Validation of Origin-Destination Matrix by Open Data

Evgeny Mayorov
Department of Transportation Organization and Management
Samara National Research University
Samara, Russia
benjamin1437@mail.ru

Oleg Saprykin
Department of Transportation Organization and Management
Samara National Research University
Samara, Russia
saprykinon@ssau.ru

***Abstract*—Many global factors can influence the socio-economic situation in the country. One such factor is the transport systems of cities. It depends on the state of the transport infrastructure of the city: business development, investment, quality of life, employment growth. However, in larger cities observed the deterioration of the situation with the transport systematic in connection with the increasing number of cars and other factors. To solve these problems resort to modelling. The dynamic modelling allows to determine the characteristics of the future project (capacity, number of road lanes, traffic light cycle, etc.) without high capital investments. Modeling of transport processes has been used in practice for a long time, but there is a problem in checking the compliance of the constructed model with the real transport situation. Existing methods, such as the manual method of collecting information about traffic flows, are expensive and labor-intensive. Authors propose the method of validation of origin-destination matrix by open data provided by Uber company.**

Keywords—Transport process, origin-destination matrix, SUMO, gravity model, Uber Movement, transport simulation, OpenStreetMap, transport analysis zones.

I. Introduction

In connection with the active development and growth of the urban population, the problem of matching the transport infrastructure with the needs of citizens' mobility is acute. The number of residents of the city is growing as well as the size of the cities. Transport systems remain one of the main problems of modern megacities. According to statistics, residents of large cities every 365 days spend 3 days waiting in traffic jams, which is an average of about 16% of the total driving time due to the poor condition of urban road networks. These statistics show the existing problem of urban road networks and the need to address it. In this regard, there are several modern approaches. The first way is to improve the road network, its reconstruction [1]. Another approach to solving the problem is to create comfortable conditions in public transport to attract citizens. This makes sense, since the desire to transfer people from personal to public transport will certainly lead to the unloading of the street and road network of the city.

Transport systems and their research become more complex over time and require precise knowledge in this area. Currently, there are many studies devoted to this area. Speaking about the transport system, it is worth noting that it consists of many interrelated parts. That is why in the world there are a large number of scientific groups engaged in research in a certain area.

Most attention should be given to the construction of origin-destination matrices (OD), which are designed to contain information on population movements throughout the city and to be able to apply these data in modelling. The data-driven approach and the model-driven approach are the most popular approaches to constructing OD matrices.

Application of GPS data for scientific research in transport systems is very popular nowadays. Therefore, the data approach is often used because it is based on the fact that the model is created based on dynamic data of residents' movement throughout the city using GPS, social networks, etc. This approach is very modern due to the widespread use of GPS navigation in our lives. In addition, the installation of GPS-Navigator in cars is now very common, and the matrix OD can be formed from the data obtained from the GPS-device. It is worth noting that the use of this method significantly reduces the possibility of obtaining an inadequate model. This area of research is relevant in our time [2-6].

Information from social networks is another source of data on population movements. Today, Foursquare and Facebook are among the best for this purpose. This method is based on the processing of geolocation data included in messages and photos. Having processed a sufficiently large amount of data, it is possible to make statistics showing the intensity of population movement in certain areas of the city at certain intervals. In this area, a lot of research is carried out on the creation and use of new social networks [7-10].

The most accurate method of obtaining data on mobility of the inhabitants were Clastres [11]. The high accuracy of the method is provided by using the data obtained from the cellular communication of residents of the city. This method provides the most up-to-date data, but processing large sets of data from different operators is very complex. The difficulties of using this method include the fact that it may be limited by the level of legislation of some countries, as well as the disagreement of subscribers with the processing of their personal data.

The model-driven approach is also popular with researchers and is used when mobile data or GPS is not available. Filippovskiy [12], which uses a universal model for mobility and migration, use the gravity method to obtain data on population movements across the country. This method makes it possible to monitor trade flows both within and between countries, migration flows and long-distance telephone calls. The gravitational method was developed by W. Reilly [13]. The idea of constructing a model of gravitational type is based on the universal law of gravity. Applying this law to the transport system, it turns out that the objects are points that generate and absorb traffic flows [14], and the mass of the object is the total volume of incoming and outgoing flow. It is worth noting that it is possible to replace the physical distance by any costs [15] associated with the movement.

