

Kohonen cards for clustering fund of the residential real-estate

*Olga Popova*¹, *Tatiana Simankina*^{2*}, and *Vitaly Lukinov*³

¹ Northern (Arctic) Federal University, Severnaya Dvina Emb. 17, Arkhangelsk, 163002, Russia

² Peter the Great Saint-Petersburg Polytechnical University, Polytechnicheskaya str., 29, St.Petersburg, , 19525, Russia

³ Moscow State University of Civil Engineering, Yaroslavskoe shosse, 26, Moscow, 129337, Russia

Abstract. The algorithm of a clustering of fund of the residential real-estate is based on a neural network simulation using T. Kohonen's maps. Self-organizing maps (SOM) divide 296 objects into 16 clusters based on 33 signs. An important result of the research is the possibility of structural analysis of housing stock which allows to form an idea of his general condition. During periodic inspection and analysis of the condition of housing stock relocation of an object in other cluster will specify the changed technical condition. The quantitative composition of similar clusters will allow to determine the necessary volume of investment of these activities.

1 Introduction

The analysis of the housing stock condition showed that main capital investments are put into new construction while participants of investment and construction activities are interested in a project with minimal payback period without consideration of complex renovation of existing developments. That leads to a considerable limitation of range of repair services, loss of industrial technologies of major overhaul, a decrease of consumer characteristics of existing residential houses. The solution for such conditions is a design of a concept of complex renovation and development of the built-up city area. This process includes monitoring of the condition of existing buildings and deciding on a list and amount of reproductive actions such as demolition of old and dilapidated buildings, designing and construction of new ones, reconstruction and modernization of existing housing facilities, their major overhaul.

Mathematical modeling techniques are widely used in the field of process control Construction of buildings or real estate economic parameters [1-7]. However, the analysis showed that the mathematical modeling has little application in the field of the technical operation of buildings.

The purpose of the work is an adaptation of clustering algorithm based on neural networks of Kohonen's self-organizing maps (SOM) implemented during the monitoring

* Corresponding author: talesim@mail.ru

and a structural analysis of the city housing facilities for complex reproduction of buildings [8].

2 Materials and Methods

Self-Organizing Map (SOM) is an efficient programming tool of neural network modeling for visualization and generalization of multidimensional data. It is suitable for solving complex problems as a process analysis, a machine perception, management and transferring of information. SOM formally can be defined as non-linear, ordered, smooth mapping of high dimension input data which to a regular array of elements of low dimension.

Let's accept that the set of input variables $\{\xi_j\}$ can be determined as a weight vector $x = [\xi_1, \xi_2, \dots, \xi_n]^T \in \mathfrak{R}^n$. There is two-dimensional ordered array of nodes (neurons) in the Figure 1, each of them has associated conjoint model, which is parametric weight vector $m_i = [\mu_{i1}, \mu_{i2}, \dots, \mu_{in}]^T \in \mathfrak{R}^n$.

Initial values of m_i can be chosen randomly, preferably from the area of input parameters. We can say that SOM is «a nonlinear projection» of probability density of function $p(x)$ of multidimensional input data vector x on a two-dimensional display. An array of nodes can represent rectangular, hexagonal and even irregular type grid (Fig. 2). Hexagonal grid is the most effective for the purpose to visually represent data.

