DOI: https://doi.org/10.30932/1992-3252-2020-18-24-40

The Problem of Data Mining in Modelling Traffic Flows in a Megapolis

Kuftinova, Natalia G., *Moscow Automobile and Road Construction State Technical University, Moscow, Russia*.*

Natalia G. KUFTINOVA

ABSTRACT

The article discusses the problems of using data mining in a transport model as a digital platform for analysing data on traffic flows in a megapolis, and prerequisites for creation in future of single data banks and an integrated environment for interaction of models of different levels as clusters of the digital economy, which will consider all modes of transport to assess transport demand and develop projects for organizing traffic in a megapolis.

The objective of the work is to study the processes of obtaining quantitative characteristics of objects of transport modelling when creating a single electronic environment by calculating the derived parameters of the transport network of a megapolis. Quantitative spatial characteristics of an object are associated with calculating the distance from a city centre and a main street and are determined using geographic information systems entailing consequent problem of data unification and efficient data storage.

As part of achieving that objective, it is shown that it is necessary to create a pre-processing and validation procedure for all primary transport data, since data sources have different formats and spatial interpolation of tracking data. For this, it is recommended to use various methods of data analysis based on GIS technologies, digital terrain modelling, topology of the road network and other objects of the transport network of a megapolis. Besides, the use of intelligent data should be preceded by formatting and grouping the source data in real time. The most common errors arise at the stage of the iterative process for obtaining quantitative characteristics of objects of transport modelling and building the optimal route in terms of travel time along a certain transport network.

The existing trends of urban growth require global digitalization of all transport infrastructure objects, considering changes in the functions of the transport environment and in intensity of traffic flows. Theis entails further development and implementation of new information technologies for data processing using neural networks and other digital technologies.

Keywords: transport system, transport flows of a megapolis, data mining, information and communication technology, assessed transit areas, GIS-technologies.

---

*Information about the author:
Kuftinova, Natalia G. – Ph.D. (Eng), Associate Professor at the Department of Automated Control Systems of Moscow Automobile and Road Construction State Technical University (MADI), Moscow, Russia, nat.gk@mail.ru.

Article received 21.04.2020, revised 23.10.2020, accepted 30.10.2020.

For the original Russian text of the article please see p. 24.

Background. The process of burst development of information and communication technologies and the implementation of 5G networks contribute to creation and accumulation of big data, rising importance of issues of high-quality data collection as well as of development of a single data source containing complete, relevant, and reliable information about any transport system. In practice, a possibility of using unstructured data (e.g., on capacity of assessed transport areas to generate and attract transit flows, on information necessary to develop matrix of inter-area trip distribution) for modelling traffic flows often meets difficulties. Obtaining such data is accompanied by signing of formal contracts and other administrative procedures. The use of geographic information systems (GIS) helps to overcome those and other problems related to the use of statistical data on current traffic flows and parking areas in the centres of gravity for the residents of the megapolis.

The study sets the *objective* of construction of models of spatial distribution of indicators of objects in the form of continuous surfaces based on discretely specified information using GIS technology to study the regularities of territorial structures that have the property of continuous distribution, the continual approximation of which depends on the degree of approximation and on the rate of reflexion of geographic patterns in that discrete information. When using a graphical tool for modelling artificial fields, e.g., artificial isolines (pseudo-isolines), careful assessment of representativeness of data, reliability of conclusions obtained and the suitability of approximation results for use with other layers of the GIS database should be performed.

To achieve this objective, the study employed *methods* for predicting speeds and for analysis of information about the topology of the road network based on time series and data mining to determine spatial coordinates. The methods for monitoring and predicting the characteristics of the traffic flow of the road network based on the spatiotemporal approach of predicting derived parameters allow through the analysis of previous periods to reduce impact of incomplete information about the current state of the traffic flow on the predicted values.

Results.

Data mining is a modern concept, which initially assumes that data can be inaccurate, heterogeneous, contain omissions and at the same time have gigantic volumes [1; 2]. The preparation and accuracy of such data depend on quality of the processes of:

- definition and analysis of requirements for data;
- collection (input) of data for storage;
- data pre-processing to ensure high-quality analysis.

