Mining Smart Card Data from an Urban Transit Network

Bruno Agard
École Polytechnique de Montréal, Canada

Catherine Morency
École Polytechnique de Montréal, Canada

Martin Trépanier
École Polytechnique de Montréal, Canada

INTRODUCTION

In large urban areas, smooth running public transit networks are key to viable development. Currently, economic and environmental issues are fueling the need for these networks to adequately serve travel demand, thereby increasing their competitiveness and their market share. Better balance between transit supply and demand will also help reduce and control operating costs.

The fact is, however, that transit operators are finding it extremely difficult to adjust the service to meet the demand, because this demand changes continuously with the time or day of travel (period of the day, day of the week, season or holiday) and other factors like weather and service breakdown. In order to enhance their service, operators need to better understand the travel demand (customer behaviors and the variability of the demand in space and time). This can be achieved only by continuously monitoring the day-to-day activities of users throughout the transit network.

Some large cities around the world take advantage of smart card capabilities to manage their transit networks by using Smart Card Automated Fare Collection Systems (SCAFCS). An SCAFCS gives travelers greater flexibility, since a single card may be used by one user at various times and on different parts of the transit network, and may support various fare possibilities (by travel, line, zone, period, etc.). For transit operators, these systems not only validate and collect fares, but also represent a rich source of continuous data regarding the use of their network. Actually, this continuous dataset (developed for fare collection) has the potential to provide new knowledge about transit use. Following the application of various pretreatments which make it possible to extract real-time activity, data mining techniques can reveal interesting patterns. These techniques are aimed at precisely describing customer behavior, identifying sets of customers with similar behaviors, and measuring the spatial and temporal variability of transit use. Patterns are extracted and analyzed to document various issues, such as identifying transit use cycles or homogeneous days and weeks of travel for various periods of the year. This information is required for a better understanding and modeling of customer behavior, and consequently better adjustment of the service to the demand. These adjustments may, for instance, lead to the restructuring of the transit network, to the adaptation of route scheduling or to the definition of new and different subscription options (fares).

Below, results from various experiments conducted with a real dataset are provided. They show the potential of data mining to provide useful and novel information about user behavior on a transit network. The data processed in the study are extracted from a system operating in a Canadian city (Gatineau, Quebec).

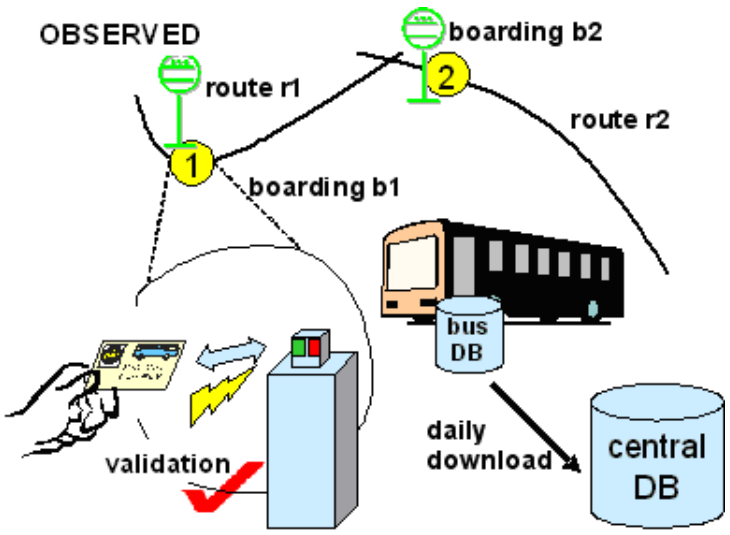
BACKGROUND

Smart Card in Public Transport

Generally, SCAFCS are composed of cards, onboard readers and a centralized information system (see Figure 1).

A smart card is simply an RFID device implanted in a transport card, and is similar to a typical credit card. Smart cards offer various advantages over traditional paper systems:

Figure 1. Smart card automated fare collection system (SCAFCS)



- Validation of the user transaction (boarding/transfer) is instantaneous and does not require any interaction with the driver.
- Complex fare systems (multiple zones, rates), common in large metropolitan areas where networks are integrated, can easily be managed.
- Elementary security procedures can be taken.
- Data are continuously collected, resulting in datasets much larger than those usually available to measure customer traffic; hence, a larger user proportion is observed.

