MONITORING OF APARTMENT PRICES IN THE CZECH REPUBLIC THROUGH PARSING A WEB ADVERTISING SERVER

Alena POZDÍLKOVÁ, Jaroslav MAREK, Marie NEDVĚDOVÁ
Department of Mathematics and Physics, Faculty of Electrical Engineering and Informatics, University of Pardubice, Studentská 95,
532 10 Pardubice,Czech Republic, tel. +420466 036 111,
E-mail: alena.pozdilkova@upce.cz, jaroslav.marek@upce.cz, marie.nedvedova@student.upce.cz

ABSTRACT

Time series of apartment prices in the Czech Republic are available only in the partial statistics of the Statistical Office. Apartment prices are presented mainly in the articles and comments from the real estate agents. Data unavailability leads to a small number of statistically oriented publications on the real estate market. The main aim of our paper is thus to introduce a software solution for parsing real estate websites. Of course, we are only able to retrieve data on demanded prices from advertisements, actual prices are not achieved. By automatic polling, we are able to get data on the floor area of advertised apartments and the asked purchase price. A Python script was written to retrieve data from sreality.cz. The MongoDB database is used to store ads. New ads are saved directly to the database. Then, daily average apartment price of 1 square meter for each municipality are calculated. The filtered data can then be displayed or exported to a file via the web interface. In the statistical analyses, we present graphs showing the development of apartment prices and the number of advertisements in various municipalities of the Czech Republic in the period of 09/2018 – 12/2019. Next, we address the issue of clustering of municipalities with regard to the similarity of relative price changes.

Keywords: *web page parsing, real estate market, time series, apartment prices, floor area, purchased price, cluster analysis*

1. INTRODUCTION

In this paper, we will present apartment prices in the Czech Republic in the period starting in July 2018 and ending in December 2019. We will not have data from real estate transfers, but we will only work with data from advertisements. Occasionally, data provided by large real estate companies appear in the media. But this information may be affected by the interests of real estate brokers and may not best map the real estate market behaviour. We can also see predictions of apartment price developments created by economic analysts.

This contribution focuses on an automatic polling of apartment prices from the largest Czech real estate advertising website. This automatic polling is based on the Python procedures with collecting data about apartment area and price from ads through municipalities in Czechia. Of course, the final aim is to construct long price time series and apartment supply time series. At first, in order to collect this everyday information into MongoDB database, all data structure have to be defined. In their general form, information on the location, area of the apartment, demanded price and also the type of apartment are obtained. This means that there is no simple algorithm capable of providing the average price of apartments in a particular municipality.

First, the price for one square meter of flat area had to be calculated. The average price per square meter was calculated from all advertisements in one particular municipality in a given day. The source of our data are ads of sreality.cz.

The output, resulting after the execution of the proposed parsing algorithms, is, of course, represented as the daily time series of prices and the daily time series of the number of advertised apartments for each municipality in the Czech Republic.

The sreality.cz website offers an open web (HTTP) API that allows you to read and manipulate advertisements on the server. Only the reading part was used for the purposes of statistical evaluation. Downloading ads uses HTTP GET request and results are passed in JSON format. Authentication is not required to access ads because the data is publicly available.

Similarly, prices for houses or land could also be monitored. However here, the averaging (more floors, incompatibility of reported areas: floor x built-up) is much more complicated and therefore we will only deal with flats. Our goal will not be to predict future price developments, but only to compare apartment price characteristics in different locations. Cf. (Kauskale [4]), (Risse [8]). A guide to our analysis was provided by the book (Winson-Geideman [10]). Real estate data is often taken from LAG models, see (Guo [2]), (Guo [3]) and (Pozdílková [7]). The issue of investing in real estate is also examined from the perspective of price bubble prediction. There are studies examining the risk of investing in real estate and comparing profitability with the stock exchange, cf. (Szumilo [9]). However, we will not perform spatial analysis or profitability analysis in this article.

2. PARSING

In this section we describe the programmed application for retrieving data from real estate servers.

