

The Analysis of Trajectories in Moscow Subway

Mariia Nekraplonna and Dmitry Namiot

Lomonosov Moscow State University, Moscow, Russia
maria.nekraplennaya@gmail.com, dnamiot@gmail.com

Abstract. Along with the continuous growth of megacities, their transportation systems have become increasingly large and complex. The use of transportation systems by passengers directly reflects the changes that occur in the urban environment - for this reason, the study of urban mobility is an important task of digital urbanism. In particular, this paper is devoted to the study of spatial patterns (repetitive routes) in transportation systems with the case study on the Moscow subway. A brief review of data mining approaches to transportation systems data in general and to the task of spatial patterns extraction, in particular, is presented. A simple method for pattern extraction is proposed and applied to the Moscow subway data. As a result of the deployment of the proposed method the list of patterns was obtained - the graph of spatial patterns of the transport system under study was constructed based on it.

Keywords: Urban mobility · Digital urbanism · Travel behaviour · Ridership patterns · Public transport · Data mining · Time series · Clustering analysis.

1 Introduction

The present paper is a continuation of a series of works devoted to the analysis of urban displacements [1-4]. Common to all the problems is that the basic elements for the analysis are the so-called origin-destination matrices (OD-matrices). Such matrices represent information about the number of movements between two points (objects) during a certain time interval. Objects can be city railroad stations, subway stations, or even geographically delineated areas. In this study, we deal with correspondence matrices for rail transport. In this case, as is clear, the routes between the two stations are pre-defined. In the case of correspondence matrices for geo-areas, the task of choosing a route appears additionally.

Such matrices allow to fully describe traffic flows in the city. In many cases, it is the construction (prediction) of such matrices that is the ultimate goal of transport systems analysis. The development of telecommunications technology (discussed in more detail in Section 2) has led to the fact that such data can be collected (measured) with the help of telecommunications operators. This means that these kinds of data, instead of predictive (computable) data, have become raw data. The task of forecasting has become irrelevant. It is pointless to predict what will be accurately measured.

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In the above representation, the correspondence matrix is a closed system. If we consider the correspondence matrix for the Moscow metro, then there is no information about, for example, that a new route from the suburbs has opened, which will increase the number of system users (passengers). There is no information in the system about any cultural events that for some time increase the total number of passengers, and so on.

The metro responds to some new passenger flows (they are reflected in the matrix), but there is no information about these new flows in the system. Specific example. Switching suburban buses from the bus station near the Tushinskaya metro station to the new bus station near the Khovrino metro station will cause (caused) a certain peak in the load in the morning on the green metro line, but this could only be registered after the fact. It was impossible to predict this in any way, since the bus route information is external to the correspondence matrix.

From this follows another important statement. The correspondence matrix obtained by the method described above is a certain measuring tool, which can be used to determine the occurrence of some events in the city or to check the results of any action. Changes in passenger behavior (changes in routes, travel times) are something that can theoretically be determined and each such change must, naturally, be somehow explained in terms of changes in the city. The effect of any mass events will also be reflected in transport movements (participants need to arrive and leave, which should correspond to the deviation in the correspondence matrix compared to the "normal" values).

We deliberately avoid using the term transport tasks, because the analysis, although related to the transport systems (processes), but its main purpose was digital urbanism [5]. In general, the analysis was used to describe the functioning of the city as a system. Movement patterns are needed, for example, not only to identify stressful places in the transport system, but also to assess the real points of attraction in the city - where jobs are concentrated, where residents go on weekends for recreation, etc. Knowledge (understanding) of the usual structure of passenger traffic makes it possible to identify abnormal behavior (peaks or dips in traffic, changes in "normal" routes). In some cases, such anomalies are easily predictable and understandable. For example, the end of a large concert (soccer match) will cause a peak in the entrance at the nearest metro stations. In this case, the definition of anomalies will only give a numerical estimation of these peaks. In other cases, the reasons for such anomalies may not be obvious at all, and here their determination will serve as just a signal for the city services that it is necessary to look for the cause of this phenomenon. In other words, the result of the analysis for traffic data will always end up being some metric for the urban system.

