MINING PUBLIC TRANSPORT USER BEHAVIOUR FROM SMART CARD DATA

Bruno Agard[1], Catherine Morency[2,3], Martin Trépanier[1,2,3]

[1]Polygistique, [2]Groupe MADITUC, [3]Centre de recherche sur les transports (CRT), École Polytechnique Montréal, C.P. 6079, succ. Centre-Ville, Montréal (QC), Canada, H3C 3A7
[bruno.agard, cmorency, mtrepanier]@polymtl.ca

Abstract: In urban public transport, smart card data is made of millions of observations of users boarding vehicles over the network across several days. The issue addresses whether data mining techniques can be used to study user behaviour from these observations. This must be done with the help of transportation planning knowledge. Hence, this paper presents a common "transportation planning/data mining" methodology for user behaviour analysis. Experiments were conducted on data from a Canadian transit authority. This experience demonstrates that a combination of planning knowledge and data mining tool allows producing travel behaviours indicators, mainly regarding regularity and daily patterns, from data issued from operational and management system. Results show that the public transport users of this study can rapidly be divided in four major behavioural groups, whatever type of ticket they use. *Copyright © 2006 IFAC.*

Keywords: Smart card, Public transportation, Information system, Data mining.

1. INTRODUCTION

While Smart Card Automated Fare Collection Systems' popularity increases in public transport, a large quantity of data is gathered each day in the existing systems. Although they are mainly aimed to revenue collection, SCAFC systems can help planners to better understand public transport user behaviours, thus helping to improve the service. Indeed, these systems contain data on each boarding in a public transport network, with exact time and some precision on location. The datasets produced by these systems grow rapidly and requires both efficient data exploration tool and discrimination ability from planning experience. It is our belief that data mining gathers useful data processing methods to enhance the analytical relevance of smart card data.

This paper tries to join the better of two worlds – data mining methods and public transport planning models— in order to obtain an improved portrait of user behaviours in a public transport system equipped with a SCAFC system. User behaviour is a quite large research field, and this paper will focus mainly on weekday's trips habits.

In the literature review, some works on smart card data analysis in public transport we be exposed, as well as advances in data mining. Nonetheless, not any previous work combining the two research fields have been found. In the methodology section, the case study and the smart card data structure will be presented. The experiments section relates the different approaches used for analysis and the hypothesis for the different clustering actions. Finally, the paper exposes results on two aspects: data mining methods spin-offs, and user behaviour key elements.

2. REVIEW

2.1 Smart card in public transport

The complex fare system that is used by many public transport authorities can be better managed with the help of a Smart Card Automated Fare Collection System. Smart cards can store more than one transport document at a time and the card is automatically validated. The need to integrate fare policies within large metropolitan areas promote smart card usage (Bonneau 2002).

However, privacy is an important issue that could retain smart card implementation. The French Council for Computer and Liberty recommends being careful with such data because one may reconstitute the personal movements of a specific person (CNIL, 2003). Clarke (2001) recalls that smart card data is not different from other individual data collection systems (credit card, road toll, police corps database).

With technological and ethical problems resolved, several advantages arise from the analysis of SCAFC data (Bagchi and White 2004): access to larger sets of individual data, possibilities of links between user and card information; continuous data available for long periods, better data of a large part of the transit users.

These authors conducted a study on the passenger transfer behaviours on the Bradford and Merseyside transit networks (UK). The absence of alighting location information was then identified as the main issue for further analysis. In a recent paper (Bagchi and White 2005), they propose different actions to avoid undermining data quality in SCAFC systems. They especially 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 the case of the Société de transport de l'Outaouais (STO), Trépanier et al. (2004) have shown the potentialities of using SCAFC data for public transport network planning with the help of a Transportation Object-Oriented Modelling.

2.2 *Data mining tools and applications*

Anand and Büchner (1998) defined data mining as the discovery of non-trivial, implicit, previously unknown, and potentially useful and understandable patterns from large data sets.