The main sources of information for development of the positional part of GIS database are differently purposed maps. General geographic maps allow creating information objects describing the topographic features of the territory, primarily the relief and hydrographic network, which are used in almost all GIS, regardless of their thematic focus. Thematic maps (population, economy, physical maps) are basic sources for initial development and georeferencing of the corresponding spatial objects. Attributive information is mostly added later, from data sources that are not necessarily spatially coordinated. It is also possible to use complex atlases. Earth remote sensing materials make it possible to create spatial objects directly, without developing maps as an intermediate information layer. Sources of attributive information are also long-term hydrological and meteorological data, statistical materials in digital form, and other text materials. The main conditions for the possibility of using such data are the existence, at the time of its input, of spatial objects to which it belongs, and the possibility of unambiguous correlation of the attribute data with a specific object or their group.

In total, four main conceptual functions of GIS can be distinguished, of which the first two (data collection and data processing) are preparatory and are implemented, most often, once (with the exception for GIS which require constant data updating; then their collection and processing are performed repeatedly, with a certain periodicity). To the greatest extent, the possibility of acquiring new knowledge is manifested in the process of analysis of spatially coordinated data, for which various conceptual and formal models are constructed using GIS.

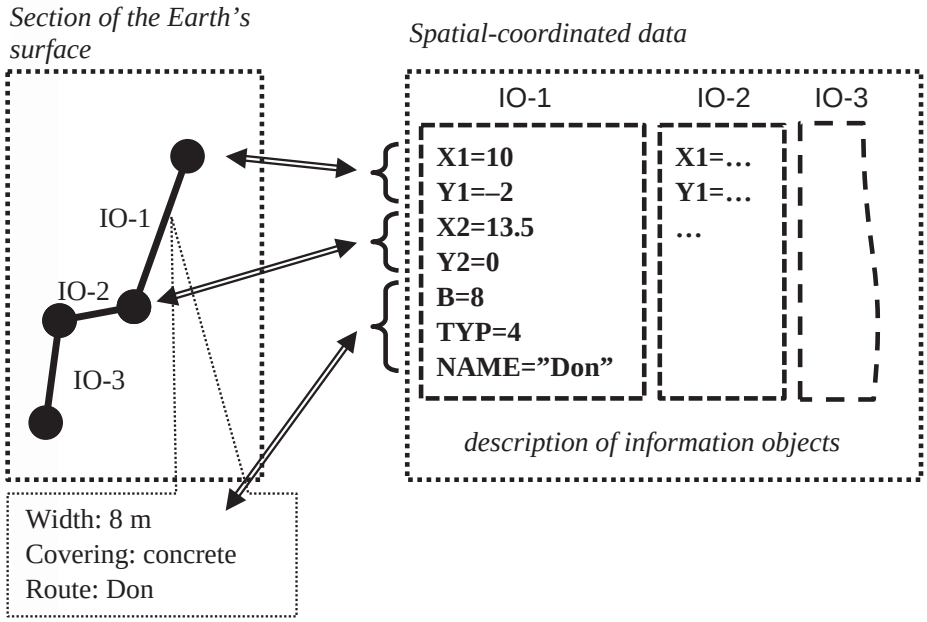
The main function of GIS is to provide information support for managerial decisions

made following the results of analysis and modelling based on spatially coordinated data. The applied functions (technological procedures) supported by the corresponding GIS software include a wide range of possibilities. Entering and editing data allow creating information objects based on both cartographic information, originally based on spatially coordinated data, and by directly entering the coordinates of the object. The support of various spatial models (grid, quadrotomic, vector models) by tools of GIS makes it possible to create different types of information objects based on the same initial data. Data storage is carried out considering the presence of a positional component in it (for which, as a rule, rather highly specialized formats are used); by contrast, attribute data uses widely accepted formats. A hierarchical or network model can be used to organize a GIS database, but relational databases are more common.