In principle, two approaches to analyzing this kind of data are possible. First, we can consider the change of the system in the time domain. How the inputs/outputs for stations change over time (see Fig. 1). For a correspondence

matrix, we summarize the data by rows (inputs to a station) or columns (outputs from a station) and analyze the change in this data over time.

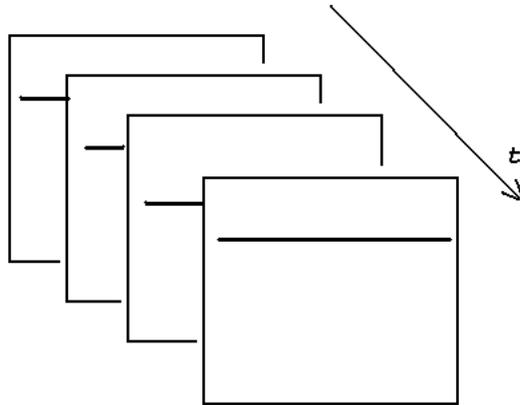


Fig. 1. Analysis in time section.

In fact, it is an analysis of how subway stations are used during the day, weekdays, and weekends. The familiar notion of "rush hour" is exactly about this type of analysis.

Another possibility is that we are investigating trajectories (spatial or spatiotemporal slice) - Fig. 2. Where do passengers go most often from station A? Are there any recurring patterns here? How are the preferred destination stations (if any) distributed by time on weekdays and weekends? If in the temporal analysis the main question was "when do passengers travel from/to the station", in the spatial analysis it is "where do passengers travel from/to the station".

The very idea of studying exactly spatial characteristics (trips, routes) came from the analysis of movements using suburban railroads [2] and works on determining anomalies in traffic [6]. For suburban railroads, there was a clear pattern of selecting the route's end station as the first station where there is an intersection with the subway line(s). And this pattern changed on weekends. As a matter of fact, the new traffic scheme on the suburban diametrical routes should have changed this pattern and suggested that passengers continue to use the railroad in the city, rather than transfer to the subway at the first opportunity.

Also, the spatial analysis leads to an estimate of the duration of movements (the duration of routes), and this, as we know, is one of the main characteristics of transport behavior. Another consideration is that the endpoints of the routes should correspond to what are commonly called points of attraction. At the same

time, such points of attraction will obviously depend on time and, perhaps, on days of the week. For example, the stations where most of the trips (routes) end in the morning are what corresponded to the working districts in the time section of the research, etc.

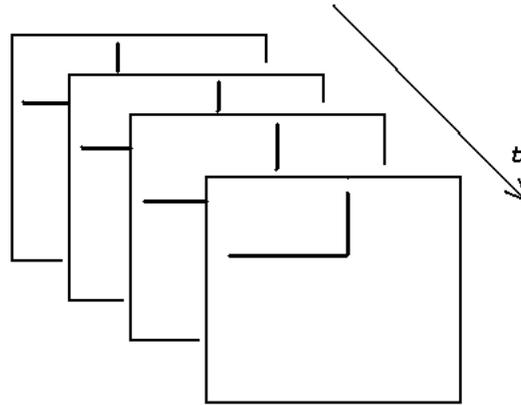


Fig. 2. Route analysis.

The first studies in terms of determining routes showed [2], for example, that in the north of Moscow on weekdays most of the passengers from Rechnoi Vokzal station go to Voykovskaya station (most experts were sure that passengers go at least to the nearest interchange) - Fig. 3. And this route remains the most popular all weekdays. At the same time, on weekends (also steadily during the month) the most popular route is to the city center. Accordingly, this division determines working points of attraction and places of rest for the northern direction.

Accordingly, in this article, we would like to test the hypothesis of whether there are stable (repeating) routes in the city's transport system or not. If such routes are not connected with the stations through which the shortest routes pass, then the analysis of trajectories (movements) will allow to highlight the transport preferences of passengers, which will reflect some existing division in the city.

