Data Mining on the Use of Railway Stations

Dmitry Namiot, Oleg Pokusaev and Vasily Kupriyanovsky

Faculty of Computational Mathematics and Lomonosov Moscow State University
GSP-1, 1-52, Leninskiye Gory, Moscow, 119991, Russia
and Center of digital high-speed transport systems
Russian University of Transport (MIIT)
Obraszova 9, bld. 9, 127994, Russia
and National Competence Center for Digital Economy
Lomonosov Moscow State University
GSP-1, 1-52, Leninskiye Gory, Moscow, 119991, Russia

Abstract. This article deals with the processing of data on the entrances and exits of passengers for railway stations in Moscow and the suburbs. Smart transport cards are used worldwide in transport applications as a payment tool. So, for railways (cities) its usage creates the big and constantly updated collections of transactions data from cards validation equipment. The deployment model for railways in Moscow region allows us to know exactly the starting and ending points of the each route. This detailed information allows us to obtain generalized information on the modes (models) of the actual use of the railway transport. The detected travel patterns could be mapped to the model of the social and economic behavior of residents of the capital region. And vice versa, we can use known artifacts of the behavior of the inhabitants of the region as the search patterns for transport data.The conclusion that mobility is one of the main characteristics and one of the key components of a smart city is a well-known fact.

Keywords: urban railways, smart card, transport cards, data mining, mobility, smart city.

1 Introduction

This paper deals with the processing of data on the check-in (entrances) and check-out (exits) of passengers for railway stations in Moscow and the suburbs. Within the framework of the existing model of moving around suburban and urban railways, each passenger presents (validates) his travel document twice: at the entrance to the railway station before the trip (check-in) and at the exit from the railway station after the end of the trip (check-out). This feature, together with the unique identity of the travel document, allows us to accurately know the starting and ending points of the route. Accordingly, it becomes possible to analyze this displacements data in order to obtain the generalized information on the modes (models) of the actual use of the railway communication. It seems to us that the information obtained during this analysis can be useful for railways

for assessing their activities and planning changes, and for city services to assess the ongoing changes in the urban environment and planning future changes.

The aim of the work is to search for patterns of travel of railway passengers and map these patterns to the model of the social and economic behavior of residents of the capital region. It is also possible to consider the inverse problem - the search (the confirmation) of known artifacts of the behavior of the inhabitants of the region in the data on the movements.

As we pointed out above, in the existing system, a railway ticket (travel document) is presented twice - at the entrance and at the exit. Accordingly, for each trip, we know the station where the ticket was used at the entrance and the station where the ticket was used at the exit. In terms of social networks, we know the pair - check-in and check-out. It means, by the way, a profitable difference, for example, from information on the validation of travel documents in the metro or buses. There (in Moscow) we have only information about the entrance. Accordingly, to obtain information about the starting and ending points of the trip, it is necessary to use any heuristic algorithm [1]. For example, let us describe from the point of view of the card validation system (e.g. the Troika card - one of the main travel cards in Moscow) a typical person's trip from home to work and back using the metro. In the morning, the passenger uses his card to enter the metro. Thus, the starting point of the route, tied to the card, appears. Further, having reached a workplace, the person is there until the evening, after which the card is validated again. This gap (a time interval between trips) lets us make a conclusion about the final point of the route: it is a first check-in station after the gap. Naturally, this will be approximate data. For data collected at railway stations, as indicated above, there is accurate information about the starting and ending points of travel. Thus, these data become in some way complete, they contain all information about trips. Note that travel tickets (disposable or reusable) are anonymous, and there is no information on the passengers themselves.

To date, detailed information on the passengers' activity at railway stations is practically not used. Perhaps, one of the few applications that could be mentioned here is the calculation of the total numbers of passengers by stations. It seems to us that the above-mentioned detailed data contain much more valuable information that reflects not only the patterns of the use of railway transport in the city but also allows us to identify some other artifacts of city life.

If we can explain the relationship of information obtained on the basis of registration data with some patterns of behavior in the city, then tracking the changes in the data on the stations (which is technically feasible on the part of the railway, for example) will allow us to determine (or predict) some changes in the life of the city. In other words, changes that can be identified in the process of constant monitoring of registration data will indicate a change in the processes in the city. Accordingly, data on passages can be used to track changes in patterns of behavior of urban residents (passengers). And vice versa, understanding the relationships will allow us to predict how changes in the city will affect the use of the railway.