2.1. Software solution

The data reserved for our research come from sreality.cz. This is some kind of web pages containing ads. Each advert was recorded separately using a Python script. Cf. (Oliphant [6]) and (McKinney [5]). At 2:30 am every day, the script tries to download all ads from each district from the website sreality.cz. The MongoDB database is used to store ads. New ads are saved directly to the database. If an ad has already been stored in the database, it will only be updated if needed (price change, etc.). Then, daily and monthly summaries for all used filters are created in the database. The filtered data can then be displayed or exported to a file after access via the web interface. Finally, the data are ready to be examined.

2.2. Architecture of data mining solution

The complete data mining solution is built using the Docker container virtualization platform. The container virtualization is a modern approach for building of scalable (web) services. The container represents a group of applications (usually only one application) that run in restricted and independent environment on a shared operating system. Container virtualization is much "lighter" virtualization method than full OS virtualization, thus is faster and requires less computing resources. Docker platform represents a tool which enables easy creation and management of containers on a single machine. Other popular container platforms alternatives include Kubernetes, Mesos and LXD. A total of 3 containers were used:

(a) applications that download and update data.,
(b) web application for presentation of results,
(c) MongoDB database server used for data persistence.

MongoDB database was chosen for persistence of data of individual advertisements. This NoSQL database provides not only means for persistence of data. Complex aggregation pipelines are widely used in the processing of downloaded data. The pipelines perform transformation, grouping and resulting calculation of results. The auxiliary aggregate values are further stored in the "cache" collection in the database. In the future, it is possible to count on the use of sharing for managing larger volumes of recorded data. Both the download and result presentation applications are programmed in Python (version 3.7) and run as separate containers. The application for downloading data uses the Python core library and the Pymongo library to communicate with the database. The web application is built over the Flask library and also uses the Pymongo, pandas and pillow libraries.

The Docker Container Platform allows the entire solution to be easily run in the target environment and offers the option of an easy transition to the Docker Swarm platform if performance is not sufficient in the future. Docker Swarm acts as an orchestrator of containers running across cluster of independent computers. Both the database and the web application can be scaled horizontally to improve performance. Database offers use of sharing and replica sets techniques for performance scaling (and for data redundancy). In (Cook [1]) are some examples of Docker usage.

2.3. Download data

Application for downloading data is basically quite simple script that optimally every day at 2:30 performs all operations. If problems occur, the download will be retried later. Data loading itself is a trivial use of the Sreality API. All districts and all advertisements are scanned. For individual advertisements, it is first tested if it is already present in the database - a combination of hash id, ad title and type of advertisement is used for verification. If the advertisement is already present, the existing record will be updated and current values will be added. If the advertisement is not in the database at all, the loaded advertisement is transformed into the required form and saved in the database.

After all advertisements are downloaded, aggregate pipelines are started in the database. In the first phase, daily summaries are created by municipality and district. In the second phase of the algorithm, monthly aggregations are created in a similar way. After the aggregations are completed, the results can be presented using a web application. Original data are preserved in the process; aggregated results can be recalculated if needed.

2.4. The web application

The application is built on a simple Flask framework. Several basic functions were included in the application during programming:

- export of selected statistics with the possibility to select and apply filtering (for example, apartments 1 + 1, 1 + 2, 1 + 3, flat in buildings with bricks, with concrete panel),
- display a map of all ads above the real map background,
- display a map of districts of the Czech Republic with average values of prices per square meter.

The basic function is to transform data from the database and present them in CSV format. The necessary data are read directly from the database and using the pandas' library, organized into the resulting table and exported to the client.

To verify the functionality of the whole solution, the support was created for displaying the map of the Czech Republic with rendering of individual ads. The Open Layers library and a map file from the OpenStreetMap service were used for the implementation. Advertisements are passed as a separate map layer above the default map background. Rendering is done using the python and pillow library. Server-side rendering of ads allows to render all (or aggregated) ads for web client. This approach requires a higher server performance but allows serving results to all types of clients.

The last functionality of the application is displaying a map of districts of the Czech Republic and information on the total price for individual months with the possibility to compare and evaluate the prices over time.

3. EXPLANATORY ANALYSIS

Through the parsing of the real estate website, we have acquired data for the past year containing the municipality identifier, the apartment price, the apartment area and the date of the advertisement. The positions of all sites with ads are rendered in the Fig. 1. Examples of such data are given in Tab. 1. On 30th April 2019 the total number of ads analysed was 16748. Overall, in the period between September 1st, 2018 and August 30th, 2019, we loaded 6,845,000 ads in 250 working days. The location of the sites with ads is shown on the map. The GPS coordinates of each settlement were obtained during parsing.