Note that for the metro (as well as for other rail transport), the travel time between two stations is fixed with sufficient accuracy. Accordingly, a trip (aka a trajectory) can be measured in units of time. This is an important characteristic specifically for urban planning and management - the time passengers spend on the road. For example, the Figure 4 shows the distribution of travel times for two cities in the United States (Boston and San-Francisco) [8]. This is an integral indicator for urban transport.



Fig. 3. 1 - real route, 2 - expected route, 3 - real weekend route.

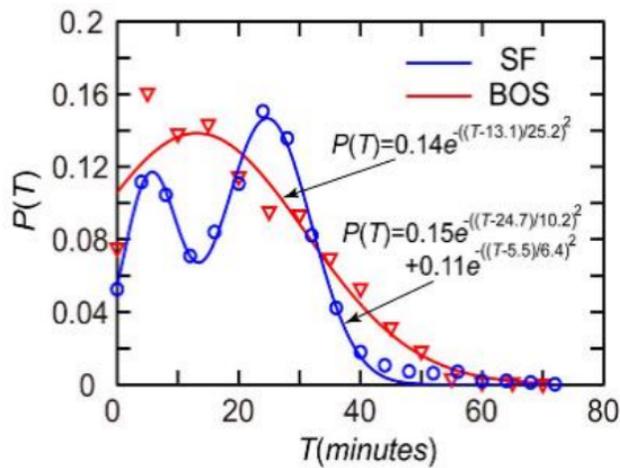


Fig. 4. Travel distribution time [8].

Possible areas of analysis are discussed in more detail in section 3.

The rest of the article is structured as follows. Section 2 describes the presentation of the data. Section 3 deals with general issues of analyzing this kind of data. In particular, the results of data analysis in the temporal domain are presented here for comparison. Section 4 is devoted to a brief analysis of works on the study of trajectories of movement in the city. Section 5 is a presentation of the first results on the study of the routes of passengers in the Moscow subway. Section 6 presents the conclusion.

2 Data and Data Representation

The Moscow Metro belongs to those few systems where smart cards are used only for entry (tap-in event) and payment is made at a single ('flat') tariff. This means that it is impossible to accurately track the passenger flow by ticket identification numbers, as in usual transport systems with double ticket attachments (tap-in and tap-out). Only some heuristic assumptions can be made that gaps in travel document usage determine the route. Fortunately, the Moscow Metro also has another important feature: mobile communication here is provided both above and below ground. Thus, an alternative way to measure passenger flows is to record the activity of their smartphones. In this case, switching from an above-ground communication tower to an underground one is equal to a tap-in event, and switching back to a tap-out event. Thus, we have an opportunity to build time series of inflow and outflow of passengers the same as in other researches using smart-card data. Obtained data are impersonal aggregate records of passenger movements with half an hour intervals for February 2018. We can neither trace the trajectory of a specific passenger nor link it to a specific phone number. The data have been previously cleared: trips longer than 4 hours have been cleared, as well as those with the start station coinciding with the arrival station. The data are provided in CSV format and contain in total about 226 044 records (207 stations during 28 days, 39 time-intervals per day). Each of the records contains two values: the number of passengers who started their trip on a certain route in the period of time and the number of passengers who finished this route. In fact, this table is a form of a record of origin-destination matrix.

3 Analysis of Transport Systems Data

The first automatic fare collection systems in public transport were introduced more than 50 years ago. During this time, quite a lot of data has been accumulated on passenger transport behavior, which is now analyzed by various scientists from different points of view. First and foremost, of course, is the prediction of the value of passenger traffic. This task has been considered in many works, among others [7].