The entropy model [16] is specific for the second law of thermodynamics and was proposed by physicists. The law states that any closed physical system tends to achieve a stable equilibrium, which is characterized by the maximum entropy of this system. The first application of the method was carried out to the transport processes of A. Wilson [17]. The principle of the entropy method is that the system of movement of residents on the road network of the city has a sufficiently large number of uncontrolled elements. Due to the fact that our study is devoted to urban transport, it is necessary to consider the transport system as closed.

Michael Ballmer compares two types of modeling in his work [18]: the micro-simulation model and the traditional destination model. The traditional distribution model is based on the use of od matrices to reflect the movement of large population flows [19]. The main advantage of the micro-simulation model is the ability to simulate the movement of each agent separately [20].

Modern transport modeling systems allow creating a visual representation of traffic flows. Currently there is a wide range of software for transport modeling, among which are more popular programs such as T7F/TSIS, SUMO [21], TRANSYT, VISUM, MATsim, Aimsun. They allow to carry out verification of transport infrastructure changes without special economic and time costs. In addition, a study of the construction of od matrices based on the results of microscopic modeling was carried out [22-24].

The main problem of modeling is the lack of data with which to compare the resulting model. The method of full-scale measurements is quite labor-intensive and takes a large amount of time. Recently, however, sources with open statistics have begun to emerge. For example, Uber has created its own statistical data system based on GIS – Uber Movement. Uber engineers took data from GPS trackers in a taxi of their company, depersonalized it, and then created a statistical database on the time spent on the movement of the city districts. These data can be used for comparative analysis of the experimental correspondence matrix.

The article is devoted to the development of methods and software for automated creation of transport microscopic models for any urban area as well as comparative analysis of the received origin-destination matrix based on open data of Uber Movement. The developed system allows getting up-to-date information from open Internet resources, process it and use it to create or update the transport model. As an example of an open resource, consider OpenStreetMap, which provides information about urban infrastructure collected by volunteers. Using OpenStreetMap information, we have identified supply and demand for transportation to build a microscopic transportation model. The proposed approach allows to obtain models with acceptable quality using only data from open sources.

II. A METHOD FOR ACTIVITY CHAINS CONSTRUCTION

One of the main objectives of this research is to develop an automated method of activity chains construction (fig. 1).

The first step in the method of activity chains construction is to obtain a map of the city of Amsterdam (one of the cities that available in Uber Movements by this day) from an open source OpenStreetMap.

The further stage is the division of the city into transport analysis zones [25]. In this study, it is necessary to take the division of the city into transport analysis zones is the same as in Uber Movement, in order to further be able to conduct a comparative analysis (fig. 2).

In this case, the division of the city is as follows: the closer to the city center, in which the number of trips as much as possible, the smaller the size of transport areas (in some cases up to the size of the quarter), and the farther from the city center, closer to the suburban areas the size of transport areas more. According to this principle, the city is divided into 181 transport areas.

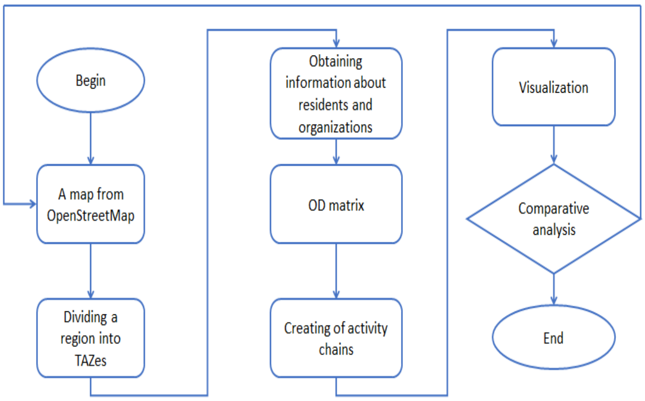

Begin
A map from OpenStreetMap
Dividing a region into TAZes
Obtaining information about residents and organizations
OD matrix
Creating of activity chains
Visualization
Comparative analysis
End

