Anomaly Detection Method for Aggregated Cellular Operator Data

Mark Bulygin, Dmitry Namiot
Moscow State University
Moscow, Russia
messimm@yandex.ru, dnamiot@gmail.com

Abstract-**According to the research of the international agency "We are social" in April 2020, there are 5.16 billion mobile phone users in the world. This is 66% of the total population. This article is devoted to the search for anomalies in traffic data received from cellular operators. The first part of the article tells about what data is collected by cellular operators, how they are processed. A brief overview of methods for analyzing these data is provided. Further, the structure of the data that is collected from cellular operators by a special department of the city of Moscow is presented. In the second part of the article, methods for anomaly detection are analyzed, taking into account the specifics of the available data, and the advantages and disadvantages of its use are revealed. Then, a proprietary method for anomaly detection is proposed, which is based on the properties of traffic data. The work of the proposed method is demonstrated in several computational experiments. Further directions of work in the field of aggregated data analysis of cellular operators are proposed.**

I. Introduction

According to the research of the international agency "We are social" in April 2020, there are 5.16 billion mobile phone users in the world. This is 66% of the total population [1]. The increase in mobile phone users is 128 million users compared to April 2019.

When used, mobile devices generate data. It can be used to solve applied problems. In particular, when making calls, connecting to the Internet, or sending SMS messages, the mobile device exchanges data with the base stations of the cellular operator. The location of the device can be established based on information from different base stations about the signal strength from the device and the signal transmission delay Fig. 1

Wang's article is devoted to a detailed description of the methods, standards, and technologies used in determining the geolocation of the subscriber [2].

Such data can be stored, aggregated, and processed by cellular operators. Non-aggregated data are more detailed and allow researchers to explore individual movement patterns. But this data is large in volume and also has a delay in receiving. Aggregated data contains information collected for specific areas and time intervals. Such data are available with less time lag and allow researchers to study traffic patterns in entire areas. The aggregation process also loses data on individual trajectories. This makes the use of such data confidential and does not violate the privacy of users.

Examples of using the cellular operator data for solving practical applied problems can be found in the article by F. Calabrese "Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome". This article describes software solutions that help researchers find answers to questions about the movement of people and transport in Rome. [3]

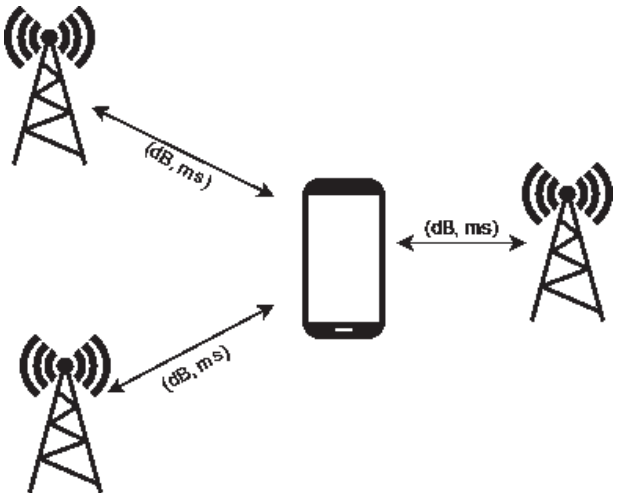

Fig. 1.Exchange of information between the device and base stations

One of the most pressing problems in the analysis of aggregated data of cellular operators is the identification of anomalies in it corresponding to social events. An article by D. Rosario et al. "Anomaly detection mechanisms to find social events using cellular traffic data" is devoted to solving such problems using wavelet analysis [4]

In Moscow and the Moscow Region, there is data, aggregated by district, on the movements of cellular operator subscribers between districts grouped half-hour intervals. Due to the presence of base stations of cellular operators in the Moscow metro, as well as the development of data analysis methods, the data presented in Table I are available for research.

The works of Russian authors [5-8] are devoted to the problems that can be solved using such data.