These systems are being used increasingly frequently, since the software and hardware tools necessary to support their implementation are now stable and accessible, with a good quality of data output (Chira-Chavala and Coifman 1996, Meadowcroft 2005). Smart cards in transportation have mainly been implemented to simplify fare collection, but could also be used to advantage to monitor the service itself. In most cases, the adoption of a smart card system by transit authorities is related to their level of funding and the level of sophistication of their other technologies (Iseki et al. 2007, Yoh et al. 2006). Cheung (2006) demonstrated the overall benefits of smart card systems in The Netherlands, but it was Bagchi and White (2004, 2005) who were the first to substantiate this potential for transit planning. Using three case studies (British networks), they illustrate the ability of smart card data to estimate turnover rates, trip rates per card and the impacts of the use of smart cards on the number of linked trips. They also discuss the complementary nature of smart card data and other data collection methods, arguing that smart cards should not replace those methods. Utsunomiya et al. (2006) presented a study based on Chicago Transit Authority data for a one-week period involving about 500,000 boarding transactions. They reported difficulties associated with the data, especially missing transactions and incorrect bus routes.

Travel Behavior Variability

In practice, many transit planners use statistics from synthetic models, onboard travel surveys or regional travel surveys to describe how their networks are used. Average statistics are constructed, describing typical customer behaviors during a typical weekday. The underlying hypothesis that all weekdays are similar is

less well accepted. Actually, many studies validate the fact that travel behaviors vary a great deal in space and time. Some studies show the importance of understanding the variations in daily peak profiles to arriving at a better assessment of demand management schemes (Bonsall et al. 1984). Others illustrate the construction of detailed results by classifying travelers in terms of similar daily activity patterns (Jun and Goulias 1997). Garling and Axhausen (2003) talk about the habitual nature of travel. Metrics are available to evaluate any similarity between days of travel using a six-week travel diary (Schlich and Axhausen 2003), and others for the spatio-temporal variability (time-space prism) of day-to-day behaviors (Kitamura et al. 2006).

Not only is there agreement that variability needs to be assessed, but also that it is hard to measure, because the available data usually rely on a single day's record of each individual's travel. Bagchi and White (2004) have argued that a SCAFCS could improve these results because they provide access to larger sets of individual data. They also offer the possibility of linking user and transit card information. Hence, continuous data are available for long periods of time, which may lead to better knowledge of a large number of transit users.

More recent results have proved the ability of smart card data to measure the demand precisely and to understand its dynamics with a view to making day-to-day predictions (Morency et al. 2007). The methods developed will soon provide automatic survey tools which will help planners perform these tasks. However, trip-end analyses require the estimation of the destination of smart card trips. Trepanier et al. (2007) propose a destination estimation model with a success rate of about 80% at peak hours.

It is important to keep in mind that a transportation smart card belongs to an individual user, and that privacy rules must be respected (Clarke 2001, CNIL 2003).

MAIN FOCUS

The case study data used in our investigation were provided by the Société de transport de l'Outaouais (STO), a transit authority which manages a fleet of 200 buses in the Gatineau (Quebec) region of Canada (240,000 inhabitants). At this time (2007), about 80% of all STO passengers have a smart card, and every STO bus is equipped with a smart card reader and a GPS capturing device.

Data Collection Process

Figure 2 shows the general data flow of the smart card information system. Each time a smart cardholder boards a bus (step A), a transaction is recorded. The reader validates the smart card with the help of basic operational information, like the route number, the direction and the fare policy for this route. The current location is punched in at the same time (GPS reading). Every night, when the bus returns to the depot, all this data on transactions is uploaded to the database server, called the *Système d'Information et de Validation des Titres* (SIVT). At this point, the operational data for the next day are downloaded to the reader on the bus (step B).

The SIVT server collects data on both user and boarding (transactions), but keeps the information separate for privacy purposes. The SIVT server receives its operational data (routes, run assignments, stop list) from the STO's service operation information system (step C).

The SIVT exchanges data with the STO's accounting system on a continuous basis (step D). The accounting system is responsible for issuing and "recharging" individual smart cards.