Westphal and Blaxton (1998) categorized data mining functions as classification, estimation, segmentation, and description. *Classification* involves assigning labels to new data based on the knowledge extracted from historical data. *Estimation* deals with filing in missing values in the fields of a new record as a function of fields in other records. *Segmentation* (or clustering) divides a population into smaller sub-populations with similar behaviour according to a predefined metric. It maximizes homogeneity within a group and maximizes heterogeneity between the groups. *Description* and *visualization* are used to explain the relationships among the data. Frequent patterns may be extracted in the form of A=>B rules with two measures of quality: the support which represents the number of times A occurs as a fraction of the total number of examples and confidence which expresses the number of times B exists in the data when A is present.

Data mining techniques offer applications in many areas, on can cite as examples Braha (2001) who propose many applications of in design and manufacturing as well as Berry and Linoff (2004) who presented numerous examples and applications of data mining in marketing, sales, and customer support. Agard and Kusiak (2004) developed a methodology for the design of product families with data mining. da Cunha et al. (2005) analyzed manufacturing quality operation for resequencing.

3. METHODOLOGY

3.1 *Case study*

Experiments were conducted at the Société de transport de l'Outaouais (STO), Gatineau, Quebec. The STO is a medium-size public transport authority operating 200 buses and servicing 240,000 inhabitants. 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.

3.2 *Smart Card Data structure*

A special information system has been developed at the STO to manage SCAFC data. This fare collection information system is made up of several subsystems as shown on Figure 1. Smart cards are first bought at emission locations and can be recharged at further locations. Then, when the user boards the buses, the smart card's fare gets validated. The validation is done following those steps: the bus system contains the planned runs for the day (a run is a sequence of stops to be deserved; it usually represents one direction of a route). The Global Positioning System (GPS) reader on the bus identifies the stop where the boarding is made. The system validates the run (correct route) at this location. Card number, date, time, validation status and stop number are stored at each boarding. This information is downloaded to the central server at each end of day.

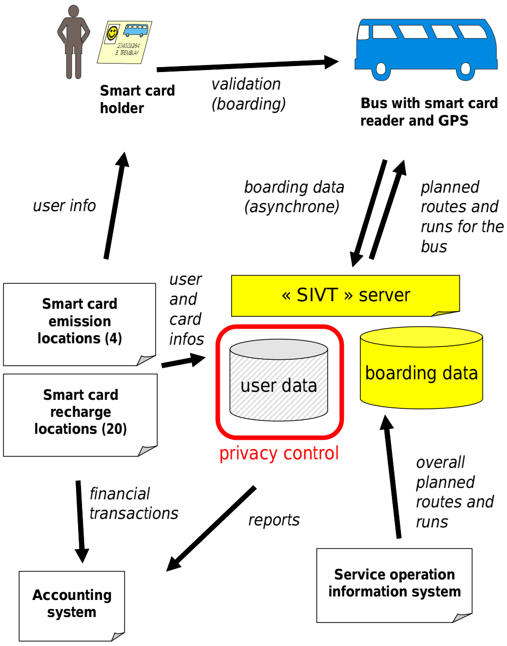
Smart card holder
validation (boarding)
Bus with smart card reader and GPS
user info
boarding data (asynchrone)
planned routes and runs for the bus
Smart card emission locations (4)
user and card infos
« SIVT » server
user data
boarding data
privacy control
Smart card recharge locations (20)
overall planned routes and runs
financial transactions
reports
Accounting system
Service operation information system

Fig. 1. The STO smart card information system

The central database management system (called SIVT server, *Système d'information et de validation des titres*) holds the information of users and boardings. The data is voluntarily separated to preserve confidentiality; individual user information will not be used in this study. The central server is also fed by service operation information system and smart card point-of-sale systems.

Figure 2 presents the simplified data model that stores boarding data. The key table ("Boardings") contains a record for each smart card validation on the network. Boarding refusals are also stored in the table, so they will be ignored for the analysis. A link is also made between boarding and the public transport network geometry ("Routes" and "Route-stops"). The link ensures the validity of the time and location of boardings. Information on cards and card recharges are also available but will not be used here.