Note that changing the data collection scheme should also change the approach to transport problems. For example, in a work typical of the traditional

approach [19], the authors try to predict the daily traffic of the entire transport network from historical data. First, the total traffic without division by stations / lines can only be useful for calculating revenue. Secondly, the article discusses the forecast, while changing the data collection scheme gives accurately measured traffic for individual stations with a step of 5 minutes. Why predict what is being measured exactly? The change in the data collection scheme also changed the transport tasks. They used to traditionally deal with forecasts. If traffic is accurately measured, then predictions become meaningless. Traffic is now another form of measurement in the city. This is what is being promoted in our article. And measurements, of course, should be more substantive. The measurement level should be shallower - a station, instead of the entire transport network. Nevertheless, apart from forecasting, an important direction is a subset of advanced knowledge extraction tasks.

We can distinguish three main directions in the analysis of traffic flows in relation to rail transport: the passenger-oriented [9, 10], the event-oriented [11, 12], and the station-oriented. The first one is aimed to identify groups of passengers with similar travel behavior. The peculiarity of this approach is the applicability only to the transport system of the tap-in-tap-out category. The second approach, event-based, clusters days (or other time intervals) depending on the similarity of the whole public transport system operating mode. For this approach, it is important that the observation period be large enough to cover all the events of interest to the researchers - for example, as many periods with different weather in the study [13]. Finally, the last approach, which was the objective of previous research [14], clusters stations on the basis of when their activity occurs, i.e. how trips made at the stations are distributed over time. It is important to note that this approach can be implemented both from a time perspective and from a spatial perspective.

The station-oriented researches with time perspective the key question is "when do passengers depart/arrive from stations?". In this case, the subject of the study is a set of different characteristics of the time series describing the passenger flow, such as peaks of passenger activity. As an example, see [13], where several clusters were identified by the time series clustering. For the Moscow subway, we have already solved such a problem in [1], the results are shown in Fig. 5.

In station-oriented studies with a spatial perspective, the subject is not the temporal characteristics of passenger traffic, but the spatial ones. In other words, the researcher asks the question "where are the existing passenger flows directed? We can talk about patterns if we see repeating routes in this study. Since such studies have not yet been conducted for the Moscow Metro, we believe it is necessary to fill in the existing gap.

Note also that we can talk about a certain combination of approaches: "where do they go at such and such periods of time." These will be time travel patterns.

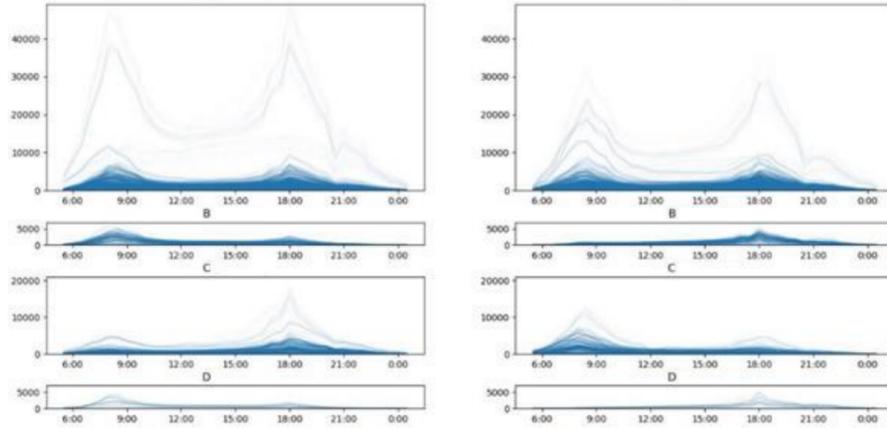


Fig. 5. Vertically there are four types of Moscow metro stations. On the left are station entrances during the day, on the right are exits.

4 Spatial-Oriented Analysis (Trajectory Analysis)

Speaking about similar works, one can cite, for example, article [14]. This article develops a comprehensive methodology for the study of prevailing trajectories in urban studies. Researchers are now using cluster analysis tools to identify different typologies of areas and principal component analysis (PCA) to determine socio-economic interactions. The work uses sequential dynamic analysis to identify trajectories.

