, may provide information about how many people use the transport system, where, when, on what route, and even more.

Most often in a transit network the data describe for each boarding, the exact time, some precision on location and some extra information about the card itself (type of card, type of ticket (fare), period of validity, geographic area of validity, and so on).

Frequently these data are used for revenue collection and stored in databases. Besides, we believe that these data contain much more interesting knowledge about the overall system and that this knowledge may help planners to better understand transit user behaviour, so as to help to improve the service.

Because of the number of passengers and the activity of each passenger, the dataset is perpetually growing. It becomes impossible for anyone without dedicated tools to have a complete view and to understand what is going on into the data. Also we believe that data mining tools may be powerful to extract such knowledge available from smart card data acquisition systems.

Then this paper proposes to take advantage both of data mining methods and public transport planning models in order to describe the regularity in user's behaviour on a transit network. Such knowledge, for important sets of users, may provide important information about every day utilization and periodic evolutions. Also seasonality information may be discovered. The focus of the study is on trips habits (in time).

The structure of the paper is the following. First, in section II, a literature review focuses on smart card data analysis in public transport and on principle data mining techniques appropriate in such type of data. Besides, not any previous work combining the two research fields has been found. Also an interest is on travel behaviour variability. Then section III describes the methodology. The case study, the data structure and the mining tools are outlined. Section IV exposes the results of an experimental analysis. It shows how the datasets are prepared and the information that is extracted for two specific users. Finally, the last section concludes the paper and proposes directions for further investigations.

II. REVIEW

This review focuses on the background elements of this project: transit smart card fare collection, data mining techniques and travel behaviour variability.

A. Smart card in public transport

Smart card technology is an "old" technology patented in 1968 by German researchers Dethloff et Grötrupp [1]. The card is nothing else than an RFID (radiofrequency identification) device embedded in a typical credit card size. In some places, the device is installed within a keychain for more convenience to the transit user. Smart cards are used mostly for fare collection in transit systems, because user boarding validation can be done instantaneously without any

Manuscript received March 8, 2006. This work was possible thanks to the support and collaboration of the *Société de transport de l'Outaouais*, who gracefully provided data for this study. The research project is also supported by the Natural Science and Engineering Research Council of Canada (NSERC), and by the Agence métropolitaine de transport of Montreal (AMT).

Prof. C. Morency is in Groupe MADITUC and in Centre de recherche sur les transports (CRT), École Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal (Québec) Canada, H3C 3A7, (e-mail cmorency@polymtl.ca).

Prof. M. Trépanier is in Polygistique, in Groupe MADITUC and in Centre de recherche sur les transports (CRT), École Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal (Québec) Canada, H3C 3A7, (e-mail mtrepanier@polymtl.ca).

Prof. B. Agard is in Polygistique, École Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal (Québec) Canada, H3C 3A7, (e-mail bagard@polymtl.ca).

interaction with the driver or the fare collection agent. It has the ability to manage complex fare systems, especially in integrated large metropolitan areas [2]. Privacy concerns arise when managing smart card data, but elementary security procedures can be taken, as smart card data is not different from other individual data collection systems (credit card, road toll, police corps database) [3].

In public transit, [4], [5] have shown the potentialities of using smart card data for transportation planning purposes. Smart cards permit to construct larger sets of data, with a continuous observation period, reaching a larger part of transit users. However, the number of attributes available to the user is much smaller in smart card systems than it could be in household surveys because of the privacy control. [4] insist on the need of implementing complementary surveys to validate SCAFC data. They also propose that the organizations prepare a 2-year settling-in period to implement such systems.

In a recent work, [6] have shown the interest of using SCAFC data to analyse bus route transfers, in order to measure the level of service (LOS) of the public transport service. Smart card systems usually do not collect data on vehicle alighting, but when GPS boarding location data is coupled to smart card collection data; it is possible to estimate the alighting location with an iterative model [6].

B. Data mining tools and applications

Data mining is a collection of techniques and tools dedicated to the discovery of non-trivial, implicit, previously unknown, and potentially useful and understandable patterns from large data sets [7].

