# An Approach to Web Advertising Server Parsing - Apartment Price Analysis in the Czech republic

Alena Pozdílková
University of Pardubice
Faculty of Electrical Engineering and Informatics
Studentská 95, 530 02 Pardubice I, Czech Republic
alena.pozdilkova@upce.cz

Marie Nedvědová
University of Pardubice
Faculty of Electrical Engineering and Informatics
Studentská 95, 530 02 Pardubice I, Czech Republic
st36806@student.upce.cz

Jaroslav Marek
University of Pardubice
Faculty of Electrical Engineering and Informatics
Studentská 95, 530 02 Pardubice I, Czech Republic
jaroslav.marek@upce.cz

***Abstract*— The real estate market is not subject to frequent analyzes in the academic field, which is due to the difficulty in obtaining data. 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. At 2:30 am every day, the script downloaded ads 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 used filters are created in the database. The filtered data can then be displayed or exported to a file via the web interface. We will present analysis of data in various municipalities of the Czech Republic in the period 09/2018 – 08/2019.**

***Keywords—real estate market; time series; apartment prices***

## I. INTRODUCTION

In this article, we will analyze apartment prices in the Czech Republic in the period starting in July 2018. 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 behavior. We can also see predictions of property price developments created by economic analysts. In our study, we will process the data we collect through long-term collection from large relative servers. By parsing web pages, you can get prices from "apartments for sale" ads. The source of our data are ads of sreality.cz.

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.

By simple calculation we get the converted price for one square meter. 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. (Risse [6]). A guide to our analysis was provided by the book (Winson-Geideman [8]). Real estate data is often taken from LAG models, see (Guo [2]) and (Pozdílková [5]). 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 [7]). However, we will not perform spatial analysis or profitability analysis in this article.

## II. PARSING

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

### *A. 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 [4]) and (McKinney [3]). 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.

### B. Architecture of data mining solution

The complete data mining solution is built using the Docker container virtualization platform. 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. Both the database and the web application can be scaled horizontally to improve performance. In (Cook [7]) are some examples of Docker usage.

### C. 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.

### D. 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),

- 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.

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 development of prices over time.

## III. 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 analyzed 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.

### A. 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.

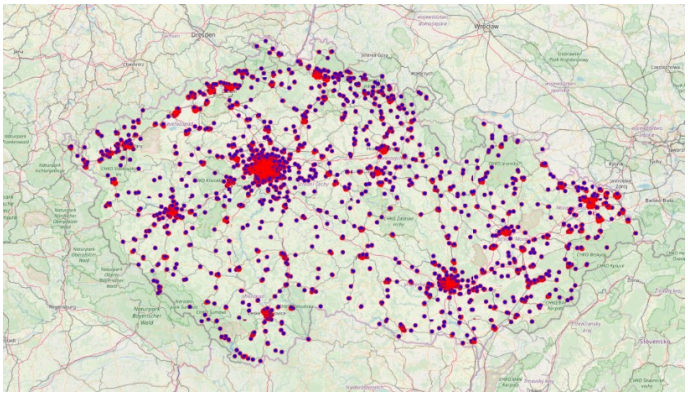

Fig. 1. Distribution of advertisement. Source: own.

TABLE I. SOURCE DATA

| Data | | | | | |
|---|---|---|---|---|---|
| | | *Date* | *Price* | *Area* | *Prize of 1 m²* |
| *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 |

[a.] Source: own.

TABLE II. AN EXAMPLE OF TEXT FEATURE VALUES

| *City* | No. of Ads | Average price of 1 m² | *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 | |

[b.] Source: own.

From the obtained data in the structure of Tab. 1, we can get basic statistical characteristics (number of ads, average daily price in a given municipality, standard deviation in a given municipality). 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 autumn of 2018 affected average prices. In Figure 2 at the top left, shows that between October 31st, 2018 and April 30th, 2019 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 m² of flat are depicted in green. The stripe of the colored squares shows the progression of prices in the districts of Prague 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. The second map at the top right demonstrates what price changes occurred between October 2018 and January 2019. The only difference is in a small number of districts, for example in 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. Of course, price fluctuations, see Figure 5 on the bottom left, can of course be influenced by only a few developers' projects, which will increase the average price, especially in small districts.

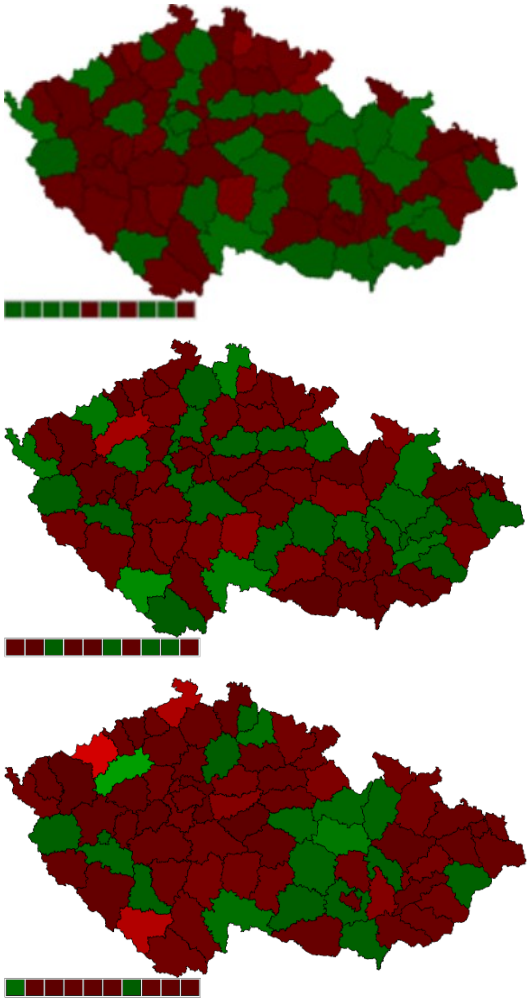

Fig. 2. Histogram of average prices for 1m² for all ads. April 2019 versus October 2018. January 2019 versus October 2018. April 2019 versus January 2019. Source: own.

