Detecting Public Transport Passenger Movement Patterns

Natalia Grafeeva(✉) and Elena Mikhailova

ITMO University, St. Petersburg, Russia
nggrafeeva@itmo.ru

Abstract. In this paper, we analyze public transport passenger movement data to detect typical patterns. The initial data consists of smart card transactions made upon entering public transport, collected over the course of two weeks in Saint Petersburg, a city with a population of 5 million. As a result of the study, we detected 5 classes of typical passenger movement between home and work, with the scale of one day. Each class, in turn, was clusterized in accordance with the temporal habits of passengers. Heat maps were used to demonstrate clusterization results. The results obtained in the paper can be used to optimize the transport network of the city being studied, and the approach itself, based on clusterization algorithms and using heat maps to visualize the results, can be applied to analyze public transport passenger movement in other cities.

Keywords: Urban transit system · Public transport · Multimodal trips · Pattern mining

1 Introduction

According to statistics, in Russia more than 80% of passengers use travel cards (monthly or longer period) instead of single tickets. In Russia, each individual trip is paid separately. Currently, most public transport operators in large cities use automated fare collection systems. These systems are based on the use of contactless smart cards, which passengers use to pay their fares when boarding ground transport or entering the subway.

The primary purpose of such systems is to simplify the interaction between passengers and the operator and the fare collection process. However, such systems also offer interesting capabilities for studying typical routes of passengers traveling within the transport network using smart cards. Upon fare payment, the server receives not only the card ID and the payment amount, but also various additional information: time of payment (transaction), card type, route number, and transport type.

This data enables the detection of patterns in the movement of public transport passengers over the course of the studied time period. In turn, public transport passenger movement patterns are quite interesting to transport operators, since they can be used to analyze the productivity of the transport network, predict passenger activity on certain dates, and estimate the balance between supply and demand of transport services. Furthermore, they enable the estimation of transport accessibility of individual city districts and more timely changes to the transport network (for example, the

Á. Rocha et al. (Eds.): WorldCIST 2020, AISC 1160, pp. 555–563, 2020.
https://doi.org/10.1007/978-3-030-45691-7_52

addition of new routes to reduce the load on transport network nodes, or the reduction of frequency of rarely used routes).

In this paper, we detect and study public transport usage patterns based on the transport data of Saint Petersburg. The following problems were considered as part of the research:

- Construction of public transport passenger profiles on the scale of one day.
- Detection of temporal clusters based on the constructed profiles.
- Analysis and interpretation of results.

2 Related Work

Some of the first papers on potential uses of the data produced as a result of smart card use were [4, 5], which discuss the advantages and disadvantages of smart card use, and also offer a number of approaches to rule-based transport data analysis. The authors also pointed out that transport data does not always contain complete information (lacking, for instance, any information about the purpose of travel), so it is best to use it in conjunction with data from specialized surveys.

The paper [1] describes a study of transport habits of passengers of Shenzhen based on four consecutive weeks' smart card data. In that paper, every transaction was described using 27 variables: card number, date, day of the week (D_1 through D_7) and 24 H_i variables, where $H_i = 1$ indicated that the card was used at least once during the i-th hour of the day. Then the authors used the k-means algorithm to clusterize the transactions, thus obtaining generalized (mean) profiles of public transport passengers.

A similar study based on similar data is described in another paper by the same authors [2], but in that paper the trips are combined into week-long transactions ($H_{i,j} = 1$ if the card was used at least once on during the j-th period of the i-th day), after which the data is passed through clusterization algorithms. This approach allowed for the detection of certain deviations in transport behavior, such as a reduction in the use of student cards during school holidays.

The methods presented in [2] were extended and improved in [13]. There, the authors discuss the advantages and disadvantages of profile clusterization methods based on Euclidean distances, and also present their own perspective on clusterization, using topic models and latent Dirichlet allocation, interpreting passenger profiles as collections of words (for example, a passenger who made a trip at 10 am on a Friday was associated with the word "Friday 10 am").

Based on the same data (and partly by the same authors as in [13]), another study was conducted [7]. Unlike [13], the paper rejects hourly aggregation of trips and criticizes it for inaccuracy. For instance, a passenger who starts their trips at 8:55 am and at 9:05 am on different days falls into different groups in case of hourly aggregation, which can make results less accurate that they actually are. The paper considers an approach based on continuous presentation of time and a Gaussian mixed model.