Different classifications of data mining tools are available. If we exclude text, sound and video mining, classical tools may be categorized as classification, estimation, segmentation, and description [8]:

- Classification is dedicated to assign labels to data based on arrangements constructed on historical data.
- Estimation evaluates missing values of a record as a function of the fields in other records.
- Segmentation (or clustering) creates subsets in a population. Elements are placed in a subset to maximize the homogeneity of the subset and to maximize the heterogeneity between the different subsets.
- Description and visualization are used to show the relationships among the data. Frequent patterns in the data are represented as association rules (with some measures of regularity). Graphical projections and representation may emphasis particular characteristics.

Data mining techniques offer applications in many areas such as design and manufacturing [9], marketing [10], manufacturing processes [11], manufacturing quality [12], plant layout [13], and business process reengineering [14].

C. Travel behaviour variability

The study of day-to-day variability of travel behaviour started more than thirty years ago. Analyses were and still are hard to conduct because of the available data which generally rely on a single day record of each individual's travel. Hence, many authors have discussed the determinants of travel behaviours in an urban area as well as the importance to understand the variations in daily peak profiles to better assess demand management schemes [15]. Some authors have used cluster analysis in order to classify travellers in groups of similar daily activity patterns: [16]-[19]. Schlich and Axhausen [20] have recently used a six-week travel diary in order to measure the similarity between days of travel. [21] examine the spatio-temporal variability (time-space prism) of day-to-day behaviours with the same data set. Garling and Axhausen [22] also discuss the habitual nature of travel.

III. METHODOLOGY

A. Case study

Data from the Société de transport de l'Outaouais (STO) smart card fare collection system were used for the study. The STO is a medium-size public transport authority operating 200 buses and servicing 240 000 inhabitants in Gatineau, Quebec. The STO operates its smart card system since 2001. Today, more than 80% of all STO passengers hold a smart card. Moreover, every STO bus is equipped with GPS reader. At each boarding, stop location and bus route are stored in the database along with a timestamp. Since the STO uses a high-level secured procedure to ensure the privacy of the data, smart card data are completely anonymous. No nominal information on user is known in any kind.

B. Smart card data collection process

The smart card data collection process at the STO is similar to other systems through the world (see Fig. 1). Every boarding is individually validated within the vehicle with the help of an on-board card reader. Validation data is stored on a local database. The content is downloaded to the central database at the end of the day, when the vehicle returns to the depot. Data is collected in an information system called *Système informatisé de validation des titres* (SIVT), which gathers information on cards, transit network elements, and validation events. For privacy reasons, personal user data is not maintained in the same system and is not available to this study.

Our focus for this study is on collected data (lower part of the figure). Each record contains information on single validation event: card number, boarding status and type (regular, transfer, refusal), date and time, transit route number and direction, stop number. Fare categories and stop XY coordinates are added with the help of other tables in the database.

C. Mining tools

In this project, the data mining techniques concerned are filtering of data, clustering (with a k-mean), and visualization. The purpose is to construct clusters of days presenting similar temporal patterns of boarding on the transit network. This will help understand if travellers have regular travel behaviours and if days differ significantly.

The dataset was extracted from a SQL Server database and was analysed within the TANAGRA freeware [23].

