Application of clusterization algorithms for building materials classification on radioactivity in R

P E Sokolov¹, S A Sentenberg²

¹Institute of Architecture and Construction, Volgograd State Technical University, 1, Akademicheskaya St., Volgograd 400074, Russia
²Motor Transport and Industrial Machinery Department, Volgograd Technological College, 83, M. Zhukov Ave., Volgograd 400107, Russia

E-mail: smist2016-2017@vgasu.ru

Abstract. The clusterization method of raw construction materials on their radioactivity with clusterization algorithms within statistical treatment environment and R analysis applied is handled in the paper. The authors make an emphasis on application of modern methods of raw construction materials radioactivity data treatment. Raw construction materials are evaluated on the basis of objective evidence, reflecting indices of specific activities and of inner natural radionuclides specific activity. As a result of net learning on the experimental data the authors acquire a cluster map provided with segmentation on the natural radionuclides effective specific activity. Consequently, both the criteria are established and the hitting of a raw material into the corresponding cluster is estimated. The application of a “decision tree” algorithm makes it possible to state rules according to which raw materials refer to these or those clusters. As opposed to traditional analysis methods, the applied technique of radioactivity estimation is based on quantitative characteristics. On the ground of the acquired results conclusions are made on a possibility and appropriateness of application of the said technique of analysis and classification of raw materials radioactivity data.

1. Introduction
Between numerous qualities of construction materials and stock their radioactivity properties stand apart [1]. This is because of the fact that radioactivity has no visible characteristics therefore, its identification requires special instrumentation to be applied [2].

Construction materials and stock are traditionally distributed into four classes on their natural radionuclides’ effective specific activity [3]. This classification was established as a result of the analysis of frequency distribution of a number of construction materials on their natural radionuclides’ effective specific activity [4].

However, there exist now quite a number of algorithms that allow radioactivity data analysis considering only the data themselves. Kohonen self-organizing maps and decision trees can be referred to such algorithms [5].

Kohonen self-organizing maps are widely used in different sciences for classifying various objects and phenomena.

Thus, in the paper [6] by means of Kohonen self-organizing maps there was revealed an interconnection between suspended particles total amount in the air and fluorides’ concentration when being exposed to some meteorological factors.
The aim of the survey [7] was to classify the monitoring data set with the use of SOM in order to discover specific aspects of ties between particles’ sizes and seasonal prevalence with respect to the air quality in the considered region. SOM application allowed realize a number of complex processes in a local (definite) region more completely.

A possibility of multi-dimensional analysis for an estimation of interconnection of natural radioactivity and radon activity with the use of artificial neuron nets (ANN) and “decision trees” is shown in the paper [8].

Self-organizing maps were used to identify the most damaged (polluted) areas, as well as to estimate a potential correlation between aerosols and meteorological factors [9]. Herein the authors emphasize that SOM is a very useful instrument for characterizing $^{232}$Th spatial distribution and establishing the pollution sources.

In the work [10] by means of SOM there was made an analysis of lead’s concentration and isotopic composition to estimate spatio-temporal regularities within rural microenvironment. Advantages of SOM analysis in comparison with traditional estimation means are shown in the paper. The use of SOM made it possible to reveal differences between samples looking outwardly similar, and consequently, have additional information.

The survey [11] is an attempt to classify the influence of fly-ash on clay strength in the course of time. SOM were applied to estimate a dependence of time influence on clays’ strength to be in the same cluster.

Artificial neuron nets were used in the work [12] to analyze $^{226}$Ra activity. The aim of the survey is to separate the anthropogenic component from the current radiation background, i.e. without considering natural radiation background.

Artificial neuron nets were used to forecast radon quantity conditioned by granulated blast-furnace slag in a cement solution [13]. The survey results showed that ANN are an alternative approach to radon concentration in a cement solution forecasting.

The clusterization method was offered in the work [14] as an instrument for sound deadening while monitoring of concrete reinforcement rusting by means of acoustic emission. It allowed increase the classification accuracy.