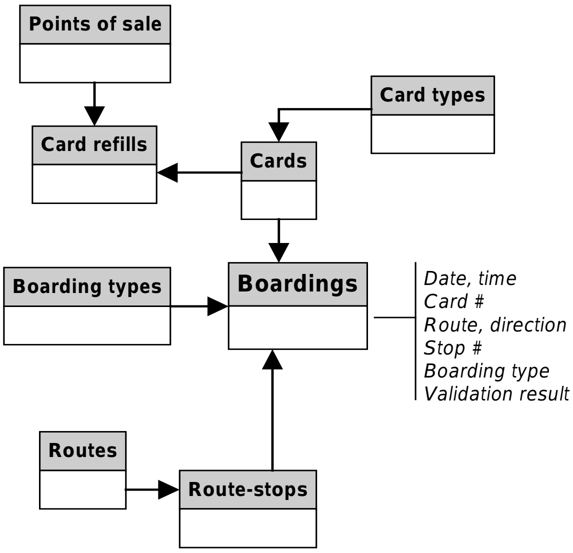

Fig. 2. Simplified database model for boarding data storage

3.3 Mining tools

Different data mining techniques and tools were used in this project:

- Filtering data
- Clusters : K-mean, HAC
- Group characterization

The dataset was extracted from a SQL Server database and was analysed within the TANAGRA free software (Rakotomalala, 2005).

3.4 Dataset

The dataset is a compilation of 2,147,049 boarding validations made by 25,452 card holders on the STO network between January 10th and April 1st, 2005. The transactions were summarized into twelve 1-week record for each card (238,895 user-weeks overall). Each record is divided into 20 binary variables, representing 5 weekdays X 4 periods per day (AM = 5:30 to 8:59, MI = 9:00 to 15:29, PM = 15:30 to 17:59 and SO = 18:00 and over). This day division is common to transport planning studies in Quebec, other countries may have different time intervals.

Table 1 presents sample records. Fields J2_AM to J6_SO represent the 20 periods (2=Monday, …, 6=Friday). For example, the card #12244 is of category "T1" (in the group "adult"). This card has been validated (user has boarded a bus at least once) on Monday morning (J2_AM) and Monday PM peak hour (J2_PM) on week "S1", on week "S2", and so on. In addition, we can assume that the holder of card #23231 has made at least one trip at each period on Monday in the first week.

Information will be extracted from that dataset in order to split the user-weeks in homogeneous groups according to their behaviours and thus explain (or reveal) the comportment of each group. In the sample dataset, customers are classified as "adults", "students" and "elderly" for commercial raisons. This classification relies on the transit card type.

Table 1: Sample dataset

Num_card	*Num_title*	*group*	*week*	*J2_AM*	*J2_MI*	*J2_PM*	*J2_SO*	**...**	*J6_SO*
12244	T1	adult	S1	1	0	1	0	…	…
12244	T1	adult	S2	1	0	1	0	…	…
1234312	T2	adult	S1	1	0	0	1	…	…
23231	T34	student	S1	1	1	1	1	…	…
…	…	…	…	…	…	…	…	…	…

3.5 Protocol

The goal is to characterize user behaviour by looking at similar trip habits during weekdays and through the weeks. The analysis will conduct to an aggregate view of the 25,452 card holders.

The first step considers the whole dataset as an ensemble and subdivides it in large homogeneous clusters on the base of the trip patterns observed on a weekly basis. The goal is to determine "natural" groupings of user-weeks. The number of groups is not fixed a priori; the Hierarchical Ascending Clustering method (HAC) is employed (Lebart et al., 2000). In order to accelerate the HAC algorithm, a first grouping is computed with a k-mean method to provide 20 groups. The result of the k-mean clustering becomes the input of the HAC. In order to build groups that have the same behaviours, the inputs of the clustering methods are constituted of all the Jx_xx columns.

In the second step, we analyse the composition of the natural grouping with respect to card type in order to see if the clustering method only reconstructs pre-known segments (adults, students, and elderly).