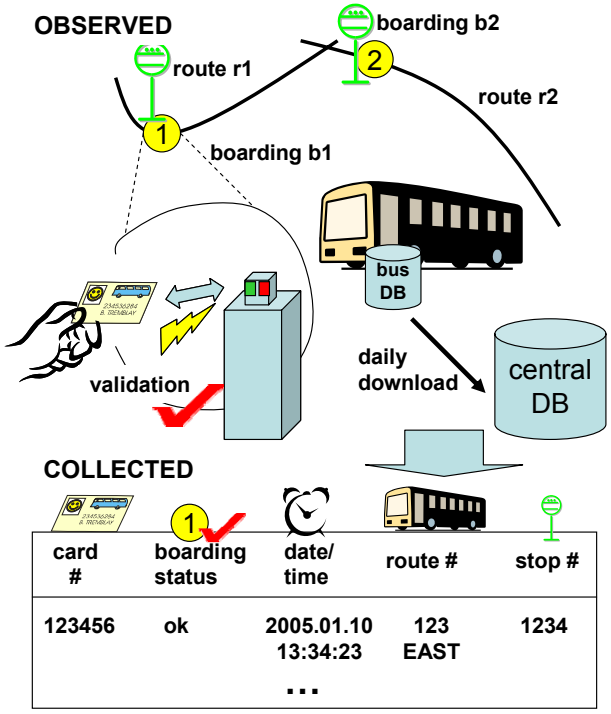

Fig. 1. The STO smart card data collection process

IV. RESULTS

A. Dataset

From an analytical viewpoint, the interest of smart card data arises from the repeatability of the observations for a single unit, in our case a smart card. The full dataset contains more than 6 200 000 transactions, collected from 43 248 smart cards between January 1th and October 4th 2005. Transactions were grouped in weeks that gives 238 895 card-week. Each transaction is represented with 27 characteristics: a card identification, a date, a type of day (D1-Monday, D2-Tuesday …) and a feature per hour (feature H08=1 means that the card was validated between 8h00min and 8h59min).

Table I illustrates the dataset. For instance, smart card_id #11 was used for two transactions on day xx (a Monday) the first one between 8h00 and 8h59 and the second one between 9h00 and 9h59.

TABLE I
EXAMPLE OF A DATASET

CARD_ID	date	DAY_TYPE	H00	H01	…	H08	H09	…	…	H23
11	xx	D1	0	0	…	1	1	…	…	…
15	xx	D2	0	0	…	1	0	…	…	…
22	xx	D2	0	0	…	0	1	…	…	…
33	xx	D7	0	0	…	0	1	…	…	…
…	…	…	…	…	…	…	…	…	…	…

In the following sections, global indicators of regularity and transit network usage are developed. Then, the behaviours of two smart cards, used during the entire period of observation (277 consecutive days), are examined in more details. Other measures of regularity are constructed with those two demonstrative cards.

B. Longitudinal analysis and regularity of travel behaviour

1) Activity rate on the transit network

A first analysis allows evaluating activity rate on the network for each card. As mentioned earlier, the range of observation is 277 consecutive days. However, not all the cards were issued at the same time; hence, the observation period ranges from 1 to 277 days. Moreover, users do not necessarily travel everyday. Activity rate (AR) is the ratio between these two periods: $AR = D_B/D_O \leq 100\%$ where:

D_B: number of days where the card was validated for boarding on a bus

D_O: number of days of observation (last day of boarding – first day of boarding between January 1st and October 4th)

$D_B \leq D_O$

The dataset used for analysis is summarised by Fig. 2. It specially shows the distribution of cards according to activity rate. On average, a card is used for travel on the transit network 58% of the days. We observe that approximately 14% of the cards belonging to the experimental sample have been used in the transit network on 60-65% of the observed days. Regarding the 7% of the cards used for boarding the transit network 95-100% of the observed days, they have, in 90% of the case, been observed for less then 10 days.

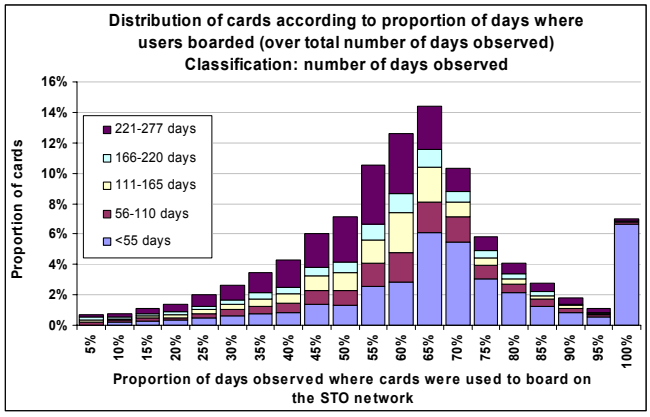

Fig. 2. Distribution of cards according to the number of observed days (on a maximum of 277 days) and the proportion of these days when they actually were used on the transit network

Details regarding the observation range (D_O) are given in Table II for the main fare types (gathering app. 80% of all the observed cards). The most interesting fact is the lower number of observed days for students and school pupils. Actually, it is plausible that these types of cards belong to users that do not use the transit network when their study activities are finished (summer). The average activity rate is comparable for the various types of cards.