Fig. 1. Schema of method for activity chains construction

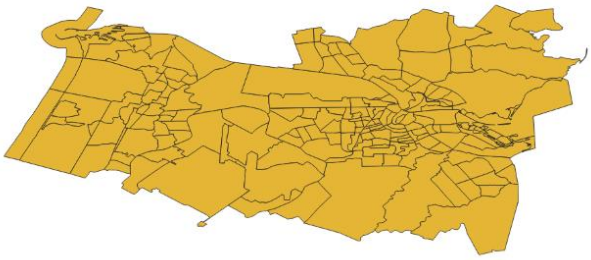


Fig. 2. Dividing the city of Amsterdam by transport analysis zones

The next stage of the developed method of activity chains construction is to collect information about the number of residents of the city and assign them to a specific transport area. In this particular case, it was possible to obtain only General information on the number of inhabitants in the city of Amsterdam.

To divide the total population of the city by transport analysis zones, it was decided to use the coefficient method. This is because the size of the transport areas in this study is different and some areas are very different from the General trends, especially those far from the center. Therefore, it is necessary to assign a certain coefficient to such areas, which will equalize the size of the area and the number of inhabitants in it.

The coefficients were assigned to the administrative division of the city as some administrative areas have a large number of small areas, namely the center of Amsterdam – has 100 and Haarlem - 22 transport areas. It was therefore necessary to assign coefficients ranging from 0 to 2 to all transport areas so that the total number of inhabitants would remain the same. The smaller the transport area, the closer its coefficient is to the value 2, and vice versa, the larger the transport area, the value is closer to 0, but 0 will never be.

The next step is to build an origin-destination matrix. A gravity model is used to distribute correspondence by calculating cost matrices representing the cost of travel between each pair of zones. The transport gravity model relates to the intensity of the flow. Between the total number of departures from the *i*-th zone and arrivals to the *j*-th zone and the cost of travel between zones *i* and *j* (1). Applying the formula of the gravitational method was obtained origin-destination matrix which describes the movement of the population in transport analysis zones:

$$T_{ij} = \frac{Q_i * D_j}{c_{ij}^2} i = 1, \ldots, N, j = 1, \ldots, M, \qquad (1)$$

where *N* is the total number of departure zones, and *M* is the total number of arrival zones. In this model, the distance between areas is considered the distance between the centres of these areas.

However, for the possibility of further comparative analysis of semi-cluded matrix with Uber open data, the data about the intensity of vehicle should be converted to traffic time (here the time spent moving from one area to another):

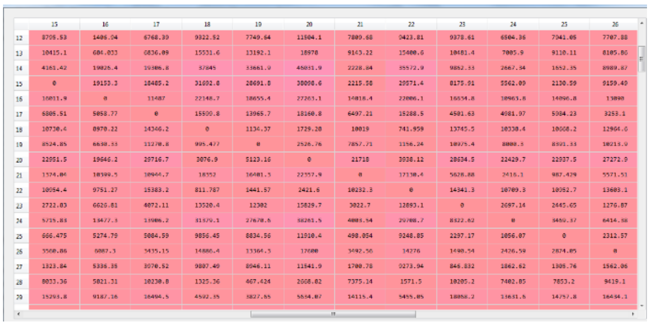
$$t_{ij} = L_{ij} * \frac{p}{T_{ij}}, \iota, j \in R, \qquad (2)$$

where *t* is the time taken to move from one area to another

L is distance between transport areas;

p is average traffic density in the city.

The figure 3 shows the transformed origin-destination matrix in time characteristics.