The paper [12] proposes and confirms that transport use patterns radically differ for different age groups, including different rush hours, distances, and destinations.

One of the key features of transport flow analysis is detecting passenger correspondences using smart cards. A correspondence is a sequence of trips with brief transfers between different transport routes. This is the problem studied in [6]. In that paper, the authors proposed two very important hypotheses which significantly simplify the construction of correspondences: most passengers begin a new trip near the station where the previous one ended; for most passengers, the destination of the last trip matches the starting point of the day's first trip.

Later, these assumptions were expanded in [16]. The authors showed that, for sequential trips with brief transfers, in the vast majority of cases the following is true: passengers don't use additional vehicles (bicycles, cars, etc.) between consecutive trips and don't take long walks between stations. Furthermore, the maximum distance between transfer stations is around 400–800 meters [16].

Constructing user behavior patterns requires identifying correspondences for which main passenger locations (home, work) can be determined. Such correspondences are sometimes called regular correspondences. For identifying regular correspondences, it is common to use so-called origin-destination matrices [15] or a statistical rule-based approach. In [8], the authors suggest to consider long breaks between consecutive trips (over a predetermined number of hours) as work activity. The home location was determined by the destination of the last trip of the day. At the same time, in [10], a similar model is proposed to estimate the frequency with which the user visits certain places. Thus, the most common stop was considered to be the home, and the second most visited, the workplace. In [3] and [11], an attempt is made to join the aforementioned approaches and combine frequency estimation with time limits. The methods in [3] were tested in London, where model accuracy reached 82% (compared to 59% for the model from [10] with the same data).

3 Dataset

As initial data for this paper, we used the correspondences of public transport passengers in Saint Petersburg. The technology for acquiring such correspondences is described in [9]. This technology is based on ideas described in [6, 16]. The total number of correspondences is about 25 million. To construct correspondences, we used trip data over a two-week period that did not contain public holidays. The amount of analyzed smart cards is over 2,250,000. Each smart card has a unique identifier, which cannot be used to determine the card owner's personal information, such as their name, sex, address, or phone number. Each correspondence is described with the following parameters:

- Card number.
- The correspondence's index (within the current day).
- The starting time of the correspondence.
- The ending time of the correspondence.
- Smart card type and name.
- Starting station ID.
- Ending station ID.

- Starting station coordinates.
- Ending station coordinates.

4 Methodology

The methodology we use to determine passenger profiles consists of the following steps:

- Primary filtering (getting rid of noise and insufficiently representative data).
- Determining home and work locations (for each passenger).
- Pattern detection (up to a day).
- Determining temporal profiles (for each pattern).
- Profile clusterization.
- Interpreting results.

Below is a more detailed description of each of the stages.

4.1 Primary Filtering

The initial dataset turned out to be rather raw (noisy), so, to simplify the study, we had to filter it and eliminate the noise. In the context of a study of the transport network, we were mostly interested in regular passengers with a sufficiently high number of trips. For this reason, we decided to exclude from consideration passengers who, over the course of the study period (14 days), used public transport on less than 7 different days and on less than 3 days in each of the considered weeks.

4.2 Determining Home and Work Locations

In this paper, we are interested in patterns of regular correspondences, in other words, of correspondences whose starting and ending locations correspond to home and work. To do that, it was necessary to determine, for each owner of a smart card, an approximate location of their home and workplace. Following the assumptions in [6] and [16], we considered only those days on which the starting point of the day's first correspondence matches the ending point of the last correspondence or is in walking distance from it, i.e. within 500 m, and the end of the first correspondence is in walking distance from the start of the last correspondence. To clarify, the days when a passenger made only one trip were excluded from consideration.

Since our data covers a period of 14 days, we used the threshold value of 7 to determine regular correspondences. That is, if such correspondences repeated at least 7 times for the same passenger, we declared them to be regular. The start of the first correspondence was declared to be the home, and the start of the last correspondence was declared to be the workplace. Non-regular correspondences were excluded from the dataset being analyzed. As a result, the dataset contained only passengers with a defined workplace location, which they visited at least 3 times a week and at least 7 times over the study period.

