Anomalies in Transport Data

O Pokusaev¹, D Namiot¹,²
¹RUT (MIIT), Center for Digital High-Speed Transport Systems, 969 Obrazcova Street, Moscow 127994, Russia
²Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, GSP-1 1-52 Leninskiye Gory, Moscow 119991, Russia

E-mail: o.pokusaev@rut.digital

Abstract. Over time, the analysis of traffic data has come down to predicting traffic flows. Building correspondence matrices describing traffic flows, forecasting traffic flows at a specific point and at a specific time interval is what transport analytics did. However, the development of technology and the economy has led to the fact that the penetration of mobile phones and the efforts of telecommunications operators provide complete information about the movements of mobile devices (their users) in modern cities. This means, in particular, that predicting flows is no longer a task. Other issues come to the fore. And one of them is just an analysis of anomalies. This is necessary, first of all, at the stage of operation of Smart City systems.

1. Introduction
The development of technology has changed, in many cases, approaches to the analysis of transport data in cities. Traditionally, the study of transport data assumed (historically) some kind of forecast [1, 2]. For example, any analysis of traffic flows in the usual sense is, ultimately, some kind of forecast. This is a forecast about traffic volumes, about the purpose of travel, their distribution in time, etc. [3, 4]. But let’s look at the technological background of these processes. For example, at suburban railway stations in the Moscow region, passengers themselves mark their travel documents at the beginning and end of the trip. Naturally, this mark is registered in some database. Accordingly, we can analyze accurately measured data for all trips between any two stations [5]. What can be predicted in this case? If we look at the metro in Moscow, then their passengers are “marked” only at the entrance. The final destination of the route must be estimated by some heuristics [6]. For example, after the “first” use of a transport card (ticket), you can try to determine at which station the same card was used for the “second” time. With high probability, this station was the ultimate in the trip [7].

But, in fact, telecommunication operators come to the rescue here. Support for mobile communications in the metro means the presence of base stations there. And this, in turn, means that the operator can record the moments when his subscriber (mobile device) switched to the underground station (that is, entered the metro) and when, on the contrary, switched from the underground station to the ground (that is, left Metro). Based on this, the operator knows the starting and ending points of any trip. Further, these data can be aggregated over time, so that the possibility of tracking a single trip is excluded, and the result is the so-called correspondence matrix (OD matrix or origin – destination matrix) [8]. Such a matrix is tied to a specific time interval (the period for which the data was aggregated) and shows the number of passengers who moved from one metro station to another.
Again, how can flows be predicted in such a system and, most importantly, why, if they are, are actually completely measured? It is one thing when the correspondence matrix is the final part of the task of analyzing traffic flows (that is, when it is necessary to understand exactly how people move in the city), and another thing is when such a matrix is the initial data [9].

Flows (trips) can, of course, somehow change during the rebuilding of the network, the emergence of new bus routes from the suburbs that bring new passengers to the final metro stations, opening transfer hubs with the city railway, etc. But information about all these events is missing in the correspondence matrix. The correspondence matrix will change under the influence of these events. This matrix will be an indicator of these events. This is the main idea of its analysis - an assessment of the effect of events that occurred in the city. Alternatively, these changes are notifications of changes in transport behavior for which you need to determine the reason. In general, both of these directions can be described as an analysis of the transport behavior of passengers in the city [10]. Transport behavior is one of the main characteristics of the Smart City [11].

The data (facts) obtained through the analysis of the correspondence matrix themselves serve as initial information for forecasting flows. For example, the operating mode of a new interchange line for the city railway is obviously determined by the characteristics of the passenger flow at the corresponding metro station, etc.

Note that, according to a similar scheme, the data of telecommunication operators allow the construction of correspondence matrices, tied simply to geographical squares. That is, you can divide the city (with the suburbs) with a geographic grid (the achievable size, depending on the density of the base stations, is about 500 x 500 m) and get the number of movements between arbitrary squares. Naturally, these movements must somehow be tied to transport, in particular, be present in the displacement matrix for the metro.

Correspondence matrices for the metro (railway) are easier to process, because here, for obvious reasons, the routes are fixed, we do not consider what happens during the movement itself, and in the models used there is only one type of allocated objects - stations.

Accordingly, the above considerations allow us to note that in modern conditions in the Smart City, traffic flows are completely measurable. They are no longer subject to prediction. There is no need to predict what exactly is measured.

If predictive analytics is no longer relevant, then other considerations come to the fore. In particular, how stable are the data (measurements) received from operators [12]? If the correspondence matrix describes transport behavior, then a change in this matrix means a change in this behavior. Such a change, of course, should be some kind of urban-related explanation (the change should be due to processes or events in the life of the city). Changes can be regular or one-time. This article is devoted to the analysis of the latter (analysis of anomalies).

2. On anomalies

In general, anomalies are defined as a departure from some standard, ordinary, expected behaviour [13]. It is clear that for the city and its transport system, anomalies can lead to the death of people or property if they are not identified and do not respond to them properly. Any forms of automatic notification of anomalies, if any, are sent at the beginning of the process or even before it begins (predicting anomalous behavior) are of great value for city residents and users of transport systems [14].

