УДК 519.853

A.V. Kozhevnikov[1], Cand. Sc. (Tech.), Associate Professor,
Ye.A. Krivosheyev[2]

1 – State Higher Educational Institution "National Mining University", Dnipropetrovsk, Ukraine, e-mail: anton_ko80@mail.ru
2 – ZAO "Visicom", Kiev, Ukraine, e-mail: evgenkr@visi.com.ua

ESTIMATION OF THE POPULATION DENSITY SPATIAL DISTRIBUTION USING CLUTTER MODEL

А.В. Кожевников[1], канд. техн. наук, доц.,
Е.А. Кривошеев[2]

1 – Государственное высшее учебное заведение „Национальный горный университет", г. Днепропетровск, Украина, e-mail: anton_ko80@mail.ru
2 – ЗАО „Визиком", г. Киев, Украина, e-mail: evgenkr@visi.com.ua

ОПРЕДЕЛЕНИЕ ПРОСТРАНСТВЕННОГО РАСПРЕДЕЛЕНИЯ ПЛОТНОСТИ НАСЕЛЕНИЯ С ИСПОЛЬЗОВАНИЕМ КЛАТТЕРНОЙ МОДЕЛИ

The algorithm of quantitative assessment of the population density spatial distribution on the rank data basis is shown. The algorithm is based on Gauss-Zaydel method with "flexible" restrictions. Source data specifications that satisfy a single-valued result are given. Calculation results in the MATLAB media are shown. It is shown that the result is independent from initial density values, but it is defined by the minimum quantitative difference between density values of clutters with neighbor ranks.

Keywords: *Clutter, population density, rank, optimization, restriction*

The term "clutter" is being commonly used in a foreign literature, serving as an applied analogy of a "cluster". In regards to the Population Density Clutter Model (PDCM) clutter is a specified landscape or urban territory with a population density rank assigned. PDCM has a practical significance for implementation of tasks closely related to GIS technologies. In particular, it is used for planning of mobile networks' base stations workload, accompanied with digital terrain models (DTM). The peculiarity of PDCM enables modeling of base stations' workload with territorial geometry, area, and population density spatial distribution taken into account at the same time. The "radioplanning" (RF-planning = radio-frequency planning) term is being used to describe the planning of mobile network systems' infrastructure over a certain area. It is commonly used in the sphere of telecommunications and information technologies by a number of companies that produce corresponding original software. The most popular software solutions and vendors are: ASSET (Aircom), Atoll (Forsk), Planet (Mentum), SignalPro (EDX).

One of the basic requirements for population density spatial distributions, used for radioplanning, is the high spatial resolution. The last one is defined by a constantly growing demand for mobile operators' service, increase of information traffic and network quality requirements, thus resulting in network density growth. Herewith, the distance between basic stations in some urban areas is about 200–300 meters nowadays. Another factor that influences much the quality of wireless networks is the intensive growth of mobile Internet services. It requires a drastic increase of a radio line bandwidth, followed by increased width of channel frequency range and the carrier frequency.

The mobile network standard of GSM-1800 with the 1.8 GHz frequency is the growing alternate of the nowadays most widely used Ukrainian mobile network standard of GSM-900 with the 0.9 GHz frequency. These mobile network standards represent wireless networks of the second generation (2G) and allow Internet access via WAP text protocol. Most of worldwide networks of the third generation (3G) which work with full-valued Internet protocols use the 2.1 GHz frequency. The shift up to 2.7 GHz frequency is possible while migration into the next generation of modern networks (4G). Hence the peculiarities of wave propagation under shown frequency ranges make it necessary to decrease the distance between base stations.

This is the reason why in most cases the direct calculation of population density gives inapplicable average results, because often the population quantity is well-measured only over relatively large administrative areas (regions). The approach of indirect measurements of population density based on the density of road network graph branches is also inapplicable for practical use.

The problem statement of population density spatial distribution calculation under the PDCM framework is shown as following [1]: source data represent population quality values for N regions, for which the calculation will be made, $\{P_i\}_{i=1,...,N}$, accompanied with population density rank values $\{r_j\}_{j=1,.,M}$ for clutters of M types and their areas in regions $\{S_{ij}\}_{i=1,.,N;\ j=1,.,M}$.