In an SCAFCS, data are designed for fare collection and revenue management. Basic statistics, such as the number of users per line or per day, are easily extracted. The hypothesis driving the work reported here is that the data hold a great deal of other information about the overall system, and that it may provide interesting and relevant knowledge. That knowledge may be useful for helping planners better understand transit user behavior, leading to improved service.

The volume of data is perpetually growing, and this represents a considerable challenge in terms of knowledge extraction. In the present case, over a 10-month period, January 1st to October 4th, 2005, 27,033 smart cards were being used on the network, and nearly 6.2 millions transactions were validated. Trepanier and Chapleau (2001) developed an object model based on the Transportation Object-Oriented Modeling approach (Figure 3). The model shows strong links between network elements (routes, stops), operational data (drivers, vehicles, work pieces) and user data (cards, transactions, fares). There are up to 36 fare types at the STO, characterized by user status (student, adult, elderly) and network use privileges (regular, express and interzone routes).

Figure 2. General data flow of the smart card information system

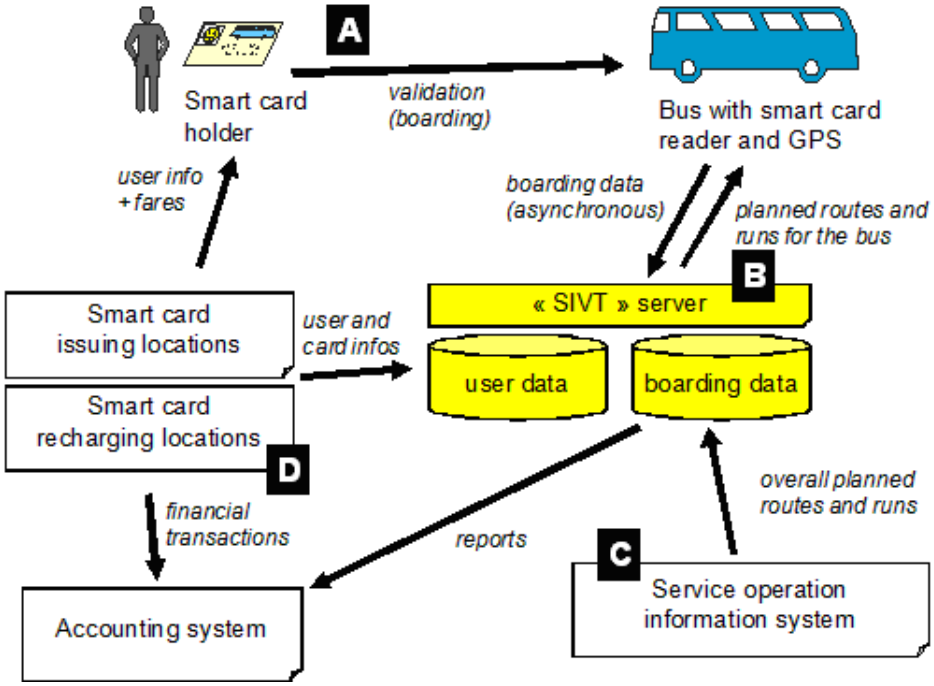
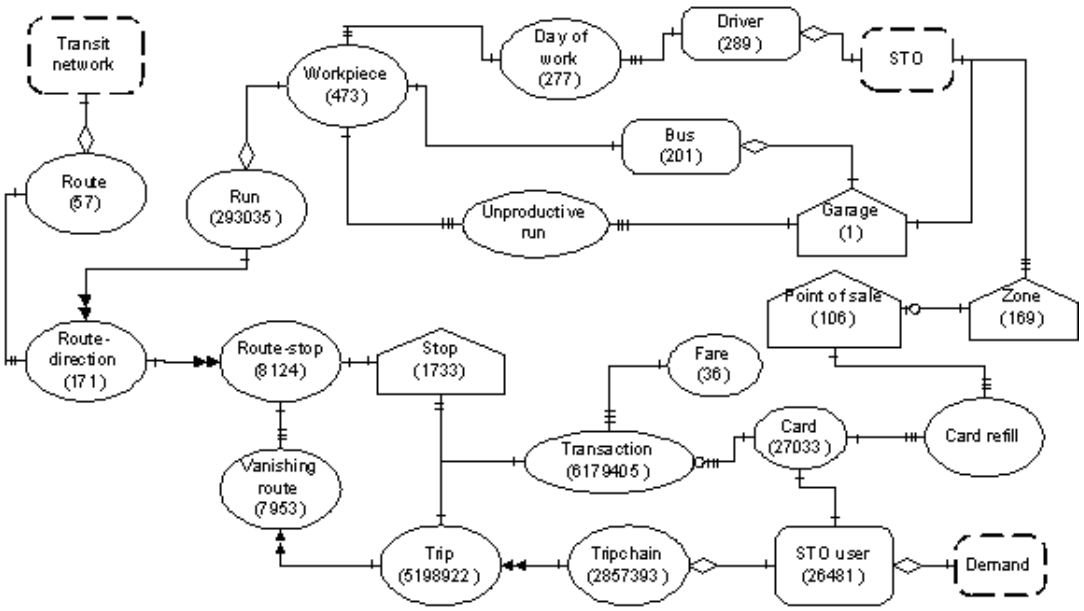