Coordinate system transformation is a commonly used GIS application. Such a need may arise already at the stage of data entry, when using several sources with different coordinate systems; in this case, data is reduced to some single unified system. Most often, geographic coordinates are converted to rectangular ones (Gauss–Krüger coordinate system), and vice versa. A geographic coordinate system is more convenient for storing positional data of spatial objects; while Gauss–Krüger coordinate system is more suitable for their visualization and implementation of spatial analysis. The main feature of transformation of coordinates from one system to another is that a curved section on the earth's surface (the surface of a spheroid) cannot be positioned («stretched») on a plane without an error. The magnitude of this error grows as the distance of the point from the origin of rectangular coordinates increases, so Gauss– Krüger system always covers only a relatively small area, within which the distance from the zero point does not exceed several hundred kilometres. At such distances, the error in the coordinates of individual points, and hence in the lengths and areas determined on their basis, remains at an acceptable level. Usually, all data used by a particular GIS is stored in a single coordinate system, and the need to transform it when analysing data rarely arises. At the same time, transformation of cartographic projections is a fairly common operation; this is due to the ambiguity of representation on the plane of the picture, which is initially located on the convex surface of the globe. The most commonly used GIS models are raster and vector models. The possibility of their mutual transformation in automatic (or as close as possible to it) mode is another important applied function of GIS. Possibilities for data analysis and transformation, due to the presence of a positional component, are implemented by a group of relevant GIS functions. It includes measuring operations and analytical geometry operations such as calculating lengths, volumes, areas, distances. Polygonal operations allow consolidating and separating areas, as well as revealing the facts of their overlapping and hitting a given area of points and lines. Spatial-analytical operations are used to analyse the proximity of objects, build buffer zones, conduct network analysis (to determine connectivity or disconnectedness of a network, identify the shortest route). Geomodelling consists in implementation with software of the laws of change in the parameters of objects, including in time, followed by modelling the behaviour of the system of objects under consideration. Digital elevation modelling is considered as a separate applied GIS function. The procedure for constructing relief is closely related to the tasks of its analysis, such as interpolation of heights and construction of isolines. The final applied function of GIS is associated with data output, including visualization of spatial objects and of their attributes, export of data to other information systems [3–6].

The problem of using data mining stands out not only in the works of domestic researchers, but also in global scientific community, in practices of the use of distributed computing technologies, and in particular of cloud computing environment, based on integration of various technologies of data mining, processing of big data, distributed computing and cloud computing.

The range of associated topics is presented in popular and research publications, particularly in those dedicated to transport policy. Thus, for example, the New York Times magazine lists nine problems associated with big data [7]. The analysis of such data is very difficult and therefore, recently, a large number of studies have been devoted to development of relevant algorithms, including mining



Pic. 1. The essence of spatial-coordinated data. (The author's drawing is based on work [15]).