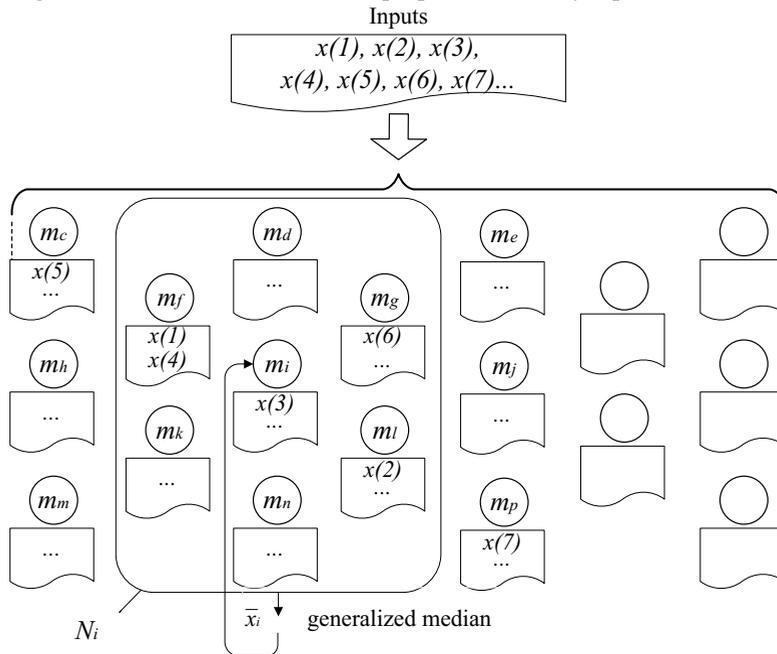


Fig. 1. Example of self-organizing process – two-dimensional sorted array

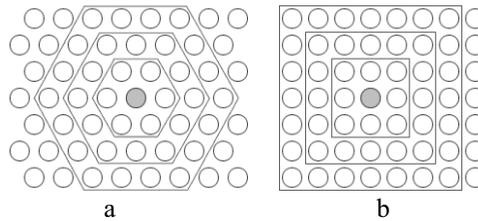


Fig. 2. The distance among neurons on the map for hexagonal (a) and rectangular (b) grids

Let's study a list of input data $x(t)$, where t is an integer index. Sometimes it is suggested that the value x should be normalized before it is fed to the considered algorithm. Normalization is not necessary but can improve the accuracy of the calculations because that way resulting reference vectors have the same dynamic range of changes of its values. Let's compare each object $x(t)$ with m_i , after that let's copy $x(t)$ in the list associated with the node for which the reference vector is most similar to $x(t)$ according to the employed common measure of distance. Often for determining the most adequate node Euclidean distance computation is used $\|x - m_i\|$:

$$\|x - m_i\| = \sqrt{\sum_{j=1}^n (\xi_j - \mu_{ij})^2} \tag{1}$$

The node's index is determined by the following expression:

$$c = \arg \min_i \{ \|x - m_i\| \} \tag{2}$$

Which is equal to

$$\|x - m_c\| = \min_i \{ \|x - m_i\| \} \tag{3}$$

After all the vectors $x(t)$ are put in the adequate lists neighbourhood N_i for the model m_i is defined. Each neighbourhood area includes grid nodes limited by some distance from the i -node. Then it is necessary to combine all the lists connected with N_i finding a pattern \bar{x}_i , located in the middle which has the smallest sum of the distances of all examples $x(t)$, $t \in N_i$. The sample \bar{x}_i is called generalized median for combining all the lists.

During the training where «the nonlinear projection» is formed, the nodes, that are topographically near the grid limited by a geometrical distance, will activate each other while learning because of it in the same enter of x . This leads to the local relaxation and smoothing effect for the weight vectors of the neuron in the observing neighbourhood, which leads to the total ordering after long training. One of the possible convergence limits for the learning process:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \tag{4}$$

Where $t=0,1,2,\dots$ – is an integer value representing discrete time, initial $m_i(0)$ can be random. In the relaxation process the function is called neighbourhood function defined on the grid points. The convergence demands the condition $h_{ci}(t) \rightarrow 0, t \rightarrow \infty$ to be fulfilled. Usually, the following correlation is carried out:

$$h_{ci}(t) = h(\|r_c - r_i\|, t), \quad (5)$$

Where $r_c \in \mathfrak{R}^2$ and $r_i \in \mathfrak{R}^2$ are vectors determining placement of the c and i nodes accordingly in the considered grid. If $\|r_c - r_i\|$ increases the condition $h_{ci}(t) \rightarrow 0$ is fulfilled.

There are two simple realizations for $h_{ci}(t)$, one of them is linear determining the neighbourhood of node c , $h_{ci}(t) = \alpha(t)$, if $i \in N_c$ and $h_{ci}(t) = 0$, if $i \notin N_c$. The value $\alpha(t)$ is determined by the value of the learning speed coefficient ($0 < \alpha(t) < 1$). As well as $\alpha(t)$, the radius of the set $N_c(t)$ usually monotonically decreases during the training process.

