ALFA BK UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGY
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
ALFATECH Journal

ISSN 1XX0-3XX1(Print), ISSN 1XXX-6XXX (Online)

Alfa BK Univerzitet

UDK: 004.6:502.3]:711.417.4
DOI: 10.46793/AlfaTech1.1.11R
Original scientific paper

Application of Big Data Analytics in Smart City for Air Quality Analysis Task

Lazar Radovanović[1], Vojkan Nikolić[2]

Abstract: *This paper explores and shows the role and importance of applying Big Data analytics in a smart city. Specifically, the main goal of this paper is the analysis of air quality in a smart city. For the purposes of this analysis, a dataset containing air quality data for the Chinese city of Beijing will be used. The dataset contains air quality data collected from Internet of Things (IoT) devices. Also, the paper will present the processes of collection, cleaning, data analysis and visualization of the obtained results. For the successful implementation of this example, the Jupyter Notebook platform and the Python programming language with associated libraries were used.*

Key words: Big Data Analytics, Smart City, Internet of Things, Python, Analysis.

1 INTRODUCTION

As globalization accelerates, cities are also developing rapidly, becoming larger and more complex ecosystems marked by rapid population growth, infrastructure demands, and escalating environmental challenges. Among the most serious problems facing these urban areas is air quality, which directly affects public health, ecosystem stability and the overall quality of life of residents. In order to meet these challenges, cities are turning more and more to innovative technological solutions. The goal of the smart city is to create sustainable ecosystems using the mentioned technologies. One of the approaches that has the most potential is the implementation of Big Data analytics in the context of smart cities [1], [2].

The rapid development of the Internet and information and communication systems, as well as the Internet of Things (IoT), have led to the creation of a new era in the generation, collection and analysis of data. These systems and devices monitor various aspects of city life, including transportation systems, energy consumption, public safety, waste management, and even environmental conditions [3], [4]. However, to address the air quality issue, such a concept entails the establishment of a vast network of sensors in the smart city that monitor air quality in real time. This IoT device-based approach generates massive amounts of data that, when analyzed with Big Data analytics, can provide broad insights into air pollution patterns, pollutant sources and the effectiveness of mitigation strategies [5].

In this paper, at the very beginning, an introduction to the topic will be presented, with a special focus on the importance and role of Big Data Analytics in the smart city, that is, the analysis of air quality data collected from the Internet of Things. After that, in the second chapter, the concept of big data analytics, which enables the processing and analysis of large amounts of data, will be explained. The next, third part will be devoted to explaining the concept of applying Big Data analytics in the task of air quality analysis. In the fourth part, the data set used in this paper will be discussed in detail, with a focus on the data sources and the key parameters found in it. In the fifth chapter, the functions of the data analysis process will be presented and described, that is, the functions used for collecting, cleaning, analyzing data and visualizing the obtained results. In the next, fourth chapter, the results of the analysis will be presented, which will identify key trends and patterns on air quality and pollution, as well as providing a deeper insight into their dynamics and the factors that influence their change over time. Finally, in the conclusion, the importance of the application of big data analytics in the smart city will be highlighted.

2 BIG DATA ANALYTICS

In today's modern world, the concept of smart cities represents an innovative technological solution for cities. This concept enables the improvement of urban life through the application of advanced technologies such as Big Data analytics. The smart city project generates huge amounts of data every day. These generated volumes of data are called Big Data. The term Big Data primarily refers to large and complex datasets. Actually, there is no single and generally accepted definition of Big Data. There are numerous different definitions that define and explain the term Big Data. When analyzing the numerous definitions of big data, it can be concluded that the term Big Data describes a situation where datasets have become so voluminous that traditional technologies and tools are unable to process them and extract useful information.

Douglas Laney is considered to be the first person who defined the concept of Big Data in 2001 by giving three basic features, i.e. characteristics of Big Data [6]. Through these features, he created a framework for understanding the concept of Big Data. He did this by defining the 3 letter V model, i.e. "3V

model". The letter V comes from the first letter of the word that describes the features, i.e. the increasing volume, velocity, and variety of data [7]. If a dataset has these three characteristics, it can be considered Big Data.