The task is to calculate population density values for clutters $\vec{d} = \{d_j\}_{j=1,..,M}$, that minimize the following objective function

$$W(\vec{d}) = \sqrt{\sum_{i=1}^{N}\left(P_i - \sum_{j=1}^{M} d_j S_{ij}\right)^2}\,, \qquad (1)$$

where the following restrictions apply

$$d_1 \geq 0;\ d_k > d_m \ where\ r_k > r_m\ \forall\ k, m\,. \qquad (2)$$

The solution of the above-mentioned task is commercial classified information of GeoImage Corporation.

If restrictions (2) are not present the task to minimize the objective function (1) is fully compliant with calculation of multiple line regression coefficients under the least-squares criterion, which may be accomplished analytically. However, with restrictions present, it should be considered as a non-linear programming task with the objective to get the quantitative assessment of population density clutters, initially given in an ordinal scale. Hence, the solution of the objective has to increase the population density clutter measurement scale. Generally the task to increase the measurement scale of an attribute is considered to be incorrect [2], because the source data may be not enough to reach the single-valued definition of the objective.

The objective of the present work is to solve the task of population density spatial distribution calculation with the use of clutter model. Objective achievement means the assessment of various possible ways to solve the task, source data which is enough to get a single-valued definition, and creation of a calculation algorithm followed by its software implementation.

Numeric methods of non-linear programming are generally divided into the following [3]: non-gradient, which use the objective function value while searching, and gradient ones, which also require calculation of partial derivatives of the objective function by independent variables (i.e. solution elements). Some authors assign methods of random search as a stand-alone class of methods, but in fact most of them are modifications of well-known non-gradient or gradient ones with use of pseudo-random numbers.

Non-gradient methods that enable a simple approach for objective restrictions (e.g. the scanning method or a blind search one) are inapplicable for the current task due to the large volume of the search area. In particular, the last one is conditional to a large number of independent variables and undefined range of their change. As the objective function has an analytical expression, its partial derivatives are simple to calculate

$$\frac{\partial W}{\partial d_j} = -\frac{\sum_{i=1}^{N}\left[S_{ij}\left(P_i - \sum_{j=1}^{M} d_j S_{ij}\right)\right]}{\sqrt{\sum_{i=1}^{N}\left(P_i - \sum_{j=1}^{M} d_j S_{ij}\right)^2}}\,, \qquad (3)$$

This enables gradient methods to be used in principal. The earliest software implementations of the designed algorithm, based on the gradient method of the most rapid slope as well as on the non-gradient one with the help of one-by-one variable change, have shown their respectively low calculation complexity. This circumstance has smoothed over the main advantage of gradient methods, e.g. the rapid optimum search. On the other hand, the easiest way to consider restrictions (2) was realized by the method of one-by-one variable change (Gauss-Zaydel method), which is the algorithm basis.

Source data of six regions of Samara city (Russia) were used, given by the CJSC “Visicom”, Kyiv [4]. The dataset was given as text files of clutter matrices, received by processing of satellite images of analyzed regions with 1 meter resolution, under the technology stated in [5]. Each matrix cell has a density clutter rank with a described area of 5x5 meters. The classification of clutters is given in Table 1.

Table 1

Clutter classification

Color	Code	Clutter Name	Clutter Description	Rank
1	2	3	4	5
	1	Open area	Open space outside the town: meadow, field, marsh	0
	2	Forest	Forest of more than 20 m average height trees	0
	3	Sea	Ocean, sea	0
	4	Inland water	Rivers, canals of more than 10 m width, lakes, reservoirs	0
	5	Residential, suburban	Country houses, 1–3 storey neighborhoods	5
	6	Urban, mean urban	Mean urban building, more than 3 and less than 7 storey height	9

Table 1 contination

1	2	3	4	5
	7	Dense urban	Dense urban building, more than 3 and less than 7 storey height	10
	8	Block of buildings	Blocks of buildings of more than 7–storey height	12
	9	Industrial and commercial areas	Industrial and commercial zones	3
	10	Villages	Villages, sanatoriums	4
	11	Open_areas in urban	Open spaces inside the town: streets, avenues, vacant, lots squares	0
	12	Parks in urban	Forest, park of less than 20m height trees, cemeteries	1
	13	Airport	Airports	2
	14	Open wet area	Marshes, swamp	0
	15	Dense residential	Dense non-regular building 1–3 storey height	6
	16	Dense urban high	Dense urban building, more than 7 storey height. Financial, business centers, malls.	11
	17	Mixed urban	Mixed urban building 1–6 storey height	7
	18	Mixed dense urban	Mixed dense urban building 1–6 storey height	8
	19	Buildings	Higher and isolated buildings, skyscrapers	13
	20	Semi open area	The territory covered by several kinds of vegetation – low grass, bush, individual trees	0

Fig. 1 shows the clutter picture of one of the analyzed city regions, with use of the rank-functional color map. Fig. 2 shows a part of the clutter matrix, which corresponds to the given Picture 1 sample.