The most frequently used option of smoothing nucleus can be defined in the Gaussian function terms:

$$h_{ci}(t) = \alpha(t) e^{-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}}, \quad (6)$$

Where $\alpha(t)$ – an another scalar coefficient of speed of the learning and the parameter $\sigma(t)$ defines the width of the nucleus which corresponds with the mentioned above radius of the set.

As well as $\alpha(t)$, $\sigma(t)$ is a monotonically decreasing function of time. The function $\sigma(t)$ can be linear, exponential or inversely proportional to t .

The ordering of the vectors m_i originates during the initial work period of the algorithm, the other steps are necessary only for more accurate mapping. After the training period the function $\alpha = \alpha(t)$ should possess small values.

Graphic representation of the results generated by SOM is called U-matrix and aims to display a distribution of the coding vectors into groups [9-11]. The essence of the method is to depict the average distance between neighbouring coding vectors in a graphical image or using shades of different colours. If the mean distance neighbouring vectors m_i is small then light shades are used and vice versa, dark shades indicate more considerable distances. «Cluster landscape» generated by SOM gives visualization of the classification.

3 Results

Let's study the use of this method for structural analysis of the housing stock of 296 buildings in the city Archangelsk. The apartments' area is 1500 sq. m. which is about 20% of the city's housing stock.

The objects of the representative sampling of dwellings of all construction periods including buildings of different types and individual houses. The indexes of the housing stock were divided into two main groups:

General indexes of an apartment house: degree of reliability and durability, volume, floor area of flats, number of floors, service life, total deterioration, replacement cost value.

Conditions of constructive elements (including foundation, walls, partitions, floor, roofing, storey, embrasure, finishing, sanitary engineering, etc): specific index of replacement cost of a constructive element which is related to 1 sq. m. of total floor area of flats, prescriptive service life of a constructive element till major overhaul according to [12], and deterioration. Besides, it was noted that according to the condition of sanitary engineering, buildings were divided into three types: well-equipped, partially equipped and ill-equipped [13].

Thirty three indexes were considered in the analysis.

SOM divided 296 objects into 16 clusters. It became a perfect compromise between the streamlined number of clusters and quality of clustering. Obtained during the training map can be displayed as a multilayer colouring. Each layer is formed by one of the input data components (Fig. 3).

In the Fig. 3 there are 12 representations of the characteristics of the objects listed above. Each hexagon is a cell of the map or a neuron of an output layer of neural network. A specific colour of gradient palette of each cell suits a pattern which is discovered among all characteristics set for the map.

A group of vectors forms a cluster if distance among vectors of this group is less than the distance to neighboring groups. When applying the SOM algorithm the structure of clusters is visualized by representation of the distance between the support vectors using unified distance matrix (U-matrix). The obtained colours of the layers altogether create an atlas which represents the components arrangement and relative placement of different components' values. In each subspace of the generated map a projection of one of the components of a multidimensional vector reflected on the surface. With the help of SOM the process boils down to receiving those projections and assessment of forming groups and clusters. According to the colour of the cell we can define an approximate meaning of the objects included in it. Each SOM representation opens in a separate subspace. On the components' projections the red colour suits the biggest value, blue suits the smallest value, intermediate values are represented by the colours gradient.