3.1. Basic analysis

For each municipality that appeared in at least one ad, we calculated the average value and the standard deviation of the price of a one square meter of apartment. The example of measured values for chosen region is given in Table I.

From the obtained data in the structure of Tab. 1, we can get basic statistical characteristics (number of ads, average daily price $\bar{x}$ in a given municipality, standard deviation s in a given municipality).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,,\quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

An example of these variables is given in Tab. 2. These values can be accumulated over a longer period. This gives us time series of prices for all municipalities with advertised apartments. This data can be studied using basic methods for time series processing. Time series in the reference period allow to assess how the change in the rules for mortgage lending in the end of 2018 affected average prices.

In Figure 2, the map at the top shows that between September 30st, 2018 and December 31th, 2018 the overall average price in about half of the districts decreased, despite an increase in the middle of this interval. In the figure, the districts where prices have risen, are shown in red. Districts with decreasing price per 1 m^2 of flat are depicted in green. The stripe of the coloured squares shows the progression of prices in the districts of Prague 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

The second map demonstrates what price changes occurred between December 2018 and March 2019, etc. The differences are especially in districts of North and South Moravia. In the first quarter of 2019, the number of red districts is increasing and the upward trend in prices is beginning to emerge.

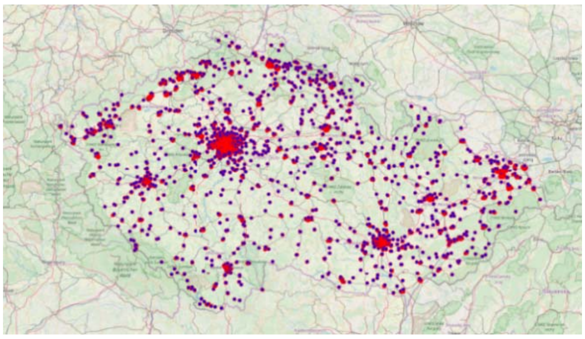


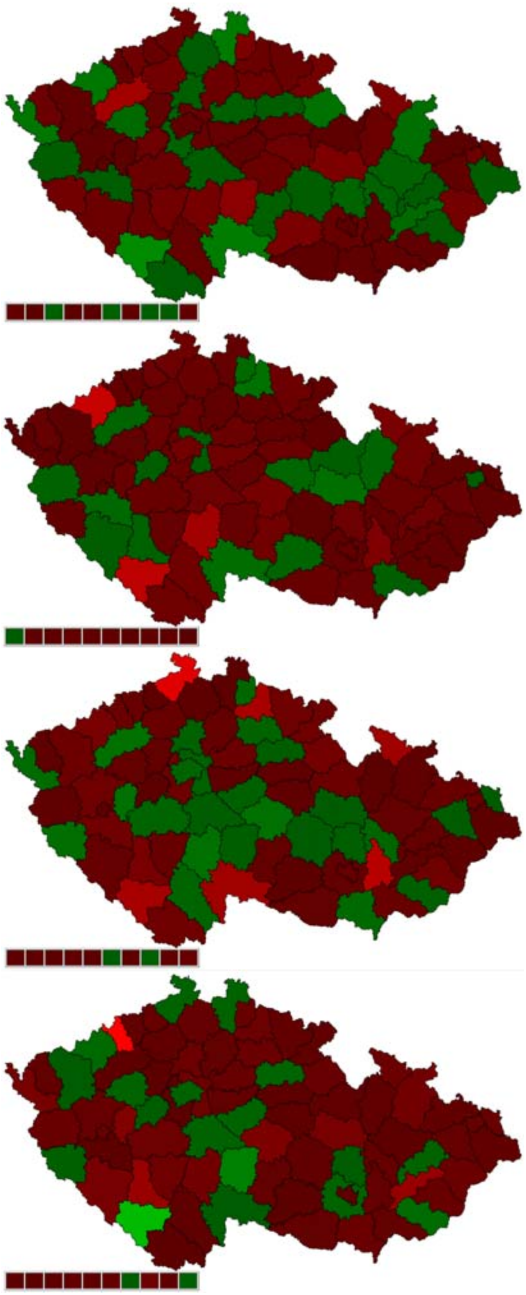
Fig. 1 Distribution of advertisement. Source: own