With regard to the analysis of trajectories, one can also note the paper [15]. It uses the fact that the trajectories in the subway can be compared in time. Here the authors split all routes into several intervals of 20 minutes and study the distribution of routes in time.

Several works performed at Moscow State University [16, 17] directly investigated the nature of the movement of passengers on suburban railways. For example, work [16] noted an asymmetry in suburban traffic on weekends - the number of people leaving by rail to Moscow was always greater than the number of those returning by rail. The same asymmetry was confirmed in [17]. Figure 6 shows a visualization using Kepler.gl [18] of real trips.

5 Analysis of Moscow Subway Trajectories

In this paper, the most popular paths were considered. Denote the number of trips from station i to station j in day n as $trips_{i,j}^n$. For station i we will call the most popular direction in day n station j such that the value of $trips_{i,j}^n$ is the largest. Next, k -th most popular direction of station i will be station j , which will be at k -th place if we order all stations in descending order according to



Fig. 6. Moscow - Tver railroad trips.

how many times they became the most popular direction for the day during the entire observation period. In Fig. 7, you can see what proportion of the 1st, 2nd, and 3rd most recurring destinations are from all of the most popular destinations for each of the metro stations for the day.

As you can see, for almost all stations, the 1st most recurring direction has been (red color) for more than half of the observation period. This suggests that the 1st most repeatable direction is indeed a spatial pattern, the deviation from which, generally speaking, should be considered more carefully. Having distinguished the spatial patterns as described above, we can construct a graph of spatial patterns of the Moscow Metro: each station has one vertex and each pattern has one directional edge (Fig. 8). An interesting observation was the existence of a large center of attraction in this graph: for some reason 82 out of 207 patterns are directed towards one station - it is Chekhovskaya station of Serpukhovsko-Timiryazevskaya subway line.

In this figure, Chekhovskaya station is circled in red. The landing of routes at this station can, in principle, be explained by its centrality in the transport graph (static characteristic of the graph, reflecting the number of the shortest routes passing through the node). Examples of other stations highlighted in green, where at least 5 routes end, are already a reflection of the characteristics of the city. From some stations, passengers go to some predefined places.

It should be noted, however, that the popular routes for the stations remained this way throughout the day. This answers the question of combining temporal and spatial analysis. At least in this study, we did not find any changes in routes over time. Routes changed only between weekdays and weekends. Technically,

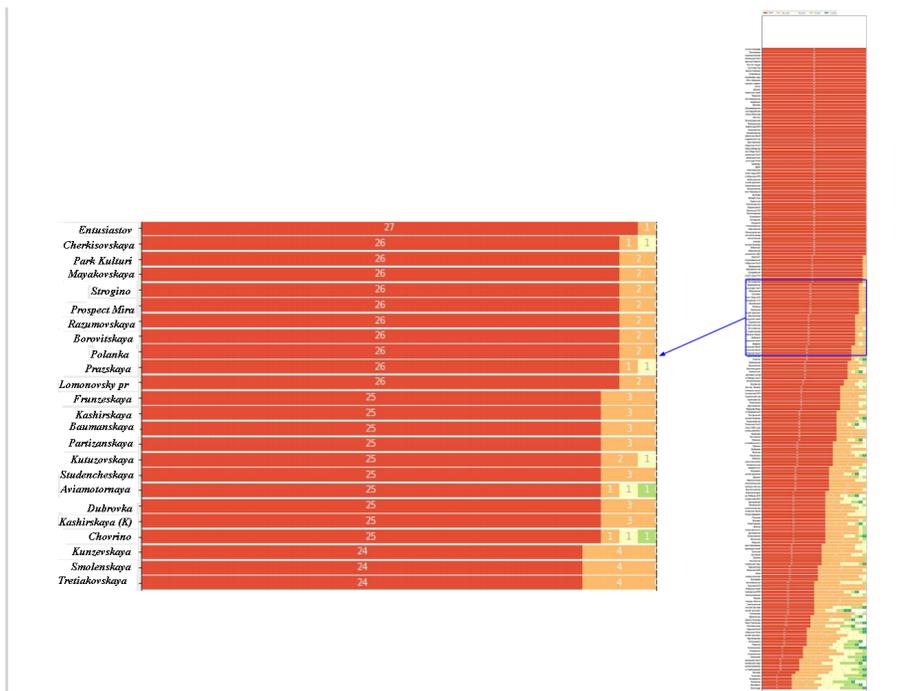


Fig. 7. First five most recurring destinations for each station.