All the above said allows make an assumption that Kohonen self-organizing maps are possible to be used to classify construction materials on the level of natural radioactivity [15].

2. Materials and methods
Kohonen maps are a variety of neuron nets where learning without a teacher is used. While the process of such learning the learning set consists of input variables’ values. Within such learning there is no comparison of output neurons with reference values, i.e. the network “learns” to understand the data structure.

Self-organizing maps of signs (SOM) represent not only an effective clusterization method, but also allow show its results as two-dimensional maps where distances between objects correspond distances between vectors inside a multi-dimensional space, and the signs’ values are shown in different colour shades. Thus, a two-dimensional map allows show tree dimensions thanks to different colour shades.

Self-organizing maps are used to solve tasks of simulation, forecasting and searching regularities within large data volumes, etc. The most widely spread application of Kohonen nets s make classification “without a teacher”, i.e. clusterization. On acquiring the new information about classes, a correction of existing rules of objects’ classification is possible.

Decision trees refer to most popular and powerful instruments of Data Mining [5], which allow solve problems of classification and regression. In the ground of decision trees there are principal rules like “if…so…” which can be formulated in a natural language. Therefore, decision trees are the most demonstrable and easily interpreted simulations.

One of the most popular algorithms of decision trees plotting is ID3 algorithm and its C4.5 modification [16]. The C4.5 algorithm starts working with all teaching examples in the tree root node.
To divide a set of examples of the root node one attribute is chosen, and for each value assumed by this attribute there is a branch plotted and a daughter node is created. Thereafter all the examples are distributed to daughter nodes corresponding the attribute values. The algorithm is repeated recursively until in the nodes there remain examples of an only class, and after that the nodes will be proclaimed leaves with attribute values.

Data obtained while raw materials survey on Volgograd region fields [17,18] were used for the analysis made. The total number of samples was 392. All the samples were subjected to gamma-ray spectrometry analysis and as a result there were obtained $^{40}$K, $^{226}$Ra and $^{232}$Th [19] specific activity values. On the ground of the obtained natural radionuclides’ specific activities data there has been made a formula evaluation of the effective specific activity [20]:

$$A_{\text{eff}} = 0.09A_K + A_{\text{Ra}} + A_{\text{Th}}$$

(1)

with $A_K$, $A_{\text{Ra}}$ and $A_{\text{Th}}$ are specific activities correspondingly of $^{40}$K, $^{226}$Ra and $^{232}$Th, Bq/kg.

3. Results and discussion

Neuron nets work only with numeric data, therefore their conversion and coding is important. A numeric code was given to all surveyed materials (see Table 1).

| Raw materials | Clay | Sand stone | Chalk | Lime stone | Sand | Dolomite |
|---------------|------|------------|-------|------------|------|----------|
| Code          | 1    | 2          | 3     | 4          | 5    | 6        |

Data prepared in such a way were imported as a text file into the R [21] statistic processing and analysis program.

Summery function having been used, so a descriptive statistics of the loaded data was obtained. The descriptive statistics results are given in Table 2.

| Row      | $^{40}$K | $^{226}$Ra | $^{232}$Th | $A_{\text{eff}}$ |
|----------|----------|------------|------------|-----------------|
| Min      | 1.00     | 26.87      | 3.71       | 3.01            |
| 1-st Qu. | 1.00     | 59.45      | 11.65      | 7.20            |
| Median   | 3.00     | 116.61     | 17.67      | 16.18           |
| Mean     | 2.92     | 342.32     | 20.08      | 21.05           |
| 3-rd Qu. | 5.00     | 631.03     | 25.00      | 34.04           |
| Max      | 6.00     | 1059.92    | 97.20      | 62.93           |

Thereafter a precompression of spatial signs was undertaken. This provided plotting more stable cluster structures. In the R environment clusterization on the main components is fulfilled in FactoMineR pack.

A usual analysis of the main components and their number choice are fulfilled by PCA() function of the pack. PCA() function work result is presented on Figure 1.
Figure 1. Original features expansion on axis of the two main components.

