Modified Method of Subtractive Clustering for Modeling of Distribution of Harmful Vehicles Emission Concentrations

Mykola Dyvak[1], Yurii Maslyiak[2], Iryna Voytyuk[3], Bogdan Maslyiak[4]

Faculty of Computer Information Technologies, Ternopil National Economic University, UKRAINE, Ternopil, 8 A. Chekhova str., email: mdy@tneu.edu.ua[1], yuramasua@gmail.com[2], i.voytyuk@tneu.edu.ua[3], bm@tneu.edu.ua[4]

Abstract: **Mathematical modeling of distribution of harmful vehicle emissions concentrations is considered in the paper. The modified subtractive clustering method for modeling is proposed. This method is characterized by its implementation simplicity due to the fact that it does not require a large sample of experimental data and does not require to set a predetermined number of clusters. An example of clustering method application for data preparation for modeling of distribution of harmful vehicle emissions concentrations is given.**

Keywords: **mathematical modeling, clustering analysis, interval analysis, interval difference operator, harmful vehicle emissions, vehicular traffic.**

I. Introduction

One of the biggest problems of large and medium cities is the pollution of the surface atmospheric layer and soils by harmful emissions from vehicles. Large amount of harmful substances is concentrated in the vehicles exhaust gases. Among them, with high concentrations: Nitrogen oxides, Carbon oxides and Sulfur oxides. Motor transport is a source of atmospheric and soils pollution, which should be considered as distributed one. For reflecting and predicting of concentrations of harmful vehicle emissions, it is expedient to use mathematical models. They can be built based on the results of selective observations of their dynamics with known boundary errors of measurements. This approach was considered in [1].

The process of harmful emissions distribution and its dynamics is considered as a mass transfer process. For its description, the difference operators (schemes) are used. Their identification is carried out using the data of measurements of harmful emissions concentrations with known boundary errors. Such data are called interval data [2,3]. As is known, the methods of difference operators identification based on the interval data analysis require a uniform measurement grid that is impossible for the real city conditions. Mostly, the measurements of harmful emission concentrations is carried out in places with intensive traffic and accumulation of vehicles. This means that measurement grid is not uniform. Thus, for building of mentioned models, it is necessary to solve three tasks related to data preparation: execute cluster analysis for defining of homogeneous vehicular traffic intensity areas; identify the discrete values of the grid step; calculate estimates of harmful vehicle emissions concentrations in the nodes of the grid. The third task is solved by methods of interpolation [4,5]. The first one and second one are the subjects for research of this work.

II. Statement of the Problem

To solve the environmental monitoring tasks, it is necessary to build models of stationary and non-stationary fields of concentrations of harmful vehicle emissions [6]. The theoretical basis for solving this type of tasks are the mathematical models of objects with distributed parameters in the form of partial differential equations. Concentration of attention on the physical properties of environment requires to significantly complicate the mathematical model. Even despite the fact that, in practice, it is impossible to verify the results of modeling with real data obtained under conditions that meet the conditions of modeling. First of all, this is related to the complexity of the measurement experiment. For example, if a mathematical model in the form of a differential equation accurately enough describes the process of transferring of chemical substances in the atmosphere in case of wind gusts or other turbulent phenomena in the atmosphere, then an integrated value of the chemical substance concentration per volume unit is established in the process of measurement. In addition, the accuracy of such measurements is low, the relative measurement error may reach 50%. Consequently, it is enough to build a mathematical model with an accuracy that is equiualent to the accuracy of the measurement experiment. At this, it is expedient to represent the experimental data in the form of intervals of possible values of the modeled characteristic:

$$[z^{-}_{i,j,h,k}; z^{+}_{i,j,h,k}],$$
$$i=1,...,I,\ j=1,...,J,\ h=1,...,H,\ k=1,...,K \quad (1)$$

where $z^{-}_{i,j,h,k}, z^{+}_{i,j,h,k}$ are the lower and upper bounds of the interval of possible values of measured concentration of harmful substances in the grid nodes with discretely given spatial coordinates $i=1,...,I$, $j=1,...,J$, $h=1,...,H$ at the dicrete time value $k=1,...,K$, respectively.

