Cluster analysis for modeling the cost of industrial land objects in Krasnoyarsk region

I A Panfilov1,2,3, Y A Alekseeva1, A A Shiryaeva1, O I Karelin2 and D V Kustov2

1 Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarskiy Rabochiy av., Krasnoyarsk, 660037, Russia
2 Siberian Federal University, 79, Svobodny pr., 660041 Krasnoyarsk, Russia
3 Krasnoyarsk State Agrarian University, 90, Mira ave, 660049, Krasnoyarsk, Russia

E-mail: crook_80@mail.ru, iu.al2012@yandex.ru, anyashiryaeva_174@mail.ru, 2935560@gmail.com, kustov.dv@gmail.com

Abstract. An assessment activity must have a legal justification. The regulatory documents are results of a land plot state assessment carried out in the Russia regions on a regular basis. In this research mathematical models for evaluating land plots in Krasnoyarsk region based on data from the state assessment of 2012 and 2017 were built. The models were supposed to formalize the mechanism for evaluating the state contractor when evaluating industrial land. Factor and cluster analysis methods were used. The factors were determined to a greater extent determining the change in the value of the objects of assessment. The data for the analysis was taken from open official sources. The simulation results showed unstable dependencies in the official results of the state contractor evaluation. Further analysis revealed errors in the official report.

1. Introduction

When assessing the value of land plots, a specialist should consider their purpose. In Russia, land plots can belong to the following categories: lands of settlements, agricultural lands, industrial and special purpose lands, nature reserves, forest lands, etc. In this study, only industrial and special purpose lands are considered. This category is also divided into six groups and several subgroups: land plots for the placement of ground-based objects of space infrastructure, land plots provided for airports, land plots for the location of power plants and related infrastructure, land plots for the placement of industrial and administrative buildings, land plots for road service facilities, land plots for the development of minerals, land plots for placing power lines, etc. Of course, each such subgroup has its own unique characteristics. However, as a real estate object, all information about them is contained in land cadastres and have common parameters: geographical coordinates, area, cadastral cost and specific cadastral value.

Lands owners often turn to appraisal agencies for revaluation of the value of such real estate objects that. To evaluate such objects, “comparative”, “income” and “cost” methods are used. Each of these methods has
its own advantages and difficulties to use. However, in any case, the appraiser has to compare his result with the official cadastral assessment made by the government contractor. A state land assessment is carried out regularly in every region of Russia. In Krasnoyarsk region, such an assessment was carried out in 2009 and 2016. The assessment resulted in two documents that were reviewed in [1]. The government contractor does not disclose the methodology it uses these estimations. The formalization of this methodology would allow the assessment to be carried out more reasonably.

2. Materials and methods

Government Decrees [1] were taken as the main source of information for this study. However, these documents themselves contain only information about the cost of objects, their category and their cadastral number. From 2009 to 2016, many objects moved from one category to another, were divided into several smaller objects, or merged into a larger one. Thus, an additional task was solved – searching for similar objects in both documents, checking whether the characteristics of similar objects and comparing their cost. In total, the documents contain information about more than 50,000 land objects. We selected 1921 objects that were present in both documents. To find the remaining characteristics of interest to us (the area of the object, the coordinates of the object), open government resources were used: the Public cadastral map and the Official website of Rosreestr. All objects are located in the yemelyanovsky district of the Krasnoyarsk region.

Table 1 show changes of cost values for each of six groups of objects. The average cost of plots assigned to the first group decreased by 0.96 times during the reporting period. While the cost of plots in the sixth group increased by 1.65 times. However, in practice, the cost per square meter varies greatly in each group.

Table 1. Mean values of cadastral cost lands of industry and other special purposes for the Emelyanovsky district of Krasnoyarsk region, RUB/sq m.

| Year | 1st group | 2nd group | 3rd group | 4th group | 5th group | 6th group |
|------|-----------|-----------|-----------|-----------|-----------|-----------|
| 2009 | 312.28    | 487.00    | 528.10    | 2.52      | 487.00    | 43.61     |
| 2016 | 299.40    | 516.11    | 631.29    | 3.11      | 516.11    | 71.85     |

Indicators were defined for each object:

1. The Cadastral number. For each object the full cadastral number is specified as the number of the district, the number of subdistrict, cadastral block, and the cadastral number.
2. The cadastral cost for each object that was specified for 2009.
3. The cadastral cost for each object that was specified for 2016.
4. The Specific indicator of cadastral cost (the cost of the object per square meter) for 2009.
5. The Specific indicator of cadastral cost (the cost of the object per square meter) for 2016.
6. The Area of the object (sq. m.).
7. The group that the object belongs to (from [1, 2]).
8. The coordinate of the object's latitude.
9. The coordinate of the object's longitude.
10. Improvements of the object. This indicator shows the presence of additional buildings or infrastructure on the location. It takes the value 0 or 1.

Unfortunately, much of the data was missing from official databases. Thus, the original database of objects was reduced to 234 objects. Objects that are not fully contained in the open Rosreestr database were excluded. For example, the coordinates of the site borders are missing.