	15	16	17	18	19	20	21	22	23	24	25	26
12	8795.53	1406.94	6768.39	9322.52	7749.64	11504.1	7809.68	9423.81	9378.61	6504.36	7941.05	7707.88
13	10415.1	604.033	6836.09	15531.6	13192.1	18978	9143.22	15400.6	10481.4	7005.9	9110.11	8105.86
14	4161.42	19826.4	19306.8	37845	33661.9	46031.9	2228.84	35572.9	9862.33	2667.34	1652.35	8989.87
15	0	19153.3	18485.2	31692.8	28691.8	38098.6	2215.58	29571.4	8175.91	5562.09	2130.59	9159.49
16	10011.9	0	11487	22148.7	18655.4	27263.1	14018.4	22806.1	16654.8	10963.8	14096.8	13898
17	6805.51	5058.77	0	15509.8	13965.7	18160.8	6497.21	15288.5	4501.63	4981.97	5984.23	3253.1
18	10730.4	8970.22	14346.2	0	1134.37	1729.28	10019	741.959	13745.5	10338.4	10668.2	12964.6
19	8524.85	6638.33	11270.8	995.477	0	2526.76	7857.71	1156.24	10975.4	8008.3	8391.33	10213.9
20	22951.5	19646.2	29716.7	3076.9	5123.16	0	21718	3938.12	28634.5	22429.7	22937.5	27272.9
21	1374.04	18399.5	18944.7	18352	16481.5	22357.9	0	17130.4	5628.88	2416.1	987.429	5571.51
22	10054.4	9751.27	15383.2	811.787	1441.57	2421.6	10232.3	0	14341.3	10709.3	10052.7	13603.1
23	2722.03	6626.81	4072.11	13520.4	12302	15829.7	3022.7	12893.1	0	2697.14	2445.65	1276.87
24	5715.83	13477.3	13906.2	31379.1	27670.6	38261.5	4003.54	29708.7	8322.62	0	3460.37	6414.58
25	666.475	5274.79	5084.59	9856.45	8834.56	11910.4	498.054	9248.85	2297.17	1056.07	0	2312.57
26	3560.86	6087.3	3435.15	14886.4	13364.5	17600	3492.56	14276	1490.54	2426.59	2874.05	0
27	1323.84	5336.35	3970.52	9887.49	8946.11	11541.9	1708.78	9273.94	846.832	1862.62	1305.76	1562.06
28	8033.36	5021.31	10230.8	1325.36	467.424	2668.82	7375.14	1571.5	10205.2	7402.85	7853.2	9419.1
29	15293.8	9187.16	16494.5	4592.35	3827.65	5634.07	14115.4	5455.05	18868.2	13631.6	14757.8	16434.1

Fig. 3. Origin-destination matrix in time characteristics

The whole developed method of constructing a correspondence matrix is universal. In order to apply this method to any other city, you only need to download a new map of the city, determine the division of the city into transport areas, as well as add new information about the number of residents in the city.

The next step is to build an activity chain. The activity chain is one of the components of visualization in the SUMO simulation program. The activity chain is obtained by converting the origin-destination matrix into the format required for modeling. To transform a matrix, you must use applications SUMO – OD2TRIPS and DUAROUTER. These programs include all the necessary configurations for accurate modeling of transport processes. OD2TRIPS determines the start and end points of the path of each simulated correspondent based on the origin-destination matrix loaded into it, but it does not include the function of building a route along the roads. And DUAROUTER already directly takes the data obtained from OD2TRIPS and converts them in accordance with the roads laid down in the downloadable map of the city. Thus, we get the connected elements of the correspondence matrix and the map of the city of Amsterdam.

The final stage of creating a model of traffic flows of the city of Amsterdam is the visualization in the program of microscopic simulation of SUMO. To do this, you need to download the correspondence matrix processed in the OD2TRIPS and DUAROUTER programs, as well as the city map in xml format.

III. Comparative analysis of OD matrix with open data

After obtaining an adequate model of transport processes in the city in the SUMO program, it is necessary to determine how the resulting model corresponds to the real situation on the roads of the city of Amsterdam.