## B. *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 September 2018 to 30th August 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 middle 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 6,845,000 ads was 55,075 CZK.

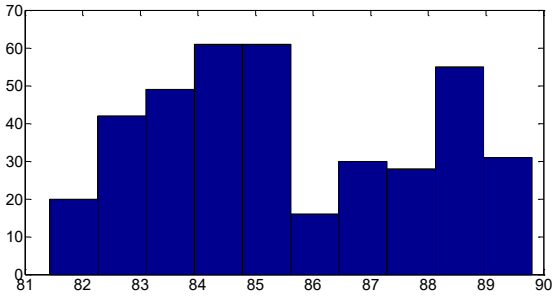


Fig. 3. Prague 10: histogram of price per square meter [thousand CZK/m$^2$] in considered period. Source: own.

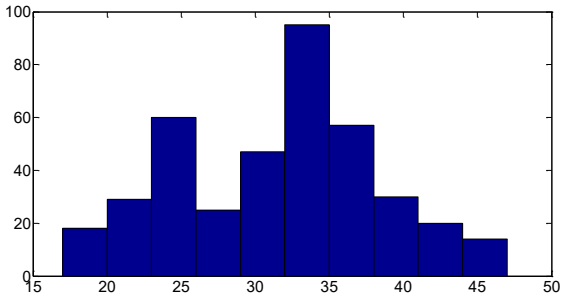


Fig. 4. Žďár nad Sázavou: histogram of price per square meter [thousand CZK/m$^2$] in considered period. Source: own.

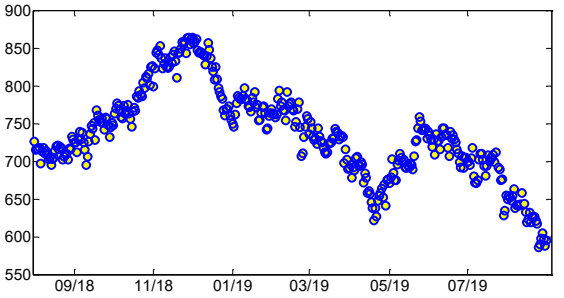


Fig. 5. The number of ads – region Prague 10. Source: own.

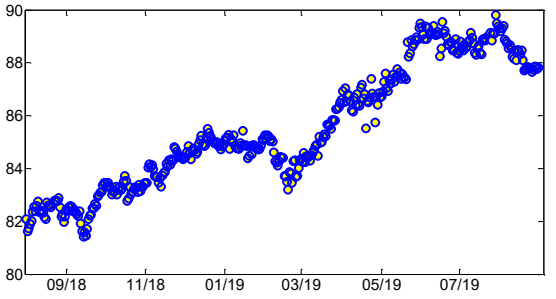


Fig. 6. Price graph [thousand CZK] – region Prague 10. Source: own.

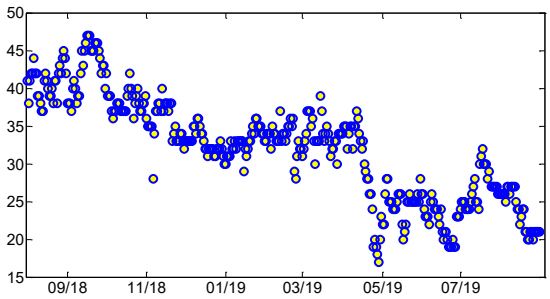


Fig. 7. The number of ads – Žďár nad Sázavou (supply decrease). Source: own.

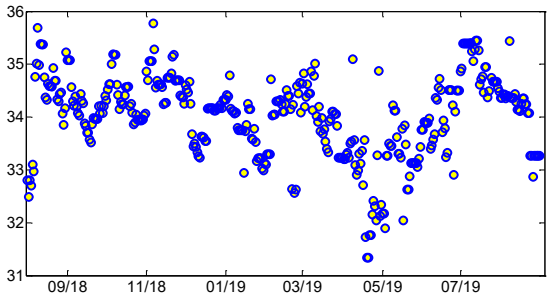


Fig. 8. Price graph – Žďár nad Sázavou. Source: own.

## IV. Concluding Remarks

The gist of article is a description of a software solution for collecting information about advertised apartments from real estate servers. 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 data leads to the opportunity of further research in this field of study.

## Acknowledgment

This research was supported by the Internal Grant Agency of University of Pardubice, the project SGS 2019 024.

## References

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

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

[3] W. McKinney, "Data structures for statistical computing in python." In: Proceedings of the 9th Python in Science Conference. June 2010, vol. 445, pp. 51-56.

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

[5] A. Pozdílková, and J. Marek, "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.

[6] M. Risse, and. M. Kern, “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.

[7] N. Szumilo, T. Wiegelmann, E. Łaszkiewicz, M. Pietrzak, M. Bernard, A.P. Balcerzak, “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.

[8] 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.