Naturally, in a modern smart city, we are dealing with different data sets. From here, in fact, the term follows - data-driven cities. Transport systems include a fairly large part of these same data. What may relate to such data:

- Information on the use of transport cards. An event is the presentation of a transport card. The attributes of such an event are the identification of the transport card, the time of its presentation and the place of its presentation, which can be expressed as a geographical location (for example, a railway station) or as a vehicle (for example, a bus, tram, etc.).
• Data from surveillance cameras. The number of cameras in cities is growing, respectively, the number of data they collect is growing.

• The trajectories. Generally speaking, trajectories are a sequence of temporary records that are labeled with coordinates. Each record can contain, generally speaking, a certain set of data related to a given trajectory and characterizing the selected point in time. Naturally, such data can be collected in different ways. For example, in modern conditions, insurance telematics assumes that the insurance company collects data on driving to calculate payments depending on the driving style. Such data collection will also be a trajectory recorder.

• Environmental data. More precisely, this could be called context data. Users can “mark” (mark their location) in the records of various social networks. Theoretically, this is also some kind of trajectory.

In practice, it is, of course, necessary to take into account the practical aspects of obtaining data and, what is equally important, the possibility of obtaining such data for analysis by the systems (services) of the Smart City.

The collection of transport card data is the basis for the functioning of the carrier’s business and, in principle, is a reliable source of data, which is usually available to city services. The main data for the description of transport behavior remain trajectories. In the classical sense, time samples of the trajectory will contain the coordinates of the vehicle. The trajectories can be grouped in the spatio-temporal aspect combining in time all the intermediate points in a certain neighborhood.

Trajectories can also be described for vehicles that use fixed routes (e.g., city railways or metro). In this case, a measurement in a time series, for example, is the name of the station (stop) and the number of passengers entering and leaving. And such data can be combined in the same way.

Now, if we return to the correspondence matrices, then they can be considered as aggregated information about the trajectories. And in this regard, correspondence matrices collected by telecommunication operators can be considered as the most accessible source of information about the trajectories of the Smart City

3. On anomaly measurement

For check-in with transport cards, collected measurements (e.g. user passages at the same railway stations) are the classic time series. The search for anomalies in the time series is a known and well-developed area.

In academic papers, a search for outliers (outline) [15] or a fixation (definition) of new objects (novelty detection) [16] are distinguished. In any case, this all applies to measurements that differ in their characteristics from other measurements. If we explain the difference between outliers and new measurements (counts) from the point of view of the currently popular approaches to machine learning, then outliers are different measurements that are present in the training sample, and new dimensions are those that are not yet available. From an urban point of view, emissions for transport data are some one-time events. They may be associated with some exclusive events. For example, a concert of a popular group at the stadium causes unusual (peak, several times higher than usual for a given time) activity at neighboring metro stations, blocking the road due to an accident causes activity on neighboring streets, etc. Measurement errors also fall into this section of the anomalies (Fig. 1).

New measurements (novelty detection), on the other hand, are the result of some new “normality”, a consequence of a change in operating modes (modes of use of the transport system- Fig. 2). For example, the opening of a bus station on the outskirts of Moscow led to a change in the use of the nearest metro station.
Here we can mention two main approaches that we used in practical problems. Firstly, these are statistical tests. These are well-known calculations:

\[ Z\text{-value} = \frac{x - \mu}{\sigma} \]  

(1)

In this definition:
- \( x \) – measured value
- \( \mu \) – mean value
- \( \sigma \) – standard deviation

The Z-value shows how many standard deviations the measured value differs from the average [17]. The idea is that, for example, for a normal distribution, most of the values will lie in a fairly limited area (measured in standard deviations). Exits abroad of this area are anomalies.

Another approach uses prediction models for time series. The idea is that if the measured value falls outside the confidence interval for the forecast, then this is an outlier [18].

To search for anomalies in the correspondence matrix, we will proceed from the following. The correspondence matrix is always built for a certain time interval. And for this time interval, this matrix describes a certain displacement pattern. To determine the anomalies, it is necessary to compare such patterns at different time intervals. For example, a certain time in the morning during the month, noon on weekdays and weekends, etc.
To compare the matrices, we will use the Frobenius norm \([19]\). The Frobenius norm, or Euclidean norm, is a special case of the p-norm for the case \(p = 2\):

\[
\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}
\]  

(2)

If we have correspondence matrices \(C_i\) for time \(T_i\) and \(C_j\) for time \(T_j\), then to compare the flows, it is necessary to measure the distance \(d(C_i, C_j)\). The Frobenius norm, in this case, will correspond to the pairwise comparison of flows for the same locations:

\[
\|C_i - C_j\|_F = \sqrt{Tr[(C_i - C_j)(C_i - C_j)^T]}
\]  

(3)

where \(T\) is the transpose of the matrix, and \(Tr\) is the trace of the matrix.

