C. Morency, M. Trépanier, B. Agard, B. Martin, J. Quashie, Car sharing system: what transaction datasets can tell us regarding the user behaviors, The 10th International IEEE Conference on Intelligent Transportation Systems - ITSC 2007, Seattle, Washington, USA, Sept. 30-Oct. 3, 2007.

Car sharing system: what transaction datasets reveal on users' behaviors

Catherine Morency, Martin Trépanier, Bruno Agard, Basile Martin, Joel Quashie

Abstract— **Car sharing systems are gaining new members every month. However, few researches are conducted to better understand how these systems are used. In this paper, typical patterns of use of the car sharing system are identified using a transaction database covering a full year of operation. Data mining techniques are used to classify users according to their temporal patterns of car use frequency, traveled distance, and week use variability.**

The experiments reveal various classes of users. With respect to number of transactions throughout the year, users are segmented in two large classes: the regular and occasional ones, the majority of users belonging to the latter. The study of average trip length leads to the identification of 5 clusters of users. Finally, 8 types of typical weeks of use are described. Information about users' patterns could help the car sharing managers to optimize the use of the cars. It can also assist users in selecting the most advantageous subscription offer.

I. INTRODUCTION

Car sharing systems balance between private and public modes of transportation. They allow individuals to use a car when needed without having to buy one for their exclusive purpose. This, obviously, comes with some disagreement such as having to reserve the car for a predefined period, having to walk to the nearest parking lot or needing to select another mode if no car is available. However, it also has a lot of advantages because it gives access to a private, flexible mode of transportation, without having the entire burden that comes with it. Hence, it is not that surprising to see that car sharing systems are becoming more and more popular and that people are willing to engage in this new mode of transportation.

Manuscript received April 2nd 2007. This work was possible thanks to the support and collaboration of Communauto, company supplying *the car sharing system in Québec*. The research project is also supported by the Natural Science and Engineering Research Council of Canada (NSERC) and the FQRNT.

Prof. C. Morency is in Groupe MADITUC and in 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 (Québec) Canada, H3C 3A7, (e-mail cmorency@polymtl.ca).

Prof. M. Trépanier is in Polygistique, in Groupe MADITUC and in 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 (Québec) Canada, H3C 3A7, (e-mail mtrepanier@polymtl.ca).

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

B. Martin is a Master student in École Polytechnique de Montréal

J. Quashie is a Master student in École Polytechnique de Montréal

This paper dives into the analysis of data issued from one year of operation of a car sharing system in Montreal. It will describe the way the system is used in order to better understand the role it plays within an overall urban transportation system. For this purpose, transaction data are examined using data mining techniques in order to unravel patterns of use and classify users in typical segments.

The structure of the paper is the following. First, in section II, a literature review focuses on car sharing systems and the few studies that have been conducted on the use of these systems. It also presents the relevant principles regarding data mining techniques for the creation of classes of users. Section III presents the study methodology and the available dataset. Next, the applied data mining techniques are described. Section IV exposes the results of the analysis. Finally, the last section proposes directions for further investigations.

II. REVIEW

This review focuses on the background elements of this project: car sharing systems and data mining techniques.

A. Car sharing systems

Car sharing is not a new transportation mode. Years ago, when cars were luxurious goods, households were uniting to buy one. Its ownership and use were shared. It was for these households the only way to access automobile.

Nowadays, car sharing responds to new needs and is provided in an organized system. People want to benefit from the car's flexibility without supporting all its inherent costs: insurance, parking, maintenance, etc. Users are also attracted to car sharing because of its good environmental image [1].

Car sharing services are managed by companies like short term car rental. Car sharing organizations (CSO) usually manage a fleet of vehicles dispatched in several predetermined parking lots. Members have access to any vehicle at any time, given that they have made a reservation in advance. Car keys are usually located in a safe-deposit box located near the parking lot. Users can keep the vehicle during a fixed period of time that can vary from one hour to several days. At the end of the rent period, the vehicle has to be returned to a parking lot (usually the one where the car was taken). Traveled distances and rent durations are recorded and used for the billing of every user (fees depend on the total duration and mileage) [2].