Figure 3. STO smart card object model for January to October 2005 data



Mining the Data

Here, we focus on the automatic extraction of user patterns on the transit network, using data mining tools. Data transformation and analysis are explained in detail. For a description of any of the data mining tools we mention, the reader may refer to the relevant chapters of the present encyclopedia.

Data Preprocessing

Each data refers to a specific spatial location and time. Every time a card is validated, a new line is created in the database. Since one card corresponds to one user, multiple lines can be created for each individual user. Each user may also have a different number of boardings, depending on his or her own activity.

Preprocessing operations have been conducted in such a way that every customer is represented on a fixed-length vector, in order to be able to extract user behavior in terms of temporal behavior. Depending on the accuracy of the analysis, two transformations are presented, both based on the same idea. In one, a day is made up of 24 periods (one hour each), and in the other a week is made up of 28 periods (4 periods per day, 7 days). Also included is a card ID and the day or week concerned (see Figure 4).

Group Behavior

Our analysis focuses on the way the individual users behave on the transit network, in terms of temporal activity. Different "groups" of users are identified. The number of groups depends on the level of granularity of interest in the study. An example, a split in 4 clusters with a k-mean, is presented in Figure 5.

Analysis of each cluster provides interesting information about the way the users who compose it use the transit network (temporal behavior). The following results are observed (see Figure 6, Agard et al. 2006): Cluster 1 represents 45.6% of the user weeks; Cluster 2: 14.8%; Cluster 3: 14.3%; and Cluster 4: 25.2%. Almost 60% of the cards from type "Adult" are in Cluster 1, while almost 80% of "Elderly" cards

Figure 4. Data transformation for the extraction of temporal user behavior

card #	boarding status	date/ time	route #	stop #
123456	ok	2005.01.10 13:34:23	123 EAST	1234
		...		

Data selection and transformation

Twenty four periods per day

Card_ID	date	Day_type	H00	H01	...	H08	H09	...	H23
2345	20/05/05	Friday	0	0	...	1	0	...	0
243	23/04/05	Saturday	0	0	...	0	1	...	0
4321	14/06/05	Tuesday	0	0	...	1	0	...	1
...	...	...	...	...	...	...	...	...	...

Four periods per day

Card_ID	week	D1_AM	D1_MI	D1_PM	D1_NI	...	D7_NI
12343	W1	1	0	1	0	...	...
12343	W2	1	0	0	1	...	...
1424	W1	0	1	1	0	...	...
...	...	...	...	...	...	...	...

