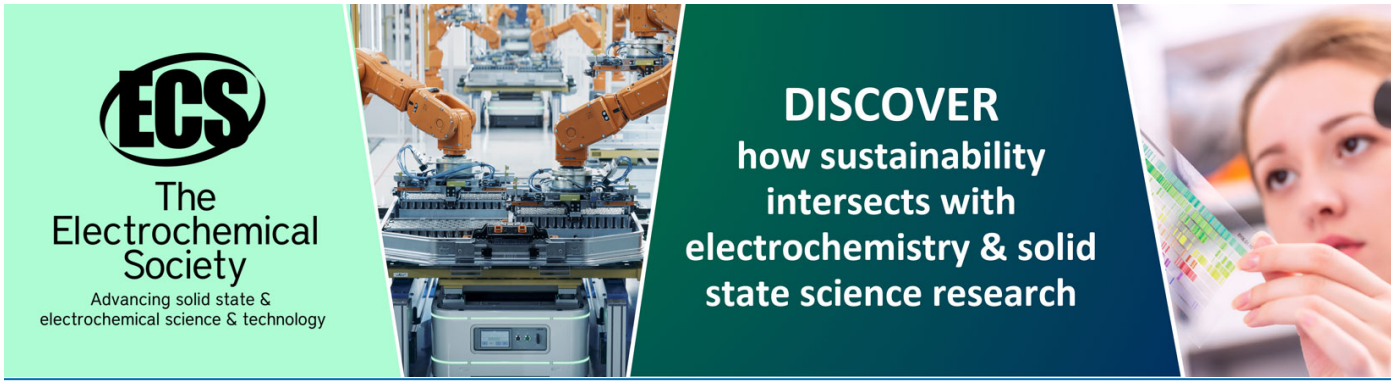
Categorical data processing for real estate objects valuation using statistical analysis



IOP Conf. Series: Journal of Physics: Conf. Series **1015** (2018) 032102 doi:10.1088/1742-6596/1015/3/032102

Categorical data processing for real estate objects valuation using statistical analysis

D S Parygin[1], V P Malikov[1], A V Golubev[1], N P Sadovnikova[1], T M Petrova[2], A G Finogeev[3]

[1] Volgograd State Technical University, 28, Lenina Ave., Volgograd, 400131, Russia
[2] Volgograd State Socio-Pedagogical University, 28, Lenina Ave., Volgograd, 400066, Russia
[3] Penza State University, 40, Krasnaya Str., Penza, 440026, Russia

E-mail: dparygin@gmail.com

Abstract. Theoretical and practical approaches to the use of statistical methods for studying various properties of infrastructure objects are analyzed in the paper. Methods of forecasting the value of objects are considered. A method for coding categorical variables describing properties of real estate objects is proposed. The analysis of the results of modeling the price of real estate objects using regression analysis and an algorithm based on a comparative approach is carried out.

1. Introduction

Prognostic modelling of the value of real estate is used in a variety of spheres of activity when carrying out transactions related to the implementation of property rights to real estate. The calculation models used in practice do not allow reacting to changes in requirements, market conditions and other factors with due diligence. Models based on the analysis of statistical data can significantly improve the accuracy and efficiency of the forecast. In this case, it becomes possible to use for the analysis a number of factors that are important for real estate valuation [1], but do not have a quantitative assessment.

Such kind of research is conducted both in Russia and abroad. For example, the authors of Grid Group [2] have been collecting and analyzing statistical data related to urban real estate for more than five years. On their website, they publish reports, have a database of articles on data processing, and also produce real estate analysis.

The article [3] describes the methodology of the "price belt". The work describes the formation of the average price for a group of real estate objects, as well as paid attention to the description of reliable indicators for evaluation. Research conducted by the author is already used for the analysis of real estate in the city of Yekaterinburg.