It is worth to note that, in the measurement experiment, we can set the lower and upper bounds based on the relative error of the measuring device: $z^{-}_{i,j,h,k} = z_{i,j,h,k} - z_{i,j,h,k}\cdot\varepsilon$ and $z^{-}_{i,j,h,k} = z_{i,j,h,k} + z_{i,j,h,k}\cdot\varepsilon$, where $z_{i,j,h,k}$ is the measured value of the harmful substance concentration; ε is relative measurement error.

Under these conditions, macromodeling is the only way to reflect the distribution of harmful emissions concentrations. The building of such macromodels is convenient to carry out based on the obtained interval data in the form (1).

In the papers of O.G. Ivakhnenko [7], the inductive approach is described for choosing of acceptable way of mathematical description of these processes. Its essence consists in defining of some difference scheme in the way of its adjustment in accordance with the experimental data. The difference scheme itself, which actually converts the values of input variables into output values, is called a difference operator. The process of adjustment of this scheme is called structural identification [8,12].

In general case, the expression of linear in parameters difference operator (DO) has the following form [2]:

$$v_{i,j,h,k} = \vec{f}^{\,T}(v_{0,0,0,0},...,v_{0,0,h-1,0},v_{i-1,0,0,0},...,v_{0,j-1,0,0},..., v_{i-1,j-1,h-1,k-1},\vec{u}_{i,j,h,0},...,\vec{u}_{i,j,h,k})\cdot\vec{g}, \quad (2)$$
$$i=d,...,\mathrm{I},\ j=d,...,J,h=d,...,H,k=d,...,K$$

where $\vec{f}^{\,T}(\bullet)$ is the vector of basis functions (nonlinear, in general case) by which, the transformation of the modeled characteristic values, as well as the input variables in the spatial grid nodes for the certain discrete moments of time is carried out; $v_{i,j,h,k}$ modeled concentration of harmful emissions in grid nodes with discretely-given spatial coordinates $i=d,...,I$, $j=d,...,J$, $h=d,...,H$ at the moments of time $k=d,...,K$; $\vec{u}_{i,j,h,0},...,\vec{u}_{i,j,h,k}$ are the vectors of input variables (controls); d is the DO order; $\vec{g}$ is the vector of unknown parameters of DO.

As a result of executing of structural identification procedure, we establish the difference computational scheme, in particular: the basis functions vector $\vec{f}^{\,T}(\bullet)$; sets and dimensionality of input variables (controls) vectors $\vec{u}_{i,j,h,0},...,\vec{u}_{i,j,h,k}$; order of the difference scheme *d*, which, as is known, is equivalent to the order of the differential equation (the analogue of the difference scheme). To realize the difference scheme, it is also necessary to set the initial conditions, that is, the value of each discrete element from the set $v_{0,0,0,0},...,v_{0,0,h-1,0},v_{i-1,0,0,0},...,v_{0,j-1,0,0},...,$ $v_{i-1,j-1,h-1,k-1}$, $\vec{u}_{i,j,h,0},...,\vec{u}_{i,j,h,k}$ (as a rule, the initial ones) and to establish the values of the parameters vector $\vec{g}$ components. If the structure of DO is known then, it remains the actual task of adjusting the parameters of DO (2) in such a way as to ensure the maximal consistency between the modeled characteristic and experimentally obtained values of this characteristic. Such a task is called the parametric identification task [9,13].