Figure 5. Decomposition into 4 clusters with a k-mean

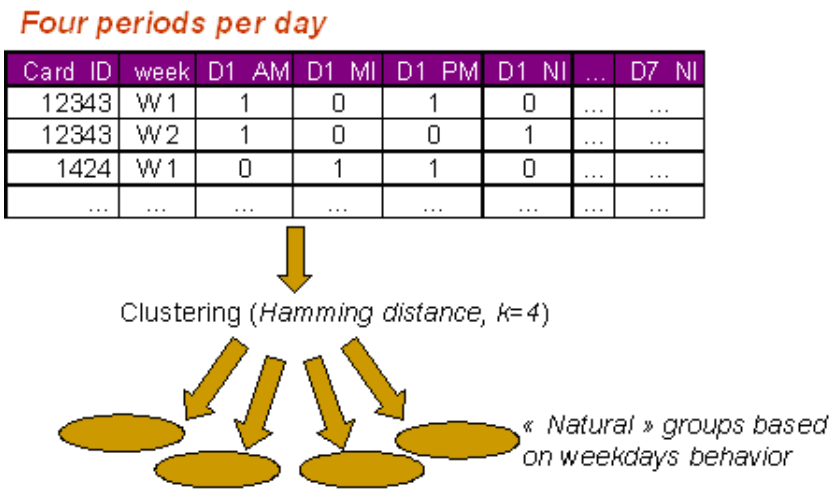


Figure 6. Composition of the clusters

Card type	Cl 1	Cl 2	Cl 3	Cl 4	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%

Card type	Cl 1	Cl 2	Cl 3	Cl 4
Adult	85.6%	62.4%	42.7%	47.7%
Student	13.9%	36.1%	55.4%	41.7%
Elderly	0.5%	1.4%	1.8%	10.6%
Total	100%	100%	100%	100%

are in Cluster 4. In contrast, Cluster 1 is composed of 85.6% of "Adults", Cluster 3 of 55.4% of "Students" and Cluster 4 of 10% of "Elderly".

Some results for the temporal behavior for each cluster (Agard et al. 2006) are summarized in Figure 7.

The users in Cluster 1 are typical workers, 79.4% of them traveling during the peak AM hours and 71.0% during the peak PM hours on weekdays. The users in Cluster 3 are the earlybirds, 77.6% of them traveling during the peak AM hours and 74.8% during the midday period. Clusters 2 and 4 show no clear patterns, but similar behavior may be observed from one weekday to another. However, Cluster 4 users are characterized by light use of the transit network.

