A New Approach to Clustering Districts and Connections Between Them Based on Cellular Operator Data

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

Abstract–According to the Digital 2021 report of the We Are Social agency, by January 2021, the population of the Earth totaled more than 7.83 billion people with a population urbanization rate of 56.4%. According to reports from previous years, the proportion of the population living in cities is steadily increasing, so it is very important to make the life of city dwellers comfortable. The article describes an approach to clustering city districts and connections between them based on features constructed according to the data of cellular operators. A demonstration of the application of this approach for clustering Moscow districts and territorial units of the Moscow region, as well as connections between them, is performed. The clustering results, as well as the approach described in the article, can be used by urbanists to build infrastructure in areas following the behavior of its inhabitants.

I. Introduction

According to the «Digital 2021» report of the «We Are Social» agency, by January 2021, the population of the Earth totaled more than 7.83 billion people with a population urbanization rate of 56.4%. According to the reports of previous years, the rate of the population living in cities increases steadily, so it is very important to make the lives of city residents comfortable [1].

One of the concepts for improving the quality of life in cities is the smart city concept. It consists of the management of city assets and city property with a high degree of use of information technology and IoT technologies. In particular, to analyze the quality of the city's transport system within the framework of this concept, the concept of transport fatigue (time on the road, travel convenience, etc.) was introduced [2].

The data of cellular operators can be used for a smart city infrastructure building. During their operation, cell phones communicate with base stations. The location of the device can be determined based on the data on the latency and signal strength from the device from different base stations. The article [3] is devoted to the standards and specifications for determining the location of cellular devices.

Device location data can be processed (aggregated by location and time, for example) by cellular operators. For the Moscow region, data aggregated by districts and half-hour intervals are available. Due to the high presence of cellular operators in the metro, data on trips using the metro can be obtained. All available data are presented in Table I.

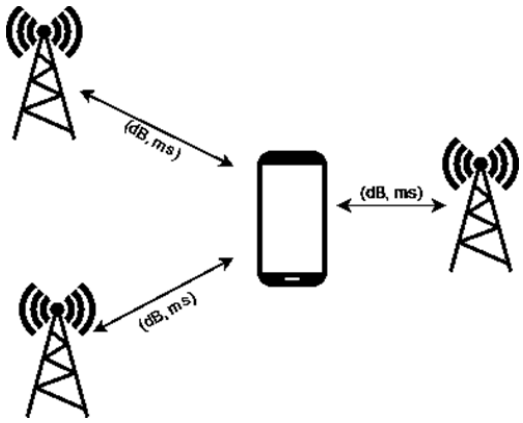

Fig. 1.Operation of cellular devices

TABLE I. Collection Dataset

Field	Description
TS	Timestamp
Departure_zid	Departure area identifier
Arrival_zid	Arrival area identifier
Customers_cnt	The total number of people who started a trip at a given half-hour interval from district departure_zid to district arrival_zid
Customers_cnt_metro	The number of people who started a trip in a given half-hour interval from district departure_zid to district arrival_zid, (the metro was used during the trip)
Customers_cnt_home_work	The number of people who started a trip at a given half-hour interval from district departure_zid (from home) to district arrival_zid (to work)
Customers_cnt_static	The number of people who are in the district for at least 60 minutes and have not made trips in a given half-hour interval

This paper aims to propose a new approach for the formal description and comparison of urban districts. Urban

infrastructure differs from district to district, and it should correspond to how these spaces are used by residents (office centers, sleeping hoods, etc.).

This article proposes an approach to feature engineering for describing city districts and the connections between them. The paper presents two feature spaces built on the data of cellular operators that describe the regions, as well as three feature spaces that characterize the connections between them. In the proposed feature spaces, the clustering of regions and connections between them was carried out. Such clustering may be of interest to specialists in the field of urban studies since the infrastructure of a city in different types of districts (i.e., districts from different clusters) should be different.

Including the type of district (the number of the cluster to which it belongs) can also help data analysts with other city data mining applications. This feature allows analysts to take into account data on the region or its connections in the form of one categorical feature.

A. On data sources in urban studies

The engineering of features that describe areas and the relationship between them, is based on the initial data on the areas. Currently, data from population censuses, data from migration services, as well as data from social surveys are used to describe districts. The population census is an expensive and rare undertaking (the last census in Moscow was carried out in 2010), while the population of the city does not have a static distribution over such long time intervals. Social surveys cover only a small portion of the population and may not be representative. Social survey data are also prone to biases, as people can misrepresent information when taking surveys.

The task of counting the number of people living in the city is relevant for a long time. One of the articles describing the methods of counting the population in the days before the widespread penetration of information technology is an article [4]. In this work, the authors propose a method that gives an estimate of the population of the district according to the formula (1)

$$P = (H * PPH) + GQ, \quad (1)$$

where H is the number of employed households, PPH is the average number of people in one household, and GQ is the group quarters population (the latter is understood as common residences such as nursing homes, student and worker dormitories, barracks, etc. - everything that does not apply to ordinary household).

The authors compare their methodology with three other methods of the time. They show that it performs better than others on many examples, but not all. The main disadvantage of this methodology is the necessity of accurate estimation of the number of employed households and the average number of people in one household. The errors of this method are quite large due to differences in household size.

Currently, there are many methods for estimating the number of residents of districts based on machine learning algorithms. In [5], satellite images of cities are used to assess the distribution and size of their population. The paper [6] presents a population size estimation method based on electricity consumption. This method offers high spatial accuracy (up to buildings) as well as high temporal resolution. These methods show higher accuracy than the method based on the number of households. These methods are guided by indirect signs, so the error is still quite large.