The rest of the paper is organized as follows, In Section 2, we describe related works. In Section 3, we describe railway data processing and discovered links with urban life.

2 Related works

Modeling urban behavior by mining geo-tagged data is a popular topic for research [2, 3]. In the first place, social network data is used to assess behavior. The check-in conception has been introduced by social networks. Technically, check-in data reveals information who spends time where and when. Also, they could be used for detecting types of activities. Obtained data can be used to describe city regions in terms of activity that takes place therein. And the next natural question is how to distinguish one region from another via the types of activity. The mathematical tools used here are mainly related to the construction of clusters based on probabilistic models. For example, a group of users who are more likely to be in the next moment in a given place or will engage in a certain type of activity [4].

For transport data, the models should look slightly different. For example, for urban railways, the routes of passengers are precisely known. Activities, naturally, are also limited. The paper [1], contains a review of smart cards data mining in Smart Cities. The typical tasks are traffic patterns detection, trips generation, and routes-based studies. In our case, traffic patterns (transit patterns) tasks are not applicable, because we deal again with the fixed railroad's routes only. As per trips generation - these problems can be partially interesting because we can use the same mathematical tools.

The paper [5] provides a rich overview of transport-related studies target behavior extraction. While the main task for most of the transport studies is still getting origin and destination pair (what is not an issue in our case), this paper enumerates also other interesting models. E.g., it is a detection that movement flow structure is polycentric; detection of power law flow distribution and negative binomial law distribution of rides; spatial and temporal pattern mining. The Hillinger coefficient could be used to measure the similarity of temporal patterns of human mobility between each pair of days and provide a base for variability analysis. As per provided studies, intra-urban trips have peak hours over a day, are different between weekday and weekend (which is almost obvious), and have a periodicity (which is not so obvious).

In our calculations, we've used the following definition for the Hillinger coefficient. Let $p_i(x)$ be distribution of probability density function ($i= 1,2, \ldots ,N$). The Hillinger coefficient among these variables is:

$$R = \sum_{x} (\prod_{i=1}^{N} p_i(i))^{1/N} \tag{1}$$

The value of R is between 0 and 1. The larger the R, the more related the probability density functions [6].

In general, for our kind of research, time-dependent analysis of urban movement patterns [7] looks a bit more suitable. E.g., paper [8] describes the temporally-based regularity of commuting measurement. The temporal patterns could be detected by the similarity of departure time and the number of traveling days.

3 Railway data processing

Data for analysis for 2016/2017 years on suburban and city stations of the Moscow region was provided by the Center for Digital High-Speed Transportation Systems of the Russian University of Transport. The data files contain the following information for each pass (input or output):

- date and time,
- type of the event: entrance or exit,
- a current station,
- a station where the passenger arrived (if this is an exit),
- type of the tariff (price characteristic): full or preferential,
- type of the ticket: one-time one-way ticket, one-time round-trip ticket, subscription (reusable ticket, travel card)

What was included in the first phase of our research? Firstly, these are the usage patterns of the stations. We can assume that there are differences in how passengers use railway stations. Moscow city (more precisely, workplaces in Moscow) is the center of attraction, and accordingly, we can expect that for the suburban stations (outside the city boundary in Figure 1) there will be a peak at the entrance of passengers in the morning hours (Figure 2).

These peaks at the entrance (see 1 in Figure 3) through the time t spent on the road should pour into the peaks at the exits from the stations placed in the city (see 2 in Figure 3).

In general, the check-ins between the two peak hours, as well as after the second peak hour, corresponds probably, to the non-obligatory activities. This is the simplest and most obvious pattern. By attenuation of the peaks to the exits, we can determine at which stations those who come to the city leave the railway and transfer to other modes of transport. Of course, for each direction of railroad, these stations will be different.

Another possible direction for future research is the analysis of this damping. At least, the primary results show that it is not absolutely stable for the chosen direction. Passengers from time to time change their habits of leaving the train in the morning. This lasts 1-2 days (not for all directions), after which everything returns to the basic scheme.

An additional finding in this connection: the detection of peaks on the exits at the stations where the geo-information system does not show connections with other modes of transport. Why do the passengers go to this station? A possible explanation is the presence of some point of attraction (business center, shopping center, etc.)