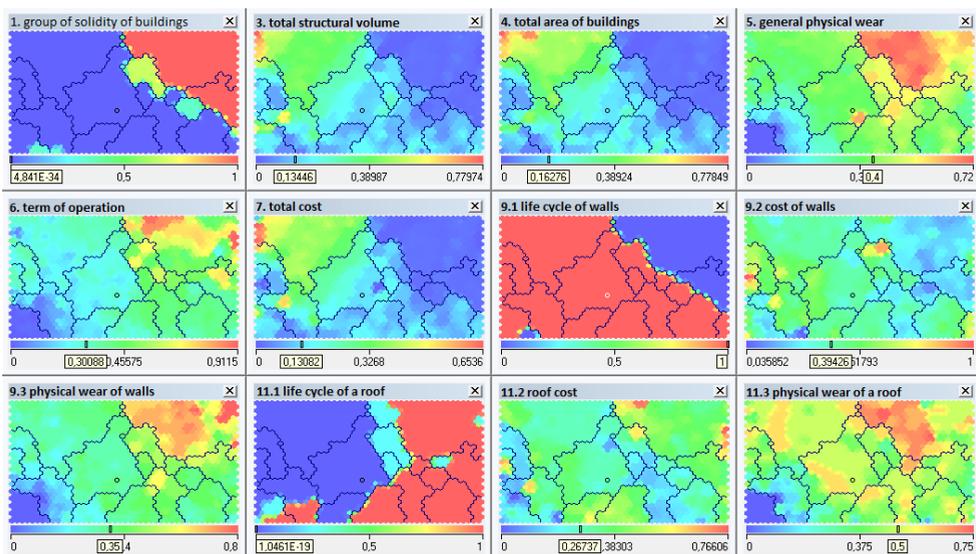


Fig. 3. SOM and the projections of the components (a map of the characteristics) for the data of the objects

The map of clusters is shown in the Fig. 4a. Clusters are groups of vectors, the distance between them is less than to the neighboring groups. In other words all the map elements included in the area of one colour (cluster) have similar characteristics. Besides, for the analysis of the whole map such visualizers as distance matrix are shown in the Fig.4b.

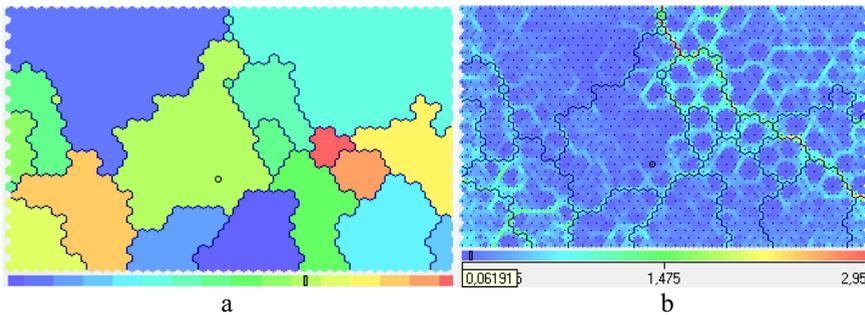


Fig. 4. a) clusters; b) distance matrix

After clustering of the representative sampling the distance matrix (unified distance matrix U-matrix) is used for visualizing of clusters structure obtained as a result of the map training. The matrix elements define the distance among neuron weight coefficients and its nearest neighbours. A big value indicates that this neuron differs significantly from the surroundings and refers to another type.

All clusters contain more than one cell (Fig. 4a). The biggest cluster contains 70 objects. The biggest part of the buildings in this cluster have the first degree of durability and reliability, the average number of floors is 9. Service life of buildings of this group is 7-41 year, total depreciation 24-44%. The next cluster consists of 52 objects which are wooden buildings of the fourth degree of durability and reliability with the considerable service life 28-103 years where number of floors is up to three. The depreciation of buildings in this group is 44-72%. The cluster is on the upper right corner of the map.

4 Discussion

After clustering of representative sampling the information about the whole city housing stock should be processed. As a result each object will be referred to its cluster [14]. Then a programme of major overhaul and maintenance is developed for a cluster which is formed based on the representative sampling [15-17]. This programme is typical for all objects of the cluster and lets forming an understanding about nomenclature, time and quantity of construction and repairing works. Then we can calculate an average payment rate for major overhauls and maintenance for groups of congeneric volumes. As a result time-labour-financial costs on development of regulatory documentation decreases significantly.

For objects, which compose clusters characterized by low quality of dwellings and high deterioration rate, it is necessary to plan other actions such as modernization, reconstruction, demolition and new construction. The number of such clusters lets defining the necessary amount of investments for the procedures.