Using data from cellular operators, an estimate of the number of people in the area for each half-hour interval can be established. The cost of obtaining data, in this case, is less than that of a population census. The received data is not biased because of false information. The information is quite accurate due to the high penetration rate of mobile phones in cities. Note that mobile phones are personal. Such an indirect indicator as the number of cellular subscribers provides opportunities for a more accurate assessment than the amount of electricity consumed or data from satellite images. Due to its high temporal resolution, this kind of data makes it possible to estimate the number of people in the district at different time intervals on weekdays and weekends. This opportunity allows researchers to build more complex and informative features describing the districts.

The data of cellular operators also contains detailed data not only on the number of people in the district but on the traffic flows between them. Previously, this information was not available, and the main task was to build estimates of traffic flows. With the advent of data from cellular operators, traffic flows between each pair of districts are measured. Also, due to the development of the infrastructure of cellular operators in the metro, the traffic flow of the metro is measured. This allows researchers to build informative features describing the connections between districts. Our previous article is devoted to the detection of anomalies in traffic flows based on such data [7].

B. Feature engineering

Aggregated cellular operator data on traffic flows and the number of citizens staying in a district for at least 30 minutes has a large volume and represents a set of time series for each connection between city districts. For the convenience of the analysis, it is necessary to extract from these data some features describing the districts and the connections between them. These features can later be used by urbanists and data analysts to solve applied problems.

The article [8] is devoted to the study of connections between districts. The authors proposed a tool for visual analysis of traffic data between districts MobilityGraphs. The data is processed and filtered before rendering. To simplify the data, a mechanism of spatial aggregation is used (districts are combined into regions, and their traffic flows are summed up). The authors apply clustering in time: the time intervals in which the city's traffic flows are similar are combined into one cluster (DBSCAN). The authors use two measurements of the strength of traffic flow: absolute traffic flow (the sum of the number of people who moved from area A to area B and from B to A) and relative (the sum of the ratio of the number of

people who moved from area A to area B to the total number of people arriving in area B and vice versa). The use of such simplifications made it possible to get rid of redundant information when visualizing traffic flows on a map, making them convenient for analysts. The features formed by the authors of this research are well suited for visualization. When solving applied problems, more complex features describing connections than absolute and relative traffic flows between regions / districts may be of interest. Also, the authors of the article do not take into account the directions of traffic flows, which do not allow researchers to identify special types of connections (donor regions, recipient regions, etc.)

Our article describes a new approach to creating features describing districts and connections between them, based on aggregated data from cellular operators. The paper presents the results of clustering of districts and connections based on the obtained features. These features provide a more complete description of the life of districts than data on population distribution and traffic flows obtained by other, both new and classical methods, since data obtained by cellular operators provide opportunities for deeper analysis

II. Proposed method

For digital urban studies, clustering of city districts and connections between them is of interest. To carry out high-quality clustering, it is necessary to construct feature spaces that describe regions (links) and then cluster in them using well-known methods such as k-means [9], DBSCAN [10], or agglomerative clustering [11].

C. District clustering

1) Clustering by the number of people staying in the districts at different times:

Information on the number of people staying in a district during each half-hour interval is available in the data of cellular operators. This information can be used for district clustering. The average number of people in the area is not informative, as people in the city are in different districts at different times. In districts with a large number of enterprises or business centers, people stay in the daytime on weekdays. In sleeping districts, people stay in the evenings and on weekends. To make the descriptions of the districts more informative, two features are used: the average number of people in the district during working hours (from 10:00 to 18:00 on weekdays) and the average number of people in the district at night (from 23:00 to 07: 00). Note that both of these features depend on the size of the population in the district. However, in the case of clustering districts according to the patterns of behavior of their inhabitants, it is worth excluding the influence of this factor. For this, minimax normalization of these two features are carried out. Thus, each district can be described by a point in two-dimensional space, where the first coordinate corresponds to the average number of people in the district during working hours (we call it working rate), and the second to the average number of people in the area at night (we call it night rate).

These features can be calculated using formulas (2) and (3).

$$working\_rate = \frac{\frac{\sum_{i=1}^{n} static[i]*[i \in work\_time]}{\sum_{i=1}^{n}[i \in work\_time]} - \min(static)}{\max(static) - \min(static)} \quad (2)$$

$$night\_rate = \frac{\frac{\sum_{i=1}^{n} static[i]*[i \in night\_time]}{\sum_{i=1}^{n}[i \in night\_time]} - \min(static)}{\max(static) - \min(static)} \quad (3)$$

where *working_rate* is the average number of people in the district during working hours, *night_rate* is the average number of people in the area at night, *n* is the total number of considered time intervals, *night_time* is the number of time intervals corresponding to the night time, *work_time* is the number of intervals corresponding to the working time, *static* is an array of length n containing values describing the number of people in the area at half-hour intervals. The recommended number of intervals for consideration is 1488, this number corresponds to the construction of features according to data for a month (31 days, 48 half-hour intervals).

These features can vary from 0 to 1. The *working_rate* is close to 1 if the average number of people in the district during working hours is close to the maximum number of people in the district. A value close to 1 characterizes a district as the district in which a relatively large number of people work. The value of this feature is close to 0 in districts where relatively few people work. The *night_rate* is close to 1 in residential districts since in them the number of people close to the maximum value is concentrated at night. A feature value close to 0 indicates that people leave the district at night, that is, it is most likely not residential.

The use of this feature space is supposed to cluster districts into different types according to the behavior of its inhabitants. Such space, for example, can be used to cluster areas into residential and work districts. The results of such clustering can be taken into account when assessing the cost of apartments in city districts using machine learning models or when calculating the sufficiency of infrastructure facilities in districts since the needs of different types of districts in such objects differ.