Table 1 Source data

Data					
		Date	*Price*	*Area*	*Prize of 1 m^2*
City	*Žďár nad Sázavou*	2019-08-05	2400000 Kč	74.0	324332.4 Kč
		2019-08-05	2400000 Kč	57.0	29122.8
		2019-08-05	2400000 Kč	63.0	28412.8
		2019-08-05	2400000 Kč	68.0	30882.4
		2019-08-05	2400000 Kč	52.0	26153.8
		2019-08-05	2400000 Kč	123.0	32113.8
		2019-08-05	2400000 Kč	71.0	25338.0
		2019-08-05	2400000 Kč	78.0	33961.54

Table 2 An example of summary data

City	*No. of Ads*	*Average price of 1 m^2*	*Standard deviation*
Žďár nad Sázavou	8	29802.2	2880.1
Velká Bíteš	2	35682.5	341.6
Nové Město na Moravě	3	49401.8	3704.6
Lavičky	5	40219.9	502.9
Rozsochy	1	15000.0	
Velké Meziříčí	1	29649.1	
Bystřice nad Pernštejnem	3	32398.8	3312.9
Osová Bitýška	1	24105.3	



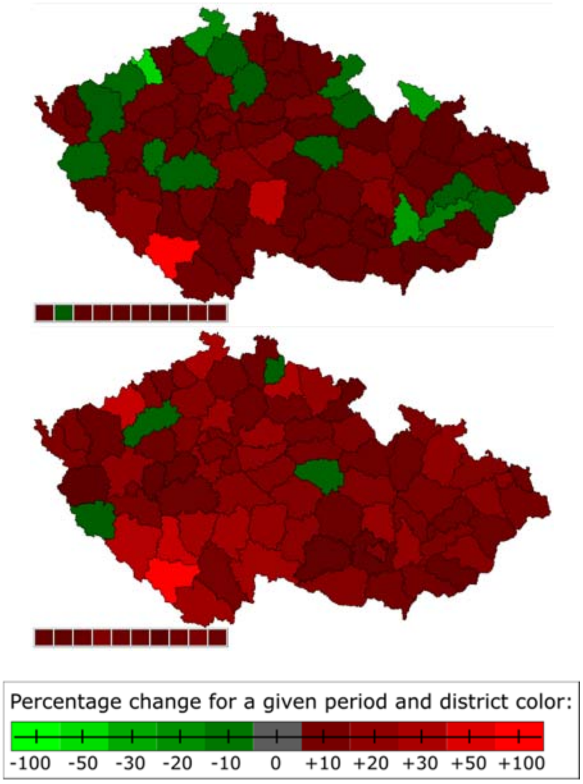


Fig. 2 Histogram of average prices for 1m^2 for all ads. September 2018 versus December 2018. December 2018 versus March 2019. March 2019 versus June 2019. June 2019 versus September 2019. September 2019 versus December 2019. December 2018 versus December 2019. Source: own

Map on the bottom monitors changes between December 2018 and December 2019. The average price decreased only in 4 districts out of 87 districts. These included the districts of Domažlice, Louny, Jablonec nad Nisou, Chrudim.

3.2. Price development

With the help of basic graphic tools, we will present to the reader some interesting facts. Figure 3 and 4 shows the histograms of average prices from 1st January 2019 to 31th December 2019 in two selected districts. It can be seen that the small district of Žďár nad Sázavou has a greater kurtosis and a smaller variability than Prague 10. This effect was also apparent in other districts.

Of course, price developments can best be grasped by showing time series. Figure 5 – 8 shows the graph of price development (in thousands of CZK) and the number of advertisements in Prague 10 and Žďár nad Sázavou. It shows how the number of ads is growing. In the beginning of the horizontal axis is November 2018, when conditions for providing mortgages were tightened. The charts show growth in the number of ads and also the growth in prices before changes. The average price then falls in Prague only for a short period of time, when the price increase is clearly related to the decline in supply. In Žďár nad Sázavou, the declining price correction remains for the first quarter of 2019.