Much attention is paid to the choice of parameters for determining the cost of real estate objects [2, 3]. In previous studies, we also solved the problem of determining the importance of parameters when evaluating real estate objects [4, 5]. In this study was used the Principal Components Analysis. The table 2 present values of new parameters of the task. This table shows that only six factors have an impact. Factor
1 explains 24.6% of the total variance, factors 1, 2, 3, 4, 5 and 6 explains 91% of the total variance. On next stage of this research we checked the case with 6 parameters.

To solve the clustering problem, we used the Two-mode-clustering, k-means clustering and Hierarchical clustering methods.

### Table 2. Total variance explained for new parameters.

| Component | Initial Eigenvalues | Total % of Variance | Cumulative % |
|-----------|---------------------|---------------------|--------------|
| 1         | 2.211               | 24.567              | 24.567       |
| 2         | 1.576               | 17.511              | 42.078       |
| 3         | 1.436               | 15.956              | 58.033       |
| 4         | 1.284               | 14.267              | 72.300       |
| 5         | 0.969               | 10.767              | 83.067       |
| 6         | 0.801               | 8.900               | 91.967       |
| 7         | 0.368               | 4.089               | 96.056       |
| 8         | 0.217               | 2.411               | 98.467       |
| 9         | 0.138               | 1.533               | 100.000      |

### 3. Results

First, we used the Two-mode-clustering method. The number of clusters is determined automatically. As a result, we get the percentage breakdown of the influence of the variable on each cluster. The result is unsatisfactory, split into just two large clusters with six variables (figure 1).

![Figure 1. Result of Two-mode-clustering method.](image)

For k-means clustering method researcher have to choose the number of clusters. In this study we varied the number of clusters from 10 to 20. Table 3 and figure 2 show the result for 14 clusters. This number of clusters present the best separation of objects.
Table 3. The result of k-means clustering method separation for 14 clusters.

| Number of Cases in each Cluster | Cluster |
|--------------------------------|---------|
| 1                              | 8       |
| 2                              | 2       |
| 3                              | 1       |
| 4                              | 3       |
| 5                              | 20      |
| 6                              | 131     |
| 7                              | 4       |
| 8                              | 2       |
| 9                              | 1       |
| 10                             | 1       |
| 11                             | 1       |
| 12                             | 49      |
| 13                             | 10      |
| 14                             | 1       |
| Valid                          | 234     |
| Missing                        | 0       |

Figure 2. Result of k-means clustering method separation for 14 clusters.

In this study we used three modification of Hierarchical clustering method: Average Linkage (Within Groups), Complete Linkage and Ward Linage. The best results were given from Ward Linkage method. For these methods, the optimal number is the number of clusters equal to the difference between the number of observations and the number of steps before the abrupt increase in the coefficient. The Ward Linkage method shows the separation already after 5th step of algorithm (figure 3). For Average Linkage (Within Groups) method and Complete Linkage method it happened after 15th and 20th steps. The methods presented do not reflect the complete division of objects into clusters. Not all objects can be uniquely assigned to one of the
clusters. Thus, the results of revaluation of the cadastral value can be considered incorrectly defined for a number of objects.

![Dendrogram using Ward Linkage.](image)

4. Conclusions
In this study, it was found that the most popular and widely used clustering methods do not provide a satisfactory result. There is no unambiguous division into clusters based on the database we have collected. In this regard, it will not be possible to identify the factors that affect the cost of objects. This result, namely poor, could be obtained for two reasons: during collecting or analyzing data, some mistakes were made that led to incorrect analysis results; the government contractor have made mistakes when determining the cadastral cost [1].

However, methods of mass assessment are losing their relevance when determining the cadastral value on the territory of Russia, since there are too many nuances for each individual case.

References
[1] Shiryaeva A A, Panfilov I A and Karelin O I 2019 Development of a real estate valuation method adjusted by valuation date Materials V Intern. Scientific. Conf “Actual Problems of Aviation and Cosmonautics” 903

[2] Dubrovsky A V, Ilyinykh A L, Malygina O I, Moskvin V N and Vishnyakova A V 2019 Analysis of price-forming factors impacting cadastral property value Vestnik of the Siberian State University of Geosystems and Technologies (SSUGT) 2 150-69
[3] Ozhegov E M, Ozhegova A and Mitrokhina E 2019 Distance in Geographic and Characteristics Space for Real Estate Price Prediction Lecture Notes in Computer Science (In: van der Aalst W. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2019) 11832

[4] Alekseeva Y A, Shiryaeva A A, Panfilov I A and Karelin O I 2016 Research of the paired assessments in real estate appraisal Materials XX Intern. Scientific. Conf “Reshetnev reading” pp 10-1

[5] Panfilov I A, Alekseeva Y A, Panfilova T A, Karelin O I and Kustov D V 2019 Correction factors application to Krasnoyarsk city real estate in the lack of information conditions Journal of Physics: Conference Series. International Scientific Conference "Conference on Applied Physics, Information Technologies and Engineering - APITECH-2019" 55092