Some studies have been conducted to estimate the potential of car sharing in urban transportation and changes in transportation behavior when a car sharing service is available [3], [4], [5] and [1]. Cervero et al. [5] state that this mode could attract users towards individual car ownership for the San Francisco area. Other authors indicate that, on contrary, CSO helps to decrease car ownership because some users will leave their individual car to enter the system. Positive impacts of car sharing on travel demand and on greenhouse gases (GHG) emissions were also reported by Fellows and Pittfield [8]. Most of these studies are based on user surveys and can hardly characterized long run use of car sharing. Barth et al. demonstrated the potentialities of using individual data collected in a multi-station shared vehicle system [10] and [2]. On a pure economic basis, CSO can be both viable and profitable [6]. Several systems are running through the world, especially in Europe (UK, Germany, Italy, Netherlands), in Canada (Toronto, Montreal, Quebec City) and in the United States (Seattle, Portland, San Francico area). In a worldwide comparative study, Enoch reports that most of the CSO are supported by the communities and the governments [7]. Support actions are related to tax incentives, starting investment, provision of parking spaces, marketing, links with transit authorities, etc.

B. Data mining tools and applications

Data mining can be defined as the process of identifying valid, novel, useful and understandable patterns from large datasets [15]. It can be used with any king of dataset from standard databases to text, audio or video streams.

When focusing on more common data sources, data mining is often practiced through four classes of tools: classification, estimation, segmentation, and description [16]:

- Classification techniques are dedicated to the marking of data with labels based on arrangements constructed using historical and contextual data.
- Estimation tools are predictive methods to evaluate missing values of a record provided a context defined by other analogous or linked records.
- Segmentation (clustering) aims at splitting a population into subsets of homogenous elements. It maximizes the similitude of each subset and the heterogeneity between the different subsets. It is a relevant way to distinguish significant or typical features of a population.
- Description and visualization tools are used to display the relationships between the data. Frequent patterns are represented as association rules which convert the raw data into understandable information. Graphical projections and representation are also used to emphasis particular properties.

III. METHODOLOGY

A. Case study

In Montreal, the car sharing system is offered by Communauto inc. Operating information are available on their website [11]. This company started his operation in 1994. It was then the first car sharing organization in North America and it is becoming more and more popular. Today, the car sharing service is also offered in Quebec City, Sherbrooke and Gatineau.

Since the beginning of the service, the number of members has increased rapidly. Six new parking lots have been implemented in the region in the last year and 2626 new members were recruited. Communauto now has about 11,000 members.

People have to pay a membership fee to finance car purchase. The fee is refundable for people who wish to quit the system. Members must choose between three types of annual membership contracts: every contract is dedicated to a particular use and proposes different annual fees and rates per traveled kilometer (Table 1).

Package	Annual fee ($)	$/km (< 100 km)	$/km (> 100 km)
A	350	0,16	0,16
B	140	0,23	0,16
C	35	0,29	0,19

Table 1. Communauto packages (Canadian dollars, March 2007)

Members have to reserve a car in advance for a predetermined period; the reservation can be done either by internet or by phone up to few minutes before the desired time, if a car is available. Users have to pick and drop the car at the same location (parking lot and space).

For long distance trips, the company offers inter-network rates that allow members to travel out of the region and to keep the car for a few days. Membership also gives access to preferential rates from commercial car rental companies for long distance trip. Monthly bills are produced and sent to the members.

B. Transaction dataset

In general, the study of travel behavior is limited by the size or temporal scale of the available datasets. In the present case, the transaction database contains all that is required to produce clients' invoices. Actually, this data is not a sample but the whole universe since it gathers all the transactions made on the system. It provides a complete review of the use of the car sharing system on a continuous basis.

Each car reservation made by phone or online, is inputted in the invoicing system. Cancellations are removed from the system. Members are billed for every reservation, even if the car was not used. In that case, the cumulated distance is null. When they do the reservation, members can choose between the various parking lots and cars available. Hence, these preferences can be observed through this selection.

Every transaction is recorded into the database. Each record contains information on a single reservation: member identification number, vehicle identification number, transaction number, time and date (beginning and end of reservation since it can cover more than one day), odometer at the beginning and end of the trip (and traveled distance, in kilometers). This transaction table can be linked to other tables containing details on members (home location, preferred parking lots, and subscription program), cars (year, model, parking id) and parking lots (capacity, location). All information remains completely anonymous and confidential to ensure privacy.