In the period under review, the average price calculated from 8,052,446 ads was 56,182 CZK.

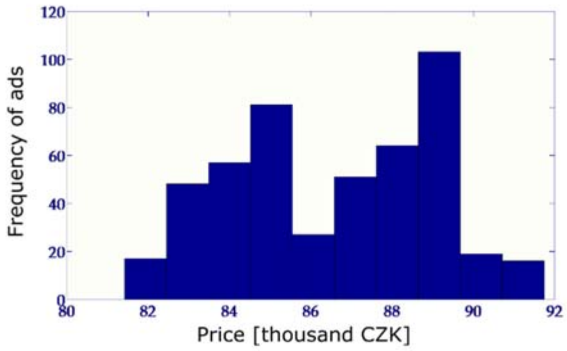


Fig. 3 Prague 10: histogram of price per square meter [thousand CZK/m^2] in considered period. Source: own

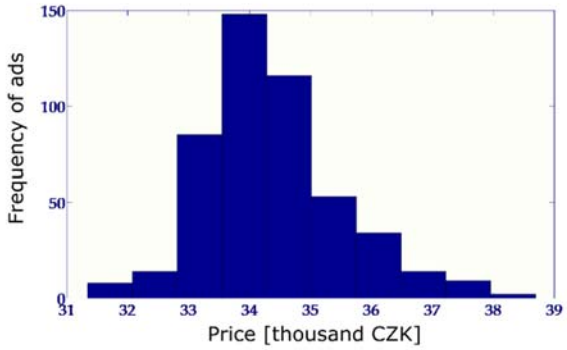


Fig. 4 Žďár nad Sázavou: histogram of price per square meter [thousand CZK/m^2] in considered period. Source: own