One of the most common approaches is the method-based comparison of the object under study with analogues having similar values of properties for which the price is known. This approach includes the following analysis algorithm [4]:

1) identify the properties by which the analogues from the initial selection of objects will be searched;
2) search for analogue objects by the desired values of properties;

3) determine the price per square meter of each analogue;
4) determine the average price per square meter of all analogues;
5) calculate the price of the object under study by its area.

Authors of publications [5-7] presented an analysis of the value of real estate in the city of Krasnoyarsk, based on regression analysis. The authors proposed a method for coding categorical variables (for example, the type of house).

The authors analysis showed that the approaches used can lead to false interpretation of the data. For example, one enters the ranking of groups, which, in fact, is missing.

In many models, categorical data are not available, which significantly limits the capabilities of the model. The more data about objects will be used, the more accurate the model will be. In this regard, it is necessary to choose a method for coding categorical data that will minimize the loss of information in the transformation and provide acceptable accuracy of the solution.

In [8], two approaches to solving this problem are considered:

1) the definition of a discrete scale of a characteristic with a fixed number of gradations;
2) One Hot Encoding

A distinctive feature of the first way of digitizing the properties of objects is the need for expert evaluation of categories. The second way that allows one to generate a numeric code without losing information [9]. Let us assume that the initial data are given by the object-characteristic matrix:

$$F = \|f_{ij}\|_{m\times n},$$

m - counts of objects, n – counts of attributes, f_{ij} - value of the j-th attribute on the i-th object, and the target vector:

$$(y_1, \dots, y_m)^T,$$

y_i - the value of the target attribute on the i-th object.

As the values of categorical attributes, one can use objects of an arbitrary nature by which a comparison operation is defined. Let us assume that $f_{ij} \in \{1, 2, \dots, n\}$ and n_j – number of different categories of the j-th attribute, $j \in \{1, 2, \dots, n\}$. Let us note that it is not necessary $\{f_{1j}, \dots, f_{mj}\} = \{1, 2, \dots, n\}$.

Let us replace the column of matrix F (the corresponding characteristic) with a binary matrix.

Despite the universality of this method, in some cases its application can be complicated due to a significant increase in the number of columns in the object-attribute matrix.

In the framework of this study, various approaches to building models for estimating the price of real estate are considered and compared.

2. Data collection

Popular Internet resources that contained information in the format of advertisements of real estate being sold or leased, as well as a federal property data register, were used as sources for data collection [10]. The structure of the data contained in the ads and lots was developed for each of the selected sources after their research. The BeautifulSoup library was used for data parsing.

As a result, a sample of data was compiled, consisting of 4000 objects [11], described by a set of 50 features. The target is to use all the attributes in the calculations. However, at this stage of the study, 6 characteristics were selected, which are represented by different types of data (Table 1): "type of apartment", "material of the walls of the house" - categorical types; "Number of rooms", "storey", "floor" - "pseudo-quantitative" types; and "area" is a quantitative type.

The application of the One-hot Encoding approach led to an increase in the number of features. The implementation of this approach is shown in Figure 1.

At the same time, for each new characteristic, their own weight coefficients are allocated. For example, instead of b_1 there will be corresponding coefficients b_{11} and b_{12}.

```
# coding method One-Hot Encoding
# input - array_coding - array of attribute categories
# output - array_coding - coded attribute category array
def coding_qualitative_properties(array_coding):
    # integer encode
    label_encoder = LabelEncoder()
    integer_encoded = label_encoder.fit_transform(array_coding)
    # binary encode
    onehot_encoder = OneHotEncoder(sparse=False)
    integer_encoded = integer_encoded.reshape(len(integer_encoded), 1)
    array_coding = onehot_encoder.fit_transform(integer_encoded)
```

Figure 1. One-Hot Encoding method.

Table 1. Selection of the data of real estate objects (fragment)

#	Type of apartment	Material of walls of the house	Number of rooms	The area of the apartment, m^2	Floor	Floors of the house	Price, thousand roubles
1	Resale accommodation	Brick	2	42,0	2	5	2000
2	Resale accommodation	Brick	2	44,8	5	5	1400
3	New building	Brick	1	38,6	9	16	1330
4	New building		1	40,2	5	5	1519
…	…	…	…	…	…	…	…
4000	Resale accommodation	Brick	1	29,0	4	5	1550

3. A regression model for calculating the value of real estate

To build a linear regression model, the authors used the Scikit-learn library for Python version 3.6. chosen as Ordinary Least Squares, OLS.