Fig. 1. Clutter picture of the city region sample

```
12 12 12 12 20 20 20 20 20 20 20 20 20 5 5 5 5 5 5 5 5 5 11 5 5 5 5 5 5 5 5 5 5
 12 12 12 20 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 11 5 5 5 5 5 5 5 5 5 5 5
2 12 12 12 12 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 11 5 5 5 5 5 5 5 5 5 5
2 12 12 12 20 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 11 5 5 5 5 5 5 5 11
12 12 12 12 12 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 5 5 11 5 5 5 5 5 11
2 12 12 12 12 20 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 5 5 11 5 5 5 11 5
 12 12 12 12 20 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 5 5 5 11 11 11 5 5
 12 12 12 12 12 20 20 20 20 20 20 20 20 20 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
2 12 12 12 12 20 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
12 12 12 12 12 12 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
12 12 12 12 12 12 20 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
12 12 12 12 12 20 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
12 12 12 12 12 12 20 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 12 12 12 12 12 12 20 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
12 12 12 12 12 12 20 20 20 20 20 20 20 20 20 20 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
12 12 12 12 12 12 20 20 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 12 12 12 12 12 12 20 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
2 12 12 12 12 12 20 20 20 20 20 20 20 20 20 20 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
12 12 12 12 12 12 12 20 20 20 20 20 20 20 1 1 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
12 12 12 12 12 12 20 20 20 20 20 20 20 1 1 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 12 12 12 12 12 12 12 20 20 20 20 20 20 20 1 20 1 11 11 11 11 11 11 11 11 5 5 5
```

Fig. 2. Part of the clutter matrix, which corresponds to the Picture 1 sample

In order to ease the data processing and decrease the volume of calculations while making an algorithm and debugging of software implementation, the following assumptions have been made:

– population density in clutters with zero rank are equal to zero;
– clutter indexes were ordered by ascending ranks;

– the population density in clutter of the first rank may not be less than a priori given value, i.e. $d_1 \ge d_{\min}$.

In the analyzed datasets clutters with codes 2, 3, 7, 10, 15, 18 were missing, and clutters with codes 1, 4, 11, 14, 20 were excluded due to zero ranks. Therefore, the following population densities were calculated only for clutters with codes 5, 6, 8, 9, 12, 13, 16, 17, 19, that after being ordered by ascending, had the following sequence: 12, 13, 9, 5, 17, 6, 16, 8, 13 and ranks $\{d_j\}_{j=1,\dots,9} = \{1,\ 2,\ 3,\ 5,\ 7,\ 9,\ 11,\ 12,\ 13\}$.

While getting of initial approximations of objective elements, it was stated that quantitative values of clutter population densities are proportional to their ranks and equal to

$$d_j^0 = r_j \frac{\sum_{i=1}^{N} P_i}{\sum_{j=1}^{M}\left(r_j \sum_{i=1}^{N} S_{ij} \right)}. \qquad (4)$$

To research the solution stability in relation to the source data before calculation of densities' initial approximation, the non-linear transformation to their ranks was applied.

$$r_j^{'} = \left(r_j\right)^p, \qquad (5)$$

where $p > 0$ is the transformation parameter.

The design of a numeric solution algorithm and its software implementation was made by three steps. The first step was to set "strict" restrictions for changing variables. They allowed the equality of solution elements with different ranks, $d_k \ge d_m \ where\ r_k > r_m\ \forall\ k,m$, but assumed the impossibility influence of the currently-changed variable onto its restriction values. The value of the currently-changed variable on the $k+1$ iteration step was calculated recursively

$$d_j^{k+1} = \left(1 \pm \delta\right) d_j^k, \qquad (6)$$

where δ is the value of relative change of a variable within the iteration step. The sign within the expression was assigned according to the direction, where the objective function is improved. Calculation results of this step have shown that the solution is not stable in regards to the source data and weak decrease of the objective function as the optimization result.