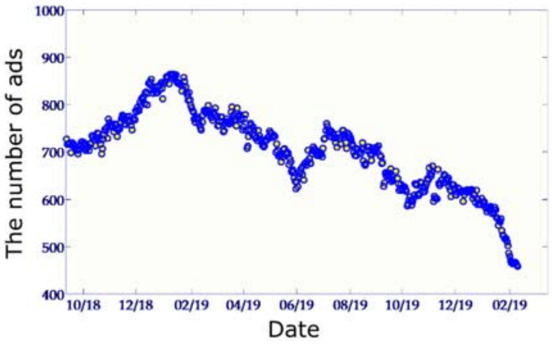


Fig. 5 The number of ads – region Prague 10. Source: own

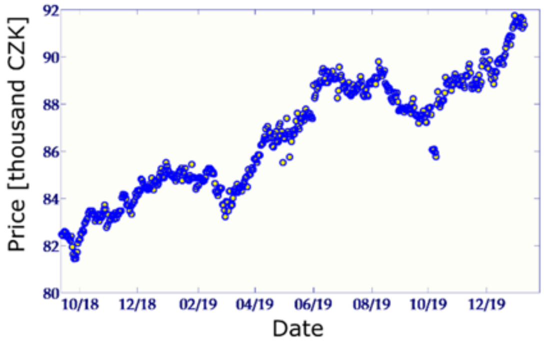


Fig. 6 Price graph [thousand CZK] – region Prague 10. Source: own

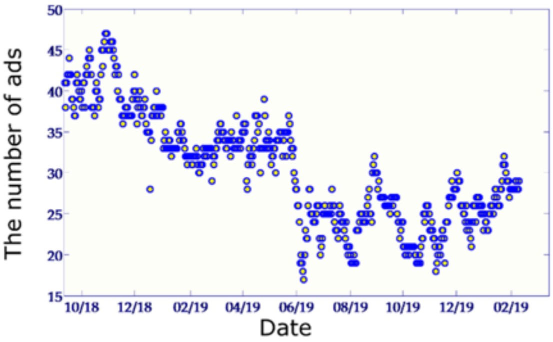


Fig. 7 The number of ads – Žďár nad Sázavou (supply decrease). Source: own

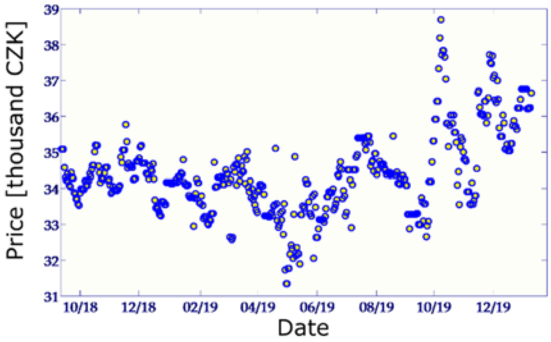


Fig. 8 Price graph – Žďár nad Sázavou. Source: own

4. CLUSTER ANALYSIS

Through the parsing of the real estate website, we have acquired data for every day in the past 18 months containing the average apartment price in every municipality. The main goal of the cluster analysis lies in a division of data into similar groups. In our study, we will perform a cluster analysis of price time series using the k-means algorithm.

4.1. K-means

The clustering is the process of splitting dissimilar data into disjoint clusters and grouping similar data into the same cluster based on some predefined similarity; k-means is the most popular method based on decomposition. First, it is necessary to determine the number of clusters for which the initial centroids (cluster centres) are determined. The next step is assigning objects to the clusters according to the calculated distance (most often Euclidean). Subsequently, the position of the cluster centroids is recalculated according to the set of objects belonging to the cluster. Since centroids change their coordinates, rechecking objects for clusters is needed. Centroids are recalculated and objects are moved between clusters until no change is witnessed. It is necessary to repeat the whole process several times, because the result always represents only a local optimum. The distance measure is computed as

$$DE(x_i, x_j) = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

4.2. Results

The clustering process results in assigning 87 districts to following clusters:

Cluster 1: Praha 5, Praha 6, Praha 8, Praha – město.
Cluster 2: Praha 3, Praha 7.
Cluster 3: Praha – východ.
Cluster 4: Semily, Trutnov.
Cluster 5: Děčín, Louny, Nový Jičín, Náchod, Ostrava – město, Strakonice, Svitavy, Česká Lípa.
Cluster 6: Chrudim, Domažlice, Frýdek Místek, Havlíčkův Brod, Hodonín, Jeseník, Kroměříž, Kutná Hora, Litoměřice, Opava, Plzeň – jih, Prostějov, Přerov, Rokycany, Tachov, Tábor, Třebíč, Ústí nad Orlicí, Šumperk.
Cluster 7: Liberec.
Cluster 8: Benešov, Blansko, Břeclav, Jihlava, Jičín Klatovy, Kolín, Pelhřimov, Plzeň – sever, Písek, Příbram, Rakovník, Rychnov nad Kněžnou, Uherské Hradiště, Vsetín, Znojmo, Český Krumlov, Žďár nad Sázavou.
Cluster 9: Jablonec.
Cluster 10: Most.
Cluster 11: Bruntál, Chomutov, Karviná, Sokolov, Teplice, Ústí nad Labem.
Cluster 12: Beroun, Brno – venkov, Cheb, Hradec Králové, Karlovy Vary, Mladá Boleslav, Mělník, Nymburk, Olomouc, Pardubice, Plzeň – město, Zlín, České Budějovice.
Cluster 13: Brno – město, Praha – západ.
Cluster 14: Prachatice.
Cluster 15: Praha 10, Praha 4, Praha 9.
Cluster 16: Kladno.
Cluster 17: Jindřichův Hradec.
Cluster 18: Vyškov.
Cluster 19: Praha 2.
Cluster 20: Praha 1.

We will not conduct a detailed discussion of the resulting clusters. This would have to include a detailed demographic and economic analysis of individual districts. But you can see that the clusters have some logic. For example, in the 1st, 2nd, 3rd and 14th cluster, the districts of Prague are being taught. In the 11th cluster there are economically weak districts.

5. CONCLUDING REMARKS

This article represents our pilot contribution to real estate market monitoring. The aim a contribution is not to observe the price development of a particular type of apartment (e.g. brick or panel) over time, but to examine the trend of the whole housing market. In this contribution, a new way of obtaining data on the real estate market was introduced. Programmed Python procedures have been successfully retrieving data from a real estate website for over 18 months. The designed and implemented platform enables automated data collection from the sreality service and their automatic processing within the long-term operation. Platform functionality can also be expanded in the future and, if necessary, vertical and horizontal scaling can be used to increase platform performance.