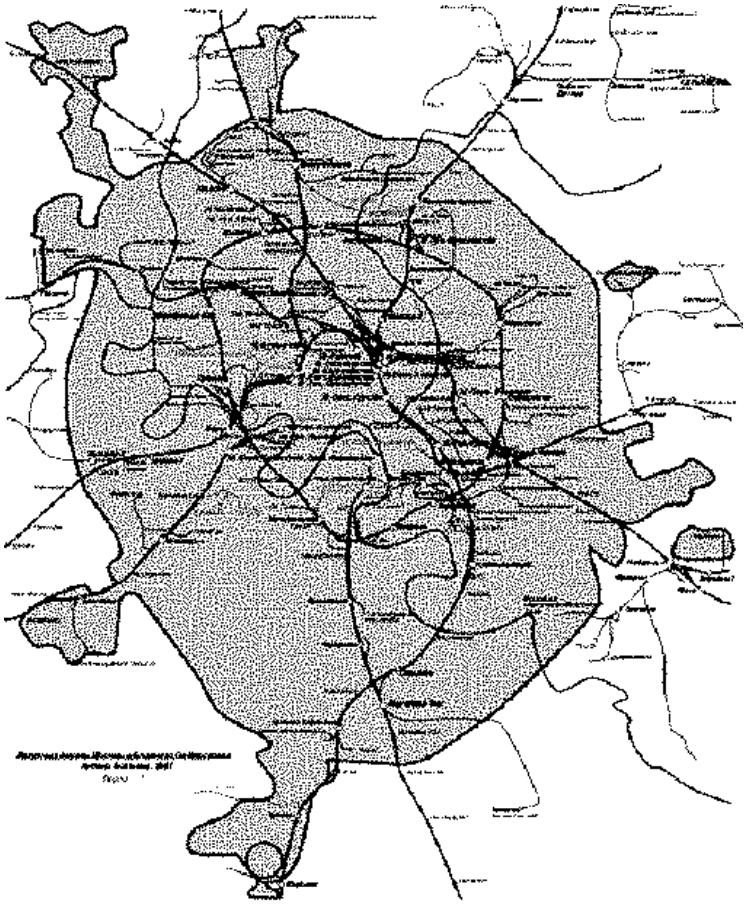

Fig. 1. Railways of the Moscow region.

According to this simple model, we should see the opposite picture in the evening. The peaks at the entrances to the "internal" (urban) stations and the peaks on the exits (with a time gap) at the "external" stations. Findings that were made here: the picture is not symmetrical. Passengers do not necessarily leave from the stations they came to (of course, we operate only with quantitative differences - there is no information on passengers). Probably, we can propose a natural explanation for this. There is some mobility upon completion of work. Outgoing traffic (large peaks) is tied to stations where there are connections to other modes of transport (where transport accessibility is better).
It corresponds, for example, with results presented in [6] for a metro. As per authors, for each metro station, the temporal trip patterns are influenced by the land uses around. For example, the homogeneity and high density of land uses around a metro station will result in obvious morning and afternoon peak hours

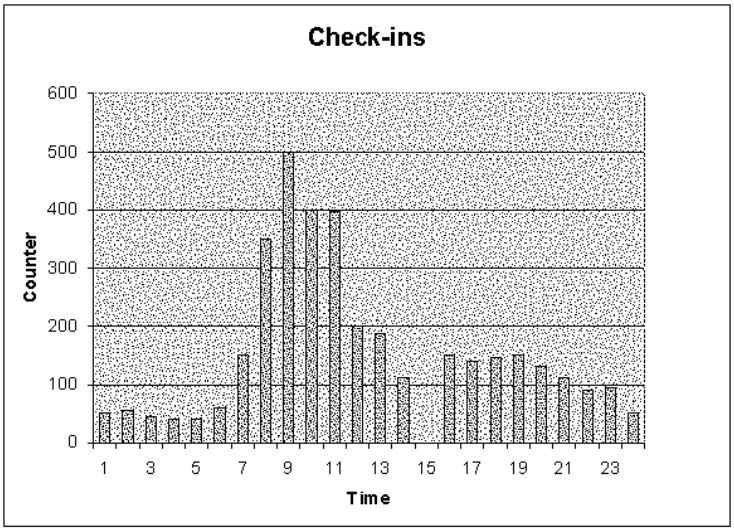

Fig. 2. Morning peak traffic (entrances to the station), associated with the working schedule.

in metro transportation. So there are some patterns over time by a station which can be mainly characterized by check-in and check-out during peak hours and working hours.
Hypotheses that require verification: asymmetry in traffic is greater in the warm season (higher mobility) and on Friday (for the same reasons).