TABLE II
NUMBER OF DAYS OBSERVED AND ACTIVITY FOR THE MAIN FARE TYPES

Fare type	AvgNbDaysObs	ActivityRate
Regular Adult	146	54.6%
Regular Student	85	50.6%
Express Adult	161	50.8%
School	74	54.3%
Loyalty Adult	187	54.7%

2) Boardings per day

Fig. 3 presents the average number of boardings per day for the main card types. We see that the students do, in average, more boardings per day during the weekdays. We also observe that the number of boardings per day clearly drops during the weekend. Finally, the behaviours seem quite regular during the 5 weekdays for all card types.

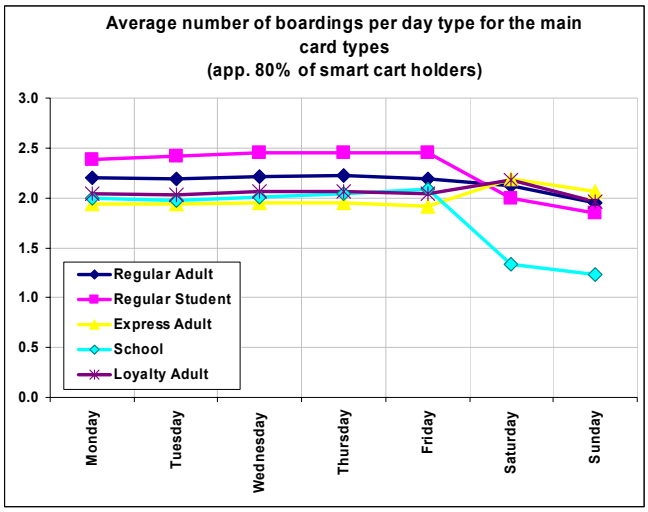

Fig. 3. Average number of boardings per day type for the 7 most frequent card types (app. 80% of smart card holders)

A study of the variability in the number of boardings for the weekdays, all over the 277 days observed, reveals that the express adults are the most regular (with a variation coefficient under 30%). In general the variability in the number of boardings is higher on weekends (between 50% and 60%) than during the week, except for the school cards.

3) Enumeration of boarding stops

It is possible to enumerate all the transit stops used for boarding over a continuous period. It is a first appraisal of the spatial regularity of the behaviour over the transit network. This type of measure is rarely possible since travel behaviour data generally observe a single day of travel. With our dataset, it is possible to cumulate all the different stops that were used by every single card and to enumerate their first use in time. Fig. 4 summarises such an analysis for all the cards. The number of new stops, the average standard deviation and the average number of stops frequented are represented as a function of D_B. The mean rate of stop renew is evaluated at 0.25 new stop per day.

Some seasonality effects appear: knowing that the first observation is around January 1st, summer season (around 170 days from the beginning) is co-extensive with a sudden growth in the average number of new stops. Also the variability grows with the number of boardings observed. That could represent the acquisition of new stops during a longer period of time.