Figure 7. Temporal behavior for clusters 1 to 4

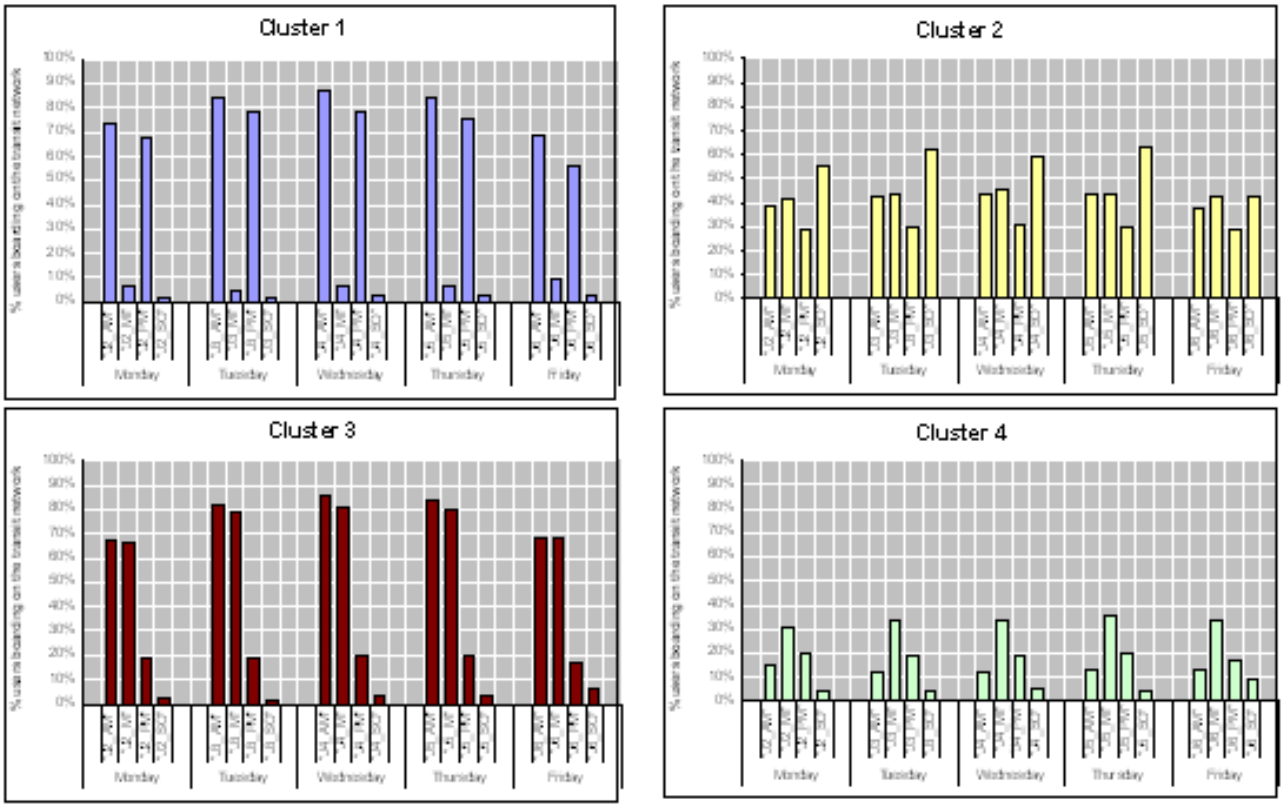
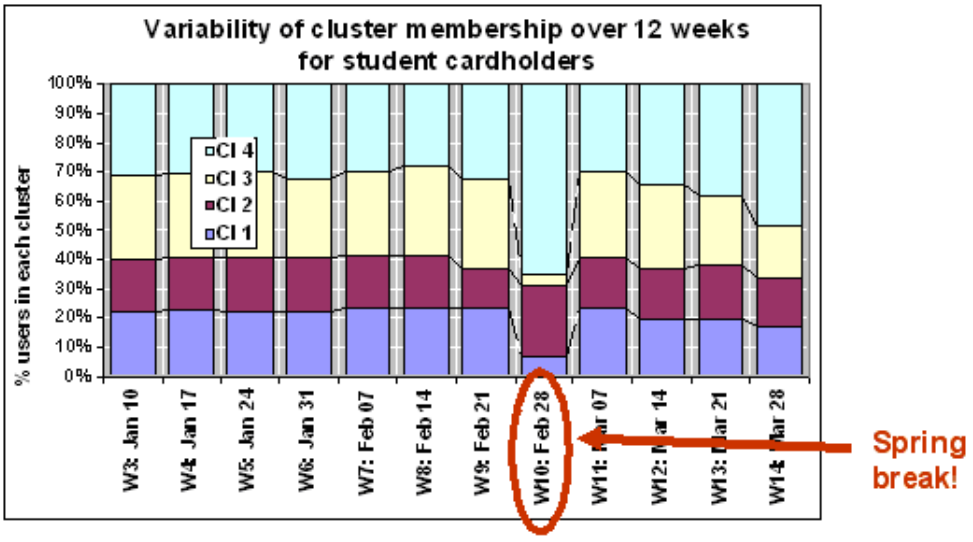


Figure 8. Variability of student behavior



If we focus on the variability (Agard et al. 2006) over 12 weeks (January to April, 2005) of a predefined population (students), we observe that the average proportion of each behavior is relatively stable, except during the spring break where Cluster 4 (light use) predominates (also, Cluster 2, with no clear pattern, becomes important during this period). All this shows a major change in student behaviors (Figure 8).

Individual Behavior

Application of the same technique (clustering on temporal description of user behavior) at various levels of granularity may reveal individual patterns (Figure 9).

Clusters are computed on the overall population in order to extract patterns which may be useful for all the datasets. From these clusters, it is relatively easy to construct one map for each user to represent the variability between clusters for each type of day (Morency et al. 2006) (see Figure 10).

The behavior of the card owner ("Regular ADULT" above) is easy to predict. For example, since 95% of this user's Sunday behavior belongs to the same cluster (no. 9), we can easily predict, from the description of cluster no. 9, which contains an hourly record of boarding times, this user's transit behavior on Sundays. Other users behave less predictably, and, in these instances, each day may be characterized by less representative clusters.

FUTURE TRENDS

Data mining of a SCAFC is a powerful tool for providing information about transit system use. Further studies will make it possible to evaluate travel behaviors in space, and to detect the existence of punctual activities involving new transit paths or established paths which are evolving.