The analysis shows the presence of stations without pronounced peaks in the morning and evening hours. At this moment, the reasonable explanation is the conclusion that the stations are not connected with work traffic. For example, for stations outside the city, this is more typical for holiday villages. In the city, it is typical for stations in the former industrial zones (where mass housing construction is only being developed). Another possible explanation for the absence of peaks is the linking of the station to some large transport (interchange) nodes, where there is always a large passenger traffic (so, working migration does not add anything significant to the constantly existing traffic).

There were no cases of the presence of one peak (regular) at the entrance and / or exit in the examined dataset. It should be noted that there is no reasonable "urban" explanation for such a hypothetical situation.

Another point related to a traffic, is the confirmation (or refutation) of the above-mentioned found move pattern on the data for the day off (e.g., Sunday, Saturday). Obviously, for stations with predominantly "working" traffic, we should see the absence of peaks in the morning and evening hours on weekends. Single outbursts are possible and connected, most likely, with some mass events held over the weekend.

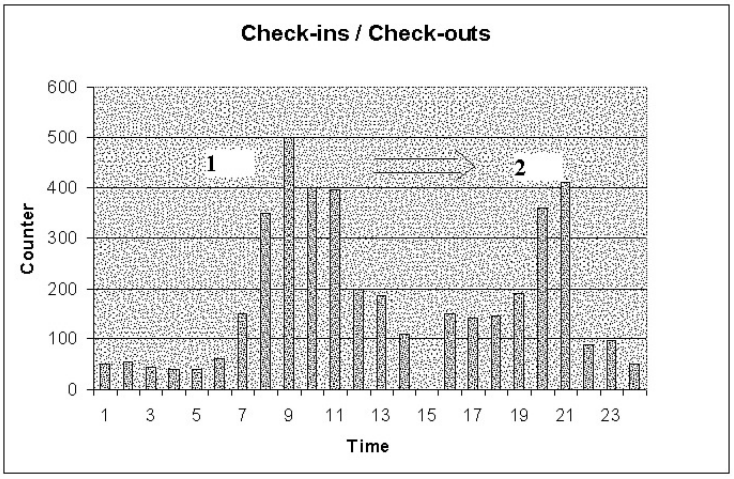

Fig. 3. Work traffic pattern.

As per classification of stations by traffic patterns, we can follow the model presented in [6]. The standard score indicates how many standard deviations the volume of the metro station is above or below the mean. It means that we can use the mean of standardized volume in two peak hours and working hour as a metric. It will explore the characteristics of railway transportation by station. Check-in (check-out) in two peak hours and working hour could be compared to the mean value of standardized volume by stations. There are three possible situations: the volume is below the mean, about the same as mean, and above the mean. The figures could be calculated, for example, for each 2 hours interval.

The next moment, which was investigated in the work - is the ratio of one-time tickets versus reusable tickets (travel cards). The idea of this comparison is based on the following fact. Reusable tickets (travel cards) are cheaper. Accordingly, those who travel constantly, will most likely use them. Therefore, a greater percentage of travel cards corresponds to more constant (robust) traffic. Tickets are bought before the trip, accordingly, it would be interesting to compare the ratio of one-time and reusable tickets at the entrance to the stations. Accordingly, stations with deviations from average ratios of ticket types were identified. So far, the explanation that is being considered here is the availability of interchange transport nodes near such stations, from where the "random" passengers for the railway arrived.

One-time tickets are of two types - one way and round-trip. The next possible step is to analyze the ratio of such tickets. Also interesting is the question of changing the ratio of one-time tickets and travel cards on the days of the week (first of all - comparing working days and days off). The increase in the number of one-time tickets at the weekends shows that these days, the railway is really "acquiring" new passengers who do not use the railroad for a week.

In addition to the above-mentioned Hillinger coefficient, we also used methods of analyzing the similarity of time series. As per [9], the measures for time series similarity could be categorized as lock-step, elastic, threshold-based, and patterns-based measures.