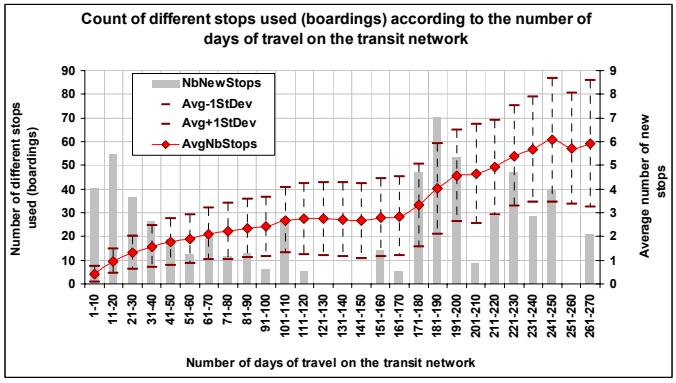

Fig. 4. Temporal evolution of the number of new stops frequented over multiple days of travel on the transit network

That kind of measure reveals that the users' behaviours are relatively variable and that they evolve during the period observed. Analyses on an exclusive day are not able to document the changing behaviours of the users. The picture made thanks to repeated observation on the same specimen nuances usual analyses based on an average day.

C. Spatio-temporal variability of the transit network usage

In this section the analyses focus on individual behaviours. Two cards (an Elderly and a Regular Adult) are selected and the regularity of their behaviours on the transit system is evaluated for the overall period observed (277 days).

The two cards presented here were selected because of their important use of the transit system (AR≈100%). Different aspects of their behaviour regularity are presented.

1) Number of boardings

The Elderly card (Resp. Regular Adult card) validated 1372 transactions (Resp. 478) during the period observed (277 days from January 1th to October 4th 2005). Transactions were made on 277 days (Resp. 274 days out of 277). It represents an average of 5 boardings (Resp. 2) per day.

Fig. 5 (Resp. Fig. 6) shows the number of boardings during the 277 days of observation for the Elderly card (Resp. Regular Adult card). That representation permits to identify some days with a particular behaviour for the Elderly card, most often Saturdays or Sundays. The number of boardings for the Regular Adult card is much more stable.

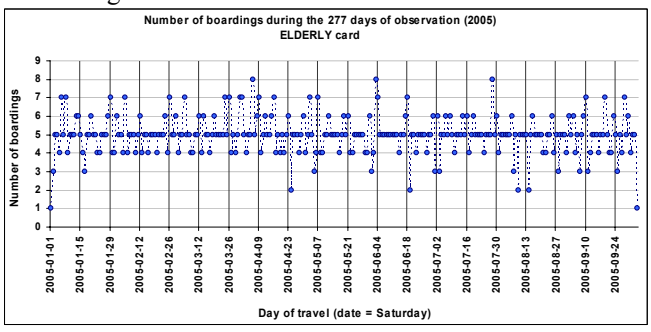

Fig. 5. Number of boardings during the 277 days of observation for an Elderly card.

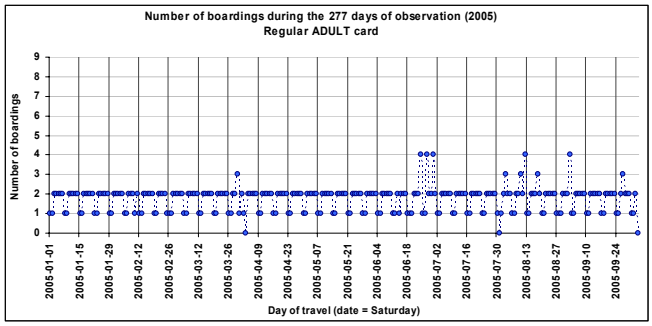

Fig. 6. Number of boardings during the 277 days of observation for a Regular Adult card.

Variability in the number of boardings is analysed on Fig. 7 for the Elderly card and on Fig. 8 for the Regular Adult card.

The Elderly card was used to validate more transactions on Saturday, and Sundays represent the day with both the minimal mean number of transactions and maximal variability (in terms of number of boardings per day).

For the period observed, the Regular Adult card was used the same number of boardings in all Saturdays. From Tuesday to Friday near 2 boardings were validated per day whereas Mondays are different from the rest of the week days. The variation coefficient is a little higher for the Regular Adult card than for the Elderly card.