Current and future research concerns include the following:

- Refinement of the alighting location estimation model (to be implemented at the STO)
- Geospatial trip behavior (using geospatial data mining techniques)
- Specific route usage over space and time (to measure user turnover)
- More detailed mining in a specific time period (using the exact time of the first boarding for each day over a one-year period, for example)
- Application of mining techniques to the measurement of the spatial variability of travel behavior (boarding locations) and linking this information with spatial statistics

Figure 9. k-mean clustering on daily activities

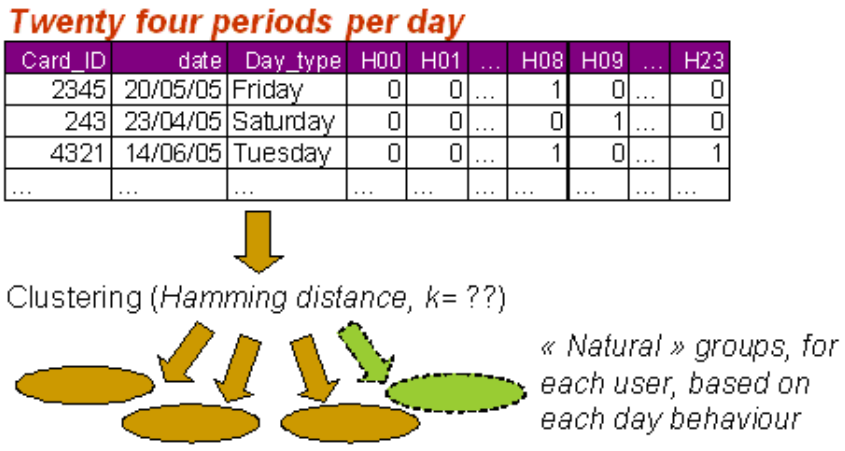
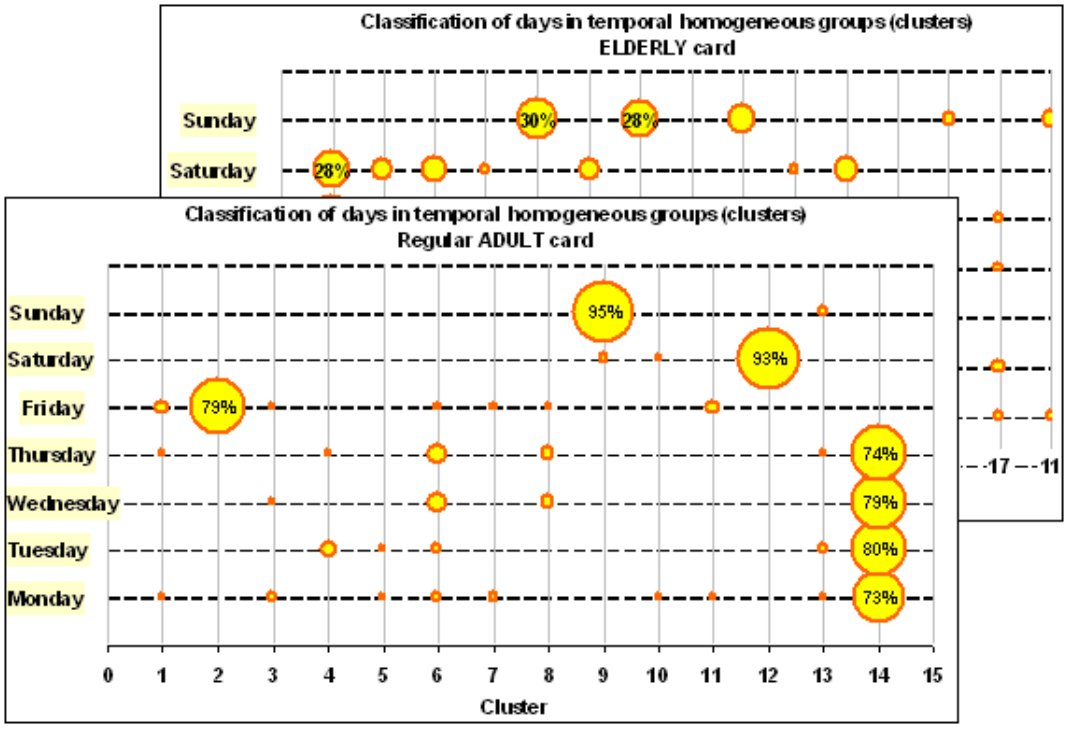


Figure 10. Examples of maps of behavior variability



- Application of subgroup discovery to reveal specific patterns in user behaviors.