As it follows from Fig.1 all the used variables have a considerable significance and can be used for further analysis. Let us use somgrid() and somt() functions from the kohonen pack for net learning. Original features set contains all the investigated variables, as far as they all are significant for clusterization (see Figure1).

May the somgrid() function create a hexagonal lattice of 6×9 cells, i.e. 392 raw materials will “self-organize” into 54 neurons of the output layer.

Let us fulfil Kohonen maps visualization with different modifiers of plot() function. “Counts” parameter indicates the number of original objects in every node of the net, and “quality” parameter shows an average distance of the node’s objects to its prototypes. The visualization result is presented in Figure 2.

Figure 2. Kohonen maps constructed with “counts” and “quality” parameters.

Now, by means of a “mapping” type map we obtain a satisfying the desired condition objects’ distribution into nodes. We also will show how shares of participation of separate original variables are distributed. The “codes”-type united map allows demonstrate the distribution of separate original variables’ participation shares correlation on the lattice (see Figure 3).
Figure 3. “Mapping” and “codes”-type maps.

As a “codes”-type map is not always well interpreted let us have “Row” and “$A_{\text{eff}}$” indices distribution on a natural scale (see Figure 4).

Figure 4. “Row” and “$A_{\text{eff}}$” indices distribution on a natural scale.

One can use activation values of each neuron on every predict for nodes grouping. Let us set the number of clusters $k = 5$ and make a hierarchical clusterization. Then plot a “codes”-type map with all variables’ participation shares distribution. Nodes’ clusters and variables are presented in Figure 5.

Figure 5. Nodes’ cluster map and variables.
Kohonen method is considered as an empiric algorithm, whereas qualitative conclusions on a data structure are made on the ground of visual analysis of the presented map. Therefore let us use the HCPC() function and construct a hierarchical clusterization. Herein a method of complete connection (complete linkage clustering) is applied, when the distance between the most remote objects is calculated [22]. Let us make a hierarchical clusterization on main components. Then plot a dendrogram with clusters shown (see Figure 6).

Figure 6. Raw materials cluster dendrogram.

As it follows from the dendrogram all the surveyed materials are divided into five clusters. With the use of rpart pack let us plot a decision tree considering all the rules according to which raw materials are divided into clusters. To form criteria of hitting of raw construction materials on their radioactivity into different clusters it is efficient to use an approach of “decision tree” plotting based on tree plotting algorithm of C4.5 classification developed by John Quinlan [16]. This algorithm is able to work with attributes presented in continuous quantities and implements iteration where tree branches with the least possible influence on the classification results are cut off, thus allowing optimize the final decision tree structure. The decision tree and decision rules’ totality are presented in Figure 7.

Because the rules’ totality is rather bulky there is a rule #1 exemplifying fragment in Figure 8.

Hereby, according to the given conditions for node 1 raw materials with the following primary divisions $^{40}\text{K} < 209.855$, $^{232}\text{Th} < 18.55$, $A_{\text{eff}} < 62.85$ and $^{226}\text{Ra} < 14.97$ Bq/kg hit into cluster 1. One to several rules correspond each cluster, and each definite raw material can be classified on radioactivity when the rules applied.

4. Summary
In the current survey an approach is offered with use of maximum objective initial quantitative data, reflecting the level of natural radioactivity of raw construction materials. Clusterization on the basis of Kohonen self-organizing maps and decision trees allows sort out groups of materials corresponding to different level of their radioactivity without applying additional subjective indices.

Kohonen self-organizing maps give an opportunity to expose and visually demonstrate any hidden interconnections and differences between radiation characteristics of raw construction materials. A taught and described Kohonen map can be used to obtain characteristics of any other raw material which has a similar input data set. The advantage of the method is that it allows work with a large number of variables and visually demonstrate the analysis results in two-dimensional form.
Figure 7. Clusters’ forming rules on the basis of a plotted decision tree.

Figure 8. Rule 1 of raw construction materials division on their radioactivity (fragment).