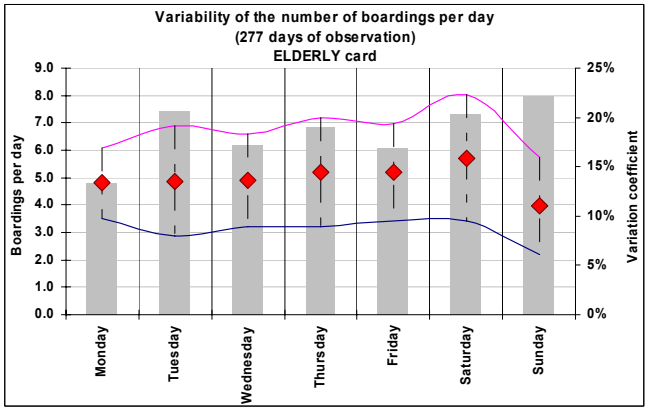

Fig. 7. Variability of the number of boardings par day - Elderly card

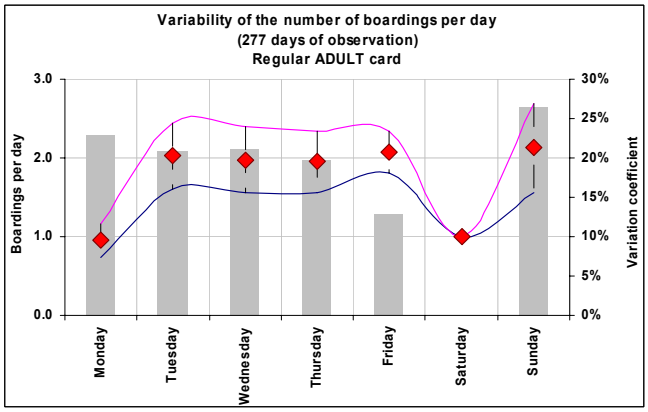

Fig. 8. Variability of the number of boardings par day - Regular Adult card

2) Boarding time

Clustering methods (k-mean) permitted to identify sets of days that are similar in terms of boarding times (from data in Table I). Two days are similar if the time when the boardings were validated is similar.

It permitted to identify 29 clusters for the Elderly card and 15 clusters for the Regular Adult card. This already shows that the Regular Adult card has a less variable temporal behaviour.

Fig. 9 and Fig. 10 show a classification that represents the proportion of each cluster (temporal behaviour) for each day (for the Elderly card, only the 15 most frequent clusters are considered, it represents 85% of the days observed).

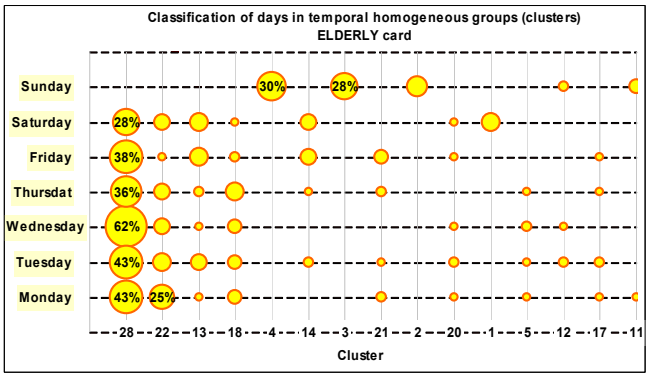

Fig. 9. Proportion of each cluster for each day - Elderly card

Wednesdays are the most regular days for the Elderly card, 62% for all the Wednesdays belong to the same cluster (n°28).

Moreover that same cluster (n°28) is the most frequent behaviour for the Elderly card whatever the day (35.4% for all the days). Cluster n°28 reveals a behaviour of 5 boardings at the following times 8h-9h, 9h-10h, 10h-11h, 11h-10h and 15h-16h.