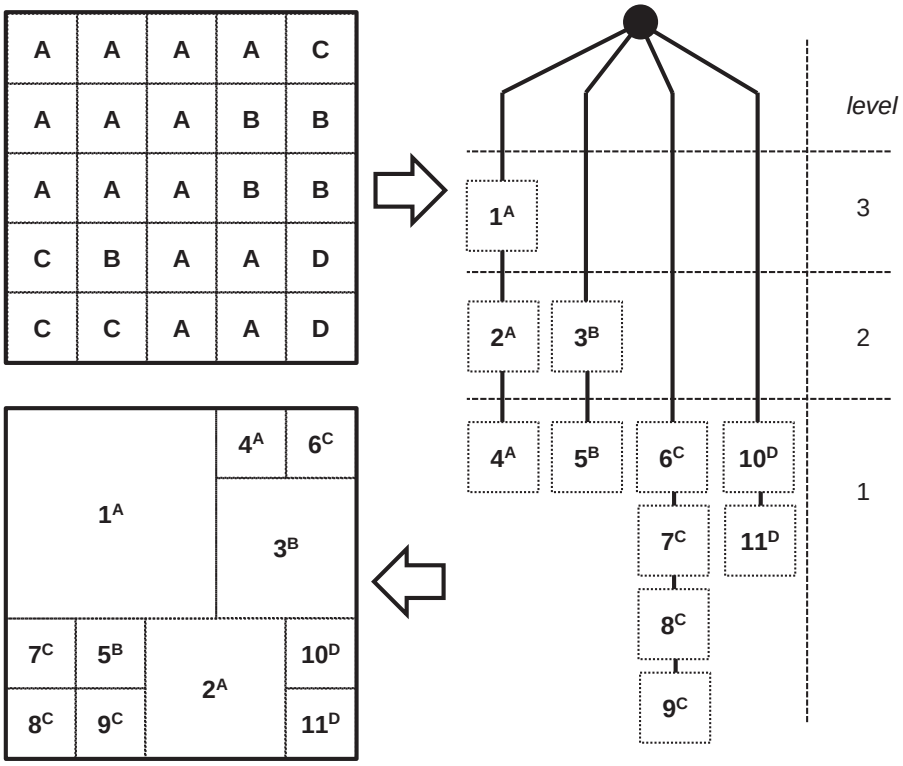
algorithms for parallel data processing in cloud computing systems, as well as for simulation of transportation processes [8–12]. For example, data mining techniques have been used to create models (classifiers) to predict severity of injury in any new accident with reasonable accuracy based on 5 973 Abu Dhabi road accident records during the six-year period from 2008 to 2013. The study aimed to establish a set of rules that can be used by the transport agencies of the United Arab Emirates (UAE) to determine the main factors influencing severity of an accident [13].

Currently, one of the most popular Data Mining tools (e.g., according to KDnuggets[1]) is RapidMiner system, developed by the company of the same name. This product is Open Source, and the version with the minimum functionality is distributed free of charge. RapidMiner implements a client-server architecture. RapidMiner Server can be used on its own to provide mining capabilities in the form of web services, thereby implementing the SaaS cloud computing model. RapidMiner implements all the necessary operations for analysis: data extracting, loading, and transforming (ETL), data pre-processing, data visualization, solving data mining problems [14]. It has an open architecture, providing the ability to extend it with new algorithms, including algorithms implemented with RWeka package.

The initial data for considered transport forecasting is provided with time series of speed and information about the topology of the road network, presented, for example, in the form of a graph. Spatial information objects are created based on spatially coordinated data (Pic. 1). In the example shown in the picture, a real physical object on the earth's surface (a road segment) is presented as a set of straight-line segments. The accuracy of such a representation, if necessary, can be made arbitrarily high by increasing the total number of smaller segments. Each segment is described by the corresponding information object (IO-1, IO-2, etc.). To describe a segment, it is necessary, first, to fix its position on the earth's surface; for this, in the example under consideration, it is sufficient to indicate the coordinates of its start and end points (X1, Y1, X2, Y2). Secondly, each road segment is characterized by certain performance indicators (such as the width of the carriageway or the type of pavement), which are also reflected while creating the information object (using the attributes B, TYP and NAME).

The data structure for creating information objects of different types will most likely also be different, but in any case, data will remain

[1] E.g.: [Electronic resource]: https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html.