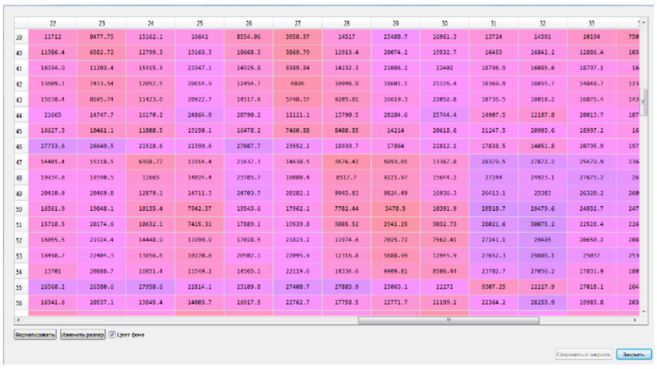
To obtain information about the origin-destination matrix with open data on the time spent moving between transport areas taken from The Uber Movement website, it is necessary to compare them. Comparative analysis in this paper is performed using regression analysis.

TABLE I. Statistical data from Uber Movement resource

source_id	dstid	hod	mean_travel_time	standard_deviation_travel_time	geometric_mean_travel_time	geometric_standard_deviation_travel_time
143	141	7	263,89	233,27	200,94	2,2
144	131	7	527,19	513,77	394,14	2,04
61	108	7	472,58	250,77	425,29	1,56
156	16	12	1723,8	335,02	1691,98	1,21
153	46	12	1468,7	354,5	1425,79	1,28
74	127	6	980,88	488,84	889,75	1,52
38	171	1	1252,2	252,61	1230,38	1,2
42	27	18	814,19	314,18	763,15	1,42
40	47	18	1087,9	420,01	1018,93	1,43
86	109	7	537,3	302,23	448,22	1,89
53	87	0	519,95	194,15	489,4	1,4
59	27	0	594,47	243,35	554,03	1,46
57	47	0	724,27	211,11	698,47	1,3

From an open source Uber Movement were taken data on the average time of movement between transport areas for the third quarter of 2018. Table 1 shows the format of the downloaded data where: "sourceid" is the number of the departure area, "dstid" is the number of the arrival area, "mean_travel_time" is the average travel time between the departure area and the arrival area, the other elements of this table are not relevant to this study. Average time obtained by statistically anonymized data with GPS the taxi industry for Uber in Amsterdam.

Then it is necessary to load data on the areas of departure and arrival, and the average time spent on correspondence. These data, as well as the origin-destination matrix should be obtained in the form of a matrix shown in figure 4. A small disadvantage of the downloaded open data is the lack of some elements of the matrix, this is due to the lack of statistical data on the movement of population from certain areas in Uber specialists.



	22	23	24	25	26	27	28	29	30	31	32	33	3
39	11712	8477.75	15162.1	16641	8554.06	3958.97	14517	23488.7	16961.3	13724	14591	10194	750
40	11986.4	6582.72	12799.3	19163.3	10668.3	3869.79	11913.4	20074.2	19532.7	16453	16842.2	12886.4	103
41	16034.9	11203.4	15315.3	21947.1	14925.8	8389.84	14232.3	21086.2	22402	18796.9	16085.6	18707.1	16
42	13609.1	7413.54	12052.5	20614.9	12454.7	4896	10996.9	18601.1	21226.4	18369.9	18855.7	14848.7	123
43	15038.4	8695.74	11423.6	20922.7	14517.8	5748.37	9205.01	16619.3	22852.8	18736.5	18818.2	16876.4	143
44	21665	14747.7	16170.2	24864.9	20790.2	12121.1	13790.5	20204.6	25744.4	14907.5	12107.8	20013.7	187
45	16627.3	10461.1	11008.5	19150.1	16478.2	7460.88	8408.85	14214	20618.6	21247.5	20903.6	18997.2	16
46	27733.6	26649.5	21528.6	22399.6	27087.7	23552.2	18939.7	17864	22812.1	17838.5	14051.8	20795.9	197
47	14401.4	15118.5	6398.77	11914.4	21637.3	14638.5	3676.42	6893.01	13367.8	28379.5	27872.2	25479.9	236
48	19934.8	19348.5	11665	14024.4	23785.7	18808.4	8517.7	9221.97	15644.2	27244	24925.1	27675.2	26
49	20410.9	20469.8	12879.1	14711.3	24703.7	20202.1	9945.83	9824.49	16936.3	26413.1	25383	26320.2	260
50	16561.9	19048.1	10135.4	7942.57	19543.6	17962.1	7782.44	3470.5	10391.9	29510.7	29479.6	24932.7	247
51	15718.5	20174.6	10632.1	7415.31	17889.1	19539.8	9085.52	2941.25	9032.73	28021.6	30075.2	22528.4	226
52	16055.5	21924.4	14448.9	11999.9	17028.5	21823.2	11974.6	7025.72	7562.41	27241.1	29485	20650.2	208
53	18498.7	22404.5	15856.6	10220.8	20502.1	22095.9	12316.8	5888.99	12945.9	27652.3	29886.1	25857	259
54	13701	20888.7	16851.4	11549.2	14565.1	22119.6	14334.6	9989.81	8586.44	23782.7	27856.2	17031.9	188
55	26568.2	26380.6	27958.6	22814.1	23189.8	27408.7	27885.9	23063.1	22272	9307.25	12227.9	17018.1	164
56	16341.6	20537.1	19845.4	14003.7	16917.5	22762.7	17758.5	12771.7	11199.1	22364.2	28253.9	19983.8	203