The k-means method is well suited for clustering in such a feature space since there is too much data for agglomerative clustering, the dendrogram for 125 Moscow districts is overloaded and difficult to understand, and DBSCAN combines clusters with fuzzy boundaries (it is assumed that such clusters exist because of the specifics data).

2) District clustering by trip type

The data of cellular operators also contains information on the number of trips from districts to work and the number of trips from districts by metro. After analyzing the morning outflow, it is possible to conclude where people get in the morning (to work or go to another area for a transfer). If people from the area go directly to work, this indicates a good transport connection of this area with the rest. If a large number of people move to another area to make a transfer,

then this may indicate existing transport problems. A large number of metro passengers can be observed in two cases: firstly, if there is a high level of development of the metro transport infrastructure in the area, and, secondly, this may indicate the general well-being of residents of the area who can get by metro due to the impossibility of purchasing cars or ordering a taxi.

These features can be calculated using formulas (4) and (5).

$$working\_percent = \frac{\sum_{i=1}^{n} \frac{work[i]}{customers[i]} * [i \in morning\_time]}{\sum_{i=1}^{n} [i \in morning\_time]} \quad (4)$$

$$metro\_percent = \frac{\sum_{i=1}^{n} \frac{metro[i]}{customers[i]} * [i \in morning\_time]}{\sum_{i=1}^{n} [i \in morning\_time]} \quad (5)$$

where *working_percent* is the average rate of people leaving the district in the morning, *metro_percent* is the number of people who go somewhere using the metro, *n* is the total number of time intervals under consideration, *morning_time* are the numbers of intervals corresponding to morning rush hours, *metro* is an array of length n containing data on departures using the metro at half-hour intervals, *work* is an array of length n containing data on departures to work at half-hour intervals.

The influence of the size of the districts is excluded since the average rates of people under consideration are already normalized by the total number of movements. This space can be used to cluster areas in terms of their transport accessibility.

D. District connections clustering

For digital urban studies, it is of interest to analyze not only the city districts themselves but also the connections between them. Clustering all connections between districts is a rather difficult task, since there are a lot of connections, and in many feature spaces it is impossible to distinguish clusters since the points are evenly distributed. It is possible to identify clusters in the links of one fixed district. In this section, three feature spaces are proposed for clustering links between regions.

1) Clustering by volume of movements and rate of movements to work

The main attributes describing a connected district are the total number of movements and the number of movements to the district for work. The total number of movements describes the tightness of connections between districts. The greatest number of movements is usually observed between neighboring districts, as people living on the borders of such districts move to the neighboring area for shops, public transport stops, sports grounds, etc. It is important to distinguish between the movement of people to work and areas containing objects of interest such as shopping malls. For this reason, clustering needs to take into account the proportion of people moving to work. High transport flow and low flow to work may indicate that the area attracts people on weekends or simply because of its proximity.

The values of the features describing the district connections (for each connected district) can be calculated by the formulas (6) and (7)

$$customers\_rate(hood) = \frac{mean\_customers(hood)}{\max(mean\_customers)} \quad (6)$$

$$work\_rate(hood) = \frac{mean\_work(hood)}{mean\_customers(hood)} \quad (7)$$

where *hood* – connected district, *customers_rate (hood)* is a normalized estimate of the flow of people to the hood area, *work_rate (hood)* is an estimate of the number of people heading to work, *mean_customers* is an array of data that stores the average number of movements from a fixed area to other areas, *mean_customers* is an array of data that stores the average the number of movements from the recorded area to other areas to work.

In this case, the number of people affects the connection, therefore, instead of the minimax normalization, division by the maximum value is used. It does not remove the influence of the strength of the connection, but at the same time brings this feature to the same scale with the average share of people moving to work, which is necessary for high-quality clustering. For clustering in such a space, it is also preferable to use the k-means method, since the features of this method are best suited for clustering in such spaces.

2) Clustering by volume of movements

As described earlier, one of the main attributes for describing a connection is the number of movements from a fixed district to a connected district. Sometimes it is necessary to divide districts according to the strength of the connection between them, without considering the characteristics of this connection. Such clustering may be necessary when revising the mode of operation of public transport, as well as the construction of new transport channels. For such clustering, it is necessary to calculate the average number of movements from a fixed district to the rest and carry out clustering in a one-dimensional space. Some of the connections, where the number of movements is very small, can be eliminated. In such a case, it is preferable to use hierarchical clustering. Such clustering is the most visual and understandable for a specialist in the subject area.

3) Clustering by volume of movements using the metro

Also, the clustering of links can be carried out according to the average rate of metro use. In this case, it is possible to identify clusters of areas with good, medium, and poor connectivity using the metro. It is also convenient to carry out such clustering using hierarchical methods.

4) On checking the quality of clustering

Before carrying out clustering, it is necessary to make sure that clusters can be distinguished in the analyzed space. Before clustering, an analysis of the value of the Hopkins statistic [12]

is performed for this. If the statistic value is close to 0, then you need to choose another space for clustering.

To check the quality of ready-made clustering without experts in the subject area, it is possible to use the analysis of silhouette coefficients [13].

With the participation of specialists, the V-measure is used to assess the quality [14]. Another option for checking the quality of clustering is to ask specialists a several questions, built according to the template: "Is it permissible to relate region A and region B to the same cluster?" and compare their responses with the clustering results.

E. Computational experiment

1) District clustering

According to the data of cellular operators, in May 2017, the clustering of Moscow districts was carried out. The first clustering was done in a feature space based on the average number of people in the area during working hours and at night. Districts of the city in such a space are represented by points, the location of which is shown in Fig. 2.