In the last step of this study, we examine the variability of the group belongings for the twelve weeks of observation. This will give a good idea of the regularity of the habits over time. It will also help identify the unusual weeks in terms of travel behaviours.

4. RESULTS

4.1 General user behaviour

The processing of the whole dataset produces 4 clusters of user-weeks with similar patterns. Two of the clusters (Group 1 and Group 3) have easily interpretable travel behaviours.

Fig. 3: The first group (45.6% of the user-weeks) clearly relates to the people with regular trips to and from constraint activities such as work since they mainly travel during the peak hours. On average, 79.4% of these users will travel during the AM peak hour and 71.0% during the PM peak hour on weekdays. The proportion of users travelling during the other periods is quite low, 6.4% during the day and 2.6% during the evening.

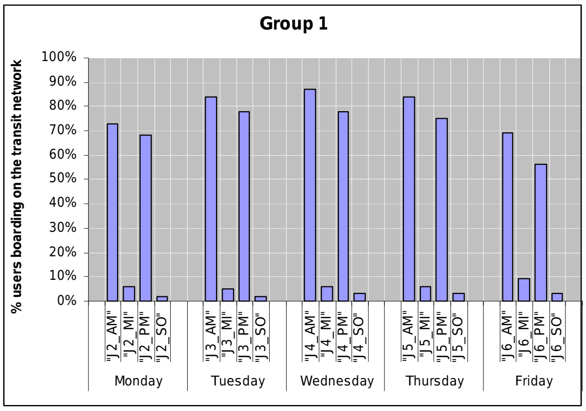

Fig. 3. General user behaviour – Group 1

Fig. 4: The third group (14.3% of the user-weeks) relates to people with regular activities in the first part of the day with, on average, 77.6% of the users travelling during the AM peak hour and 74.8% during the day.

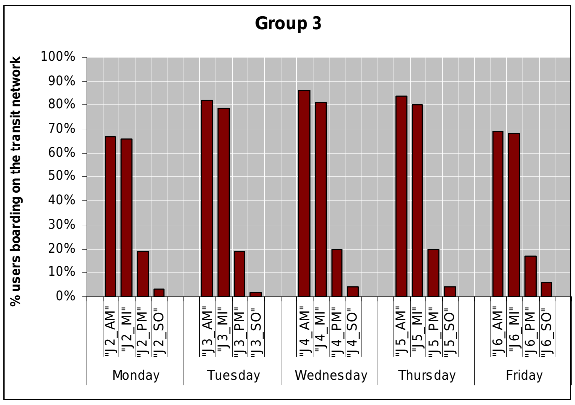

Fig. 4. General user behaviour – Group 3

The two other clusters (Group 2 and Group 4 with respectively 14.8% and 25.2% of the user-weeks) gather users with similar patterns of non-travelling. Actually, no clear travel pattern can be observed for Group 2 while Group 4 gathers users with lower use of the network. At most, 36% of the users of this group will use the transit network during the day period.