Fig. 4. Preprocessed Uber Movement data

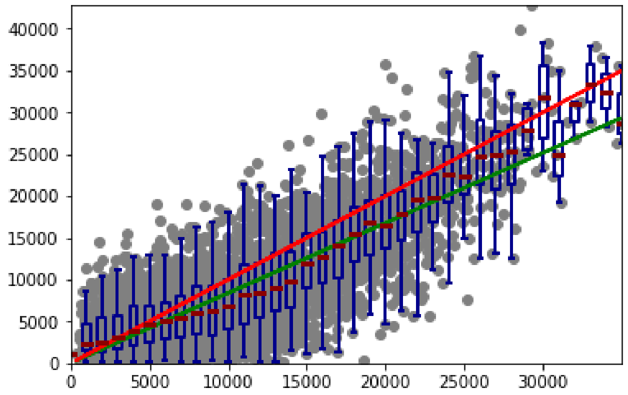


Fig. 5. Comparison modeled data with real data using regression model

The linear regression is used, the result of which will be a two-dimensional graph. The linear regression method was also calculated using the Python programming language.

Since this study uses only one comparison parameter, the linear regression formula will look like this:

$$Y = bX \tag{3}$$

where X is an independent variable, in this case it is a matrix obtained from an open source;

Y is dependent variable, in this case it is a matrix of correspondence calculated by the gravitational method and transformed into a time matrix.

To obtain a more accurate model it is necessary to conduct machine learning by loading the test data X and Y. The test data were the last 200 values of each temporary matrix. After the test, a linear regression of all remaining data can be run. The obtained two-dimensional graph of the linear regression presented in figure 5. Coefficient of the model is 0.84. R-squared metric is equal to 0.6.

IV. Conclusion

After a comparative analysis of the obtained model of the correspondence matrix with open data taken from Uber Movement, we can say that the resulting model is quite suitable for creating a simulation model of any city. The deviation from the ideal model is that there are no data on points of attraction, namely data on jobs, the number of students in schools and universities. When these data are obtained, it is possible to create an ideal model of any city.

Further stages of work on this project is the search for additional open data, as well as the creation of its own principle of dividing the city into transport areas, not based on data from Uber.

References

[1] "Bundesministeriumfür Verkehr. Bauund Wohnungswesen" (BMVBW).: Neubauvon Bundesautobahnen, 2004.

[2] A.V. Gasnikov, S.L. Klenov, E.A. Nurminskiy, Y.A. Kholodov and N.B. Shamrai, "Vvedeniye v matematicheskoye modelirovanie transportnykh potokov," 2010, MFTI. (In Russian).

[3] L. Moreira-Matias, "Time-evolving O-D matrix estimation using high-speed GPS data streams," Expert Systems With Applications, 2016, 44 pp.275–288.

[4] J.S. Russell, M. Ye, B.D.O. Anderson, H. Hmam and P. Sarunic, "Cooperative Localisation of a GPS-Denied UAV in 3-Dimensional Space Using Direction of Arrival Measurements," IFAC, vol. 50, 2017, issue 1 pp.8019-8024.

