Passenger Traffic Forecasting for New Urban Railways in Moscow

Dmitry Namiot
Faculty of Computational Mathematics and Cybernetics
Lomonosov Moscow State University; Russian University of Transport
Moscow, Russia
dnamiot@gmail.com

Oleg Pokusaev[1], Alexander Chekmarev[2]
Center for High-Speed Digital Transport
Russian University of Transport
Moscow, Russia
o.pokusaev@rut.digital, a.chekmarev@rut.digital

Abstract. **This article is devoted to the description of the forecasting model of passenger traffic of new urban rail lines. It is a question of new through railway routes, which will be organized in Moscow. They will pass through the city, connecting together the stations located in the suburbs. The schedule of such trains, comfort, and pricing policy will undoubtedly make them a popular transportation tool that will compete with the Moscow metro. This popularity will result in an increase in passenger traffic on existing fragments of these railway lines, the assessment of which is the task considered in this article. The article looks at the existing methods of measuring passenger traffic, and suggests a model for assessing its possible change.**

Key words: **time series; traffic forecast; digital urbanism.**

I. INTRODUCTION

In this paper, we discuss the problems of forecasting of passenger traffic for new lines of urban railways in Moscow. In 2017, Moscow announced plans [1] to create links between the existing radial rail routes within the city, it is so-called Moscow railways diameters (Fig. 1).

Fig. 1. The planned radial rail routes in Moscow

New trains are launched on existing lines of the railway and only the connectivity of the system is changed. There is a through motion, which was absent earlier. This through movement connects two branches of the railway that existed before. Before the introduction of such a single scheme of traffic, various branches of the railway were connected by means of public transport already in Moscow. In addition to increasing the connectivity of the system, the new lines will increase the frequency of movement. Also, high-quality trains will be used here, to which a convenient form of payment will be added (a combined smart transport card with city passenger transport). In fact, it looks like a new metro (the main type of public transport in Moscow). Naturally, one can expect the growth of passenger traffic for the railway lines, which formed the basis for new through-routes. The evaluation (estimation) of this increase is the task considered in this paper. The need for such an assessment is caused both by economic considerations (it is necessary to understand the economics of the new project) and technical considerations; it is necessary to answer the question whether the existing infrastructure (stations, transitions for urban public transport systems, etc.) will be sufficient to serve a new flow of passengers.

The rest of the work is structured as follows. Section 2 examines the available means of measuring passenger traffic. In Section 3, we discuss our model, which is put in the assessment of the increase in passenger traffic after the launch of new radial lines.

II.ON TRAFFIC MEASUREMENTS

At present, for the railways in Moscow, the existing passenger traffic can be measured quite accurately. Under the existing scheme, at most stations passengers must present (validate) travel documents at the entrance and at the exit. In terms of social networks, this corresponds (is analogous) to check-in marks (at the input) and check-out (on exit) [2]. This increases the value of information on the use of railway stations comparing with other transport data. From the validation information, we can immediately restore the passenger's route. Traditionally, most travel documents in Moscow (in the Moscow region) are validated only at the entrance (check-in). Accordingly, to restore the actual route,

we must use some heuristic algorithms. A review of such algorithms is, for example, in our paper [3]. The idea is that the breaks (gaps) in the use of travel documents are just the delimiters of trips. For example, if we noted the use of a travel card (in Moscow, for example, it is a Troyka card) at point A, and then after a relatively long break at point B, then we can assume that the passenger traveled from A to B, then, for example, was at work and after its completion makes a new trip. The process of route restoring based on the data of telecommunication operators also relies on heuristics [4]. For example, the place where calls are made in the morning and in the late evening is considered "home", in the daytime is considered "work" and so on. Of course, for such classification, we can impose additional conditions too. For example, we can set a minimum total time of stay for a month to recognize the place as a working place. And in the case of information on the validation for railway stations, the route is known [5].

Technically, data on validations (checking/inspection of travel documents) are presented as separate text files, each of which describes passes for a particular station in one month. One entry (a line in the file) corresponds to one pass (to the entry or exit). The data is completely anonymous and does not contain any personal information. There may be a type of travel document used (for example, it is a preferential tariff or not, it is a one-time or reused travel document, etc.), but there is basically no identification of documents at all. The size of each such log depends, of course, on the use of a particular station in a particular month and varies between 20 and 70 Mb.