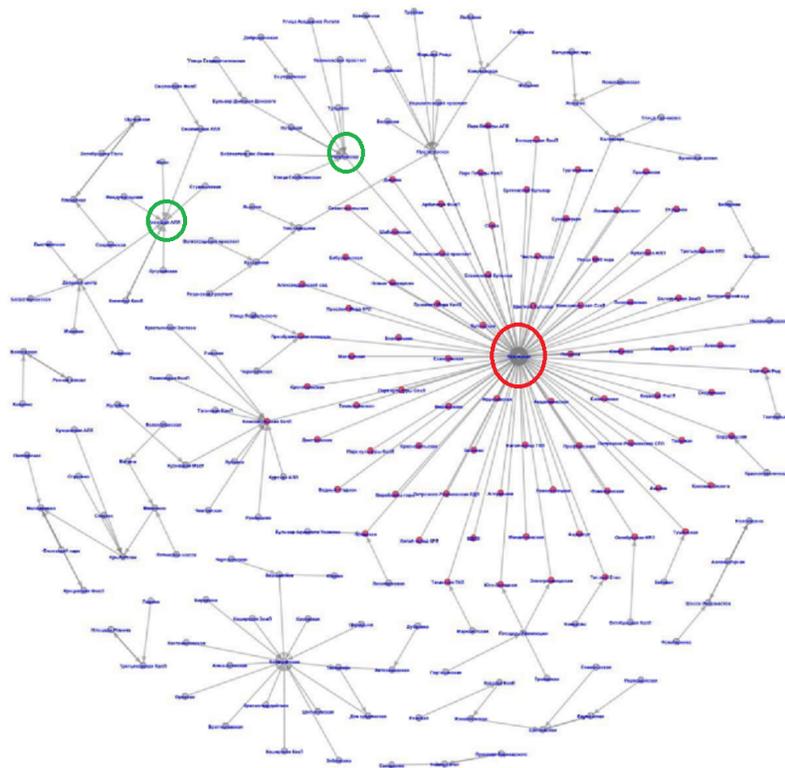


Fig. 8. On spatial patterns of the Moscow Metro.

this should mean that the patterns found fit well with the points of attraction of the city - where passengers travel on weekdays, and where - on weekends.

Explaining such routes is already the task of urbanists. For example, one of the route hypotheses in Figure 2 is that passengers arriving from the north of Moscow do not want to spend a lot of time commuting to work. And the station of attraction (the end of the routes) Voikovskaya - there is the first station where there is a sufficient number of offices and business centers (jobs). The stations before her belong more to the rescue areas. But this, of course, is only a hypothesis. What is important here is that trajectory analysis allows you to identify such areas in the city, that is, it is some kind of urban metric. Accordingly, the change in routes will serve as an indicator of some changes in the city (opening and settling of office centers, settling in housing estates, etc.).

6 Conclusion

This paper is devoted to the study of spatial patterns (repetitive routes) in transportation systems with the case study on the Moscow subway. A brief review of data mining approaches to transportation systems data in general and to the task of spatial patterns extraction, in particular, is presented. A simple method for pattern extraction is proposed and applied to the Moscow subway data. As a result of the deployment of the proposed method, the list of patterns was obtained, and the graph of spatial patterns of the transport system under study was constructed based on it. The result is the first time confirmation (finding) of stable routes in the metro transport system. And such routes are not associated with the centrality of the transport system graph (that is, with stations through which more shortest routes pass). The fact that the movements between the designated stations remain constant during the work week and constant during the weekend is precisely an indicator that these are some routes due to the structure of the city. Understanding the routes of passengers on the Moscow metro serves to determine their transport behavior. The latter is one of the basic characteristics of the transport system in a Smart City.