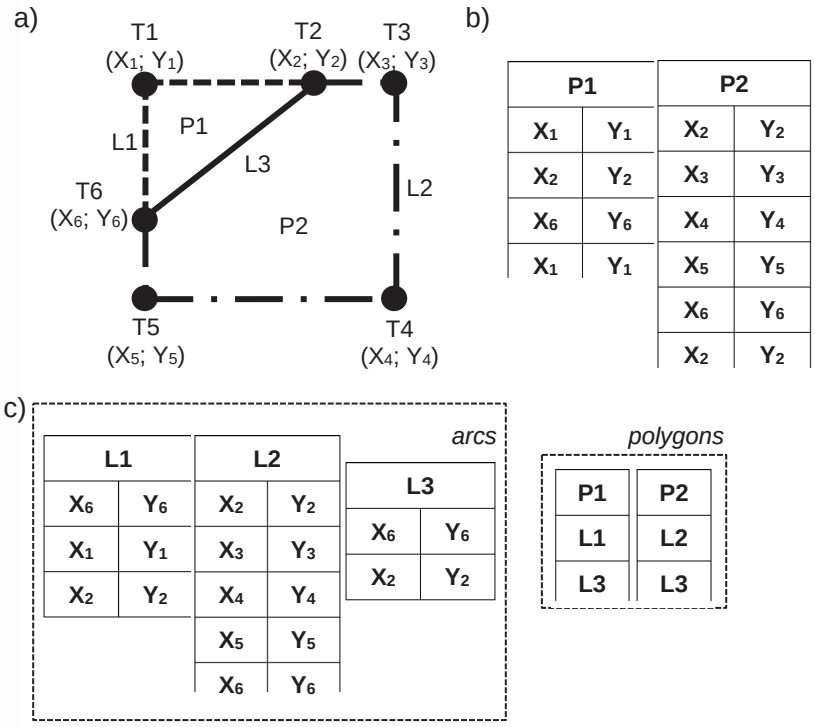
Pic. 2. Construction of a quadrotomy-based data model (the author's drawing is based on work [15]).

spatially coordinated. All technological procedures when working with spatially coordinated data (its input, editing, transformation, storage, transmission, displaying) have significant features in comparison with other types of data. These features are primarily due to the need to maintain a constant linkage between positional and attribute data of specific information objects. In addition, the presence of positional information leads to the emergence of fundamentally new opportunities, and therefore of technological procedures for their implementation: reflection of information objects on the screen (their visualization), organization of search for objects depending on their location, including relative to each other, etc. Another common model of spatial objects is using quadrotomy (Pic. 2). When building it, the so-called quadtree is used, which is formed starting from the highest possible level of object aggregation. In the example in Pic. 2 the highest level of aggregation is the third one; so nine objects with A attribute can be combined into a single 3 x 3 element.

The enlarged object is marked as 1A (here the object number and attribute value are combined), which is reflected in the quadtree and in the final layout of objects.

Since there are no more aggregation possibilities at level 3, the transition to level 2 is made. Here it is possible to create two enlarged objects, marked as 2A and 3B, after which we go to the lowest level, at which objects that do not have the possibility to be enlarged are marked and reflected in the quadtree and on the layout diagram. As a result of constructing a quadrotomy model in Pic. 2, it was possible to reduce the total number of objects from 25 to 11 without losing the attribute information. The reduced number of objects, and the resulting compactness of data storage, are the advantages of the quadrotomy model. In addition, the quadtree, created at the stage of model building, can subsequently provide the fastest possible search for objects by a given attribute. The model is distinguished by a rapid increase in resolution with increasing quadtree levels number; it also allows variable spacing when downsizing elements across levels.

To describe three-dimensional objects, an octotomic model (octotree) is used, which forms objects in the form of cubes of variable size.

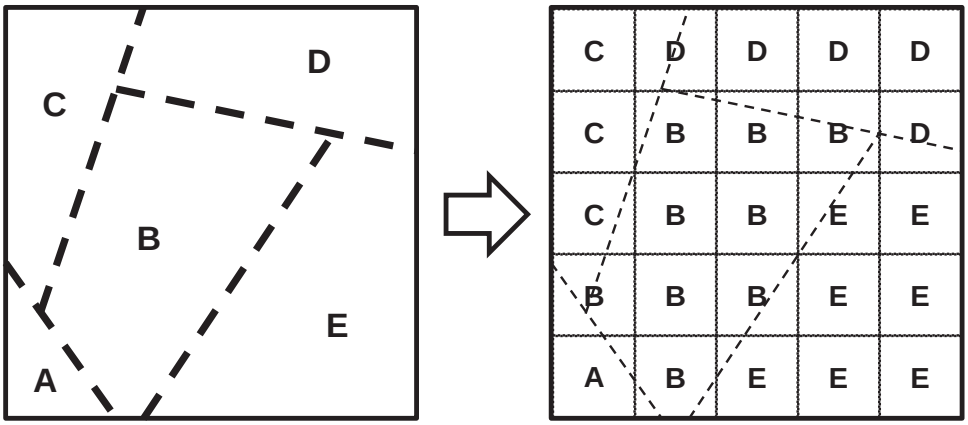


Pic. 3. Vector data models: a) initial polygons; b) non-topological model; c) topological model (author's drawing is based on the work [15]).