The purpose of this article is to propose a method for detecting anomalies in such data. Note that the task of the work is to detect anomalies at a given time interval, and not to predict anomalies The traffic flow data collected by cellular operators is an indicator (metric) for events in the city that

affect the movement of people. Based on the traffic flow data, we cannot predict these social events with a high degree of accuracy. Changes in traffic flows are a reflection of these events. Accordingly, for the anomalies found, city services (transport administration, for example) can determine the cause and, most importantly, its nature. Based on this data, they can understand whether this is a one-time event (for example, a concert that caused peak traffic at the nearest metro stations), or is it a new normal state (for example, a new business center has been opened and morning and/ evening traffic at nearby metro stations has changed)

We also note that in the proposed method we explicitly use information about the nature of measurements. So, "normal" traffic flow for time *t* is calculated not as averaging (smoothing) traffic for times preceding $t_1 \dots t_n$, but as averaging (smoothing) traffic for the same times *t*, but on days of such of the same type (weekdays, weekends, etc.) preceding this one (Fig..2). City traffic flow, for example, at 12 noon on weekdays is much more stable than traffic in the 8-12 hour interval on a particular Monday. It does not change much if there are no important social events in the city. The task of our work is to identify such events.

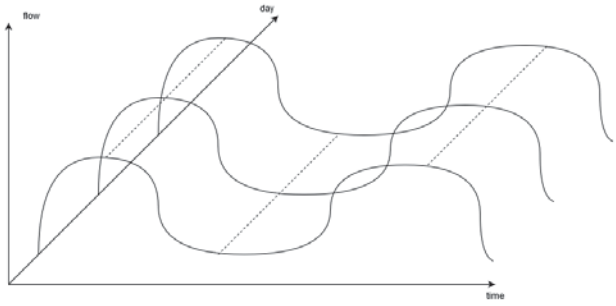

Fig. 2. Traffic flows

Note that the proposed approach automatically takes into account seasonal trends. The data, for example, for June is compared with the days close to them.

It is in this representation of "normality" that the novelty of the work consists. This form of analysis is possible precisely because of the way in which the data is collected.

TABLE I. COLLECTION DATASET

Timestamp	**Departure district**	**Arrival district**	**Count of customers**	**Count of metro customers**
Count of customers, going to work from home	Count of customers, going to home from work	Count of customers going back	Count of customers finishing a trip	Count of customers staying in a district

It is important to emphasize that this is a new approach to transport problems. There is no task of predicting traffic - it is now accurately measured by cellular operators.

It makes no sense to predict what is now being measured. But these measurements now serve as a reflection of "urban" processes in the city. And the study of the stability of these indicators becomes a new transport task. "Violations" this stability - these are social events.

At the beginning of this work, the features of the application of existing methods for time series anomaly detection are considered. The advantages and disadvantages of their application to the traffic flow time series are revealed. Then an algorithm for detecting anomalies is proposed, which takes into account the features of the time series describing traffic flows. The capabilities of its customization are described. The main advantages of its application over existing methods are revealed. Further, a computational experiment is carried out, which shows the possibilities of using this algorithm.

II. MAIN PART

A. Application of existing methods for anomaly detection in data

Having fixed the district of departure and the district of destination, it is possible to obtain the time series of displacements for each pair of traffic flows. The work [9] is devoted to a review of methods for detecting anomalous observations in time series.

In this paper, the author identifies three main groups of methods for anomaly detection in data.

The first group consists of methods based on predictive models. The most popular predictive models for time series are ARIMA models, models based on classical machine learning (random forests, boosting, etc.), as well as models based on neural networks.

Traffic data have significant autocorrelation between current and past observations, as well as weekly observations. This autocorrelation breaks down on holidays, so ARIMA models give false signals of anomalies. For this reason, false signals are possible both a week after the holidays and a week after anomalies. To solve this problem, SARIMAX models can be used, where data on anomalies, holidays, and other events are taken into account as additional features. The process of constructing a SARIMAX model is also complicated, It is necessary to bring the time series to a stationary form, mark up anomalous observations, and also reconfigure the model when novelty appears in the data. Predictive models based on random forests and gradient boosting have the same disadvantages, excluding the need to bring the series to a stationary form. The main advantage of this group of methods is the construction of a predictive model that can be used to solve other problems. In the case of working with aggregated data of cellular operators, this advantage is not important, since real values are available with a low time delay.