The obtained time series allowed us to draw price graphs and perform selected statistical analyses. They also included clustering of districts of the Czech Republic according to similar development of apartment prices. The obtained data leads to the opportunity of further research in this field of study.

ACKNOWLEDGMENTS

This research was supported by the Internal Grant Agency of University of Pardubice, the project SGS 2020.

REFERENCES

[1] COOK, J.: Docker. In: Docker for Data Science. Apress, Berkeley, CA, 2017.

[2] GUO, J. – QU, X.: Spatial interactive effects on housing prices in Shanghai and Beijing. Regional Science and Urban Economics, 2018, no. 1, pp. 1–14.

[3] GUO, K. – WANG, J. – SHI, G. – CAO, X.: Cluster analysis on city real estate market of China: based on a new integrated method for time series clustering. Procedia Computer Science, vol. 9, 2012, pp. 1299–1305.

[4] KAUŠKALE, L. – GEIPELE, I.: Integrated Approach of Real Estate Market Analysis in Sustainable Development Context for Decision Making. Procedia Engineering, vol. 172, 2017, pp. 505–512.

[5] MC KINLEY, W.: Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. June 2010, vol. 445, pp. 51–56.

[6] OLIPHANT, T. E.: Python for scientific computing. In: Computing in Science & Engineering. May-June 2007, vol. 9, no. 3, pp. 10–20

[7] POZDÍLKOVÁ, A. – MAREK, J.: Spatial lag model for apartment prices in Pardubice region. In: D. Szarkova, D. Richtarikova, P. Letavaj, J. Gabkova(Eds.), Proceedings of 17th Conference on Applied Mathematics Aplimat 2018, Bratislava: Slovac University of Technology Bratislava, February 2018, pp. 867–875.

[8] RISSE, M. – KERN, M.: Forecasting house-price growth in the Euro area with dynamic model averaging. North American Journal of Economics and Finance, vol. 38, 2016, pp. 70–85.

[9] SZUMILO, N. – WIEGELMANN, T. – LASKIEWICZ, E – PEITRZAK, M. – BERNARD, M. – BALCERZAK, A. P.: The real alternative? A comparison of German real estate returns with bonds and stocks. Journal of Property Investment & Finance, 2018, vol. 36, no. 1, pp. 19–31.

[10] WINSON-GEIDEMAN, K. – KRAUSE, A. – LIPSCOMB, C. A. – EVANGELOPOULOS, N.: Real Estate Analysis in the Information Age: Techniques for Big Data and Statistical modelling. Routledge, Abingdon: Oxon. 2017.

Received February 17, 2020, accepted March 30, 2020

BIOGRAPHIES

Alena Pozdílková studied applied mathematics (MSc) at Palacký University in Olomouc and graduated in 2005. She completed her doctoral studies at the University of Hradec Králové, Faculty of Informatics and Management. She defended her PhD in the field of Information and knowledge management in 2014; her thesis title was “Optimization of selected decision problems using extremal algebras“. Since 2014 she is working as a tutor with the Department of Mathematics and Physics, at Faculty of Electrical Engineering and Informatics, University of Pardubice. Her scientific research is focusing on linear algebra, applications of mathematics in economics.

Jaroslav Marek study numerical mathematics at Palacky University Olomouc, doctoral study of applied mathematics at Palacky University Olomouc. His Ph.D. thesis from 2007 — Statistical problems of decreasing of metrological uncertainties of the type B — was focused on certain problems of geodesy. Lecturer at the Department of Mathematical analysis and mathematical applications, Faculty of Science, Palacky university Olomouc from 1998-2009. Since then, he works at the Department of Mathematics and physics, Faculty of Electrical Engineering and Informatics, University of Pardubice 2010-present. Author’s research interests are applied mathematics, mathematical statistics, especially with applications in geodesy.

Marie Nedvědová in 2017 finished the study of Computer Science at Faculty of Electrical Engineering and Informatics, University of Pardubice. Her master thesis was devoted to the application of mathematics to digital art. Her scientific research is focused on computer graphics, digital art, applications of mathematics. At present, she undergoes the doctoral study program Electrical engineering and informatics at University of Pardubice.