4.3 Pattern Detection

Regular correspondences were further used to determine each passenger's work schedules (for example, 5 working days and 2 days off; 3 working days and 2 days off; etc.) For that purpose, we encoded passenger behavior as sequences of digits 1 and 0, which indicated working days and days off, respectively. A working day was determined by the presence of a regular correspondence from home to work and back, and a day off, by the absence of one. We call such a character sequence a schedule. In turn, a pattern is defined as the shortest repeating substring that satisfies the following requirements:

- the pattern always begins with 1;
- the pattern always end with 0, except when it consists of only one letter.

Table 1. Examples of schedules and their respective patterns.

Schedule	Pattern
11110001111000	1111000
11111001111100	11111001111100
11001100110011	1100
10101010101010	10

Table 1 shows examples of schedules and their respective patterns. When processing schedules it is important to keep in mind that the data may not be entirely accurate, since on one day a passenger might, due to unforeseen circumstances, skip work or take a taxi. Thus, a weakened assumption was made that a pattern is the shortest repeated substring that matches the schedule with up to one error. Levenshtein distance was used to determine errors. Identifying patterns allowed us to qualitatively group passengers for later determination of temporal profiles and more detailed study. The most numerous groups of passengers are shown in Fig. 1.

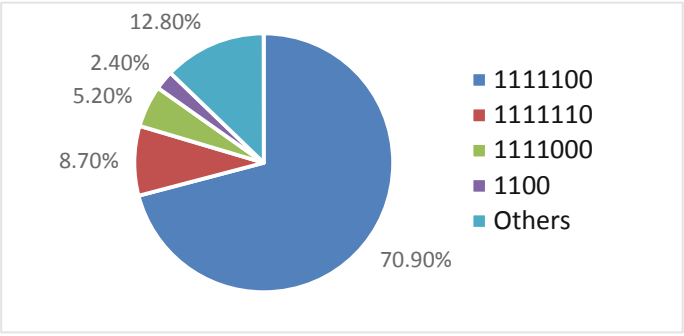

Fig. 1. Distribution of St. Petersburg passengers by pattern.

4.4 Determining Temporal Profiles

After grouping passengers by pattern, we ask quite a reasonable question: what distinguishing features does each group have and how exactly do they use public transport? At this stage of the study, the goal is to detect subgroups within groups based on the temporal habits of the group's passengers. For that, we collected all regular correspondences of each passenger into a single profile, describing the distribution of their trips by each hour (0 to 23) of each day of the group's pattern. In other words, each profile P is represented as a multidimensional vector whose elements describe the total number of correspondences made by the passenger on the first day of the pattern from midnight to 1 am, then from 1 am to 2 am, and so on. Note that if the number of days comprising the pattern is D, then the resulting vector's length is $24 \times D$. Such a multidimensional vector can be illustrated using a heat map (Fig. 2).

day / hour	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1								1	1		1					1			1					
2									2							1			1	1				
3									2							2								
4									2							2								
5									2							1	1							
6																								
7																								

Fig. 2. Example of a temporal profile.

4.5 Profile Clusterization

Then we clusterize the profiles. Profiles were interpreted as points in a multidimensional space, and the distance between them was calculated as Euclidean distance in a multidimensional space. In [1, 2, 14] it was mentioned that the k-means algorithm is well suited for clusterization of passenger profiles. That is the algorithm that we used in our study.

4.6 Interpreting Results

To demonstrate clusterization results and their interpretation, we used heat maps. Each cluster was used to produce its own heat map. The map was divided into cells corresponding to time intervals and days of the week. The larger was the percentage of passengers who used the transport network during a particular day of the week, the darker is the color of the corresponding cell. As an example, we provide the heat map visualization of the most numerous (and the most predictable) group, which corresponds to the pattern 1111100.

This group can be split into 7 temporal clusters, as shown in Fig. 3. The largest clusters, 1 and 2 (22.69% and 18.19%), describe passengers working from 8 am to

Cluster 1: 22.69%

D \ H	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1																								
2																								
3																								
4																								
5																								
6																								
7																								

Cluster 2: 18.19%

D \ H	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1																								
2																								
3																								
4																								
5																								
6																								
7																								

Cluster 3: 16.53%

D \ H	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1																								
2																								
3																								
4																								
5																								
6																								
7																								

Cluster 4: 15.70%

D \ H	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1																								
2																								
3																								
4																								
5																								
6																								
7																								

Cluster 5: 9.46%

D \ H	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1																								
2																								
3																								
4																								
5																								
6																								
7																								

Cluster 6: 10.33%

D \ H	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1																								
2																								
3																								
4																								
5																								
6																								
7																								

Cluster 7: 7.11%

D \ H	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1																								
2																								
3																								
4																								
5																								
6																								
7																								

Fig. 3. The visualization of the temporal clusters of the group corresponding to pattern 1111100.