Huge amounts of data generated by various systems and devices in a smart city need to be processed, i.e. collected and analyzed. These amounts of data are impossible to collect and analyze with classical methods and tools. For the purposes of analyzing this data, it is necessary to apply Big Data analytics, which has systems and technology for this task. Also, for the term Big Data analytics, there is no single and generally accepted definition. According to Microsoft, the term Big Data analytics refers to methods, tools, and applications for collecting, processing, and extracting insights from high-volume, high-velocity data [8]. Data analysis can be understood as a process that includes collection, cleaning, data analysis and a later one visualization of the obtained results. In other words, Big Data analytics provides opportunities to discover hidden trends, patterns and correlations in large sets of raw data in order to obtain information that allows decision makers to make decisions based on them.

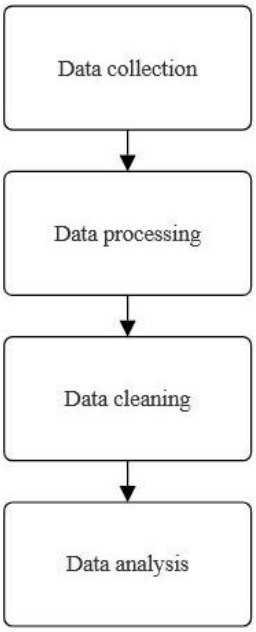

Figure 1 Big Data Analytics process [9]

Tableau, as one of the leaders in this field, defines Big Data analytics through four key steps (Figure 1). These steps enable efficient management of large amounts of data and extraction of useful information from it. The four key steps in Big Data analytics according to Tableau are: Data collection, data processing, data cleaning and data analysis [9].

3 CONCEPT OF ANALYSIS OF AIR QUALITY

The primary aim of this paper is to demonstrate the role and importance of applying Big Data analytics in a smart city. More precisely, the methodology of this paper focuses on the application of Big Data analytics for the task of air quality analysis in the context of a smart city. The approach is structured to use the vast amount of data generated by Internet of Things (IoT) devices used in urban environments to monitor and assess air quality in real time. The data set chosen for this task refers to the Chinese city of Beijing. The main steps of the methodology are data collection, data pre-processing and cleaning, data analysis, visualization of the obtained results. The dataset chosen for this task pertains to the Chinese city of Beijing. In this paper, the analysis of the mentioned dataset and the visualization of the obtained results will be performed in order to show the importance and role of Big Data analytics in this city. The analysis of the dataset will be performed in the Jupyter Notebook platform [10] by using the Python programming language and its libraries. The Python programming language version used is 3.10.9. [11]. Libraries used in this paper are Pandas and Matplotlib. The Pandas library will be used for importing the dataset and for data cleaning and analysis [12], while the Matplotlib library will be used to visualize the results [13]. These two libraries are extremely important for the data analysis process.

4 DATASET

In order to achieve the condition for the application of Big Data analytics in a smart city for the task of analyzing air quality data, a dataset is required. The dataset used in this work was taken from the Kaggle site [14]. In this dataset there are parameters of various air pollutants. The data for this dataset was collected from an IoT device stationed at the Olympic Park in the Chinese city of Beijing. Data were collected for the period from March 2013 to February 2017. The name of this measuring station where the IoT device is located is "Olympic Park" or "Aoti Zhongxin". This sensor, i.e. IoT device continuously collected data on various air pollutants, such as particulates (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO) and ozone (O3). The reason for choosing this data set is the lack of recent, publicly available air quality data in this city.

5 DATA ANALYSIS PROCESS FUNCTIONS

In this chapter, the functions that were used for the process of collection, cleaning, data analysis and visualization of the obtained results are presented. The functions used belong to the Python libraries Pandas and Matplotlib, and they are:

- import – this function is used to import Python libraries.
- pd.read_csv() – the function is used to import, i.e. load a set of data from the source csv file and convert it into a DataFrame file, suitable for analysis.
- head() – serves to display the first 5 rows of the dataset.
- info() – provides information about columns, i.e. variables, data types found in the DataFrame.
- pd.to_datetime() – converts one or more columns of string type into a date format column.
- index.year() – filters the dataset based on the year.
- isna().sum() – this function performs a search of the dataset to find missing values. Determines and

calculates the total number of missing values and displays the total value by column.