The vector spatial data model is characterized by maximum flexibility (since it does not contain restrictions on the shape of objects, their size and location), which simultaneously leads to the need to fully specify all positional data for each object. The vector model can be presented in two forms: non-topological, and topological. The non-topological variation (the «spaghetti» model) implies an independent indication of the boundaries for each object; this leads to the fact that data on the boundaries of neighbouring objects is duplicated and, as a result, the total amount of used data increases. In addition, when changing the boundaries of an object, it is necessary to identify and synchronously change the boundaries of all objects adjacent to it. The topological version of the model includes a set of individual polylines (arcs) that are the boundaries of objects, as well as information on belonging of these arcs to specific spatial objects. In this case, there is no duplication of information, and when editing the boundary of adjacent objects, the changes will be automatically reflected for both of them. An example of two varieties of the vector model is shown in Pic. 3, which shows two polygons P1 and P2 having an adjacent boundary L3.

The non-topological model contains merely instructions about the coordinates of the boundary points of each polygon (three points for P1 and five points for P2; the starting point is indicated a second time at the end of the list, as a sign that the boundary polyline is closed). The topological model first describes three arcs L1, L2 and L3, after which it indicates which arcs each of the polygons consists of. The topological model, in comparison with the non-topological one, is more complex, and for this reason it is used less often, mainly at the stage of forming spatial objects (entering their boundaries). At the stage of analysis and modelling, higher performance is provided by a non-topological model. Building of more complex objects based on elementary spatial elements is carried out within the framework of a certain model of spatial data. One of the simplest, and, at the same time, one of the most frequently used, is grid-type model, an example of which is shown in Pic. 4.

When using a grid model, the territory is divided into identical sections (cells), each of



Pic. 4. Raster grid data model (author's drawing is based on the work [15]).

which is an independent spatial object. The cell size determines the spatial resolution of the model. Attribute data for each object is defined independently of each other. Pic. 4 shows definition of one of five possible attribute values. A convenient feature of this type of model is that the spatial position of a cell is determined only by its number, which eliminates the need to specify the full positional data of each cell (coordinates of all corner-points). This enables compact storage of information. For example, in Pic. 4 objects can be described as a sequence of 25 attribute values (from A to E each). In this case, if necessary, the coordinates of all corner-points of each object can be easily restored by its number. Due to the presence of contiguous cells with the same attribute value, an even more compact description becomes possible, e.g., C4DC3BDC2B2E3B2EAB3E (where, 4D is used instead of DDDD, saving 2 characters). In grid models, it is possible to use cells of any regular shape, including curvilinear. It is also allowed to use different resolutions and different cell shapes for different attributes. In the case of rectangular cells, the only attribute of which is their colour, the model turns into a raster one. For example, to construct the optimal travel route in terms of travel time along a certain transport network, it is required to predict speed of vehicles within the segments. In this case, the forecasting horizon should be no less than the typical time of a single trip, so that when building a route before leaving, one can consider speed of movement at segments of the transport network close to the purpose of the trip.

Conclusion.

Besides using algorithms described above, for the purpose of transport modelling it is necessary to solve a series of problems referred to collection and processing of initial data.