"Flexible" restrictions were introduced on the second step. In this case if the current variable j was increasing, the maximum value of variable's index q, which could change, was determined from condition $d_q^k \le d_j^{k+1}$. Then all variables with indexes from $j+1$ to q the value of d_j^{k+1} was assigned. If the objective function was decreasing, changes of independent variables were ultimately assigned. If it was missing, the step of changing of the currently-changed variable was fragmented. Same types of operations were applied if the value of the currently-changed variable was decreasing. Calculations have shown that the solution is stable in regards to the source data and it has a significant (up to 30 percent) decrease of the objective function as the optimization result. However, the calculation result has shown only two values of population density, formed for common clutters with ranks 1–3 and 5, 7, 9, 11, 12, 13 correspondingly. The practical inapplicability of the such version algorithm was shown.

An addition to "flexible" restrictions, next statement was introduced on the third step. The minimum allowed difference between population densities of clutters with neighbor ranks is limited by a "clearance" of the C size

$$d_{j+1} - d_j \ge C d_{j+1},\ or\ d_{j+1} \ge \frac{d_j}{1-C}. \qquad (7)$$

The value of C was constant in hereinafter calculations, however, its value may be set individually for each pair of clutters. Taking this restriction into account, the rank correction was used before calculating of initial approximations of expression (4).

$$r_{j+1}^{'} = \frac{r_j}{1-C}\ \ where\ r_{j+1} < \frac{r_j}{1-C}. \qquad (8)$$

If the current variable j was increasing, the maximum value of variable's index q from condition

$$d_q^k - C\sum_{s=j+1}^{q} d_s^k \le d_j^{k+1} \qquad (9)$$

was determined. Then all variables with indexes from $j+1$ to q the value of d_j^{k+1} was assigned.

$$d_s^k = d_j^{k+1} + C\sum_{t=j+1}^{s} d_t^k. \qquad (10)$$

If the objective function was decreasing, changes of independent variables were ultimately assigned. If it was missing, the step of changing of the currently-changed variable was fragmented. Same types of operations were applied if the value of the currently-changed variable was decreasing.

Software implementation of the shown algorithm is made in MATLAB. Calculated values of the objective function and population density of clutters for different algorithm parameters are given in Table 2. Fig. 3 and 4 show corresponding plots of values of the objective function W and population densities of 9 clutters $\{d_j\}_{j=1,\dots,9}$ versus the number n of the optimization step for different values of the parameter C with $p=1$. In this case the term "step" defines the full variation cycle of a single independent variable. Hence result values of population density of clutters with minimum and maximum ranks differ greatly, density plots are shown using the logarithmic scale of the ordinate axis.

Table 2

Calculated values of the objective function and population density of clutters for different algorithm parameters