Fields that are contained in the records:

- date and time;
- price characteristic (full or reduced ticket);
- type of benefits (federal, local, etc.);
- ticket type (one-way ticket, one-way ticket, one-way ticket, subscription, etc.);
- a starting station;
- an end station.

This kind of information allows to analyze summary data on the stations (total entries/exits) and, if necessary, analyze the distribution of passengers on the entries/exits too. For example, the following pictures show hourly distributions of entries and exits at a suburban station near Moscow. This is a typical picture with working traffic, where we have a peak at the entrances in the morning and a peak at the exits in the evening (in the morning people go to Moscow, in the evening they return back from work) (fig. 2, 3). Note that this picture is not necessarily for all stations.

To analyze the data, a cloud tool from Google Collaboratory [6] was used. Technically, this is a cloud implementation of Jupyter notebook. It does not require any software installation, everything is accessible through the browser. The data files were stored on Google Drive.

Thus, the processing of validator data allows receiving a correspondence matrix: for each pair of stations, specify the number of passengers traveling between the two stations. And this kind of matrix can be built in any time slice: for an hour, a day, a month, etc.

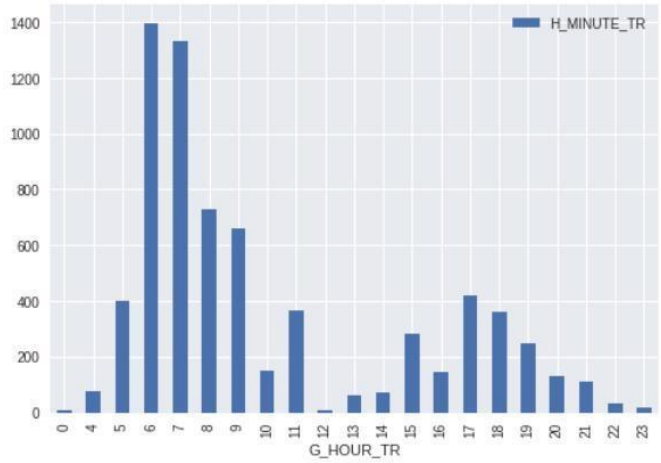

Fig. 2. Entries in suburb

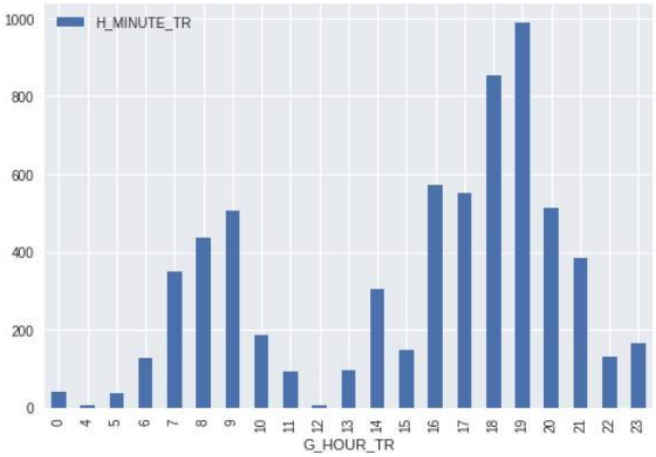

Fig. 3. Exits in suburb

This set is an objective data on the existing passengers of the railway. What can serve as a basis for assessing the potential demand for such a service?

In our case, we used data collected by mobile operators. Telecommunication operators in order to bill users for the services they provide constantly collect information about the activity of users (mobile devices). This is not related to any tracking of subscribers, this is purely an economic event (they need to know the duration of the calls made in order to calculate their cost, the numbers for which calls were made, since this can also affect their cost, etc.). In the same log, system software puts information about mobile tower which is currently serving a specific mobile device. Accordingly, the fact of changing this information is reflected too. A cellular tower (base station) has a very specific location (specific geographical coordinates). Accordingly, the location of the mobile device (change of location) can be associated with geographic coordinates. In fact, it looks more

complicated than described here (it is necessary to simulate radio signal propagation, know the location of other base stations, antenna patterns, etc.), but the result will be exactly this; we can estimate the geographic position of the mobile device.