However, the most important result of the research is an opportunity for structural analysis of the housing stock, which allows forming an understanding about its general condition. In doing so, systematic monitoring of the housing stock lets tracing the direction and a dynamic of growth of changes, which in turn allows us to react accordingly to the changes. In other words periodic examination and analysis of the condition of an apartment house will indicate the changed technical condition. Transferring of objects to a cluster with better characteristics will indicate a high quality of service life while transferring of the objects with worse characteristics will represent aging of the house and the necessity to renovate the building. To define the optimal way to plan such works it is suggested that we should use index of reduced costs [18], which equals a multiplication of the specific index of the estimated cost of construction and repair works and deterioration of the structural element during the service life before the next renovation.

$$P = C \cdot k^\gamma \quad (7)$$

Where C is a specific index of total estimated costs on construction and renovation of a structural element during the service life

k^γ – deterioration rate (before the next repair) of a structural element

γ – index connecting technical condition rate with expenses of construction and renovation works; for major renovation $0 < \gamma \leq 1$ (for maintenance – $1 < \gamma \leq 2$); if $\gamma \rightarrow 0$ then deterioration rate increases and there is necessity for construction and repair; if $\gamma > 1$ then costs of major renovation soar because of an increase of frequency of the works. Based on the analysis of the deterioration characteristics we got the optimal value $\gamma = 0.33$ which indicates 70% deterioration. Optimization of that index allows us to find the best correlation between costs (repair expenses) and quality (technical condition rate).

Minimal value of reduced costs is an optimal correlation between the cost rate of construction and repair works and value of physical deprivation of construction until the next repair.

Conclusions

SOM approach is a very effective tool for an analysis and assess of technical condition and construction characteristics and comparing of housing facilities, therefore their vast use for an examination and monitoring of the housing stock is promising [19]. Change in the structure of the housing stock can be considered by temporary subspaces for each research group and by adaptation of existing SOMs to the new data about the housing stock structure. The signals of changes of the housing stock structure can be traced by comparing later examinations with values generated by previously created SOM.

Acknowledgments

The reported study was funded by RFBR according to the research project No. 16-00-00001

References

1. V. Urbanavičienė, A. Kaklauskas, E. Zavadskas, J. Šliogerienė, J. Naimavičienė, N. Vatin, *International Journal of Strategic Property Management* 18(3) (2014)
2. S. Eraslan, *Economic Modelling*, 58, 482-493 (2016)
3. L. Shukurov, I. Khalilov, L. Shafigullin, *Research Journal of Pharmaceutical, Biological and Chemical Sciences*, 7 (5), 1390-1397 (2016)
4. I. Arribas, F. García, F. Guijarro, J. Oliver, R. Tamošiūnienė, *International Journal of Strategic Property Management*, 20 (1), 77-87 (2016)
5. J. Golej, M. Pánik, D. Špírková, *International Multidisciplinary Scientific GeoConference, SGEM*, 3, 441-452 (2016)
6. S. Canas, F. Ferreira, I. Meidutė-Kavaliauskienė, *International Journal of Strategic Property Management*, 19 (4), 368-380 (2015)
7. J. Vorel, *Acta Universitatis Carolinae, Geographica*, 49 (1), 83-97. (2014)
8. V. Rakovskij, O. Popova, *Industrial and Civil Engineering*, 10, 19-21 (2012)
9. T. Kohonen. *Self-organizing maps* (Springer – Verlag, 2001)

10. G. Deboeck, T. Kohonen, (eds.) *Visual Explorations in Finance with Self-Organizing Maps* (London, 1998)
11. O. Popova, Vestnik of Irkutsk State Technical University, **10**, 171-177 (2013)
12. VSN 58-88(r) (Moscow: Stroiizdat, 1990)
13. T. Simankina, O. Popova, Construction of Unique Buildings and Structures, **7**, 12 (2013)
14. T. Simankina, O. Popova, Appl. Mech. Mater., **584-586**, 1944-1950 (2014)
15. O. Popova, A. Lukin, Industrial and Civil Engineering, **11**, 32-34 (2013)
16. O. Popova, Modern problems of science and education, **1** (2014)
17. O. Popova, T. Simankina, Industrial and Civil Engineering, **7**, 40-50 (2013)
18. O. Popova, PhD Thesis (St. Petersburg, 2014)
19. T. Simankina, O. Popova, Appl. Mech. Mater., **725-726**, 1057-1062 (2015)