References
[1] UNSCEAR 2006 Report to the General Assembly with Scientific Annexes 2009 Effects of ionizing radiation. Sources-to-effects assessment for radon in homes and workplaces (United Nations Scientific Committee on the Effects of Atomic Radiation vol II Annex E) New York: United nations p 142
[2] Sokolov P E, Sidelnikova O P and Kozlov Y D 1995 The need for control of radioactivity of building materials Stroitelnye materialy 9 18–19
[3] GOST 30108-94 The construction materials and products Determination of specific effective activity of natural radionuclides (Moscow)

[4] Krisyuk E M 1989 Radiation background of dwellings (Moscow: Energoatomizdat) p 120

[5] Kohonen T 2001 Self-Organizing Maps third ed. (New York: Springer) p 501

[6] Zhong B, Wang L, Liang T and Xing B 2017 Pollution level and inhalation exposure of ambient aerosol fluoride as affected by polymetallic rare earth mining and smelting in Baotou, north China Atmos. Environ. 167 40–48

[7] Tsakovski S, Simeonova P and Simeonov V 2011 Classification and modeling of different fractions of aerosol monitoring data J. Environ. Sci. and Health A 46(2) 157–169

[8] Forkapic S, Maletić D, Vasin J, Bikit K, Mrdja D, Bikit I, Udovičić V and Banjanac R 2017 Correlation analysis of the natural radionuclides in soil and indoor radon in Vojvodina, Province of Serbia J. Environ. Radioact. 166 403–411

[9] Wang L, Zhong B, Liang T, Xing B and Zhu Y 2016 Atmospheric thorium pollution and inhalation exposure in the largest rare earth mining and smelting area in China Sci. Total Environ. 572 1–8

[10] Deljanin I, Antanasijević D, Vuković G, Urošević M A, Tomašević M, Perić-Grujić A and Ristić M 2015 Lead spatio-temporal pattern identification in urban microenvironments using moss bags and the Kohonen self-organizing maps Atmos. Environ. 117 180–186

[11] Göktepe A, Sezer A, Sezer G and Ramyar K 2008 Classification of time-dependent unconfined strength of fly ash treated clay Constr. Build. Mater. 22(4) 675–683

[12] Varley A, Tyler A, Smith L, Dale P and Davies M 2016 Mapping the spatial distribution and activity of $^{226}$Ra at legacy sites through Machine Learning interpretation of gamma-ray spectrometry data Sci. Total Environ. 545-546 654–661

[13] Jang H-S, Shuli X, Lee M, Lee Y-K and So S-Y 2016 Use of Artificial Neural Network for the Simulation of Radon Emission Concentration of Granulated Blast Furnace Slag Mortar J. Nanosci. Nanotechnol. 16 (5) 5268–5273

[14] Calabrese L, Campanella G and Proverbio E 2012 Noise removal by cluster analysis after long time AE corrosion monitoring of steel reinforcement in concrete Constr. Build. Mater. 34 362–371

[15] Sokolov P E 2017 Use of Kohonen maps for assessing the radioactivity of building materials Bulletin of science and education of the North-West Russia 1(3) 70–78

[16] Quinlan J R 1996 Improved Use of Continuous Attributes in C4.5 J. Artif. Intell. Res. 4 77–90

[17] Sokolov P E and Drik V Y 2017 Natural radioactivity of clay rocks used for manufacture of construction materials International Research Journal 66 (12) 148–151

[18] Sokolov P E and Sentenberg S A 2018 Heat treatment influence evaluation on radon-222 emanation factor in building material International Research Journal 6-1(72) 48–57

[19] Radiation protection 112 1999 Radiological Protection Principles Concerning the Natural Radioactivity of Building Materials p 16

[20] Norms of radiating safety (NRB - 99/2009) 2009

[21] Kabacoff R I 2015 R in Action. Data analysis and graphics with R (New York: Manning) p 579

[22] Shitikov V K and Mastitsky S E 2017 Classification, regression and other data mining algorithms using R Togliatti (London) p 351