Algorithm parameters		Objective function value	Ranks of population densities of clutters r_j								
			1	2	3	5	7	9	11	12	13
C	p	$W, \cdot 10^4$	Values of population density of clutters $d_j, \cdot 10^{-3}$ m^{-2}								
0	0,5	1,37	0,8	0,8	0,8	4,6	4,6	4,6	4,6	4,6	4,6
0	1,0	1,37	0,8	0,8	0,8	4,6	4,6	4,6	4,6	4,6	4,6
0	1,5	1,37	0,8	0,8	0,8	4,6	4,6	4,6	4,6	4,6	4,6
0,05	0,5	1,44	0,8	0,8	0,8	4,3	4,5	4,7	5,0	5,2	5,5
0,05	1,0	1,44	0,8	0,8	0,8	4,3	4,5	4,7	5,0	5,2	5,5
0,05	1,5	1,44	0,8	0,8	0,8	4,3	4,5	4,7	5,0	5,2	5,5
0,1	0,5	1,53	0,8	0,8	0,9	3,9	4,3	4,8	5,3	5,9	6,6
0,1	1,0	1,53	0,8	0,9	1,0	3,9	4,3	4,8	5,3	5,9	6,5
0,1	1,5	1,53	0,8	0,8	0,9	3,9	4,3	4,8	5,3	5,9	6,6
0,2	0,5	1,75	0,5	0,7	0,8	3,1	3,9	4,9	6,1	7,7	9,6
0,2	1,0	1,75	0,5	0,7	0,9	3,1	3,9	4,9	6,1	7,6	9,6
0,2	1,5	1,74	0,6	0,7	0,9	3,1	3,9	4,9	6,1	7,6	10,2
0,3	0,5	2,01	0,4	0,5	0,8	2,3	3,3	4,8	6,8	9,7	25,4
0,3	1,0	2,01	0,4	0,5	0,8	2,3	3,3	4,7	6,8	9,7	25,6
0,3	1,5	2,01	0,4	0,5	0,8	2,3	3,3	4,7	6,8	9,7	25,6
0,4	0,5	2,42	0,4	0,6	1,0	1,7	2,8	4,7	7,8	13,1	21,8
0,4	1,0	2,42	0,4	0,6	1,0	1,7	2,8	4,7	7,8	13,1	21,8
0,4	1,5	2,41	0,4	0,6	1,0	1,7	2,8	4,7	7,8	13,0	21,6
0,5	0,5	2,95	0,1	0,3	0,5	1,1	2,2	4,4	8,8	17,5	35,1
0,5	1,0	2,95	0,1	0,3	0,5	1,1	2,2	4,4	8,8	17,5	35,1
0,5	1,5	2,95	0,1	0,3	0,5	1,1	2,2	4,4	8,8	17,5	35,1

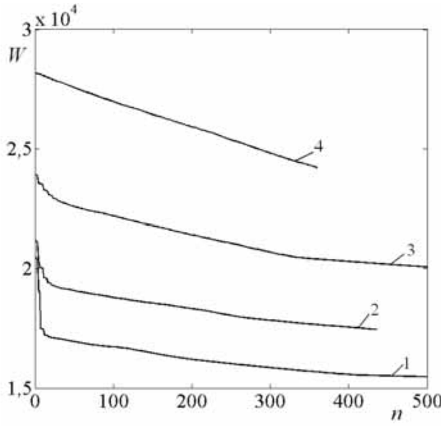

Fig. 3. Plot of values of the objective function W versus the number n of the calculation step (for different values of the parameter C): 1 – C=0.1; 2 – C=0.2; 3 – C=0.3; 4 – C=0.4

The data shown state the result is independent from initial density values, with their changes being guaranteed by modification of the transformation parameter p. Besides, an independence of the solution from the variables change order was established. Parallel curves of change of independent variables in logarithmic scale during the process of optimization prove that the approach of "flexible" restrictions with "clearance" was the right one to be implemented. At the same time, the search process and the obtained result significantly depend on the C parameter. Therefore, in order to calculate the practically-valued assessment of population density under the current algorithm framework, the source dataset should include values of this parameter (e.g. derived from the expert evaluation method).

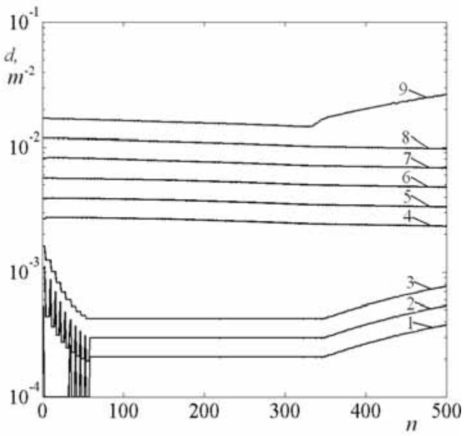

Fig. 4. Plots of values of population densities of clutters d versus the number n of the calculation step (for clutters with different ranks r): 1 – r=1; 2 – r=2; 3 – r=3; 4 – r=5; 5 – r=7; 6 – r=9; 7 – r=11; 8 – r=12; 9 – r=13; with value of parameter C=0.3

Thus, the current paper shows the solution algorithm of assessment of the population density spatial distribution with the use of clutter model and calculation results, obtained by its software implementation. It is shown that the result is independent from initial density values, but it defined by the minimum quantitative difference between density values of clutters with neighbor ranks. If this data is supplied, the implemented software may be used for calculation of practically-significant assessments of population density spatial distribution.