References

1. Nekraplonna, Mariia, and Dmitry Namiot: Metro correspondence matrix analysis. *International Journal of Open Information Technologies* 7.7 (2019): 68-80.
2. Namiot, Dmitry, et al.: Where and when-about one approach to traffic analysis in the city. *International Journal of Open Information Technologies* 9.3 (2021): 44-49.
3. Misharin, A., D. Namiot, and O. Pokusaev: On Processing of Correspondence Matrices in Transport Systems. In: 2019 International Multi-Conference on Industrial Engineering and Modern Technologies, FarEastCon. IEEE (2019).
4. Namiot Dmitry, and Oleg Pokusaev: On mobility patterns in Smart City. In: CEUR Workshop Proceedings, (2019).
5. Kitchin, Rob: The real-time city? Big data and smart urbanism. *GeoJournal* 79.1 (2014): 1-14.

6. Bulygin, Mark, and Dmitry Namiot: Anomaly Detection Method For Aggregated Cellular Operator Data. In: 2021 28th Conference of Open Innovations Association (FRUCT). IEEE (2021).
7. Zhang, Kai, et al. A framework for passengers demand prediction and recommendation. 2016 IEEE International Conference on Services Computing (SCC). IEEE, 2016.
8. Wang J. et al. Vulnerability analysis and passenger source prediction in urban rail transit networks //PloS one. – 2013. – T. 8. – №. 11. – C. e80178.
9. El Mahrsi, M.K., C^ome, E., Oukhellou, L., Verleysen, M.: Clustering smart card data for urban mobility analysis. IEEE Transactions on intelligent transportation systems 18(3), 712– 728 (2017)
10. Goulet-Langlois, G., Haris, N.K., Jinhua, Z.: Inferring patterns in the multi-week activity sequences of public transport users. Transportation research part C: emerging technologies, 64(3), 1–16 (2016)
11. Yang, C., Yan, F.F., Xu, X.D.: Clustering daily metro origin–destination matrix in Shenzhen, China. Applied mechanics and materials 743, 422–432. (2015)
12. Yang, C., Yan, F., Xu, X.: Daily metro origin-destination pattern recognition using dimensionality reduction and clustering methods. Intelligent transportation systems 2017 IEEE 20th International Conference, 548-553 (2017)
13. Zhao, X., Wu, P.-Y., Ren, G., Ji, K., Qian, W.-W.: Clustering analysis of ridership patterns at subway stations: a case in Nanjing, China. Journal of urban planning and development, 145(2) (2019)
14. Li, Yuchen, and Yichun Xie. A new urban typology model adapting data mining analytics to examine dominant trajectories of neighborhood change: a case of metro Detroit. Annals of the American Association of Geographers 108.5 (2018): 1313-1337.
15. Shen, Ping, et al. Cluster and characteristic analysis of Shanghai metro stations based on metro card and land-use data. Geo-spatial Information Science 23.4 (2020): 352-361.
16. Misharin, A., D. Namiot, and O. Pokusaev. On Passenger Flow Estimation for new Urban Railways. IOP Conference Series: Earth and Environmental Science. Vol. 177. No. 1. IOP Publishing, 2018.
17. Medvedenko, Stepan, and Dmitry Namiot. "Visual analysis of railway passenger traffic data." International Journal of Open Information Technologies 9.6 (2021): 51-60.
18. He, Shan. From Beautiful Maps to Actionable Insights: Introducing Kepler. gl, Uber's Open Source Geospatial Toolbox. Uber Engineering Blog (2020).
19. Ivanov, V. V., and E. S. Osetrov. Forecasting daily passenger traffic volumes in the Moscow metro. Physics of Particles and Nuclei Letters 15.1 (2018): 107-120.