The second group of methods, highlighted in [9], consists of methods based on clustering. The most popular method for

this group is to use DBSCAN Clustering with a variable threshold. The main disadvantage of this method, as well as other methods of this group, is the impossibility of processing values that may be typical for one interval, and at the same time anomalous for another interval. For example, a normal weekday rush hour value might be anomalous for a weekend morning interval. Thus, to use the algorithms of this group, it is necessary to construct features to take into account the date and time of observation. The advantages of this group of methods are that there is no need to build a high-quality predictive model and the ability to find anomalous values that are not extreme.

The third group of methods is statistical profiling. The methods of this group give the most controllable and easily interpreted results. Also, the advantages of the methods of this group include the fact that the calculated statistics, as a rule, have a low computational complexity. These statistics can be easily recalculated when novelty appears, for example, when a new transport channel opens. The disadvantages of the methods of this group include the need for the correct choice of statistics. When analyzing data on traffic flows and using the (100-*a*)%-percentile as a statistics, some of the observations will be recognized as anomalous, although they are not. Anomalous values that are not extreme will be ignored when using such statistics. In the case of using the mean or median with some threshold without grouping, some anomalies may also be missed. Traffic flow on weekends for residential districts is less than on weekdays. When taking an average, a value is obtained that is not typical for either a weekday or a weekend. This value is anomalous, but will be skipped by the algorithm.

Summarize the existing advantages and disadvantages of using these groups of methods to detect anomalies in traffic data in the Table II.

B. Proposed Method

Traffic flow data between districts has properties that can be used to detect anomalies. Such time series have a strong autocorrelation: observations of days of the same type (weekdays, weekends) have a strong connection.

Data on intervals with the same time stamps and type of day in the absence of anomalies are not highly oscillatory and are concentrated in the area of average values due to the constancy of traffic flows. An algorithm belonging to statistical profiling is proposed to detect anomalies in time series corresponding to traffic flows.

At the input, the algorithm takes the value of the time series for which the anomaly is checked, the day of observation, the time of observation, the day type of observation, the parameter that regulates the sensitivity to the novelty of the data, the historical data on the values of the time series, as well as the threshold that regulates the sensitivity of recognizing the anomalies. The output is 0 if the observation is typical and 1 if it is anomalous.

The essence of the algorithm is that the modulus of the difference between the checked value and a certain value calculated from the historical data is compared with the threshold passed as a parameter. If this module is greater than the threshold, then the value is considered abnormal, otherwise typical. The scheme of the algorithm is shown in the Fig.2.

TABLE II. ON ANOMALIES DETECTION METHODS

Group of methods	Advantages	Disadvantages
Methods based on predictive models	The predictive model makes it possible to predict future observations, the ability to take into account additional features	The complexity of building a high-quality predictive model, the need to refit predictive models when novelty appears in the data
Methods based on clustering	Does not require data tagging and complex predictive model building, can find anomalous observations that are not extreme or are close to the mean	The complexity of accounting for signs of date and time. Do not cope in cases where the same value can be both typical and anomalous depending on the date and time
Statistical profiling	Simplicity and clarity of the results obtained, usually statistics are fairly easy to calculate	They do not cope in cases where the same value can be both typical and anomalous depending on the date and time, do not cope with anomalous values that are not extreme

The essence of the algorithm is that the modulus of the difference between the checked value and a certain value calculated from the historical data is compared with the threshold passed as a parameter. If this module is greater than the threshold, then the value is considered abnormal, otherwise typical. The scheme of the algorithm is shown in the Fig.3.

More formally, the algorithm can be presented in two steps.

At the first step, for the checked interval by day, time, day type, historical data, and the count of days taken into account, the average is calculated over intervals with the same time and type of observation day. The calculation is carried out according to the formula (1).

$$hist\_est(day, time, type, delay, hist) = \frac{\sum_{i=delay}^{i=1} [get\_type(i) = type] * hist(i, time)}{\sum_{i=delay}^{i=1} [get\_type(i) = type]} \quad (1)$$

where *day* is the observation day for which the anomaly is checked, *time* is the half-hour observation interval for which the anomaly is checked, *type* is the type of observation day for which the anomaly is checked, *delay* is the historical data review period in days, and *hist* is the historical data, and *get_type* - a function that returns the type of day (weekday or weekend), *hist_est* - the calculated value.