Naturally, this estimate is made with some accuracy. This accuracy, obviously, is related to the number of base stations that participate in the calculations. Therefore, we can say that in urban conditions (high density of base stations) this accuracy will be higher than, for example, in a province where fewer people live and fewer base stations deployed. In general, under current conditions, we can expect the accuracy of several hundred meters (for example, a square of 500×500 meters). Note that locating with some kind of rigidity corresponds also needed to observe the interests of privacy of mobile subscribers. If movements are issued for all mobile subscribers from a certain geographical square, then it is not possible to track the movements of any individual subscriber.

As a result, the data of mobile operators make it possible to build a correspondence matrix (origin-destination matrix) for geographical squares. For each pair of squares (say, 500×500 meters), you can get the number of people moving from one square to another. Again, it could be done in any time frame: hour, day, month, etc. Of course, the total size of such a geographic grid can vary and the total size of the data depends on the grid. In our case, it covered a sufficient area in the suburbs of Moscow. Sufficiency, in this case, it assumes either the direct availability of a railway running through Moscow or the possibility of getting by transport to the station of this road at a reasonable time.

Information collected on geographical squares (in any time frame) can be aggregated, so that we can obtain information on the movements for administrative areas. Usually, for administrative regions, we have the population statistics. It means that we can use this statistics for verification of collected data.

The result is an objective picture of the movement of mobile devices (subscribers) from areas adjacent to railway stations. From consideration, we can immediately exclude moving to close distances (we assume that these are pedestrians). The remaining movements will be transport movements. Strictly speaking, transport modeling starts here. We suppose that we are considering a specific area in a suburb of Moscow. All these moving mobile subscribers must be somehow sorted by mode of transport. It is known exactly how many they used the railway. If there is information on the use of buses in the suburbs (for example, validation of transport cards), the difference between the total number of subscribers and the number of passengers of the railway and buses will give the number of those who will get to Moscow by cars. This is the current picture. After opening the urban railway, it will change. Flows should be redistributed between modes of transport (in fact, the main idea is that some of the motorists will transfer to rail transport), there may be new passengers who did not travel before because of the fact that getting to the city was uncomfortable.

III.On traffic models

Firstly, from the analysis of operator data on movements, for example, for a month, you can determine the nature of displacements by the type of "home-work." This is a standard approach, which was used in the earliest works on the analysis of such data [7]. The idea is to relate the time of day and user activity. If the mobile subscriber is constantly active (calls, writes SMS, etc.) from a certain region in the evening, then this region can be recognized as his "home". Similarly, the constant activity from a particular region during the daytime leads to the classification of this region as a "work". The term "permanently" defines here the time required for the classification. For example, it could be 7 days during a month. This value is given empirically. In paper [8], the algorithm assumes the location from which a user departs in the morning and to which he or she returns at night is home. It also infers that the location of the longest recurring stays during weekday daytime hours is the user's workplace.

After this classification, two things become available. Firstly, we can describe the displacement in terms of home-work. This can be a more stable pattern of displacement. Secondly, it becomes possible to estimate the actual population by geographical objects (administrative units, if aggregate information from squares).

Estimating the number, in this case, is the first point in the assessment of passenger traffic. This is obviously the upper estimate of possible railroad traffic. This is the theoretically possible number of passengers. There are simply no more of them.

At the same time, we assume that the distribution of "new" passengers by stations will be exactly the same as the current distribution of railway passengers by stations. In other words, each station will receive its proportional increase in passengers.

From a practical point of view, in most areas (stations), the figures obtained exceeded the physical capacity (capacity) of the respective stations.

There is one important observation about the classification of movements by home-work pattern. It is not necessary that both sites for a particular subscriber will be classified. For example, subscribers with traveling related work. They do not stay long in one place to classify their activities there. But at the same time, they create a load on the transport system. On the other hand, there may be subscribers who do not really move anywhere. In this case, the home area for them will be determined; there will be really registered telephone activities at the place of residence.