Based on the requirements of ensuring the mathematical model accuracy within the bounds of the measuring experiment accuracy, the conditions of consistency between experimental data, represented in the interval form (1), and data obtained based on the mathematical model in the form of DO (2), can be formulated in such form:

$$v_{i,j,h,k} \subset [z^{-}_{i,j,h,k}; z^{+}_{i,j,h,k}],$$
$$\forall i=d,...,I,\forall j=1,...,J,\forall h=d,...,H,\forall k=d,...,K \quad (3)$$

Based on the results of conducted analysis, we can state that for ensuring of conditions of the given accuracy (3) of the macromodel in the form of linear DO (2) during solving the task of its parametric identification, the application of interval data analysis methods [9] is substantiated.

Let's assume that the vector of parameters estimates $\hat{\vec{g}}$ in the DO (2) is obtained based on the interval data analysis. Substituting the vector of parameters estimates $\hat{\vec{g}}$ of DO instead of their true values $\vec{g}$ in expression (2) together with given initial interval values of each element in the set $[\hat{v}_{0,0,0,0}],...,[\hat{v}_{0,0,h-1,0}],[\hat{v}_{i-1,0,0,0}],...,[\hat{v}_{0,j-1,0,0}],...,$ $[\hat{v}_{i-1,j-1,h-1,k-1}]$ and given vectors of the input variables $\vec{u}_{i,j,h,0},...,\vec{u}_{i,j,h,k}$, we obtain an interval estimate of the harmful substance concentration $[\hat{v}_{i,j,h,k}]$ in the nodes with discretely given spatial coordinates $i=1,...,I$, $j=1,...,J$, $h=1,...,H$ at discrete moments of time $k=1,...,K$:

$$[\hat{v}_{i,j,h,k}]=[\hat{v}^{-}_{i,j,h,k};\hat{v}^{+}_{i,j,h,k}]=\vec{f}^{\,T}([\hat{v}_{0,0,0,0}],...,[\hat{v}_{0,0,h-1,0}], [\hat{v}_{i-1,0,0,0}],...,[\hat{v}_{0,j-1,0,0}],...,[\hat{v}_{i-1,j-1,h-1,k-1}], \vec{u}_{i,j,h,0},...,\vec{u}_{i,j,h,k})\cdot\hat{\vec{g}}, \quad (4)$$
$$i=1,...,I,\ j=1,...,J,h=1,...,H,k=1,...,K$$

Thus, the mathematical model of stationary and non-stationary fields of harmful emissions concentrations for the task of environmental control will be described by a DO in general form (4). Taking into account that all calculations in equation (4) are carried out using interval arithmetic rules [2], the difference operator (4) is called an interval difference operator (IDO).

The conditions of consistency of experimental data, represented in interval form (1), with the data obtained based on macromodel in the form of IDO (4) are formulated as follows:

$$[\hat{v}^{-}_{i,j,h,k}\hat{v}^{+}_{i,j,h,k}] \subset [z^{-}_{i,j,h,k}; z^{+}_{i,j,h,k}],$$
$$\forall i=1,...,I,\forall j=1,...,J,\forall h=1,...,H,\forall k=1,...,K \quad (5)$$

Let's substitute in expressions (5), instead of interval estimates of the harmful substance concentrations $[\hat{v}^{-}_{i,j,h,k};\hat{v}^{+}_{i,j,h,k}]$, its interval values calculated using IDO (4), together with taking into account the given initial interval values of each element from the set

$$[\hat{v}_{0,0,0,0}]\subseteq[z_{0,0,0,0}],...,[\hat{v}_{0,0,h-1,0}]\subseteq[z_{0,0,h-1,0}],$$
$$[\hat{v}_{i-1,0,0,0}]\subseteq[z_{i-1,0,0,0}],...,[\hat{v}_{0,j-1,0,0}]\subseteq[z_{0,j-1,0,0}],... \quad (6)$$
$$[\hat{v}_{i-1,j-1,h-1,k-1}]\subseteq[z_{i-1,j-1,h-1,k-1}]$$