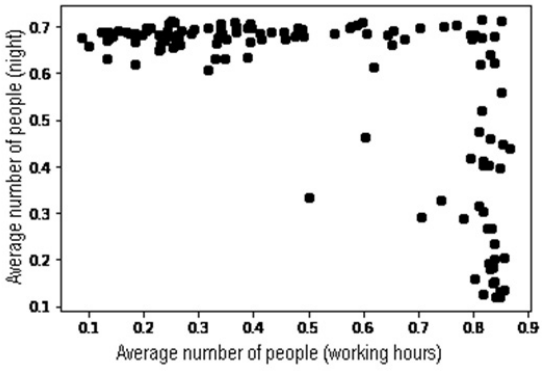

Fig. 2. Districts of the city are presented in feature space

The vertical axis shows the average number of people staying in the area at night (minimax normalization), and the horizontal axis shows the average number of people staying in the area during working hours (minimax normalization). As can be seen from the graph, there are no districts in Moscow where there are relatively few people both during working hours and at night. When displaying the territorial units of Moscow and the Moscow region (Fig. 3), in the same feature space, territorial units are noticeable, where it is not crowded both at night and during working hours.

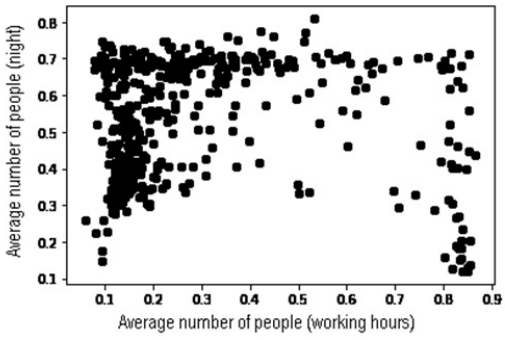

Fig. 3. Districts of the city and region are presented in feature space

These territorial units represent the locations of summer cottages and garden plots.

Hopkins statistics both for Moscow and for Moscow and the region show high values (over 0.85), so it is possible to use clustering methods. The optimal number of clusters is found by analyzing the silhouette coefficients.

In the case of clustering data for Moscow, the admissible values of silhouette coefficients are obtained by dividing the data into 2, 3, 4, or 5 clusters. Visual data on clustering and values of silhouette coefficients are presented in Figures 4-7.

With the number of clusters from 2 to 5, all clusters have objects for which the silhouette coefficient is greater than the average for all clusters. The spread of silhouette coefficient values is small between different clusters.

Negative values of silhouette coefficients are observed only in the case of clustering into 5 clusters, there are few of them relative to the total number of objects. The resulting clusters smoothly pass from one to another, which indicates that there is no clear division of districts by type. Let's take a closer look at clustering with the number of clusters equal to five (Fig. 7).

The zero cluster includes districts that are more residential than working. In them, the night_rate is close to 0.65, and the working_rate is close to 0.4. Typical representatives: Yaroslavsky, Ryazansky, Kuntsevo.

The first cluster is a working cluster, characterized by night_rate close to 0.2, working_rate close to 0.85, typical representatives are Arbat, Tverskoy

The second cluster is a cluster of districts that are equally residential and working. Both indicators are close to 0.7. Typical representatives: Matushkino, Akademichesky, Ramenki.

The third cluster is formed by residential districts: in them, the working_rate is less than 0.3. This cluster includes Kuzminki, Vykhino-Zhulebino, Yuzhnoye Butovo, Maryino, etc.

The fourth cluster is a cluster of districts that are more working than residential. The cluster representatives are characterized by a working_rate close to 0.8, as well as a night_rate close to 0.45. Typical representatives: Kapotnya, Voikovsky, Yuzhnoportovy, etc.

Clustering is carried out in the second feature space. In it, one feature is the rate of movements using the metro during the morning rush hour. and the second is the rate of movements to work during the morning rush hour. The graph (Fig. 8) shows the territorial units of Moscow and the region in this feature space.

With the number of clusters from 2 to 5, all clusters have objects for which the silhouette coefficient is greater than the average for all clusters. The spread of silhouette coefficient values is small between different clusters.

Negative values of silhouette coefficients are observed only in the case of clustering into 5 clusters, there are few of

them relative to the total number of objects. The resulting clusters smoothly pass from one to another, which indicates that there is no clear division of districts by type. Let's take a closer look at clustering with the number of clusters equal to five (Figure 7).

The zero cluster includes districts that are more residential than working. In them, the night_rate is close to 0.65, and the working_rate is close to 0.4. Typical representatives: Yaroslavsky, Ryazansky, Kuntsevo.

The first cluster is a working cluster, characterized by night_rate close to 0.2, working_rate close to 0.85, typical representatives are Arbat, Tverskoy

The second cluster is a cluster of districts that are equally residential and working. Both indicators are close to 0.7. Typical representatives: Matushkino, Akademichesky, Ramenki.

The third cluster is formed by residential districts: in them, the working_rate is less than 0.3. This cluster includes Kuzminki, Vykhino-Zhulebino, Yuzhnoye Butovo, Maryino, etc.

The fourth cluster is a cluster of districts that are more working than residential. The cluster representatives are characterized by a working_rate close to 0.8, as well as a night_rate close to 0.45. Typical representatives: Kapotnya, Voikovsky, Yuzhnoportovy, etc.

Clustering is carried out in the second feature space. In it, one feature is the rate of movements using the metro during the morning rush hour. and the second is the rate of movements to work during the morning rush hour. The graph (Figure 8) shows the territorial units of Moscow and the region in this feature space.