[5] S. Singh and P.B. Sujit, "Landmarks based path planning for UAVs in GPS-denied areas," IFAC, vol 49, 2016, issue 1 pp.396-400.

[6] B. Kim, D. Kim, S. Park, Y. Jung and K. Yi, "Automated Complex Urban Driving based on Enhanced Environment Representation with GPS/map," Radar, Lidar and Vision, IFAC, vol. 49, 2016, issue 11 pp.190-195.

[7] A.A. Kheir, "Intra-urban movement flow estimation using location based social networking data," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-1/W5, 2015 International Conference on Sensors & Models in Remote Sensing & Photogrammetry, Kish Island, Iran.

[8] Z. Zhang and Z. Wang, "The data-driven null models for information dissemination tree in social networks," Physica A: Statistical Mechanics and its Applications, 2017, pp.394-411.

[9] C. Yu, B. Xiao, D. Yao, X. Ding and H. Jin, "Using check-in features to partition locations for individual users in location based social network," Information Fusion, 2017, pp.86-97.

[10] Z. Jiao, L. Ran, J. Chen, H. Meng and C. Li, "Data-Driven Approach to Operation and Location Considering Range Anxiety of One-Way Electric Vehicles Sharing System," Energy Procedia, 2017, pp.2287-2294.

[11] K. Friso, "Enriching the transport model of the Rotterdam region by cell phone data," Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2015.

[12] F. Simini, "A universal model for mobility and migration patterns," Macmillan Publishers Limited, 2012.

[13] W.J. Reilly, "The law of retail gravitation," New York, 1931.

[14] R. Wiedemann, “Simulation des Straßenverkehrs flusses,” PhD-thesis, University of Karlsruhe, Germany, 1974.

[15] M. Beckmann, C.B. McGuire and C.B. Winsten, “Studies in the economics of transportation,” RM 1488. Santa Monica: RAND Corporation, 1955.

[16] S.C. Fang, J.R. Rajasekera and H-S.J. Tsao, “Entropy optimization and mathematical programming,” Kluwer Academic Publisher, 1997.

[17] A.G. Wilson, “Entropy in urban and regional modelling,” London: Pion, 1970.

[18] M. Balmer, “Generating Day Plans Based on Origin-Destination,” 2005.

[19] T. Roughgarden and E. Tardos, “How bad is selfish routing,” Journal of the ACM, 2002.

[20] D. Krajzewicz, J. Erdmann, M. Behrisch and L. Bieker, “Recent Development and Applications of SUMO – Simulation of Urban Mobility,” International Journal On Advances in Systems and Measurements,2012.

[21] C. Hofer, G. Jäger, M. and Füllsack, “Generating Realistic Road Usage Information and Origin-Destination Data for Traffic Simulations: Augmenting Agent-Based Models with Network Techniques,” In: C. Cherifi, H. Cherifi, M. Karsai and M. Musolesi (eds) Complex Networks & Their Applications VI. COMPLEX NETWORKS 2017. Studies in Computational Intelligence, 689, pp.1223-1233. Springer, Cham, 2018.

[22] O. Saprykin and O. Saprykina, “Multilevel Modelling of Urban Transport Infrastructure,” In Proceedings of the 1st International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS-2015). Portugal, Lisbon: SCITEPRESS, 2015 pp.78-82.

[23] A.E. Gorev, K. Buttger, A.V. Prokhorov and R.R. Gizatullin, ”Transport modeling fundamentals,” St. Petersburg, Kosta, 2015, 168 p.

[24] O. Saprykin and O. Saprykina, “Validation of Transport Infrastructure Changes via Microscopic Simulation: A Case Study for the City of Samara,” Russia, In Proceedings of the 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS 2017) in Naples, Italy, 26-28 June, 2017 pp.788-793.

[25] L.M. Martinez, J.M. Viegas and E.A. Silva, “A traffic analysis zone definition: a new methodology and algorithm,” Springer Science+Business Media, LLC, 2009.