In the second step, the observation anomalousness is checked using the value found in the first step. The modulus of the difference between this value and the tested *value* is found, and then it is compared with the threshold value *th*. Based on the comparison result, a conclusion is made about the

abnormality of the observation. These actions are described by the formula (2)

$$abnormality(value, hist\_est, th) = [|hist\_est - val| > th] \quad (2)$$

where *value* is the value for which the anomaly is checked, *hist_est* is the value calculated in the previous step, *th* is the threshold value, *abnormality* is the anomaly label of the value

When using this algorithm, two parameters are selected: *th* - the threshold, and also *delay* - the number of days to view historical data.

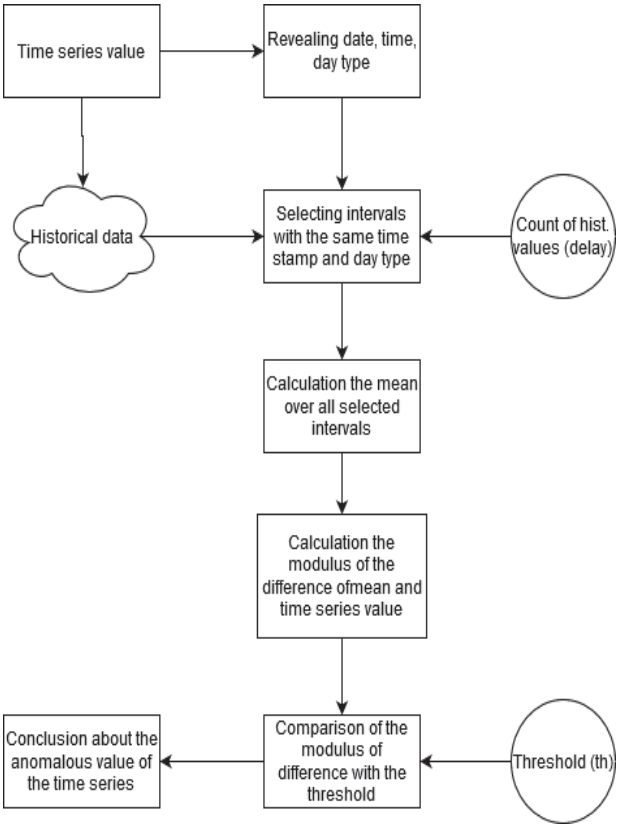

Fig. 3. Scheme of the algorithm

The *delay* parameter should be selected greater than or equal to 7, since otherwise there may not be days off in the considered period of historical data. In the process of finding the historical value, in this case, division by 0 may occur. The recommended value of this parameter is 30, which corresponds to the length of calendar months. The larger the given value of the parameter, the smaller the impact of one observation on the mean. If this parameter is selected equal to 7, the average value is based on the data of the last week, the novelty in the data, which appeared a week ago, is considered typical. Algorithm readings with this value of the parameter quickly adapt to new conditions. This can be a significant advantage of the algorithm if anomalies are estimated in districts where new transport channels are being actively introduced. In areas with a formed transport system, this parameter should be selected equal to 30 or more. In this case, individual observations associated with short-term anomalies, such as major road accidents or road repairs, have less impact on the mean, and the algorithm does not adapt to them, detecting false anomalies.

Adding new day types is another way to customize the algorithm. The basic version of the algorithm uses two types of days: weekdays and weekends. Distinguishing Saturdays as special types of days allows researchers to recognize anomalies better in districts with large educational institutions, since Saturday is also a workday in it. During significant social events, such as fireworks and major concerts, planned deviations in traffic flows can occur. If specialists need to receive data on anomalies taking into account such events, then such types of days as "Concert Day", "Fireworks Day" and others should be distinguished. Data for special types of days should be stored separately. This allows researchers not to increase the delay parameter up to large values.