5 pm or 6 pm, respectively; clusters 3 and 6 depict the behavior of passengers who go to work from 7 am to 5 pm or 6 pm; cluster 4 contains passengers who go to work later than usual (to 9 am); cluster 7 consists of passengers who start working earlier than usual (6 am); and, finally, cluster 5 has no defined time of either the beginning or the end of work. It is easy to see that, in general, this group's passengers work standard eight-hour days (most likely, with a lunch break). An interesting fact is the "boot" that is clearly observable in most clusters around the end of the work day on the fifth day out of seven: it perfectly matches the common understanding that on Fridays people prefer to leave work earlier. One of the reasons for this is the shortened working day on Fridays in most public institutions.

5 Conclusion

Smart cards provide curious capabilities for studying the transport network. In this paper, we processed and analyzed a huge array of transport data over a two-week period, allowing us to determine and qualitatively interpret groups of passengers with similar habits of public transport use. We hope that the results we obtained will influence the optimization of the transport network of the city we studied, and the approach itself, which is based on clusterization algorithms and visualizing clusterization results using heat maps, will find its use in analyzing the movement of public transport passengers in other cities.

References

1. Agard, B., Morency, C., Trepanier, M.: Analysing the variability of transit users behaviour with smart card data. In: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, pp. 44–49 (2006)
2. Agard, B., Morency, C., Trepanier, M.: Mining public transport user behavior from smart card data. IFAC Proc. Vol. **39**, 399–404 (2006)
3. Aslam, N., Cheng, T., Cheshire, J.: A high-precision heuristic model to detect home and work locations from smart card data. Geo-spatial Inf. Sci. **22**(11), 1–11 (2018)
4. Bagchi, M., White, P.R.: The potential of public transport smart card data. Transp. Policy **12**, 464–474 (2005)
5. Bagchi, M., White, P.R.: What role for smart-card data from bus systems? Municipal Eng. **157**, 39–47 (2004)
6. Barry, J., et al.: Origin and destination estimation in New York city with automated fare system data. Transp. Res. Rec. **1817**, 183–187 (2002)
7. Briand, A.-S., et al.: A mixture model clustering approach for temporal passenger pattern characterization in public transport. Int. J. Data Sci. Anal. **1**, 37–50 (2015)
8. Devillaine, F., Munizaga, M., Tr´epanier, M.: Detection of activities of public transport users by analyzing smart card data. Transp. Res. Rec. **2276**, 48–55 (2012)
9. Graveefa, N., Mikhailova, E., Tretyakov, I.: Traffic Analysis Based on St. Petersburg Public Transport. In: 17th International Multidisciplinary Scientific GeoConference: Informatics, Geoinformatics and Remote Sensing, vol. 17, no. 21, pp. 509–516 (2017)
10. Hasan, S., et al.: Spatiotemporal patterns of urban human mobility. J. Stat. Phys. **151**, 1–15 (2012)

11. Huang, J., et al.: Job-worker spatial dynamics in Beijing: insights from smart card data. Cities **86**, 83–93 (2019)
12. Huang, X., Tan, J.: Understanding spatio-temporal mobility patterns for seniors, child/student and adult using smart card data. In: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XL-1, pp. 167-172 (2014)
13. Mahrsi, M.E., et al.: Understanding passenger patterns in public transit through smart card and socioeconomic data: a case study in Rennes, France. In: The 3rd International Workshop on Urban Computing, New York (2014)
14. Bouman, P., van der Hurk, E., Li, T., Vervest, P., Kroon, L.: Detecting activity patterns from smart card data. In: 25th Benelux Conference on Artificial Intelligence (2013)
15. Zhao, J., Rahbee, A., Wilson, N.H.M.: Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. Comput. Aided Civil Infrastruct. Eng. **22**, 376–387 (2007)
16. Zhao, J., et al.: Spatio-temporal analysis of passenger travel patterns in massive smart card data. IEEE Trans. Intell. Transp. Syst. **18**(11), 3135–3146 (2017)