CONCLUSION

This chapter shows that data mining is a powerful tool for knowledge extraction from a SCAFCS. Used for more than fare collection, an SCAFCS has the potential to reveal hidden patterns which are useful for addressing operational concerns.

The analysis of continuous data makes it possible to extract user transit behaviors and their variability. Data mining tools allow us to obtain interesting information about the overall system. That knowledge will help planners better understand the behaviors of transit users, with a view to improving service.

Trip behavior is an important topic in the transportation literature, and understanding user behavior helps in the design of better networks offering better services.

In the study presented in this chapter, the data generated for fare collection were used in different ways to evaluate the actual use of a public transit system, making it possible to:

- define typical customers and measure their travel behaviors, and
- analyze the variability of use of the system with respect to day, week or season.

With these results, it is possible to improve a transit system by adjusting the balance between customer behaviors and the level of service provided on each route. It is also possible to propose variable fares, which encourage temporal or spatial shifts towards less congested runs and routes.

Smart card data are proving to be richer than the usual travel data that focus on a single day of travel for a single individual. They also constitute a supplementary source for analysis.

REFERENCES

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

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

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

Bonsall, P., Montgomery, F., Jones, C. (1984) Deriving the Constancy of Traffic Flow Composition from Vehicle Registration Data, *Traffic Engineering and Control*, 25(7/8), 386-391.

Cheung, F. (2006) Implementation of Nationwide Public Transport Smart Card in the Netherlands, *Transportation Research Record*, no. 1971, 127-132.

Chira-Chavala, T., Coifman, B. (1996) Effects of Smart Cards on Transit Operators, *Transportation Research Record*, no. 1521, 84-90.

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

Gärling, T., Axhausen, K.W. (2003) Introduction: Habitual travel choice, *Transportation*, 30(1), 1-11.

Iseki, H., Yoh, A.C., Taylor, B.D. (2007) Are Smart Cards the Smart Way to Go? Examining the Adoption of Smart Card Fare Systems Among U.S. Transit Agencies, *Transportation Research Board Meeting*, Washington, 22p.

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*, no. 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*, no. 40, 259-279.

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, *The 9th International IEEE Conference on Intelligent Transportation Systems – ITSC 2006*, Toronto, Canada, September 17-20.

Morency C., Trépanier M., Agard B. (2007) Measuring transit use variability with smart card data, *Transport Policy*, 14(3), 193-203.

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

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., Tranchant, N. (2007) Individual Trip Destination Estimation in Transit Smart Card Automated Fare Collection System, *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations,* 11(1), 1-15.

Utsunomiya, M., Attanucci, J., Wilson, N. (2006) Potential Uses of Transit Smart Card Registration and Transaction Data to Improve Transit Planning, *Transportation Research Record*, no. 1971, 119–126.

Yoh, A.C., Iseki, H., Taylor, B.D., King, D.A. (2006) Interoperable Transit Smart Card Systems: Are We Moving Too Slowly or Too Quickly? *Transportation Research Record*, no. 1986, 69–77.

KEY TERMS

Activity Pattern: Sequence of activities made by a single person within a day. Each activity is bounded by the trips made before and after.

Board: To go onto or into a transportation vehicle.

Smart Card: Electronic device the size of a credit card which can store a small amount of data. It acts essentially like a radiofrequency identification tag (RFID).

Smart Card Automated Fare Collection System: System used to collect and validate fare payment aboard public transit vehicles.

Transit Mode: Public mean of transportation like bus, subway, commuter train.

Travel Behavior: In the field of transportation research, refers to the trip habits (frequency, purpose, time of departure, trip-end locations, etc.) of individual users on each transportation mode.

Urban Transit Network: Mass transit system which serves an urban population. It comprises modes and the provision of the service is guaranteed in the form of fixed schedules.