The *th* value can be used to adjust the sensitivity of the algorithm. The *th* value can be a constant specified by a subject matter expert. In this case, the intervals are recognized as anomalous, the values in which differ from the mean over intervals with the same timestamp and day type by more than this constant. This concept of anomaly can be easily grasped by a subject matter expert. Districts vary in size. For some districts, a deviation of 100 people from the mean in outgoing traffic is not significant, while for other districts it is a significant anomaly. In such cases, the researcher can express the threshold value through *hist_est*. For example, if the threshold is chosen equal to *0.3 * hist_est*, then this means that half-hour intervals are recognized as anomalous, in which the observed values are 30% higher than the mean for intervals with the same timestamp and day type. Note that when analyzing metro data at night, there are nonzero observations, while the mean close to zero. Such observations are rare and most often associated with the internal work of the metro employees. If urbanists do not need to recognize these works as anomalies, then some constant should be added to the threshold value.

Historical data values *hist_est*, calculated using formula (13), can be used to solve other applied problems of urban studies. By calculating the *hist_est* values for all 48 half-hour intervals of a weekday and a weekend from the month data, researchers can store this data and use it to quantify the change in traffic between districts in different months. With the help of such averaged data on weekdays and weekends, for example, it is possible to estimate the level of self-isolation of the population by city districts during the COVID-19 pandemic.

The algorithm described above has advantages over the existing algorithms for detecting anomalies in time series. Compared to methods based on predictive models, this method is easier to implement, since it does not require building a predictive model, and also adapts to the novelty in the data. For the algorithm to work, the marking of anomalous observations for all districts is not required. The results of the algorithm can be easily interpreted and explained to a subject matter specialist. An important advantage over the use of statistical profiling and clustering-based methods is the ability to find anomalous values that are anomalous only in the context of date and time. The proposed profiling method

correctly signals such anomalies, since it is applied on a sample of data with the same time stamps and day type. It should be noted that the application of methods based on clustering on the same samples can also give a successful result. The solution based on profiling is chosen because of the broader possibilities in the interpretation and explanation of the results obtained. The advantage of clustering-based methods for detecting anomalies with values close to the mean is not significant due to the specifics of the data.

C. Application of the algorithm and analysis of the results

To check the correctness of the proposed algorithm for detecting anomalies, a computational experiment is carried out. In the course of this experiment, the proposed algorithm for detecting anomalies is fed into the input data on total traffic flows from all districts of Moscow, as well as metro traffic flows for May 2017. To assess the quality of detecting anomalies, the algorithm's response to known anomalies was checked, such as the festive events of May 1, 2017, the Immortal Regiment campaign, which took place on the territory of Begovoy and Tverskoy districts on May 9, 2017, and holiday fireworks. Since in most areas traffic flows are quite stable, the quality of detecting anomalies is also checked visually using graphs. For a more detailed assessment of the quality of the algorithm, as well as its tuning, it is necessary to involve experts in the subject field.

After the calculations, the results are analyzed. Below is the result of finding anomalies in the total traffic flow from the Begovoy district to the Tverskoy district by two methods of statistical profiling: the proposed algorithm and an algorithm based on the calculation of two statistics: 95%-percentile and 5%-percentile.

In Fig. 4 and Fig. 5, the half-hour intervals of May 2017 are marked on the horizontal axis. The vertical axis shows the number of people moving from the Begovoy district to the Tverskoy district. dots on the graphs mark the anomalies found using the proposed algorithm (Fig. 4), as well as those found using the method based on comparison with the 5% percentile and 95% percentile (Fig. 5).

The proposed algorithm coped with the task better. All peaks corresponding to a known anomalous event are highlighted, the start of the observation anomaly signals corresponds to the start date of the event, while when applying the percentile-based profiling algorithm, the signals are generated at the peak of the anomaly. Also, the proposed algorithm does not detect false anomalies in low values of night traffic. Let's consider the results of the functioning of the proposed algorithm for other districts.