C. *Mining tools*

In this project, the following data mining techniques are used: filtering of data, clustering (with k-mean), and visualization. The k-mean divides the whole dataset into subsets so that each of them has a specific meaning and profile. It reveals the global behavior pattern of a subset of users. The comparison of such profiles can then help understand the key facts and the patterns of use of the car sharing system. Unlike traditional statistical analysis, data mining approaches not only focuses on global and aggregated phenomenon but can also extract singular behaviors and characteristics. The analysis begins with the identification of the most significant user profiles. Further characterization follows using the subset of data tagged with valuable and accurate description profiles. The dataset was extracted from a SQL Server database. Analyses were conducted with the TANAGRA freeware [17].

IV. RESULTS

A. *Key facts*

In 2004, there were 119,900 trips done using the shared cars of Communauto in Montreal. These transactions where made by a total of 4,345 members. During the year, the 258 cars available traveled a total of 6.8 millions kilometers, an average of 26,555 km per car.

A particular attention must however be paid to these average results because the car sharing activity is continuously increasing. In January 2004, there were 1,855 active members (who did at least one trip during the month) and 184 available cars. In December 2004, the number of active members was 2,551 (+37.5%) and there were 235 cars available (+27.7%). The number of monthly transactions rose from 8,015 in January to 12,350 in December.

Fig. 1 presents the stacked evolution of the number of active members during the year of 2004. These data can be used to estimate the survival of members through the months by linking them to the first month where they were observed in the system. For instance, from the 1,855 members doing at least one trip in January 2004, 1,172 are still active in December (survival rate of 63.2%). Similar figures are estimated for every month and illustrated in Fig. 2. We can see the survival rate after a given number of months, depending on the initial month of activity. It is noteworthy that the survival rates observed for the January group must be interpreted with caution since it gathers all the members that were active during this first month of observation, whether or not it was their first month of activity in the system (some users could have been active in 2003 or previously). This figure shows that the survival patterns are quite similar for the various months. The survival rate ranges between 59.1% and 74.6% after one month. It more or less stabilizes after 4 to 5 months, ranging between 45% and 60%.

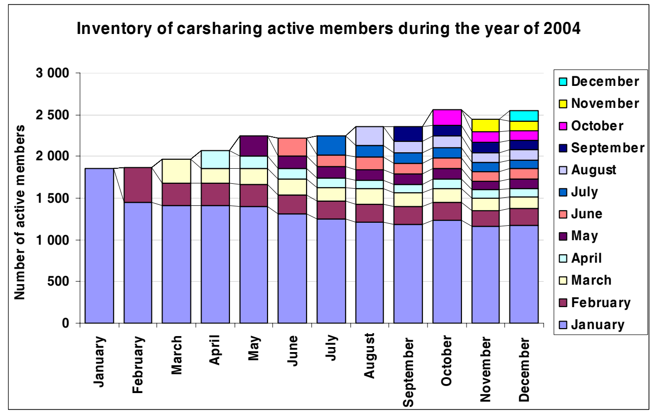

Fig. 1. Inventory of car sharing active members during the year of 2004

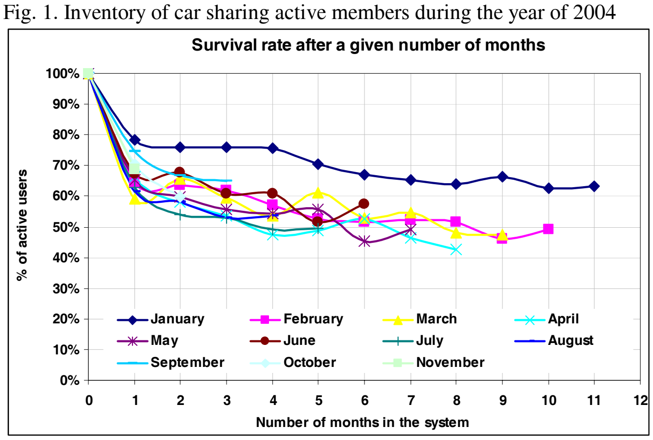

Fig. 2. Survival rates of users in the car sharing system after a given number of months

This continuously evolving state of the membership during the year led us to retain a subset of the dataset, trying to have a stable group of users for the study of the behaviours over 365 days of operation. Actually, to understand the patterns of use throughout the year, we selected the 1,170 members that were active both in January and December (the January group in Fig. 1). The patterns of use of the system are examined using this subset of data.