The fourth cluster is a cluster of districts that are more working than residential. The cluster representatives are characterized by a working_rate close to 0.8, as well as a night_rate close to 0.45. Typical representatives: Kapotnya, Voikovsky, Yuzhnoportovy, etc.

Clustering is carried out in the second feature space. In it, one feature is the rate of movements using the metro during the morning rush hour. and the second is the rate of movements to work during the morning rush hour. The graph (Figure 8) shows the territorial units of Moscow and the region in this feature space.

The horizontal axis shows the rate of metro passengers in the total traffic flow from the districts during the morning rush hour. The vertical axis shows the rate of people leaving the districts for work. In this space, territorial entities with a close to zero rate of metro passengers are distinguished, these are territorial units of the Moscow region. These territorial units can be subjected to clustering according to one feature - the rate of people leaving for work during the morning rush hour. In this clustering, only Moscow districts are considered (Fig. 9)

The horizontal axis shows the share of metro passengers in the total traffic flow from the area during the morning rush hour. The vertical axis shows the proportion of people leaving the district for work. When calculating the Hopkins statistic, a value close to 1 was obtained, so the data is suitable for clustering. The analysis of silhouette coefficients shows that clustering can be carried out with the number of clusters equal to five (Fig. 10).

All clusters contain objects with silhouette coefficient values above average. The spread in the values of this coefficient between clusters is small. Negative silhouette coefficient values correspond to atypical data. The clustering results are shown in the following Fig. 11.

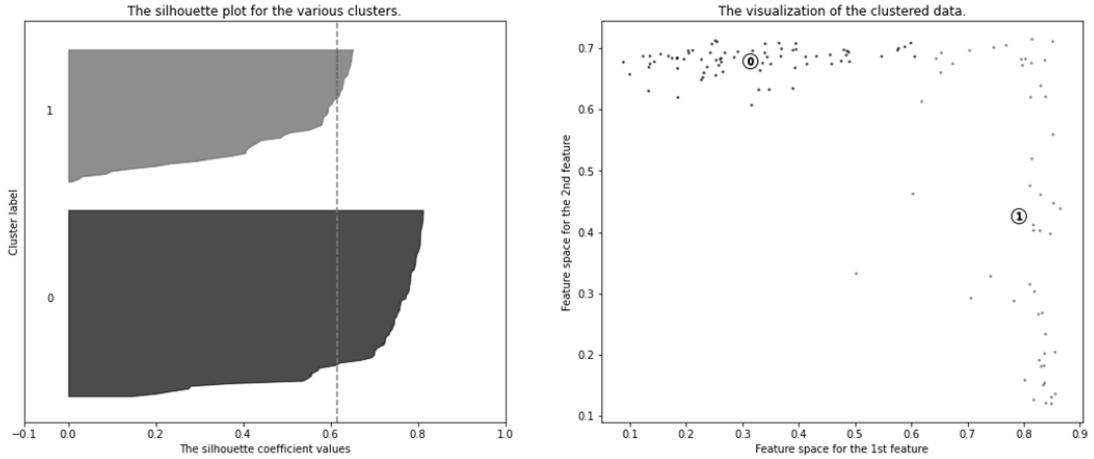

Fig. 4 Silhouette coefficients and clusters (number of clusters equals 2)

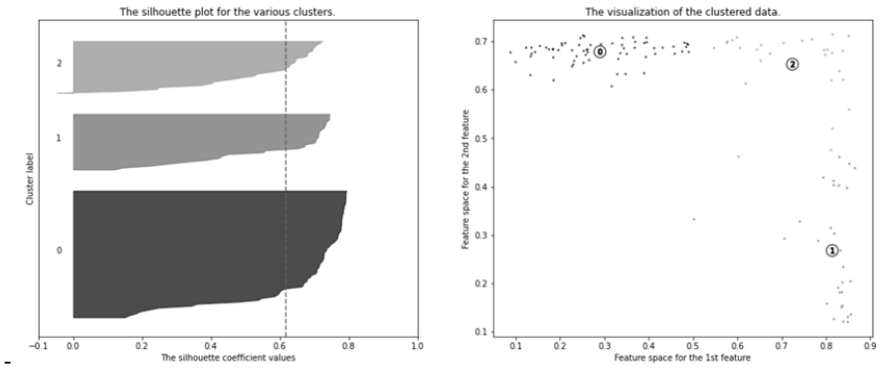

Fig. 5. Silhouette coefficients and clusters (number of clusters equals 3)

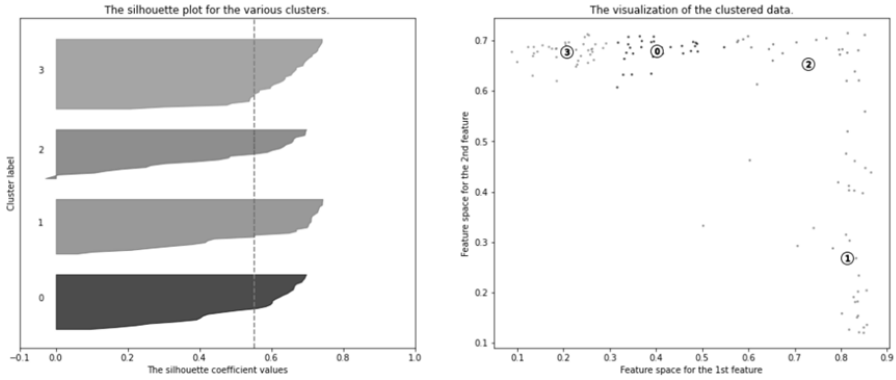

Fig. 6. Silhouette coefficients and clusters (number of clusters equals 4)