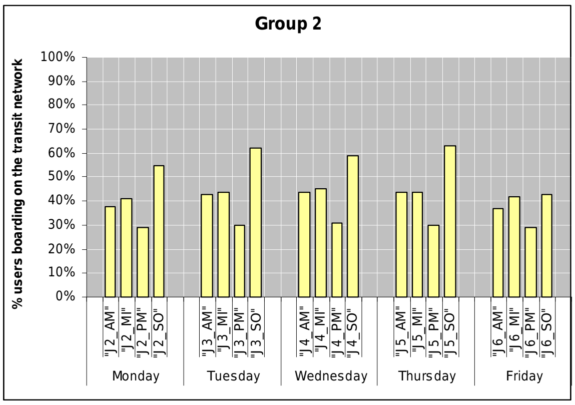

Fig. 5. General user behaviour – Group 2

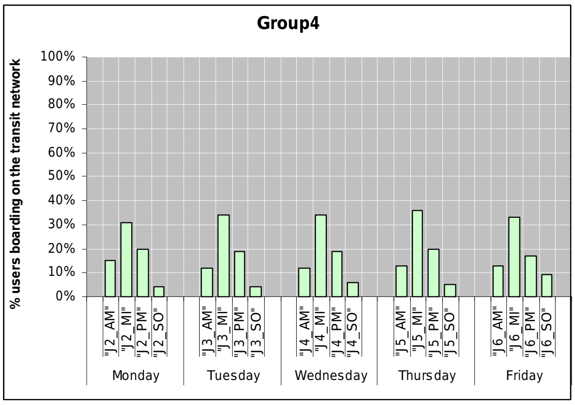

Fig. 6. General user behaviour – Group 4

4.2 Composition of the natural groups

The following tables summarize the composition of the four clusters created by the data mining method.
On the first hand, Table 2 presents the distribution of the user-weeks in the four groups according to the type of card hold by the traveller. For instance, it shows that almost 80% of the weeks of travel of elderly card holders are classified in Group 4. This is no surprise since this Group has a low level of mobility. It also confirms that the symmetrical movements observed for user-weeks of group 1 are partly explained by the plausible level of constraint of the activities of the adults, with almost 60% of the adult card holders being in this group.

Table 2: Distribution of user-weeks in the four clusters according to card type

Card type	Gr1	Gr2	Gr3	Gr4	TOT
Adult	**58,8%**	13,9%	9,2%	18,1%	100%
Student	21,0%	17,7%	26,4%	34,8%	100%
Elderly	6,2%	6,4%	7,9%	**79,5%**	100%

Table 2 also shows that the weeks of travel of student card holders are spread over the four natural groups. This is quite intriguing. It may be caused by the fact that some students automatically receive transit passes due to their status without actually needing it for their main travel needs. What is then observed is a very low subset of their travelling behaviours on weekdays.
On the second hand, Table 3 presents the composition of the four natural clusters in terms of card type (of the travellers for which user-weeks are examined). We see that the adult card holders represent 85% of the user-weeks of Group 1. This also explains the regularity of the travel patterns on weekdays of this Group. This table shows that student are the main segment composing Group 3 which shows regular travel patterns in the first part of the weekdays.

Table 3: Composition of the four clusters regarding card type

Card type	Adult	Student	Elderly	Total
Gr1	85,6%	13,9%	0,5%	100%
Gr2	62,4%	36,1%	1,4%	100%
Gr3	42,7%	55,4%	1,8%	100%
Gr4	47,7%	41,7%	10,6%	100%

What these tables show as well is that some card types are linked to regular behaviours such as the elderly cards (app. 80% of the users with this type of card belong to group 4) but that other types can reveal variable travel behaviours over several weeks.

4.3 Variability over 12 weeks

A better understanding of these two groups is obtained by the study of the variability of belonging of the users over the 12 weeks.

For the student case, a study of the belonging among these groups over the 12 weeks, presented in Figure 7, clearly spots one irregular week of travel. The atypical distribution observed on week 10 is due to the school break where students temporarily adopt other activity patterns.

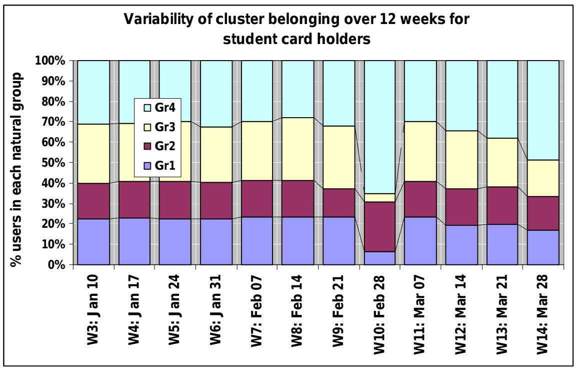

Fig. 7. Variability of cluster belonging over 12 weeks of the student card holders