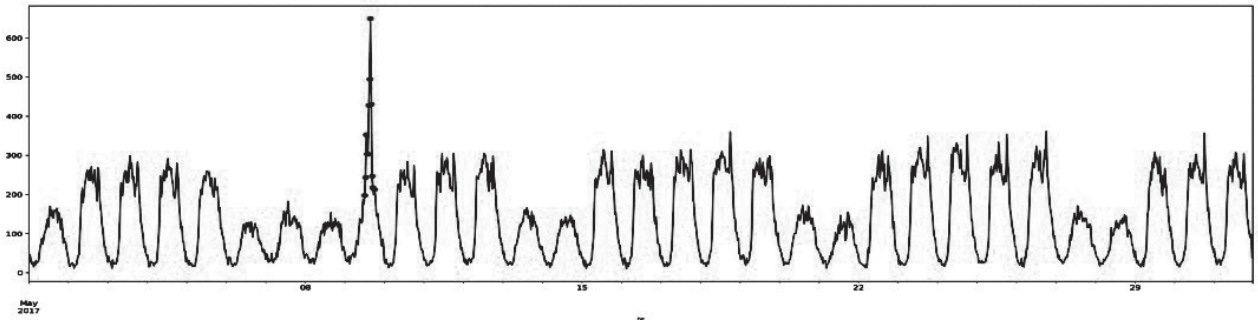

Fig. 4. Anomaly detecting by the proposed algorithm

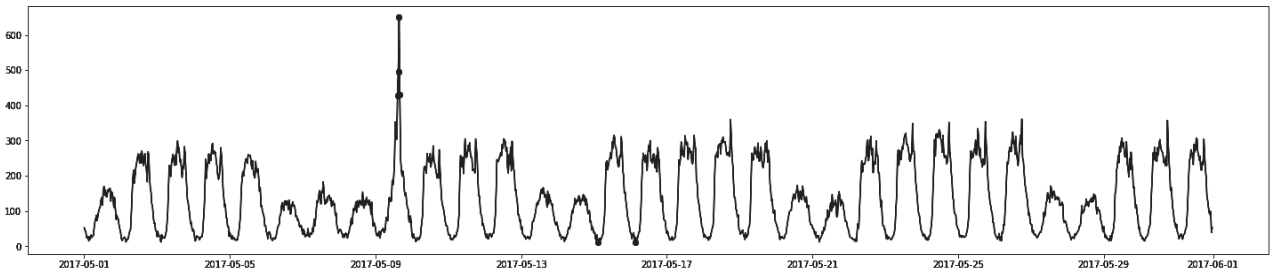

Fig. 5. Anomaly detecting by the algorithm based on comparison with 95% percentile

The result of detecting anomalies in the flow from the Maryino district to the Tverskoy district is presented (Fig. 6).

The Fig.6 shows that corresponding to the celebrations on May 1st and 9th anomalies are highlighted. No false positives are observed. It is important to note that values found to be anomalous are the only abnormal in the context of date and time. For example, the first anomalous value is not anomalous for a weekend, but is anomalous for that particular time period. Algorithms that do not take into account the relationship of data with time, for example, methods based on clustering, cannot cope with identifying anomalies of this type.

One of the events that attract many people is the May 9th fireworks. One of the best platforms for viewing fireworks is the observation deck located on Vorobyovy Gory in the Ramenki district. Anomalies are also observed in the time series corresponding to movements from neighboring districts to the Ramenki. An example is the identification of anomalies

in the flow from the Akademichesky district to the Ramenki district (Fig. 7).

On this graph, the horizontal axis shows the half-hour intervals of May 2017. The vertical axis shows the number of people moving from Akademichesky to Ramenki. Dots on the graphs mark the anomalies found using the proposed algorithm. The anomalies corresponding to the fireworks are identified correctly.

Dots on the graphs mark the anomalies found using the proposed algorithm. The anomalies corresponding to the fireworks are identified correctly. Several anomalies have also been identified, corresponding to high levels of movement between districts on Saturdays. Perhaps these levels are due to the fact that Moscow State University is located in the Ramenki district, where Saturday is a workday. To avoid responding to events that are not abnormal, the researcher can either change the threshold setting or introduce a new type of day, "Saturday". This day type may be relevant for the analysis of districts where large educational institutions are located, since Saturday is a workday in it. In this case, no modification of the algorithm is required; the researcher may have to change the sensitivity settings of the algorithm.