and given vectors of input variables $\vec{u}_{i,j,h,0},...,\vec{u}_{i,j,h,k}$. We obtain such interval system of non-linear algebraic equations (ISNAE) [3]:

$$\begin{cases} [\hat{v}^{-}_{0,0,0,0};\hat{v}^{+}_{0,0,0,0}]\subseteq[z^{-}_{0,0,0,0};z^{+}_{0,0,0,0}],..., \\ [\hat{v}^{-}_{i-d,j-d,h-d,k-d};\hat{v}^{+}_{i-d,j-d,h-d,k-d}]\subseteq[z^{-}_{i-d,j-d,h-d,k-d};z^{+}_{i-d,j-d,h-d,k-d}]; \\ [\hat{v}_{i-1,j-1,h-1,k-1}]=\vec{f}^{\,T}([\hat{v}_{0,0,0,0}],...,[\hat{v}_{0,0,h-1,0}],[\hat{v}_{i-1,0,0,0}],..., \\ [\hat{v}_{0,j-1,0,0}],...,[\hat{v}_{i-d,j-d,h-d,k-d}],\vec{u}_0,...,\vec{u}_{k-1})\cdot\hat{\vec{g}}; \\ z^{-}_{i,j,h,k}\le\vec{f}^{\,T}([\hat{v}_{0,0,0,0}],...,[\hat{v}_{0,0,h-1,0}],[\hat{v}_{i-1,0,0,0}],...,[\hat{v}_{0,j-1,0,0}],..., \\ [\hat{v}_{i-d,j-d,h-d,k-d}],\vec{u}_0,...,\vec{u}_k)\cdot\hat{\vec{g}}\le z^{+}_{i,j,h,k}; \\ i=d,...,I,d=2,...,J,h=d,...,H,k=d,...,K. \end{cases} \quad (7)$$

So, the ISNAE (7) is obtained by substituting the interval estimates of the output characteristic (given in the form of initial conditions and predicted using expression (4) in the remaining nodes of the grid) into conditions (5). Therefore, the task of parametric identification of IDO (4) under conditions (5) is the task of solving ISNAE in the form (7). Methods for estimation of solutions of the obtained ISNAE are described in [10].

The analysis of the proposed scheme for building of mathematical model of harmful vehicle emissions distribution showed that before its implementing, it is necessary to obtain a uniform grid of measured concentrations (3) in its nodes and vectors of influences on them $\vec{u}_{i,j,h,0},...,\vec{u}_{i,j,h,k}$. The main among them, is the vehicular traffic intensity. This task is solved using modified method of subtractive clustering of data [11] on the traffic intensity.

III. Modified Subtractive Clustering Method

As the basis for method of clustering of vehicular traffic distribution, it is expedient to use the "mountain" clustering method with subtractive algorithm of its implementation. This method does not require a large sample of experimental data and does not require to set a predetermined number of clusters that significantly reduces the time for its implementation. It is also worth to note that the number of clusters based on this method is regulated by the only parameter which is the cluster radius [11].

According to the clustering method, in the beginning, we form the potential cluster centers from the rows of data matrix for the clustering of input variables and calculate the potentials of identified cluster centers using the expression:

$$P_h\left(c_h\right)=\sum_{k=1}^{K}\exp\left(-\alpha\cdot\left\|\vec{c}_h-\vec{x}_k\right\|\right), \quad (8)$$

where $\vec{c}_h=(c_{1h},c_{2h},...,c_{Kh})$ is a potential center of *h-th* cluster; *a* is a positive constant; $\left\|\vec{c}_h-\vec{x}_k\right\|$ is a distance between potential center of *h-th* cluster $\vec{c}_h$ and input experimental data $\vec{x}_k$, *k*=1,…,*K*, $h=1,...,H$; *H* is a number of possible clusters.