The analysis of passenger activity by stations shows a very high level of constancy. If, for example, we build a one hour aggregated graph for the entrances to a typical Moscow suburban station (see Fig. 2), then such time series for different days of the week will be very similar. The degree of similarity of this kind of time series, measured by different metrics, is quite large. For example, in our work, we have successfully used a shape-based similarity measure, it is Angular Metric for Shape Similarity (AMSS) [9].

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.

The cosine similarity metrics minimize the influence of outliers in similarity computation, where outliers are defined as much bigger or smaller data points than their immediate neighbors [9].

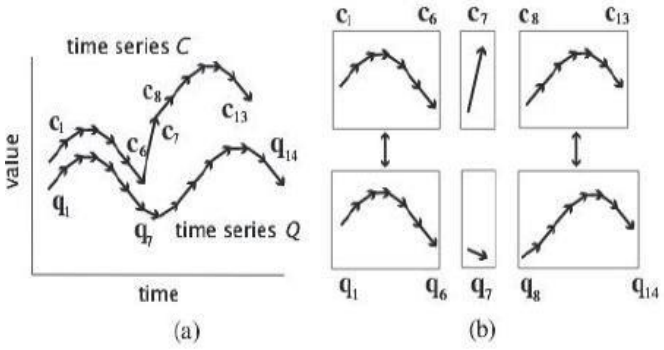

Fig. 4. Angular Metric for Shape Similarity [9]

This means that the railway traffic is fairly stable. The railway has its regular passengers who use it. And the fluctuations here (before the launch of new traffic conditions) are rather small.

The telecom data is here the only available information on the behavior of passengers. Historical data for this kind of projects are not available for obvious reasons. There are also no survey estimates. Accordingly, several heuristics were used to estimate the future passenger traffic, based on the characteristics of the current and the maximum possible passenger traffic [10]. Let's look at these heuristics.

A. We estimate the share of railway passengers for each of the stations of the new route. In other words: what percentage of all those moving from the area (according to the statistics of the mobile operator) is actually transported by the railway (according to the railway station validation data).

Determine the maximum percentage across all stations. And after that, we assume that such a percentage (coverage) will be at each station. So, we can recalculate traffic for all stations based on a new coverage.

Just for the explanation: 100% of the movements will not transfer to the railway. The maximum percentage of coverage by the railway is a real achievable share. Thus, we determine the new number of passengers (entrances) for the stations, and then these passengers are distributed to the other stations in the same proportion as it exists for the real data.

B. We can estimate the share of new passengers for railway stations according to seasonal trends. In summer, the number of residents who travel to the city from suburbs increases at the expense of those townspeople who live in dachas, but travel to the city.

Accordingly, we will get some increase in the number of people in the suburbs, which moves (according to the telecom operator) from the area to the city and back. Of these trips, some share will have to rail transport, which can be objectively determined according to the validators (according to the real data). The share of new "residents" who chose rail transport can be extrapolated to the entire population of the district and thus calculate the possible share of the railway in the transportation of residents in the area.

A possible explanation: the new "residents" do not have stable preferences in comparison with permanent residents, and their choice will reflect the objective correlation of transport shares.

C. Most of the inhabitants of the suburbs who travel at least once used the railroad. This assumption seems reasonable, since the railway, in any case, is the simplest way to travel to Moscow. So, the appearance of new passengers at the railway is possible only due to the fact that those who have traveled rarely will travel more often. Accordingly, in addition to counting the number of trips, it is necessary to calculate their frequency distribution. Technically, this is a grouping of data on the movement between areas (territories, cells) by the user ID. Data from operators contain some kind of subscriber identification (for example, IMEI or hash code, which replaces this identifier to preserve privacy). If we calculate the number of user movements per month from the area to the area, we can get useful frequency information too. E.g., it is some like this:

> *N* movements were made by *X* subscribers
> *N-1* displacements were made by *Y* subscribers *N-2* movements made by *Z* subscribers etc.

For reasons of privacy, we can also cluster the data by the number of displacements. For example, in steps of 5:

> from 5 to 10 movements made by *X* subscribers;
> from 10 to 15 movements made by *Y* subscribers;
> from 15 to 20 movements made by *Z* subscribers, etc.