The training of this model was made using a training sample, which was made up of the initial sample. Records from the original sample were selected randomly. In this case, their number is 75% of the amount of initial data. The rest of the records were used as the test data of the regression model.

It should be noted that some features, which at first glance are quantitative, may also be categorical. For example, "level" is not actually an absolute quantitative characteristic of an apartment, because the first level of an apartment building in terms of valuation is closer to the uppermost floor than to the second. Therefore, it makes sense to perceive such features of the object as pseudo-quantitative.

Accordingly, to use these data in the model, they must also be encoded. For this, the maximum value is determined for each numerical characteristic. This value is the dimension of the common interval of the given characteristic.

Also, for each characteristic, a step interval is determined. It turns out that the total number interval is divided into groups. Further, for each sample object, it is determined whether it belongs to a particular group. If the object belongs to a certain group, then, in accordance with the binary coding

approach, the value of the object in this group is 1, otherwise - 0. To perform this coding, a method has been developed, the implementation of which is shown in Figure 2.

```
# coding method for pseudo-quantitative properties
# input
# array_coding - array of attribute categories
# step – value of the step interval
# output
# array_coding - coded array of pseudo-quantitative attribute
def coding_quantitative_properties(array_coding, step):
        for i, tmp in enumerate(array_coding):
                array_coding[i] = tmp // step + 1
        max_value = max(array_coding)
        num_group = 0
        for i, tmp in enumerate(array_coding):
                arr = [0 for j in range(max_value)]
                num_group = tmp
                arr[num_group-1] = 1
                array_coding[i] = arr
        return array_coding
```

Figure 2. Pseudo-quantitative properties method.

With the help of this method, the signs "number of rooms", "level" and "house number" were coded. In this case, the value of the sign "area of the apartment" was not coded, since the price of the property directly depends on the area of the object. As a result, the number of columns in the matrix "object-attribute" increased to 68 (Table 2).

Table 2. The "object-attribute" matrix after character coding (fragment)

#	New building	Resale accommodation	Block	…	Panel	1 room	…	10 rooms	The area of the apartment, sq.m	1 level	…	25 level	1- storey	…	25- storey	Price, thousand roubles
1	0	1	0	…	0	1	…	0	42,0	0	…	0	0	…	5	2000
2	0	1	0	…	1	0	…	1	44,8	0	…	0	0	…	5	1400
3	1	0	0	…	0	1	…	0	38,6	0	…	0	0	…	16	1330
4	1	0	0	…	0	1	…	0	40,2	0	…	0	0	…	5	1519
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
4000	0	1	0	…	0	1	…	0	29,0	0	…	0	0	…	5	1550

The realized model on this set of data on the researched real estate objects gave the results presented in Table 3. The root-mean-square error of this model of the training and test samples has the following values: 0.87 and 0.86, respectively.

Table 3. Results of forecasting the value of objects by different approaches

Object	The projected price using a regression model	The projected price using a comparative approach
Flat.1.3/5.40	1’924’536	1’803’028
Flat.2.2/5.40	1’750’847	1’633’328
Flat.3.5/16.80	4’384’998	4’212’010

4. The comparative approach for calculating the value of real estate

A comparative approach for the valuation of real estate objects is based on the comparison of the investigated object with other objects having similar characteristics, and the price of which is known. This method includes the following analysis algorithm:

1. determination of the parameters by which the analogues from the initial data sample will be searched;
2. collection of information on the selected parameters;
3. comparison of each found analogue with the object under study for distinguishing distinctive properties;
4. averaging the price per square meter of the analogues found.