To achieve maximum digitization performance, it is necessary to consider errors associated with interpretation of the results of data analysis, e.g., of analysis of readings of vehicle registration plates (number plates). Obviously, the forecast depends on time, day of the week and season of travel. External factors include weather conditions. In addition, there are nontrivial regularities between speeds of movement on different edges of the road network graph [16; 17]. For example, a very low speed in a certain area (traffic jam, congestion) can cause a decrease in speed on some ribs (congestion spreading) and at the same time an increase in speed on other edges due to the fact that the number of vehicles entering them sharply decreases (a sort of «shielding» effect). The number of posts for data collection and survey to create a transport model is assigned based on capacity of an observer or video camera to record the maximum possible number of events for monitoring traffic and pedestrian flows. But there are several shortcomings in relation to the observers and counting devices, such as recording heterogeneous information (simultaneous recording of the number of incoming and outgoing flows of vehicles or pedestrians [bi-directional counting]). Regarding the use of video recording as a tool of monitoring the centres of gravity, it is

worth noting some disadvantages associated with videorecording in the dark, manual video data processing (data digitization), difficulty to assess the number of passengers in private cars in case the vehicles' windows are even lightly tinted. The listed problems entail a sharp decrease in quality of collected and analysed data for further transport modelling.

Those problems can be solved thanks to improvement of methods and technical tools to collect initial data, but also through development of new information technology of their processing, particularly involving neural networks.

REFERENCES

1. GOST R [State Standard] 56670-2015 Intelligent transport systems. Subsystem for monitoring the parameters of traffic flows based on the analysis of telematic data of urban passenger transport [*GOST R56670-2015 Intellektualnie transportnie sistemy. Podsistema monitoring parametrov transportnykh potokov na osnove analiza telematicheskikh dannykh gorodskogo passazhirskogo transporta*]. [Electronic resource]: http://docs.cntd.ru/document/1200125977. Last accessed 20.07.2020.

2. Vaksman, S. A. Information technologies in management of urban public passenger transport (tasks, experience, problems) [*Informatsionnie tekhnologii v upravlenii gorodskim obshchestvennym passazhirskim transportom (zadachi, opyt, problemy)*]. Ed. by S. A. Vaksman. Yekaterinburg, Publishing house of AMB, 2012, 250 p. [Electronic resource]: http://www.waksman.ru/Russian/Criticism/Vaksman/Begin.pdf. Last accessed 20.07.2020.

3. Kuftinova, N. G. Modelling the dynamics of traffic flows using cluster analysis [*Modelirovanie dinamiki avtotransportnykh potokov s pomoshchyu klasternogo analiza*]. *Collection of scientific works of IV international scientific and practical conference «Transport planning and modelling»*, St. Petersburg, 2019, pp. 106–108. [Electronic resource]: https://interactive-plus.ru/e-articles/545/Action545-470172.pdf. Last accessed 20.07.2020.

4. Kuftinova, N. G. Mathematical modelling of traffic flows on the basis of macro- and micro-approaches of the urban transport system [*Matematicheskoe modelirovanie transportnykh potokov na osnove makro- i mikro-podkhodov transportnoi sistemy*]. *Collection of scientific works of III international scientific and practical conference «Transport planning and modelling. The digital future of transport management»*. Ed. by D.Sc. (Eng), Professor S. V. Zhankaziev. Moscow, MADI publ., 2018, pp. 67–76. [Electronic resource]: https://www.elibrary.ru/item.asp?id=37057419. Last accessed 20.07.2020.

5. Kuftinova, N. G. Intelligent transport infrastructure of a megalopolis based on geoanalysis and geo-modelling of motor transport systems [*Intellektualnaya transportnaya infrastruktura megapolisa na osnove geoanaliza i geomodelirovaniya avtotransportnykh sistem*]. *Logistic audit of transport and supply chains: Materials of an international scientific and practical conference*, Tyumen, TIU publ., 2018, pp. 76–82. [Electronic resource]: https://docplayer.ru/45169767-Udk-informacionno-logicheskaya-model-transportnoy-seti-megapolisa-kuftinova-n-g.html. Last accessed 20.07.2020.