In our case, if the only property of a cluster which is the number of vehicles $u_{xi,yj,k}$ in the point with coordinates x_i, y_j at a discrete moment of time *k* is taken into account, the expression for estimation of potentials of given cluster centers, has such form:

$$P_h\left(x_h,y_h,k\right)=\sum_{i=1}^{I}\sum_{j=1}^{J}\exp\left(-\alpha\cdot\left\|\overline{u}_{x_h,y_h,k}-u_{x_i,y_j,k}\right\|\right), \quad (9)$$

where $P_h\left(x_h,y_h,k\right)$ is potential of a point (center of cluster with coordinates x_h, y_h at moment of time *k*); $\overline{u}_{x_h,y_h,k}$, $u_{x_i,y_j,k}$ are the numbers of motor vehicles in a point of potential cluster center x_h, y_h, k and in points x_i, y_j, k with defined traffic intensity and measured concentrations, respectively.

The illustration of the potentials distribution is represented as a surface in the form of a mountainous relief (Fig. 1), whose peaks have the highest potential values and are pretenders to be the centers of the formed clusters.

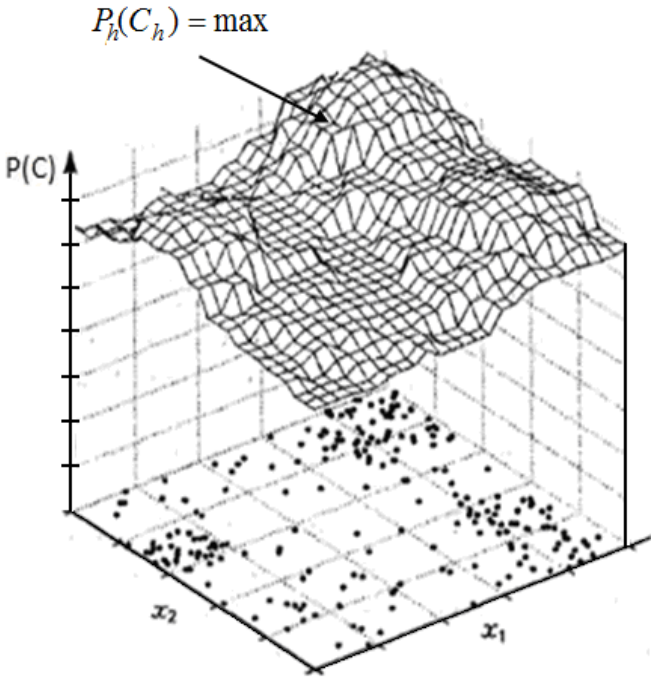

Fig. 1. The illustration of potentials distribution based on the mountain clustering method.

As we can see in the Fig. 1, one "mountain peak" is surrounded by other peaks that causes the problem of building of very similar data clusters with the corresponding centers. This does not provide the high quality clustering results.

As centers of the clusters, we choose the coordinates of "mountain peaks", that is, the center of the cluster is the point on the city map with the highest value of potential:

$$(x_h,y_h,k)=\arg\max_{h=1,...,H}P_h\left(x_h,y_h,k\right). \quad (10)$$

In order to avoid the building of similar clusters, we must recalculate the potential values for the remaining possible cluster centers:

$$P_{h+1}\left(x_{h+1},y_{h+1},k\right)=P_{h+1}\left(x_h,y_h,k\right)-P_h\left(x_h,y_h,k\right)\cdot \exp(-\beta\cdot\left\|\overline{u}_{x_{h+1},y_{h+1},k}-\overline{u}_{x_h,y_h,k}\right\|),h=1,\ldots,H, \quad (11)$$

where $P_h\left(x_h,y_h,k\right)$ is a potential center of *h*-th cluster on *h*-

th iteration; $P_{h+1}(x_h, y_h, k)$ is a potential of center of h-th cluster on h+1 iteration; β is a positive constant, $\left\|\bar{u}_{x_{h+1},y_{h+1},k} - \bar{u}_{x_h,y_h,k}\right\|$ is a distance between potential center of h+1 cluster and center of found h-th cluster.