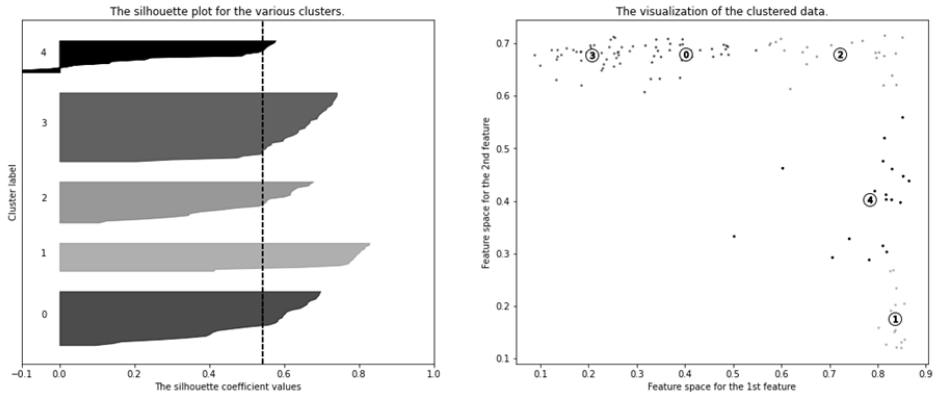

Fig. 7. Silhouette coefficients and clusters (number of clusters equals 5)

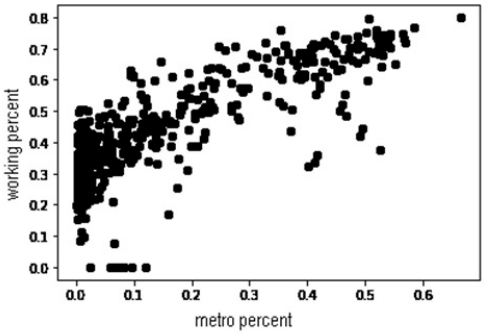

Fig. 8. Districts of the city and region are presented in feature space

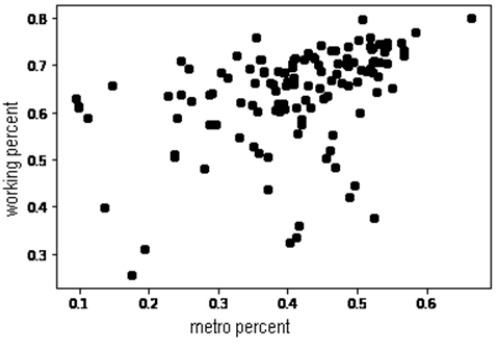

Fig. 9 . Districts of the city are presented in feature space

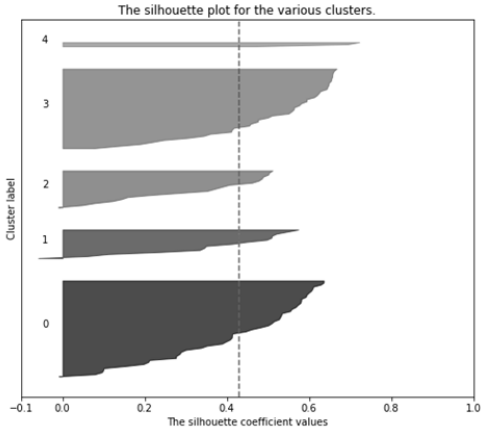

Fig. 10. Silhouette coefficients visualization

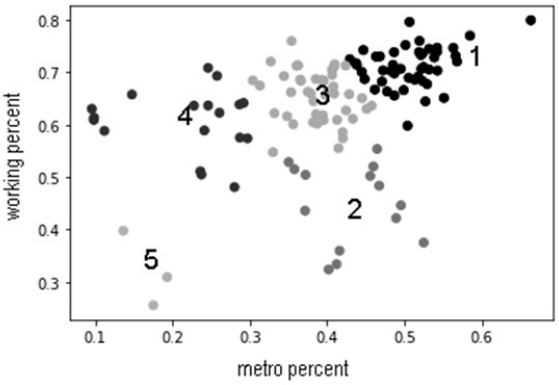

Fig. 11. Clustering results

A cluster with a high rate of movements on the metro and a high rate of movements to work is allocated as a result of clustering. This cluster is marked as first. It includes such districts as Kuzminki, Izmailovo, Otradnoye, Tushino. From these districts, people mostly go to work in the morning with a high degree of metro use. These are districts on the outskirts of Moscow.

The second cluster is a cluster with a low rate of commuters but average metro use. These are districts of the city center, such as Arbat, Khamovniki, Begovoy. These districts are marked as cluster 2.

The third cluster is a cluster with approximately the same metro utilization, but with a higher rate of people commuting to work. It is marked as cluster 3. It includes such districts as Ramenki, Alekseevsky, Ostankino, Sokolniki, Voikovsky.

The fourth cluster is a cluster with low metro usage and average commuting. Such areas include Yaroslavsky, Kryukovo, Kurkino, Rostokino, etc..

The fifth cluster is small and consists of three districts: Vnukovo, Kapotnya, and Molzhaninovsky. The first and the last are located outside the Moscow Ring Road. The second has poor transport accessibility. No one of the districts has a metro.

The latter two clusters can be examined by urban scientists for transport problems. There are quite a few commutes in the fourth cluster, but very low metro utilization. Perhaps the residents of these areas use a large number of personal vehicles, which contributes to the creation of a difficult situation on the roads.

2) .District Connection Clustering

To cluster connections, data on the total flow and metro flow for Dolgoprudny for May 2017 is taken.