So-called lock-step measure in this classification is well-known Euclidean distance. It is defined as the square root of the sum of the squared differences between corresponding data points in two time series data. As it is mentioned in the all statistical papers, the main problem of the Euclidean distance is the need to have the same length for time series. It is not a problem in our case, because we can, for example, divide the day into five-minute intervals and thus construct the same length-of-time time series for the number of inputs and outputs.

So-called elastic measures use dynamic programming to align sequences with different lengths. The typical example is so-called DTW [10]. Threshold system assumes that we have a user-provided threshold T and converts sequence data to so-called threshold crossing. They are treated as points in two-dimensional space composed only of data points above the introduced threshold T. Pattern-based measures first find some representative matching segments (called local patterns), in a time series by focusing on amplitude and trajectories (up or down). Actually, this approach takes into account such factors as the number of local patterns, gap bound, time shifting factor, amplitude shifting factor, time scale factor, and amplitude scale factor [9].

In our work, we've successfully used a shape-based similarity measure, so-called Angular Metric for Shape Similarity (AMSS) [11]. This approach treats a time series as a vector sequence and focus on the shape of the data and compares data shapes by employing a variant of cosine similarity. It is illustrated in Fig. 4.

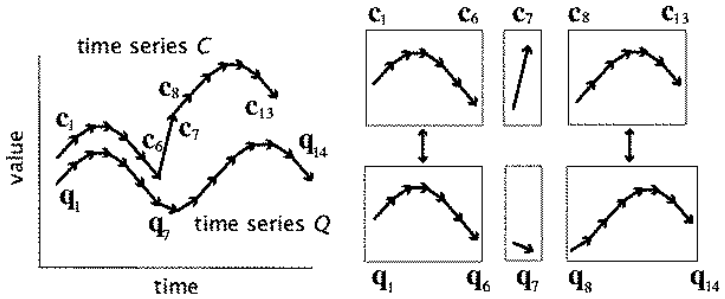

Fig. 4. AMSS [11].

4 Conclusion

The paper analyzes data on the entrances and exits of passengers of railway stations in the Moscow region. The main tasks that were considered in this work were the building of mapping models of the results of processing check-in/check-out data on the socio-economic aspects of the life of the inhabitants of the region. In the article, methods of detection (extraction) of usage patterns of railway stations linked with work traffic are considered. We classified the railway stations according to the received usage patterns. Also, an approach is proposed for assessing how changes in the city (for example, the construction of former industrial zones) will be reflected (respectively, can be tracked) in the modes of use of railway stations. Analyzing similarity distributions and methods for measuring the similarity for time series were used as analysis tools. The results of the work are of practical use in the development of the system of urban railways in Moscow.

References

1. Namiot, Dmitry, and Manfred Sneps-Sneppe. "A Survey of Smart Cards Data Mining" http://ceur-ws.org/Vol-1975/paper33.pdf Retrieved: Apr, 2018
2. Zhang, Chao, et al. "Gmove: Group-level mobility modeling using geo-tagged social media." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp.1305-1314.
3. Namiot, Dmitry, and Elena Zubareva. "Data-driven Cities." International Journal of Open Information Technologies 4.12 (2016): 79-85.
4. Eelikten, Emre, Graud Le Falher, and Michael Mathioudakis. "Modeling urban behavior by mining geotagged social data." IEEE Transactions on Big Data 3.2 (2017): 220-233.
5. Yue, Yang, et al. "Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies." Travel Behaviour and Society 1.2 (2014): 69-78.
6. Gong, Yongxi, et al. "Exploring spatiotemporal characteristics of intra-urban trips using metro smartcard records." Geoinformatics (GEOINFORMATICS), 2012 20th International Conference on. IEEE, 2012, pp.1-7.
7. Yue, Yang, et al. "Mining time-dependent attractive areas and movement patterns from taxi trajectory data." Geoinformatics, 2009 17th International Conference on. IEEE, 2009, pp.1-6.
8. Ma, Xiaolei, et al. "Understanding commuting patterns using transit smart card data." Journal of Transport Geography 58 (2017): 135-145.
9. Ding, Hui, et al. "Querying and mining of time series data: experimental comparison of representations and distance measures." Proceedings of the VLDB Endowment 1.2 (2008): 1542-1552.
10. Namiot, Dmitry. "Time Series Databases." DAMDID/RCDL. 2015. pp.132-137.
11. Nakamura, Tetsuya, et al. "A shape-based similarity measure for time series data with ensemble learning." Pattern Analysis and Applications 16.4 (2013): 535-548.