The procedure of cluster centers calculation is carried out until all the rows of the input variable matrix X, which is represented by the set (3), are excluded.

The above procedure is based on the subtractive clustering algorithm, which is based on the following steps.

Step 1. Forming of potential cluster centers. They are all points of measured harmful emission concentrations and corresponding intensities of vehicular traffic.

Step 2. Calculation the potential of possible cluster centers based on (9).

Step 3. Selecting the data point with the maximal potential for representation of the cluster center based on (10).

Step 4. Excluding the influence of the found cluster center in the way of recalculating the potentials for other possible cluster centers by (11).

Step 5. Identifying the next cluster and the coordinates of its center. If the maximal value of the cluster center potential exceeds some predetermined threshold which is the cluster radius, that is $P_h(x_h, y_h, k)$, then proceed to *step 4*, otherwise, the algorithm is completed.

The iterative procedure for identification of cluster centers and the recalculation of potentials is repeated until all points in the space of input experimental data are located within the neighborhoods of the radius of sought cluster centers.

As a result of the clustering algorithm implementation, we obtain h clusters, $h = 1,\ldots,H$, with the corresponding centers. The next step is the identification of a uniform grid nodes for homogeneous parts of a cluster. The discrete values of the grid nodes coordinates are equal to the cluster diameter, and the value $\bar{u}_{x_h,y_h,k}$ is the number of vehicles in the point of a cluster center x_h, y_h, k. To assign the number of vehicles at k-th moment to the grid nodes, it is enough to analyze, what cluster the node is in. If the node belongs to the h-th cluster, then, the number of vehicles in the node is $\bar{u}_{x_h,y_h,k}$.

IV. Example of Clustering Method Application

Let's consider the application of the developed clustering method for obtaining of uniform grid of nodes on an example of Ternopil city.

The fragment of map of central part of Ternopil city with the marked points of the measured vehicular traffic intensity for one discrete time (one hour) is shown in the Fig. 2. As we can see, the traffic is distributed not uniformly over the territory. Therefore, it is advisable to measure its intensity at some selected points, where this intensity is the highest, for example, as it is shown on the map of Ternopil city.

The points of measurement of traffic intensity are colored red on the map.

The application of cluster analysis for determination of areas with specific vehicular traffic intensities under condition of identification of cluster centers that are located on a certain uniform grid gives a possibility to define the discretization step for building a DO. In our case, the cluster is the set of points of a certain area of the city with similar values of current vehicles number.

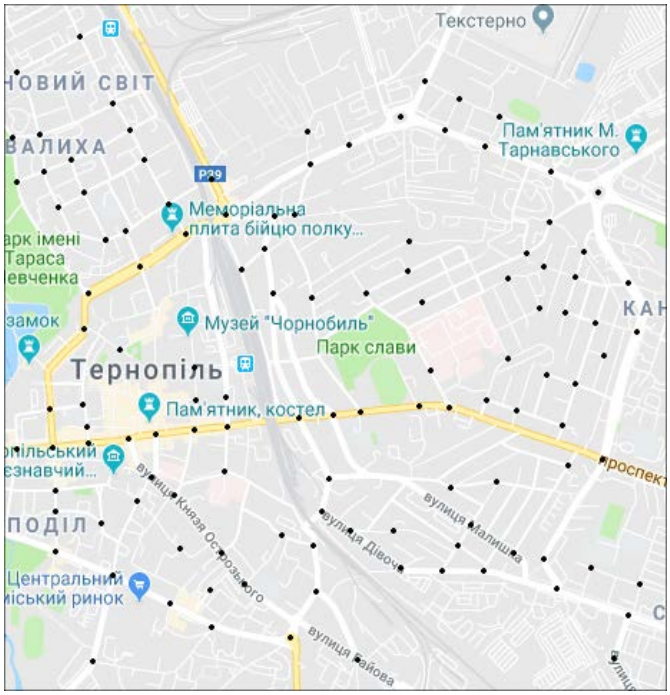

Fig. 2. Points of measurements of vehicular traffic intensity on the example of Ternopil city.