B. *Variability of the number of transactions per day*

The first action is to examine the average number of transactions made by each user for all day of the year. The systematic analysis of these disaggregate behaviors using data mining techniques results in the identification of two groups of users:

- C1: 8.70% of the users perform an average of 0.2 to 0.7 transactions per day; we will call them “frequent users”.
- C2: the majority of the members do around 0.1 to 0.2 transactions per day; we will call them “occasional users”.

Fig. 3 illustrates, for these two clusters, the average number of transactions per member for every day of the year (2004). We see that it varies within days, weeks and months and that a generalized decline appears during the summer months (July and August). Hence, it must be said that the variability of these transaction rates is very high for both groups with coefficient of variation (standard deviation / average) around 125% for C1 and 280% for C2.

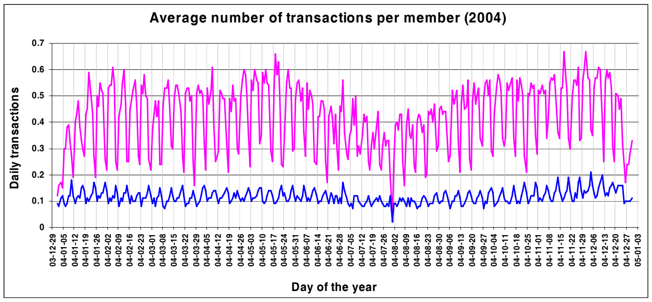

Fig. 3. Variability of the average number of transactions per member throughout the year of 2004

Fig. 4 synthesizes these same numbers by type of day. It shows that frequent users (C2) have a higher number of transactions during the weekdays. On the other side, occasional users (C1) have a slightly higher number of transactions on weekends, but it remains lower than group C2. The study on average distance traveled by transaction and week will help understand these facts.

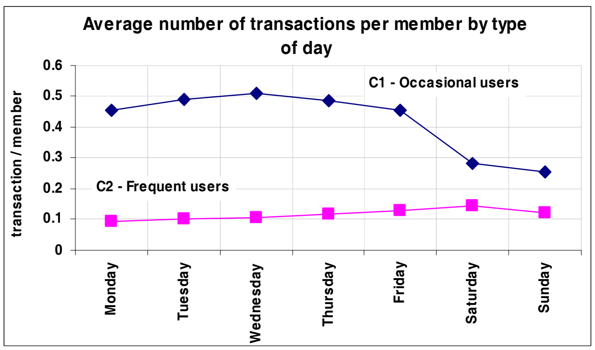

Fig. 4. Average transaction rate by type of day for the two types of users

C. *Study of traveled distance*

The same dataset is used to understand the behaviors with respect to traveled distances. Since the transaction dataset is first fit to assist the production of monthly bills, it needs to precisely measure the kilometers traveled every time a car is borrowed. Currently, every member needs to state the odometer level at the beginning and end of a trip. Data are cross-checked in order to assure coherence between successive loans. Some cars are now being equipped with GPS systems. It helps to validate the declared information and eventually will lead to the automatic saving of the spatio-temporal circumstances of every trip.

These data are processed in order to measure the distribution of traveled distances. Data mining techniques are applied to a dataset representing, for every member, the distribution of their trips in classes of 10-kilometers (see example dataset in Table 2). Because frequency of use varies a lot between members, the dataset is normalized in proportion of trips belonging to every class.

Member ID	0-10 km	10-20 km	20-30 km	30-40 km	40-50 km	50-60 km	60-70 km	70-80 km	80-90 km	90-100 km	100 km +
55	24	34	16	6	4	6	2	0	0	0	8
67	42	8	8	17	0	0	0	8	8	0	8
264	17	21	24	21	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	0
467	0	22	11	44	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	22

24% of the trips made by member 55 are between 0 and 10 km long

Table 2. Example of the dataset used to study the distribution of traveled distance per transaction

Data mining techniques help to classify the users in groups of similar behaviors with respect to the distance traveled when a car is borrowed. The process leads to the creation of multiple relevant classifications of 2 to 5 clusters depending on the level of granularity expected. The average distributions of each cluster for two of these are discussed (see Fig. 5). Table 3 presents the scale of these various clusters.