If we have a sequence of correspondence matrices \(C = \{C_1, C_2, \ldots, C_n\}\) for a sequence of time intervals \(T = \{T_1, T_2, \ldots, T_n\}\), then the grouping (clustering) of such matrices should give time intervals with similar flows. Using the norm described above, we can obtain the distance matrix \(DF\),

The element \((i, j)\) of this matrix is \(d(C_i, C_j) = \|C_i - C_j\|_F\)  

(4)

This matrix can be normalized \(\text{norm}(DF) = (DF - \text{min}(DF))/(\text{max}(DF) - \text{min}(DF))\), where \(\text{min}(DF)\) and \(\text{max}(DF)\) are, respectively, the minimum and maximum values in \(DF\).

The distance matrix obtained in this way is the basis for building clusters. Known approaches can be used for this, for example, DBSCAN and others [20].

4. Conclusion
The paper deals with the search for anomalies in transport data. If there is data on the movements of mobile users collected by telecommunication operators, the task of forecasting traffic flows in its classical form no longer exists. All traffic flows become known (measurable). Fully measurable flows play the role of metrics (indicators) in the Smart City. These flows change in accordance with some events (changes) in urban life. On this basis, the search for anomalies in such data becomes an important point in understanding the processes taking place in the Smart City. The paper presents an anomaly search algorithm based on data from correspondence matrices collected by mobile operators.

5. References
[1] Nagy A M & Simon V 2018 Survey on traffic prediction in smart cities Pervasive and Mobile Computing 50 148-163
[2] Chen X, & Chen R 2019, October A Review on Traffic Prediction Methods for Intelligent Transportation System in Smart Cities In 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) pp 1-5 IEEE
[3] Pan G, Qi G, Zhang W, Li S, Wu Z, Yang L T 2013 Trace analysis and mining for smart cities: issues, methods, and applications IEEE Communications Magazine 51(6) 120-126
[4] Castro P S, Zhang D, Chen C, Li S, & Pan G 2013 From taxi GPS traces to social and community dynamics: A survey ACM Computing Surveys (CSUR) 46(2) 1-34
[5] Namiot D, Pokusaev O, Lazutkina V 2018 On passenger flow data models for urban railways International Journal of Open Information Technologies 6(3) 9-14
[6] Misharin A, Namiot D, & Pokusaev O 2018, August On Passenger Flow Estimation for new Urban Railways In IOP Publishing Vol 177 1
[7] Namiot D, Sneps-Sneppe M 2017 July A Survey of Smart Cards Data Mining In AIST
(Supplement) pp 314-325

[8] Misharin A, Namiot D & Pokusaev O 2019, October On Processing of Correspondence Matrices in Transport Systems In 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon) pp. 1-6 IEEE

[9] Sánchez L E A, González-Enrique J, Ruiz-Aguilar J J, Moscoso-López J A, Turias I J 2019, October OD Mobility Estimation Using Artificial Neural Networks. In International Congress on Engineering and Sustainability in the XXI Century pp 643-652 (Springer, Cham)

[10] Benevolo C, Dameri R P, D’auria B 2016 Smart mobility in smart city In Empowering Organizations pp. 13-28 (Springer, Cham)

[11] Papa E, Lauwers D 2015 Smart mobility: opportunity or threat to innovate places and cities In 20th international conference on urban planning and regional development in the information society (REAL CORP 2015) pp 543-550

[12] Namiot D, Nekraplonna M, Pokusaev O, Chekmarev A 2020 OD-matrix and passenger flow analysis International Journal of Open Information Technologies 8(4) 25-30

[13] Djenouri Y, Belhadi A, Lin J C W, Djenouri D, Cano A 2019 A survey on urban traffic anomalies detection algorithms IEEE Access 7 12192-12205

[14] Zhang M, Li T, Yu Y, Li Y, Hui P, Zheng Y 2020 Urban Anomaly Analytics: Description, Detection and Prediction IEEE Transactions on Big Data

[15] Gupta M, Gao J, Aggarwal C C, Han J 2013 Outlier detection for temporal data: A survey. IEEE Transactions on Knowledge and data Engineering 26(9) 2250-2267

[16] Pimentel M A, Clifton D A, Clifton L, Tarassenko L 2014 A review of novelty detection Signal Processing 99 215-249

[17] Pang L X, Chawla S, Liu W, & Zheng Y 2011, December On mining anomalous patterns in road traffic streams In International conference on advanced data mining and applications pp 237-251 (Springer, Berlin, Heidelberg)

[18] Yu Y, Zhu Y, Li S, Wan D 2014 Time series outlier detection based on sliding window prediction Mathematical problems in Engineering

[19] Werman M & Weinshall D 1994, October Similarity and affine distance between 2D point sets In Proceedings of 12th International Conference on Pattern Recognition Vol 1 pp 723-725 IEEE

[20] Birant D & Kut A 2007 ST-DBSCAN: An algorithm for clustering spatial–temporal data Data & knowledge engineering 60(1) 208-221