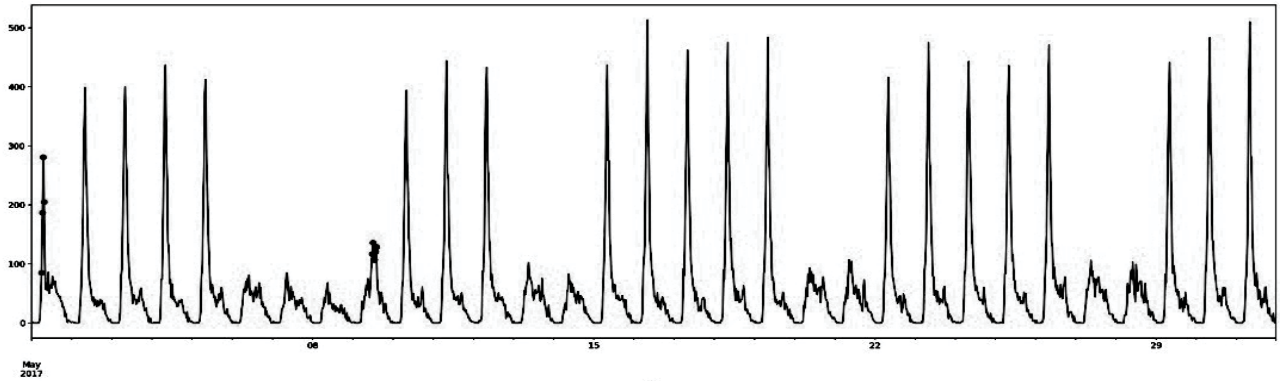

Fig. 6. Anomaly detecting in the flow from the Maryino district to the Tverskoy district

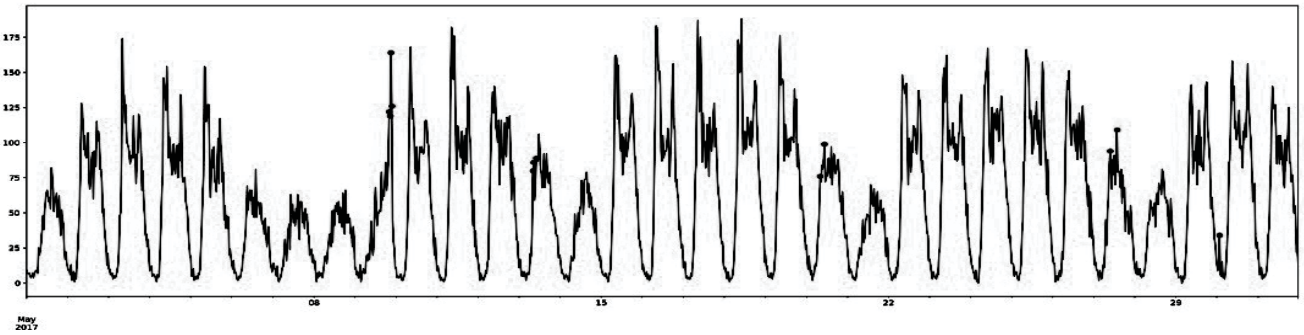

Fig. 7. Anomaly detecting in the flow from the Akademichesky district to the Ramenki district

In general, the algorithm demonstrates its correctness. All identified anomalies can be explained, but for a complete understanding, it is necessary to involve experts in the subject area. To identify anomalies in all areas, the same settings are used, the length of the considered period is 30 days, the threshold is the value calculated from historical data multiplied by 0.3 and increased by 5. The last increase in the value calculated from historical data is necessary. so that anomalies in the movement of the metro at night are not detected (the average for such intervals is close to 0).

It is necessary to obtain anomaly labeling from a specialist in the field of urbanism to estimate the quality of algorithms for detecting anomalies, in which 0 corresponds to typical values, and 1 to anomalous, for several time series. Further, using this labeling as ground truth data, the accuracy, precision, recall, and F1-measure can be calculated. These metrics can be used to assess the quality of the algorithm with the specified parameters, select the best parameters, and also compare with other algorithms for detecting anomalies.