Users	**1170**				
2 Clusters	*C1*	*C2*			
	345	**825**			
	29.5%	***70.5%***			
5 Clusters	*C1*	*C2*	*C3*	*C4*	*C5*
	238	**334**	**423**	**110**	**65**
	20.3%	***28.5%***	***36.2%***	***9.4%***	***5.6%***

Table 3. Members per cluster for two classifications

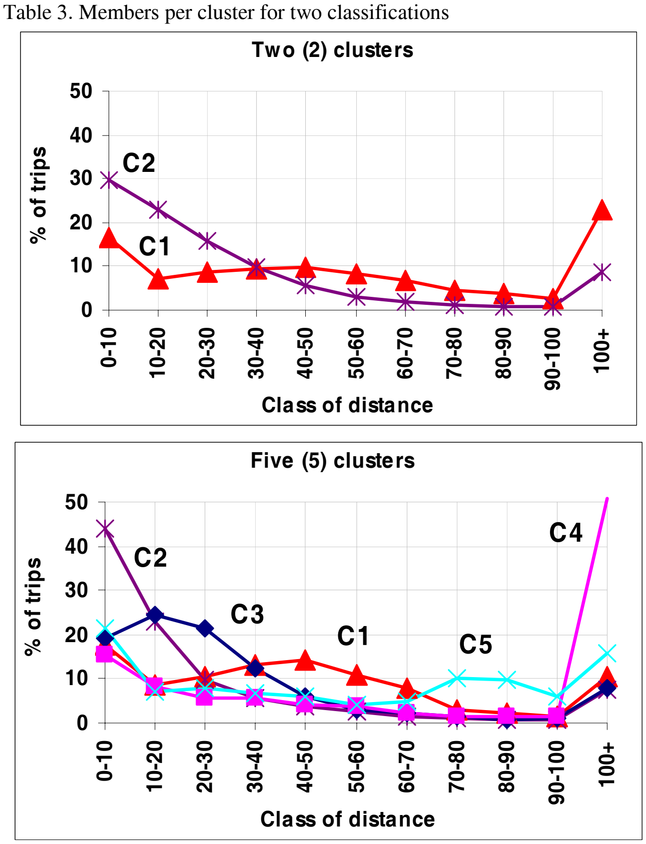

Fig. 5. Average normalized distribution of trip length for typical clusters of users for two classifications

From these, we observe that:

- The 2 clusters segmentation splits between users mainly doing short trips (C2 – more than 52% of their trips are less than 20 kilometres long) and users with almost one-quarter of their trips being very long (almost 25% of more than 100 kilometres)
- The 5 clusters segmentation first enhances the differentiation between people doing mainly short trips (C2 – more than 40% of less than 10 km trips and two-third of less than 20 km trips) and people doing mainly very long trips (C4 – 50% of more than 100 km trips). Moreover, it identifies three different types of intermediate behaviours, with more dispersed distribution but each with a small peak located around 10-20 km for C3, 40-50 km for C1 and 70-90 km for C5.

This analysis reveals that some users traveled very long distances with the shared cars. Actually, part of the fleet can be borrowed to perform intercity trips, covering very long distances, over multiple days.

D. *Temporal patterns of use*

This dataset was reformatted to allow the study of the temporal patterns of use during the year. The objective is to test whether typical patterns of use can be detected. The week is the analytical unit of observation. All the transactions made by the 1,170 users active during the year of 2004 were converted into weeks of travel. Hence, a maximum of 52 records of information are created for every user, expressing its pattern of use during every week of this same year. Every record of information indicates if a car was used each of the seven days of every week. Table 4 gives an example of the dataset that was processed using data mining techniques. Every record relates to a user-week. The resulting dataset represents 32,864 user-weeks of travel.