This will give the distribution of residents (those who travel) on the frequency of travel. And this distribution shows the reserves of growth in travel. In this distribution, we will be interested in the low-frequency "tail". In fact, we are interested in aggregated data again (the number of trips) but only for a part of the inhabitants (for low-frequency travelers). We proposed to sum these low-frequency trips and then conduct something like a stress test. For the number of trips increases by 25% (50%, 100%) recalculate this percentage in thousands of passengers. Then we can share this amount of new passengers by the stations using the existing distribution of passengers (calculated by real validation data). And for each growth figure, it will be possible to assess whether there are enough trains (capacity of platforms, the throughput of turnstiles, etc.) and what will

be the increase in passengers at metro stations (when passengers change from railways to metro).

D. For areas where it is possible to count motorists (as a difference between all movements and validations of travel documents on the railway and public transport), we can use historical data on the launch of urban railways abroad. The idea of replacing motorists on the railway, in fact, was always one of the main for urban railways. In the literature, it was noted that the railway to the cities is the only transport today that can move without traffic jams and, at the same time, provide a fairly high level of comfort. It is noted that the city railway is the only transport today, for which motorists are transplanted. Because, for example, using a bus often means meet the same traffic jams as on a personal car, but only with a lower level of comfort (as opposed to a personal car). But it is obvious that 100% of the replacement still does not happen. The figure that occurs in the literature is up to 20–25% of the users of vehicles that switched to the railway. This was determined, of course, after the projects were put into operation. That's exactly this figure - 20% of the number of motorists can be used to calculate the increase in the number of passengers in the railway.

IV.CONCLUSION

In this paper, the models of an estimation of passengers' traffic for new lines of a city railway projected in Moscow are considered. The initial data for building the forecast was information on the current load of the railway transport, data on the use of public transport (in some areas outside Moscow) and measurements based on the data of mobile operators and describing the movement of mobile subscribers in the Moscow region. The model offers a number of heuristic approaches to traffic estimation, which are combined with real data of existing rail traffic.

V. ACKNOWLEDGMENT

Some of the ideas presented here were previously condemned by us at data analysis conferences (AIST 2018, DAMDID 2018). We are grateful to the reviewers of these conferences. Their criticism and constructive comments allowed us to seriously improve our work.

VI. REFERENCES

[1] Shneps-Shneppe D. On Digital Signaling for Moscow City Railways // International Journal of Open Information Technologies. 2018. Vol. 6, N 6. P. 28–37.

[2] Namiot D. Sneps-Sneppe M. Customized check-in procedures // Smart Spaces and Next Generation Wired/Wireless Networking. Berlin; Heidelberg: Springer, 2011. P. 160–164.

[3] Namiot D., Sneps-Sneppe M. A Survey of Smart Cards Data Mining // Supplementary Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017) Moscow, Russia, July 27-29, 2017. Moscow, 2017.

[4] Steenbruggen J., Borzacchiello M.T., Nijkamp P. et al. Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities // GeoJournal. 2013. Vol. 78, N 2. P. 223–243.

[5] Ratti C., Frenchmann D., Pulselli R.M. et al. Mobile landscapes: using location data from cell phones for urban analysis // Environment and Planning B: Planning and Design. 2006. Vol. 33, N 5. P. 727–748.

[6] Google Collaboratory https://research.google.com/colaboratory/unregistered.html.

[7] Wang J., Li J., He K. et al. Vulnerability analysis and passenger source prediction in urban rail transit networks // PloS one. 2013. Vol. 8. N 11. P. e80178.

[8] Mearian L. Ford, MIT use Bostonians' cellphone location data for traffic planning // Computerworld. URL: https://www.computerworld.com/article/3112845/cartech/ford-mit-use-bostonians-cellphone-location-data-for-city-trafficplanning.html.

[9] Nakamura T., Taki K., Nomiya H. et al. A shape-based similarity measure for time series data with ensemble learning // Pattern Analysis and Applications. 2013. Vol. 16, N 4. P. 535–548.

[10] Намиот Д.Е., Покусаев О.Н., Лазуткина В.С. О моделях пассажирского потока для городских железных дорог // International Journal of Open Information Technologies. 2018. Vol. 6, N 3. P. 9–14. [Namiot D., Pokusaev O., Lazutkina V. On passenger flow data models for urban railways // International Journal of Open Information Technologies. 2018. Vol. 6, N 3. P. 9–14.