The result of the proposed method of cluster analysis application is schematically shown in Fig. 3. As we can see, during the clustering process, H clusters with different vehicular traffic intensity and radius r were defined. So, the discrete values of grid nodes coordinates are equal to the cluster diameter.

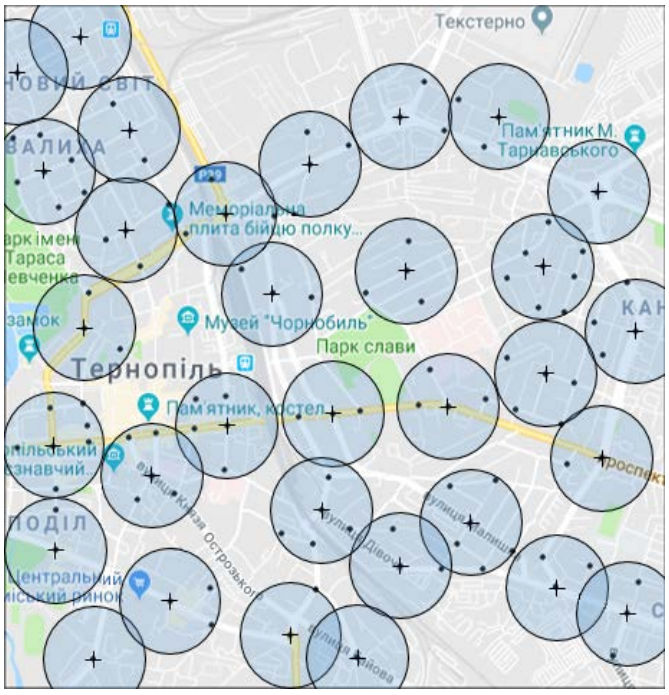

Fig. 3. The result of clustering of vehicles quantity distribution on the Ternopil city map.

Obtained grid for building of distribution model of harmful vehicle emission concentrations in the form of IDO (4) is schematically shown in Fig. 4.

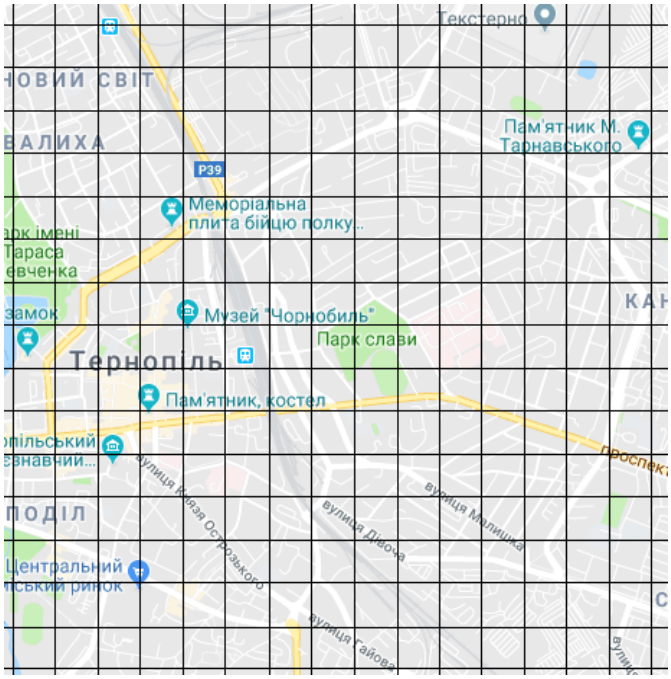

Fig. 4. The uniform grid with known spatially distributed vehicular traffic intensities.

For building of mentioned model in the form of IDO (4), it is enough to execute interpolation and identify the pollution concentrations in the grid nodes.