Member ID	Week	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
55	1	0	0	0	0	0	0	1
55	2	0	0	0	0	1	1	0
55	3	0	0	0	0	0	1	0
55	4	1	0	1	0	0	0	0
55	5	0	0	0	0	0	1	0
55	6	1	0	0	[illegible]	0	0	1
55	8	1	0	0	[illegible]	[illegible]	[illegible]	[illegible]

"1" indicates that a shared car was used Sunday of the 4th week by member 55

Table 4. Example of the dataset used to study the temporal patterns of use throughout the weeks

The classification process resulted in the creation of eight (8) significantly different clusters representing typical weeks of travel using shared cars. This decomposition was obtained in a two step strategy: the first step considers computing an important number of clusters (50), with a k-mean method; then the centroïds of each cluster are considered for hierarchical ascending classification, resulting dendogramme reveals that eight (8) clusters provide the most pertinent results. Then 8 clusters are computed with the k-mean. Fig. 6 illustrates these typical weeks that really relate to different temporal patterns of use. Those typical weeks can be synthesized as follow:

- C1 (12.9% of user-weeks): no trips on Mondays or Sundays, always on Wednesdays; other occasional trips during the other days
- C2 (15.4%): never on Mondays or Wednesday, very few often during the week-end (Friday, Saturday, Sunday) and approximately half of the times on Tuesdays and Thursdays.
- C3 (3.7%): regular trips on Wednesdays and Sundays
- C4 (19.9%): always a trip Saturday but never Monday or Wednesday
- C5 (13.7%) : always a trip on Monday and sometimes all the other days
- C6 (8.9%): mainly week trips (Monday to Thursday) by few trips on week-ends
- C7 (11.7%): always a trip on Fridays and very few often on the other days
- C8 (13.7%): a trip Sunday and very few often during the other days.

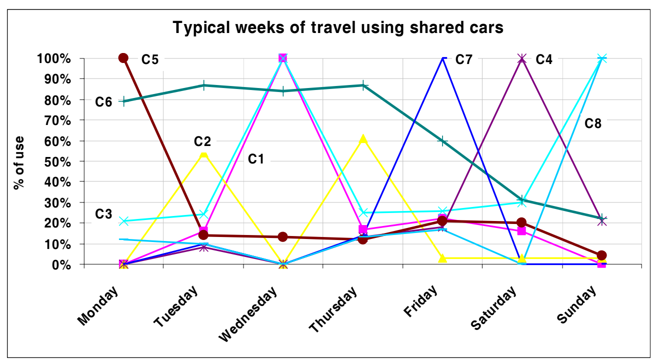

Fig. 6. Typical weeks of travel for the eight clusters of user-weeks

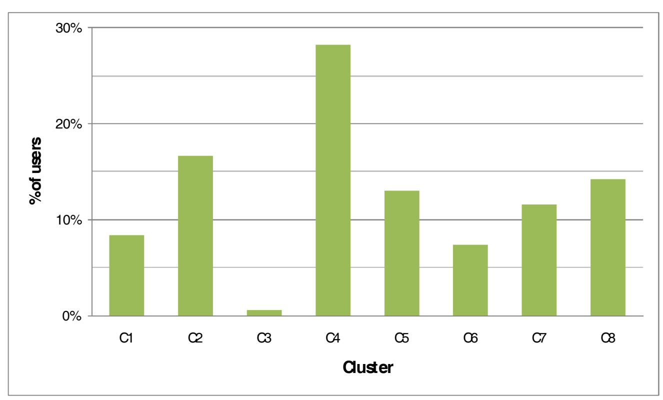

Fig. 7. Distribution of users according to their dominant cluster

Fig. 7 presents the distribution of users according to the cluster in which they belong the most often during the 52 week period. Weeks where users did not use the car sharing service are not considered. It shows that cluster C4, occasional week-end use, is very popular among members.

Table 5 presents statistics on user variability, distributed by dominant cluster, according to the following indicators:

- No. of users. The number of users for which this cluster is dominant.
- Cluster dominance. The average proportion of active weeks for which the user belongs to the dominant cluster.
- Avg. no of clusters. The average number of different clusters the users belongs to.
- % of active weeks. The average proportion of weeks where users are active in the system (over 52 week period).

We can see that cluster C6 (weekday trips) has a very strong dominance on its users, and that these users are more active than those belonging to other clusters. People using the service on Fridays (cluster C7) are less active and remain more often in the same clusters. Overall statistics show that people will use the service about half the weeks in a year and will vary its use (high average of 6,1 over 8 different clusters).