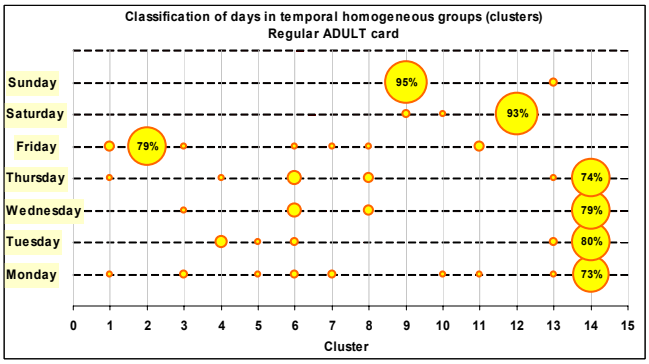

Fig. 10. Proportion of each cluster for each day - Regular Adult card

The Regular Adult card shows a concentration of the behaviours in 4 clusters (n°2, 9, 12 and 14). Each cluster is representative of a day for a minimum of 73% (and a maximum of 95%!).

The weekend days are particularly homogeneous, 93% and 95% of the behaviours belong to the same cluster.

Week days have also a stable behaviour (the same cluster from Monday to Thursday, and another important principal cluster for Friday). It outlines that this card is used regularly, in terms of boarding time.

Fig. 11 (Elderly card) and Fig. 12 (Regular Adult card) expose for each day (in the 277 day period) the cluster that is concerned.

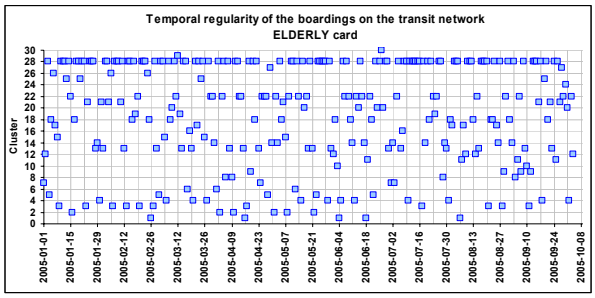

Fig. 11. Temporal regularity of the boardings - Elderly card

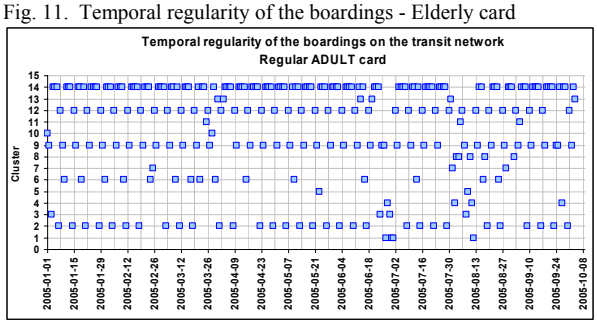

Fig. 12. Temporal regularity of the boardings - Regular Adult card

V. Conclusion and Perspectives

Smart card data have the ability to reveal various aspects regarding routine and variability of travel behaviours. In this respect, they are richer than usual travel data focusing on a single day of travel for each individual. Obviously, they cannot replace travel survey data since they only observe transit users but they shed more light on the monitoring of behaviours over time and definitely help qualify the classical measures of mobility regarding the average week day.

This paper has demonstrated various measures of regularity, some of them using data mining techniques. The enumeration of all the transit stops used for boarding by all the cards proves that travel behaviours vary in space, either due to the existence of punctual activities involving new transit paths or evolving regular paths. Hence, it showed that the number of stops used does not stabilise in time. A segmented study per card type will help further this analysis. The study of two demonstrative cards with high activity rates validated the usability of smart card data and data mining techniques to appreciate the temporal variability of use of the transit system. Systematic application of these techniques to every single card will help generalise the results and estimate regularity index for card types. Thus, a better understanding of the variability of the transit demand could help the transit operators to move towards a more customised supply for types of users that would match the observed variability.

Future experiments include the study of the spatial variability of behaviour over contiguous and compatible (Mondays, Tuesdays …) days of travel as well as measures combining both spatial and temporal perspectives. In this regard, spatial clustering methods and activity prisms will be examined.

References

[1] M. Shelfer and J.D. Procaccino, Smart card evolution, *Communications of the ACM*, July 2002, Vol. 45, No. 7, pp. 83-88.

[2] W. Bonneau and editors, The role of smart cards in mass transit systems, *Card Technology Today*, June 2002, p.10.