V. Conclusions

The modified method of subtractive clustering and interval analysis for modeling of distribution of harmful vehicle emissions concentrations and vehicular traffic intensity under conditions of non-uniform measurement grid were proposed and substantiated.

Acknowledgment

This research has been supported by National Grants of Ministry of Education and Science of Ukraine “Mathematical tools and software for control the air pollution from vehicles” (0116U005507) and “Mathematical tools and software for classification of tissues in surgical wound during surgery on the neck organs” (0117U000410).

References

[1] A. Veremchuk, A. Pukas, I. Voytyuk and I. Spivak, "Mathematical and software tools for modeling objects with distributed parameters," *2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET)*, Lviv, 2016, pp. 149-152.

[2] I. Voytyuk, M. Dyvak, V. Nemish “Method and genetic algorithm for structure identification of interval differentice operators in the tasks of environmental monitoring”, *Collected Works of Donetsk National Technical University Series "Information, cybernetics and computer science."*, 2011, Vol. 14 (188), pp. 8-17.

[3] M. Dyvak “Mathematical Modeling Tasks of Static Systems with Interval Data”, Ternopil: TNEU Publishing House "Economic Thought", 2011, 216 p.

[4] M. Dyvak, I. Voytyuk, T. Dyvak, A. Pukas “Application of the interval difference operator for approximation of fields of harmful emissions concentration from vehicles”, *Measuring and Computing Devices in Technological Processes*, 2011, No. 1 (37), pp. 44-52.

[5] Kvyetnyy R. N., Dementiev V. Yu., Mashnitsky M. O., Judin O. O. “Difference methods and splines in multidimensional interpolation problems”, Vinnitsa: UNIVERSUM, 2009, 87 p.

[6] N. Ocheretnyuk, I. Voytyuk, M. Dyvak and Y. Martsenyuk, "Features of structure identification the macromodels for nonstationary fields of air pollutions from vehicles," *Proceedings of International Conference on Modern Problem of Radio Engineering, Telecommunications and Computer Science*, Lviv-Slavske, 2012, pp. 444-444.

[7] A.G. Ivakhnenko “Inductive method of self-organizing of models of complex systems”, Kyiv: Scientific thought, 1981, 296 p.

[8] M. Dyvak, I. Voytyuk, T. Dyvak, A. Pukas “Application of the interval difference operator for approximation of fields of harmful emissions concentration from vehicles”, *Measuring and Computing Devices in Technological Processes*, 2011, No. 34 (110), pp. 86-94.

[9] T. Dyvak “Parametric identification of interval difference operator on the example of macromodel for distribution of humidity in the drywall sheets in the process of drying”, *Information Technologies and Computer Engineering: international Scientific Journal*, 2012, Vol. 3, pp. 79-85.

[10] M. Dyvak, N. Porplytsya, Y. Maslyak, M. Shynkaryk “Method of Parametric Identification for Interval Discrete Dynamic Models and the Computational Scheme of Its Implementation,” *Advances in Intelligent Systems and Computing II: Selected Papers from the International Conference on Computer Science and Information Technologies, CSIT 2017*, pp.101- 112, 2018.

[11] Shtovba S. Introduction to the theory of fuzzy sets and fuzzy logic. Access mode: http://matlab.exponenta.ru/fuzzylogic/book1/index.php

[12] Porplytsya, N., Dyvak, M., Dyvak, T., Voytyuk, I. “Structure identification of interval difference operator for control the production process of drywall.” *Proceedings of 12th International Conference on the Experience of Designing and Application of CAD Systems in Microelectronics, CADSM’2013*, pp. 262-264 (2013).

[13] Fliess, M., Sira-Ramirez, H. “Closed-loop parametric identification for continuous-time linear systems via new algebraic techniques.” *H. Garnier & L. Wang. Identification of Continuous-time Models from sampled Data, Springer*, pp. 362–391, 2008.