Consider the clustering of the 25 most connected with Dolgoprudny territorial units in a space in which each district is described by the total number of movements from Dolgoprudny to this district. The agglomerative clustering method is chosen as a clustering method. A visualization of the dendrogram is shown in the following Fig. 12. This dendrogram shows a cluster with one urban district of Khimki. It is the main connected territorial unit since Dolgoprudny and Khimki are geographical neighbors. NPO Energomash and MKB Fakel are located in Khimki, in which residents of Dolgoprudny work. Territorially, Khimki includes country houses, where residents of Dolgoprudny rest. There is also a large shopping center in Khimki, attracting residents of Dolgoprudny.

A cluster stands out, consisting of the urban district of Lobnya, the Fedoskinskoe settlement, the Severny district, and the urban settlement of Mytishchi. The northern region is geographically close to one of the central streets of Dolgoprudny - Pervomayskaya. Also in Dolgoprudny, there are stations from which the train leaves for Moscow, which can be used by residents of the Northern District. For the settlement of Fedoskinskoye, Dolgoprudny is the most accessible (in terms of the availability of roads) and the closest territorial unit with urban infrastructure. Some residents of Lobnya and Mytishchi work at the enterprises of Dolgoprudny. Also, these settlements are neighbors, which also contributes to high traffic between areas. Of the districts that are not neighboring, the densest is the connection with the Tverskoy district, where some residents of Dolgoprudny work.

A dendrogram built on the feature of movement data from Dolgoprudny (using the metro) is shown in Fig. 13.

This dendrogram shows that the Tverskoy district, where residents of Dolgoprudny get to work by metro, as well as other districts of the city center, is allocated into a separate cluster. Note that the Zamoskvorechye district is more closely connected with Dolgoprudny than Arbat, although it is located further from Dolgoprudny. This is due to the fact that in the Zamoskvorechye district there are stations on the green line of the metro, at one of the stations of which (Khovrino) bus 368 arrives from Dolgoprudny. For the same reason, there are quite strong connections with the Airport and Begovaya districts.

Consider clustering in the feature space, where each of the districts connected with Dolgoprudny is described by the number of movements (relatively to the maximum, that is, a movement to Khimki) and the rate of people moving to work.

Connection clustering uses the k-means method in such a space. The clustering results are presented in Fig. 14. The X-axis shows the relative number of movements, and the Y-axis shows the rate of people who commute to work. In this

clustering, the urban settlement of Khimki is also distinguished for the reasons described earlier. In another cluster (shown as cluster 1), Mytishchi, Fedoskinskoe, Lobnya, and the Severny district are highlighted - in these areas of Moscow and the Moscow region, residents of Dolgoprudny move quite often, not only to work (cluster 2). A movement to the rest of the areas is rather small; they are clustered mainly by the rate of people who travel to these regions to work. The third cluster contains points corresponding to connections with the city center districts, such as Tverskoy and Arbat. People travel to these districts to work.

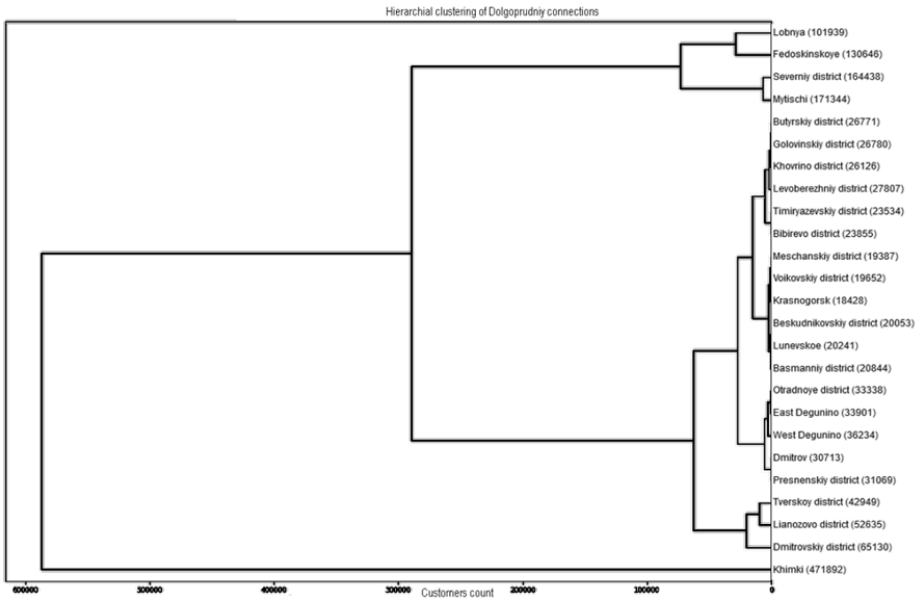

Fig. 12. Dendrogram of connections (Dolgoprudny, total flow)