References

1. GeoImage – Population maps – GEOpop [Electronic resource]: GeoImage Corporation official site – Mode of access: http://www.geoimage.fr/images/ stories/products/ geopop.pdf – Title from the screen.
2. Гайдышев И. Анализ и обработка данных. Специальный справочник/ Игорь Гайдышев. – СПб.: Питер, 2001. – 752 с. – ISBN 5-318-00220-X.
3. Бояринов А.И. Методы оптимизации в химической технологии/ А.И. Бояринов, В.В. Кафаров – М.: „Химия“, 1975. – 576 с.
4. Візіком – геоінформаційні системи, радіопланування, GPS навігатори, LBS проекти [Електронний ресурс]: офіційний сайт ЗАТ „Візіком“ – Режим доступу: http://www.visicom.ua – Назва з екрану.
5. 3D модели для планирования и оптимизации беспроводных сетей связи. – К.: ВИЗИКОМ, 2010. – 15 с.

Розглянуто алгоритм отримання кількісних оцінок просторового розподілу щільності населення на основі рангових даних. Алгоритм грунтується на методі Гаусса-Зайделя з „гнучкими“ обмеженнями. Сформульовано вимоги, які пред’являються до вихідних даних, достатні для отримання однозначного рішення. Представлено результати розрахунків у середовищі MATLAB. Показано, що результат не залежить від початкових щільностей, але визначається мінімальною кількісною різницею між щільностями клаттерів із сусідніми рангами.

Ключові слова: *клаттер, щільність населення, ранг, оптимізація, обмеження*

Рассмотрен алгоритм получения количественных оценок пространственного распределения плотности населения на основе ранговых данных. Алгоритм основывается на методе Гаусса-Зайделя с „гибкими“ ограничениями. Сформулированы требования, предъявляемые к исходным данным, достаточные для получения однозначного решения. Представлены результаты расчётов в среде MATLAB. Показано, что результат не зависит от начальных плотностей, но определяется минимальным количественным различием между плотностями клаттеров с соседними рангами.

Ключевые слова: *клаттер, плотность населения, ранг, оптимизация, ограничения*

Рекомендовано до публікації докт. техн. наук Б.С. Бусигіним. Дата надходження рукопису 28.02.11

УДК 004.8

М.В. Назаренко,
Л.В. Саричева, канд. фіз.-мат. наук, доц.

Державний вищий навчальний заклад „Національний гірничий університет“, м. Дніпропетровськ, Україна, e-mail: sarycheval@nmu.org.ua

АЛГОРИТМ КЛАСТЕРИЗАЦІЇ НА ОСНОВІ НЕЧІТКИХ МНОЖИН

M.V. Nazarenko,
L.V. Sarycheva, Cand. Sc. (Phys.-Math.) Assoc. Prof.

State Higher Educational Institution “National Mining University”, Dnipropetrovsk, Ukraine, e-mail: sarycheval@nmu.org.ua

CLUSTERING ALGORITHM BASED ON FUZZY SETS

Запропоновано математичну модель і метод кластеризації, що враховує інтуїтивне уявлення про групування даних, не накладаючи апріорних припущень щодо структури даних. Метод кластеризації FuzzyCluster розроблено на основі нечіткого опису, здатного функціонувати в умовах апріорної невизначеності щодо структури даних, а також такого, що враховує інтуїтивне уявлення про групування даних. Порівняння результатів кластеризації алгоритмом FuzzyCluster з алгоритмами k-means та c-means на п’яти стандартних наборах даних показує його переваги.

Ключові слова: *кластеризація, нечітка множина, міра близькості, функція належності, нечітке відношення*

Вступ. Кластеризація даних – процес групування елементів даних на класи так, що елементи в одному класі є якомога близькими, а елементи різних класів є настільки різнорідними, наскільки це можливо.

У жорсткій кластеризації дані розділені на окремі кластери, де кожен елемент даних належить одному з кластерів. У нечіткій кластеризації елементи даних можуть належати до більш ніж однієї групи і з кожним елементом множини пов’язана функція належності до кожного кластеру. Вона вказує на силу зв’язку між цим елементом даних і конкретною групою. Нечітка кластеризація є процесом присвоєння цих мір належності та їх використання для визначення складу кожного з кластерів.

Основні відомі алгоритми кластеризації (наприклад, модифікації алгоритмів K-Means, Expectation Maximization, реалізовані, у тому числі, у Microsoft Analysis Services 2005) накладають обмеження на гео-