D. Future research

After detecting anomalies in the time series describing traffic flows, it becomes possible to study them. The clustering of the found anomalies may be of practical interest for urbanists. For anomalies of various types, the features of appearance, development, and also their completion can be identified. Such data can help improve procedures for responding to important social events.

During the operation of the anomaly detection algorithm, typical traffic flow values are calculated for each pair of districts at half-hour intervals. Analysis of changes in these values when creating new transport channels can help to quantify the effectiveness of its work. At the same time, there can be observed both an increase in the flow for all time intervals (people began to move more often between the study districts), and its shift in time with the same quantities (people began to move between areas faster, so the start time of the trip shifted to a later time)

Aggregated data of cellular operators on traffic flows can be used to find out the distribution of citizens leaving or entering the area. Changes in these distributions can occur as a result of the construction of new places of residence, social attractors, or new transport channels. Urbanists need to respond to these changes in order to timely and correctly change the city's transport network.

Data on traffic flow from cellular operators can be used not only for solving urban problems. This data contains a lot of information about the districts of the city and the connections between them. Such information can be useful in solving many applied problems such as real estate appraisal, choosing a location for opening an office, and so on. However, such aggregated data is too large and has too many features, therefore, by itself, in its raw form, it is not suitable for inclusion in machine learning models. From this data, it is necessary to extract features that can be useful in solving an applied problem.

The available data allows researchers to estimate the number of people working in the district, for example, as the average number of people staying in the district for at least an hour from 10 to 18 hours on weekdays. The number of people living in an area can be estimated as the average number of people in the area at night. It is cheaper to obtain such an estimate than in a census, and data may be available for each half-hour interval of weekdays and weekends if necessary.

Based on the available data, it is possible to carry out clustering of areas. The cluster of a region can be used as a feature for solving applied machine learning problems. It is important to note that the clustering of the observations themselves is of no practical interest. On the basis of the data, it is necessary to form features that describe the areas, and then cluster in the space of these features using existing methods.

III. Conclusion

In the process of the study, an algorithm is proposed for anomaly detection in traffic data. This algorithm takes into account the features of the data, which gives it advantages over the use of already known machine learning methods. The resulting algorithm is simple enough to set up and allows researchers to adjust the sensitivity, and also automatically adapts to the novelty in the data. Further research will focus on the study of traffic distributions, which can help select the best statistics and thresholds for the existing profiling method. Another area of further research may be the study of the found anomalies themselves, an understanding of their types, processes of origin, development, and disappearance.

References

[1] Report of the international agency “We are social” Web: https://digitalreport.wearesocial.com/ Retrieved: Nov, 2020
[2] Wang, S., Min, J., and Yi, B. Location based services for mobiles: Technologies and standards. In IEEE ICC. Beijing
[3] Calabrese, Francesco, et al. "Real-time urban monitoring using cell phones: A case study in Rome." IEEE Transactions on Intelligent Transportation Systems 12.1 (2010): 141-151.
[4] Garroppo, Rosario & Niccolini, Saverio. (2017). Anomaly detection mechanisms to find social events using cellular traffic data. Computer Communications. 116.10.1016 / j.comcom.2017.12.009.
[5] Namiot, Dmitry, and Manfred Sneps-Sneppe. "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. 2017
[6] Namiot, D. E., O. N. Pokusaev, and V. S. Lazutkina. "On models of passenger flow for urban railways." International Journal of Open Information Technologies 6.3 (2018)
[7] Namiot, Dmitry, Oleg Pokusaev, and Vasily Kupriyanovsky. "Data Mining on the Use of Railway Stations." AIST (Supplement). 2018.
[8] Nekraplennaya, M.N., and D.E. Namiot. "Analysis of metro correspondence matrices." International Journal of Open Information Technologies 7.7 (2019).
[9] Effective Approaches for Time Series Anomaly Detection. Web: https://towardsdatascience.com/effective-approaches-for-time-series-anomaly-detection-9485b40077f1 Retrieved: Nov, 2020