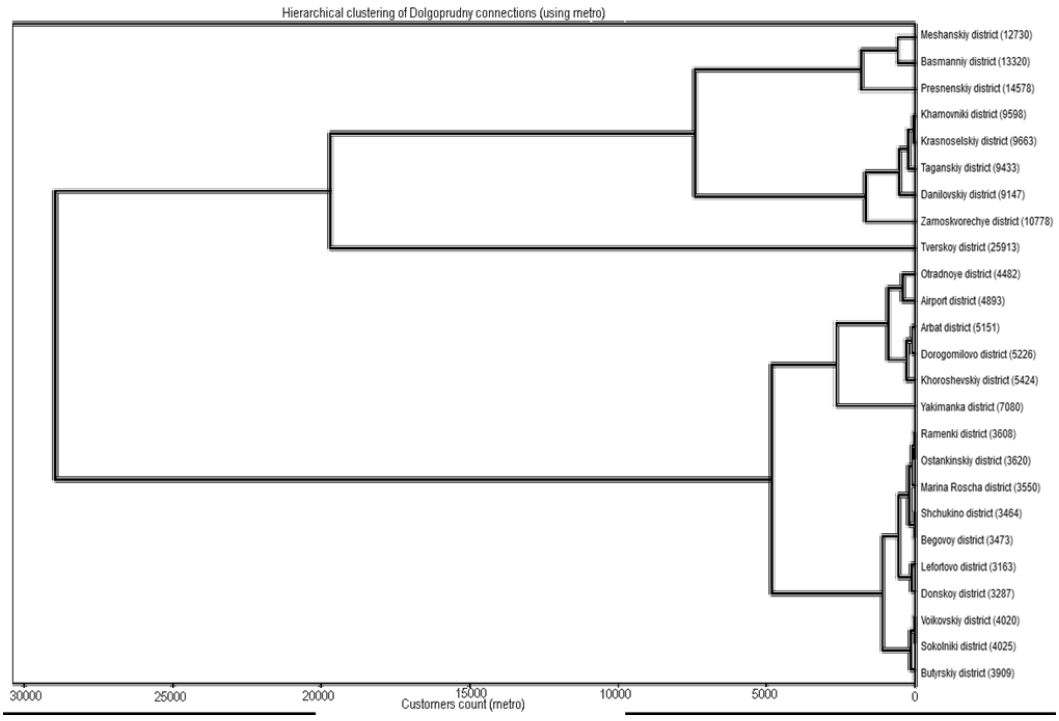

Fig. 13. Dendrogram of connections (Dolgoprudny, metro flow)

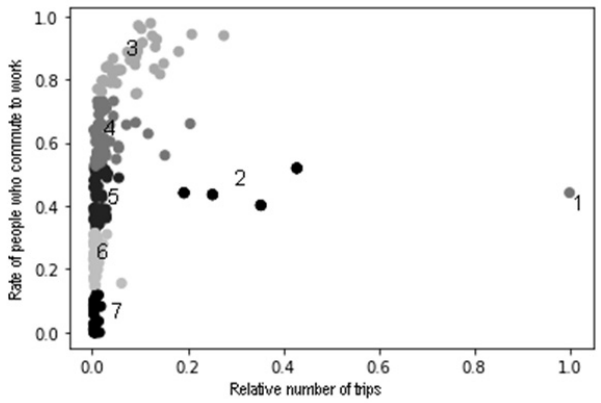

Fig. 14. Clustering of areas connected with Dolgoprudny

III. Conclusion

This paper proposes a new approach to clustering city districts and connections between them based on data from cellular operators. Five feature spaces are presented, two of which describe the districts, and the rest - the connections between them. In these spaces, districts are clustered, the quality of clustering is assessed. An explanation of the resulting clusters is given. The approaches, methods, and results described in this paper can be used by urbanists and data analysts to solve transport problems and plan urban infrastructure.

References

[1] Report of the international agency “We are social” “Digital 2021”, Web: https://wearesocial.com/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital

[2] Namiot, Dmitry, et al. "On the assessment of socio-economic effects of the city railway." International Journal of Open Information Technologies vol.6, Jan, 2018, pp. 92-103.

[3] Wang, S., Min, J., and Yi, B. Location based services for mobiles: Technologies and standards. In IEEE ICC. Beijing

[4] Smith, Stanley & Mandell, Marylou. “A Comparison of Population Estimation Methods: Housing Unit Versus Component II, Ratio Correlation, and Administrative Records”, Journal of the American Statistical Association, vol. 79, 1984, pp.282-289

[5] Lin W. Population “Estimation in a Desert City Using Parameters from Satellite Imagery”, State University of New York at Binghamton, 2019.

[6] Peter Wei, Xiaofan Jiang. “Data-Driven Energy and Population Estimation for Real-Time City-Wide Energy Footprinting”, Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation ,Association for Computing Machinery, New York, NY, USA, pp. 267–276.

[7] Bulygin M., Namiot D. “Anomaly Detection Method For Aggregated Cellular Operator Data”, 28th Conference of Open Innovations Association (FRUCT), IEEE, 2021, pp. 42-48.

[8] T. von Landesberger, F. Brodkorb, P. Roskosch, N. Andrienko, G. Andrienko and A. Kerren, “MobilityGraphs: Visual Analysis of Mass Mobility Dynamics via Spatio-Temporal Graphs and Clustering,”, Transactions on Visualization and Computer Graphics, Jan. 2016, vol. 22, no. 1, pp. 11-20, 31.

[9] MacQueen J. et al., “Some methods for classification and analysis of multivariate observations, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967, vol. 1. no. 14, pp. 281-297.

[10] Ester M. et al., “A density-based algorithm for discovering clusters in large spatial databases with noise”, Kdd, 1996. vol. 96, no. 34., pp.. 226-231.

[11] Zhambyu M., “Hierarchical cluster analysis and correspondences”, Finance and Statistics, 1988, 342 p.

[12] Banerjee A., Dave R. N., “Validating clusters using the Hopkins statistic”, IEEE International conference on fuzzy systems, 2004, vol. 1. pp. 149-153.

[13] Rousseeuw P. J., “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”, Journal of computational and applied mathematics, 1987, vol. 20, pp. 53-65.

[14] Rosenberg A., Hirschberg J., “V-measure: A conditional entropy-based external cluster evaluation measure”, Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 2007, pp. 410-420.