6. Kuftinova, N. G. General characteristics of transport models for assessing the road network in urban areas [*Obshchaya kharakteristika transportnykh modelei dlya otsenki dorozhnoi seti na gorodskikh territoriyakh*]. *Collection of scientific papers of III international scientific conference «Scientific Discoveries», Karlovy Vary–Moscow*. Moscow, 2018, pp. 47–60. [Electronic resource]: https://www.elibrary.ru/item.asp?id=34934683. Last accessed 20.07.2020.

7. Marcus, G., Davis, E. Eight (No, Nine!) Problems with Big Data. [Electronic resource]: https://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html?_r=0. Last accessed 20.07.2020.

8. Chenyang Xu; Changqing Xu; Trieu-Kien Truong. Mining the spatio-temporal pattern using matrix factorisation: A case study of traffic flow. *IET Intelligent Transport Systems*, 2020, Vol. 14, Iss. 10, pp. 1328–1337. DOI: http://dx.doi.org/10.1049/iet-its.2019.0705. Last accessed 23.10.2020.

9. Alam, O., Kush, A., Emami, A., Pouladzadeh, P. Predicting irregularities in arrival times for transit buses with recurrent neural networks using GPS coordinates and weather data. *Journal of Ambient Intelligence and Humanized Computing*, 2020. DOI: https://doi.org/10.1007/s12652-020-02507-9. Last accessed 23.10.2020.

10. Guerreiro, G., Figueiras, P., Silva, R., Costa, R., Jardim-Goncalves, R. An architecture for big data processing on intelligent transportation systems. An application scenario on highway traffic flows. *2016 IEEE 8th International Conference on Intelligent Systems (IS)*, 2016, pp. 65–72. DOI: 10.1109/IS.2016.7737393. Last accessed 20.07.2020.

11. Kohan, M., Ale, J. M. Discovering Traffic Congestion through Traffic Flow Patterns Generated by Moving Object Trajectories. *Computers, Environment and Urban Systems,* March 2020, Vol. 80, Article 101426. DOI: 10.1016/j.compenvurbsys.2019.101426. Last accessed 20.07.2020.

12. Kumar, B. A., Vanajakshi, L., Subramanian, S. C. A Hybrid Model Based Method for Bus Travel Time Estimation. *Journal of Intelligent Transportation Systems*, 2018, Vol. 22, Iss. 5, pp. 390–406. DOI: 10.1080/15472450.2017.1378102. Last accessed 20.07.2020.

13. Taamneh, M., Alkheder, S., Taamneh, S. Data-Mining Techniques for Traffic Accident Modeling and Prediction in the United Arab Emirates. *Journal of Transportation Safety & Security*, 2017, Vol. 9, Iss. 2, pp. 146–166. DOI: https://doi.org/10.1080/19439962.2016.1152338. Last accessed 20.07.2020.

14. Norris, D. RapidMiner – a potential game changer. November 15, 2013. [Electronic resource]: https://www.bloorresearch.com/2013/11/rapidminer-a-potential-game-changer/. Last accessed 17.07.2020.

15. Yakubovich, A. N., Kuftinova, N. G., Rogova, O. B. Information technologies on vehicles: Study guide [*Informatsionnie tekhnologii na avtotransporte: Ucheb. posobie*]. Moscow, MADI, 2017, 252 p. [Electronic resource]: http://www.lib.madi.ru/fel/fel1/fel17E429.pdf. Last accessed 17.07.2020.

16. Mousa, S. R., Mousa, R. M., Ishak, S., Radwand, L. Modeling Speed-Density Relation for Highways in Developing Countries with No Lane Discipline: A Case Study in Egypt. *2017 ASCE India Conference*, 2018, pp. 725–735. DOI: 10.1061/9780784482025.074. Last accessed 17.07.2020.

17. Kong, X., Das, S., Jha, K., Zhang, Y. Understanding Speeding Behavior From Naturalistic Driving Data: Applying Classification Based Association Rule Mining. *Accident Analysis and Prevention*, September 2020, Vol. 144. DOI: https://doi.org/10.1016/j.aap.2020.105620. Last accessed 23.10.2020. ●