Dominant cluster	No. of users	Cluster dominance	Avg. no of clusters	% of active weeks
C1	98	36.2%	6.49	57.5%
C2	195	37.4%	6.03	48.4%
C3	7	34.7%	6.71	88.2%
C4	330	40.4%	6.06	54.0%
C5	152	35.4%	6.38	57.5%
C6	86	50.6%	6.29	77.1%
C7	136	36.7%	5.69	44.3%
C8	166	33.6%	6.19	50.0%
All	*1170*	*38.2%*	*6.13*	*54.0%*

Table 5 : Statistics on user variability

V. Conclusion and Perspectives

The focus of the study was to analyze the day to day usage of a car sharing system using a transaction dataset. It is important to understand users' behaviors for this kind of alternative transportation to better assess the possible benefits of these systems for cities environment,

The results show that two subsets of customers could be identified (frequent and occasional users). These two sets of users relate to different behaviors in terms of average number of transactions per day as well as favorite periods of use (days of the week versus week end). Results also show that users can be classified in clusters according to trip length, helping to identify short distance users from long-run (inter-network) users. Next, the variability of users' behaviors can be detected by classifying user-weeks in clusters of typical weeks. Strong patterns can be deducted, especially for weekday users and week-end only users.

In perspective, it seems that data mining techniques can be used to help the analysis of travel behaviors in car sharing systems. But some methodological issues remain to be clarified in the near future. First, it will be interesting to see at what extent the clustering techniques could help to better plan the different fare policies in car sharing systems, given that the policies can influence the use itself. Next, parking locations may also have an influence on variability and use indicators. Finally, system performance measure like car availability at reservation, refusal rate (car unavailability) and car time use could be added to the analysis.

References

[1] K. W. Steininger, C. Vogl and R. Zettl, Car-sharing organizations -- The size of the market segment and revealed change in mobility behavior, *Transport Policy*, 3 (1996), pp. 177-185

[2] M. Barth, J. Han, & M. Todd (2001). Performance evaluation of a multi-station shared vehicle system. *Paper presented at the 2001 IEEE Intelligent Transportation Systems Proceedings*, Oakland, CA

[3] S. A. Shaheen, Commuter-based carsharing: Market niche potential, Transportation Research Record (2001), pp. 178-183.

[4] A. T. H. Chin and D. S. W. Lee, Impact of the Car Sharing Scheme on mode choice, *Proceedings of the 1998 Conference on Traffic and Transportation Studies - ICTTS*, ASCE Reston VA USA, Beijing, China, 1998, pp. 508-517

[5] Cervero, Robert; Creedman, Nina; Pai, Madhav; Pohan, Muhammad, 2002. *San Francisco City CarShare: Assessment of Short-Term Travel-Behavior Impacts.* Institute of Urban & Regional Development IURD Working Paper Series (University of California, Berkeley).

[6] S. A. Shaheen and J. Wright, The Carlink II pilot program: Testing A commuter-based carsharing model, *2001 IEEE Intelligent Transportation Systems Proceedings*, Oakland, CA, 2001, pp. 1067-1072.

[7] Enoch, Marcus, 2002. *How public and private organisations support city car share clubs : A worldwide review . Integration into Urban Planning Work Package.* MOSES (Mobility Services for Urban Sustainability) Project European Commission, DG TREN. Energy and Environmental Research Unit. The Open University. Milton Keynes, U.K. 28 p.

[8] N. T. Fellows and D. E. Pitfield, Economic and operational evaluation of urban car-sharing, *Transportation Research, Part D: Transport and Environment*, 5 (2000), pp. 1-10.

[9] Barth, M., Li, W., & Todd, M. (2004). Interoperability options for shared-use vehicle systems. *Transportation Research Record (1887)*, 137-144.

[10] M. Barth and M. Todd, Intelligent transportation system architecture for a multi-station shared vehicle system, *IEEE Conference on Intelligent Transportation Systems, Proceedings - ITSC (2000)*, pp. 240-245

[11] Communauto: www.communauto.com (page visited February 20th, 2007)

[12] Barth, M., & Shaheen, S. A. (2002). Shared-use vehicle systems: Framework for classifying carsharing, station cars, and combined approaches. *Transportation Research Record (1791)*, 105-112

[13] Barth, M., Todd, M., & Shaheen, S. A. (2003). Intelligent transportation technology elements and operational methodologies for shared-use vehicle systems. *Transportation Research Record (1841)*, 99-108.

[14] Bolognin, A. (2005). Car sharing in Venice. *Public Transport International*, 54(2), 29-30.

[15] O, Maimon, L. Rokach (2005) *The data mining and knowledge discovery handbook*, Springer.

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

[17] 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.

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