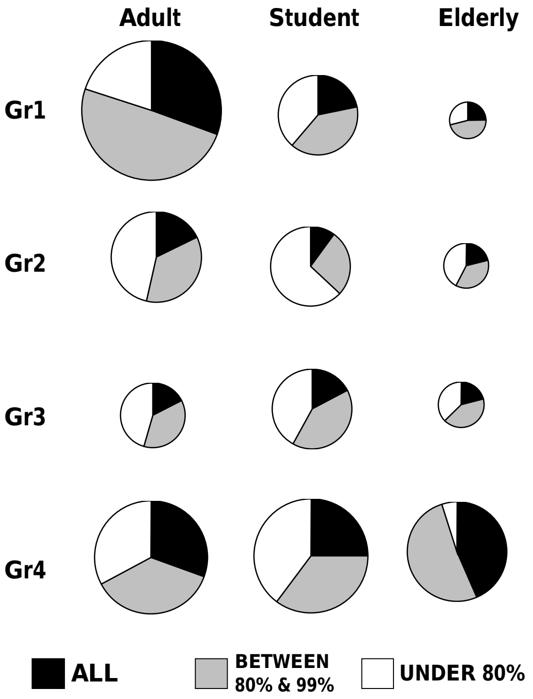

Fig. 8. Proportion of users regarding their cluster belonging over weeks (week # 10 removed)

In the figure 8, the black sectors of the pie charts indicate users that belong to the same group for all weeks (gray is for most of the weeks). The size of the charts is proportional to the number of users. We can see that a large majority of the adults of group 1 have a regular behaviour. It is less the case for adults of other groups. Students have less regular behaviour. That could be related to 1) irregular college schedules and 2) students are sometimes accompanied to school by their parents and will not use public transit. Elderly people of group 4 are concentrated in Group

4, but let us remind that this group is atypical with a low level of mobility.

5. CONCLUSION

This study is only a first impression about mining data from Smart Card Automated Fare Collection systems. First results tell that data mining techniques help to identify and characterize market segments among public transportation users. It also demonstrates the wide set of statistics that can be calculated from the dataset; much work is still to come to pinpoint the most important statistics for the STO and the user behaviour researchers.

Further studies will be conducted to better characterise both supply and demand on the STO public transport network. Typically, the following elements may be examined:

- Geospatial trip behaviour (with the help of geospatial data mining techniques).
- Specific route usage over space and time (to measure the turnover of the users).
- More detailed mining in time period (for example, using the exact time of the first boarding of each day over a year period).

REFERENCES

Anand, S.S. and Büchner, A.G. (1998), Decision Support Using Data Mining, Financial Times Pitman Publishers, London, UK.

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

Berry, M.J.A. and Linoff, G. (2004), Data Mining Techniques: For Marketing, Sales, and Customer Support, New York: Wiley.

Bonneau, W. and editors (2002). The role of smart cards in mass transit systems, Card Technology Today, June 2002, p.10.

Braha, D. (2001), Data Mining for Design and Manufacturing, Kluwer, Boston, MA.

Chapleau, R. (1986). Transit Network Analysis and Evaluation with a Totally Disaggregate Approach, Selected proceedings of the World Conference on Transportation Research, Vancouver.

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

CNIL –Commission nationale de l’informatique et des libertés (2003). Recommandation relative à la collecte et au traitement d'informations nominatives par les sociétés de transports collectifs dans le cadre d’applications billettiques, CNIL, Délibération N° 03-038.

Lebart, L., Morineau, A., Piron, M., (1998), Statistique exploratoire multidimensionnelle, Ed. Dunod, 2000.

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

Trépanier, M., Barj, S., Dufour, C., Poilpré, R. (2004), 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" du congrès annuel de 2004 de l'Association des transports du Canada à Québec (Québec), p.4, 10-14.

Westphal, C. and Blaxton, T. (1998), Data Mining Solutions, John Wiley, New York.

Agard, B. and Kusiak, A. (2004) Data-Mining Based Methodology for the Design of Product Families, International Journal of Production Research, **42** (15), pp. 2955-2969.

C. Da Cunha, C., Agard, B., and Kusiak, A. (2005), Improving manufacturing quality by re-sequencing assembly operations: a data-mining approach, 18th International Conference on Production Research – ICPR 18, University of Salerno, Fisciamo, Italy, July 31 – August 4.

ACKNOWLEDGEMENTS

The authors wish to acknowledge 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).