[3] R. Clarke, Person location and person tracking: Technologies, risks and policy implications, *Information Technology & People*, 14 (2), 2001, pp. 206-231.

[4] M. Bagchi and P.R. White, The potential of public transport smart card data, Transport Policy12, 2005, p. 464-474

[5] M. Trépanier, S. Barj C. Dufour and R. Poilré, Examen des potentialités d'analyse des données d'un système de paiement par carte à puce en transport urbain, Exposé préparé pour la séance sur "Utilisation des systèmes de transport intelligents (STI) à l'appui de la gestion de la circulation", *congrès annuel de 2004 de l'Association des transports du Canada à Québec*, Québec, p.4, 10-14.

[6] M. Hoffman and M. O'mahony, Transfer Journey Identification and Analyses from Electronic Fare Collection Data, *Proc. of the 8th Intl. IEEE Conf. on Intelligent Transportation Systems*, Vienna, Austria, Sept 13-16, 2005, p. 825-830M.J.A.

[7] S.S. Anand and A.G. Büchner, *Decision Support Using Data Mining*, Financial Times Pitman Publishers, London, UK, 1998.

[8] C. Westphal and T. Blaxon, *Data Mining Solutions*, John Wiley, New York, 1998.

[9] D. Braha, *Data Mining for Design and Manufacturing*, Kluwer, Boston, MA, 2001.

[10] M.J.A. Berry and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*, New York: Wiley, 2004.

[11] B. Agard, A. Kusiak, Data Mining in Selection of Manufacturing Processes, in O. Maimon and L. Rokach (Ed.), *The Data Mining and Knowledge Discovery Handbook*, Springer, 2005, pp. 1159-1166.

[12] C. da Cunha, B. Agard, and A. Kusiak, Improving manufacturing quality by re-sequencing assembly operations: a data-mining approach, *18th Intl. Conf. on Production Research – ICPR 18*, University of Salerno, Fisciamo, Italy, July 31 – August 4, 2005.

[13] B. Agard, C. da Cunha, Manufacturing plant layout supported with data mining techniques, *6th Intl. Conf. on Integrated Design and Manufacturing in Mechanical Engineering – IDMME 06*, Grenoble, France, May 17-19, 2006.

[14] C. da Cunha, B. Agard, Business Process Reengineering with Data Mining in Real Estate Credit Attribution: a Case Study, *Intl. Conf. on Information Systems, Logistics and Supply Chain – ILS 2006*, Lyon, France, May 15-17, 2006.

[15] P. Bonsall, F. Montgomery and C. Jones, Deriving the Constancy of Traffic Flow Composition from Vehicle Registration Data, *Traffic Engineering and Control*, Vol. 25, No. 7/8, 1984, pp 386-391.

[16] E.I. Pas and F.S. Koppelman, An examination of the determinants of day-to-day variability in individuals' urban travel behavior, *Transportation*, no 13, 1986, p.183-200.

[17] E.I. Pas, A flexible and integrated methodology for analytical classification of daily travel-activity behavior, *Transportation Science*, Vol.17, No.4, 1983, p.405-429.

[18] E.I. Pas, Weekly travel-activity behavior, *Transportation*, no.15, 1988, p.89-109.

[19] M. Jun and K. Goulias, 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*, no.24, 1997, p. 309-331.

[20] R. Schlich, K.W. Axhausen, Habitual travel behaviour: Evidence from a six-week travel diary, *Transportation*, no.30, 2003, p. 13-36.

[21] R. Kitamura, T. Yamamoto, Y.O. Susilo and K.W. Axhausen, How routine is a routine? An analysis of the day-to-day variability in prism vertex location, *Transportation Research Part A*, no.40, 2006, p. 259-279.

[22] T. Gärling, K.W. Axhausen, Introduction: Habitual travel choice, *Transportation*, no.30, 2003, p. 1-11.

[23] R. Rakotomalala, "TANAGRA : un logiciel gratuit pour l'enseignement et la recherche", *in Actes de EGC'2005*, RNTI-E-3, vol. 2, 2005, pp.697-702.