Based on the proposed algorithm, a method for estimating the real estate object (apartment) was developed on the basis of a comparative analysis, the implementation of which is shown in Figure 3.

```
# method that implements a comparative approach
# input
# example – object under study
# for_train – array of values for the attribution of records to the #           test sample
# output
# approximate_price - forecast price
def comparative_approach(example, for_train):
analog_list = []
for i, tmp in enumerate(dataSet):
        if for_train[i] == False:
        if tmp['countRooms'] == example['countRooms']:
        if tmp['type'] == example['type']:
        if abs(int(tmp['floor'])/int(tmp['countFloor']) - int(example['floor'])/int(example['countFloor'])) <
0.35:
                if tmp['material'] == example['material']:
                analog_list.append(tmp)
approximate_price = 0
for tmp in analog_list:
                approximate_price += int(tmp['price'])/int(tmp['square'])
approximate_price /= len(analog_list)
approximate_price *= int(example['square'])
return approximate_price
```

Figure 3. Comparative approach.

The results of calculating the cost of the objects under study are given in Table 3. The methods analyzed give similar results, but the accuracy of the method based on the comparative approach depends significantly on the availability of a sufficient number of analogues.

5. Conclusion

The conducted research is necessary for the solution of a problem of forecasting cost of the real estate. In the course of it, a software solution was developed that allows an evaluation of real estate objects. The main methods of encoding categorical data were considered and the One-Hot Encoding approach was chosen. The analysis of the results of modelling the price of real estate objects using regression analysis and an algorithm based on a comparative approach is carried out.

Experimental studies were carried out using a large sample of data and showed the adequacy of the model constructed.

At this stage of the study, only the internal properties of the objects were used, which are the type of real estate, its area, the quality of the building structure. However, external factors, such as geographic location (distance from the center, proximity to hospitals, schools, etc.), the ecological condition of the area and many other things are important in the assessment. Their impact on the value of real estate is the goal of the further research.

6. Acknowledgments

The reported study was funded by RFBR according to research projects No. 17-37-50033_mol_nr, No. 16-07-00388_a, No. 16-07-00353_a, No. 18-010-00204, No. 18-07-00975.

References

[1] Golubev A, Chechetkin I, Parygin D, Sokolov A, Shcherbakov M 2016 Geospatial data generation and preprocessing tools for urban computing system development *Procedia Computer Science 5th International Young Scientist Conference on Computational Science (YSC 2016)* Elsevier **101** 217–226

[2] Statistics of the real estate market, URL: http://www.statrn.ru/

[3] Panasenko N 2015 Price zone: old form - new content, URL: http://www.estimatica.info/assessment/real-estate/24-tsenovoj-poyas-staraya-forma-novoe-soderzhanie

[4] Gryaznova A G Fedotova M A 2001 Appraisal of real estate *Finances and Statistics*

[5] Patsuk E B, Korshakevich I S, Elistratova A A 2015 Modeling the cost of apartments in Krasnoyarsk in 2012 *Actual problems of aviation and astronautics* **11** 597–599

[6] Yuzaeva A G, Savchenko L M 2016 Modeling the cost of housing in Krasnoyarsk for 2013 *Actual problems of aviation and cosmonautics* **12** 123–125

[7] Yuzaeva A G, Savchenko L M 2016 Modeling the cost of housing in Krasnoyarsk for 2014 *Actual problems of aviation and cosmonautics* **12** 126–128

[8] Anisimova I N, Barinov N P, Gribovskiy S V 2004 Accounting for different types of pricing factors in multidimensional models of real estate appraisal *Valuation questions* **2** 2–15

[9] Dyakonov A G 2014 Methods for solving classification problems with categorical features *Applied Mathematics and Informatics* **46** 103–127

[10] Parygin D, Sadovnikova N, Kalinkina M, Potapova T, Finogeev A 2016 Visualization of data about events in the urban environment for the decision support of the city services actions coordination *(SMART 2016)* IEEE 283–290, URL: http://ieeexplore.ieee.org/document/7894536/

[11] Avito - The Internet site for placing ads about goods and services from individuals and companies, URL: Avito.ru