- fillna() – fills the missing values in the DataFrame by taking a known value that is above the missing value.
- groupby() - the function groups certain parameters for the purpose of analysis.
- mean() – calculates the mean value.
- subplots() - the function is used to create various shapes and elements on the diagram, it is also used to determine the size of the elements and the size of the diagram itself.
- bar() – this function is used to create bar charts.
- set_title() – sets the name of the title on the axes.
- set_xlabel() this function defines the labels that will be found on the X axis.
- set_ylabel() – the function defines the labels that will be found on the Y axis.
- set_xticks() – used to set ticks on the X axis.
- set_xticklabels() - used to set the label elements on the X axis. In this example, a parameter of 45 degrees is entered. The label name will be displayed according to the 45 degree angle.
- grid() - the function adds an auxiliary two-dimensional grid to the diagram, which enables easier observation and understanding of the results.
- legend() – describes the elements that are on the diagram.
- set_facecolor() – defines the background color of the diagram.
- tight_layout() – adjusts the defined parameters in the diagram.
- show() – displays the elements on the diagram.

6 RESULTS OF THE ANALYSIS

After the presentation and description of the functions used in the process of collection, cleaning, data analysis and visualization of the obtained results, in this chapter the description and presentation of the achieved results of the analysis, i.e. their visualization, was carried out. The analysis is divided into three parts. The main goal of the first part of the analysis is to calculate the mean value of PM10 and PM2.5 particles by month for the year 2016 for the Chinese city of Beijing. The reason why 2016 was chosen is that it is the last year with complete data for all months, while for 2017 there is only data for January and February. The result of this analysis is shown according to μg/m3 measurement unit, as can be seen in the following pictures (Figure 2 and Figure 3).

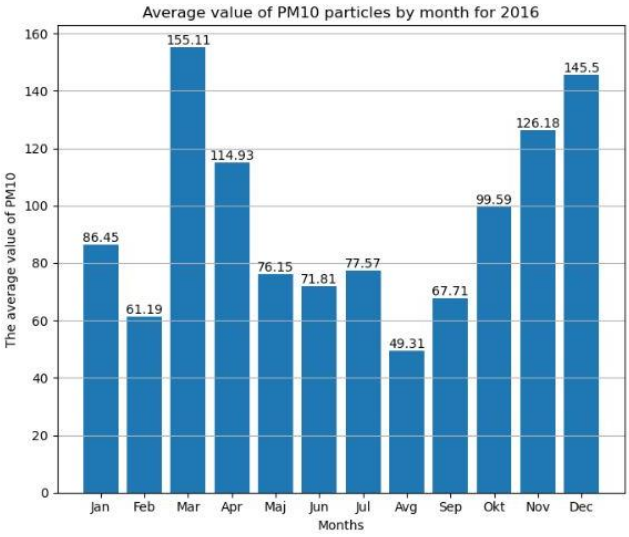

Figure 2 Display of PM10 particles

On the Figure 2 shows the visualization of the analysis results for PM10 particles. Based on the analysis of the image, that is, the diagram, certain patterns and trends can be observed. Based on the diagram, it can be seen that the month of March and then the month of December are the periods with the highest average values of the concentration of PM10 particles. This phenomenon can be attributed to certain meteorological and climatological conditions, but also to other factors such as increased industrial production and the use of fossil fuels. Also, it can be observed that during the summer, the mean values of the concentration of PM10 particles are the lowest, while already, as autumn begins, there is a gradual increase of PM10 particles. The highest concentration of PM10 particles based on the diagram is during the winter and spring periods. This can be attributed to certain seasonal causes, because in winter and spring the degree of heating season is the highest, as well as due to low temperatures that influence the air to be denser, which leads to the retention of PM10 particles in the air. While in the summer period the temperature is higher, the air is not retained, but allows much more PM10 particles, which reduces the level of air pollution.

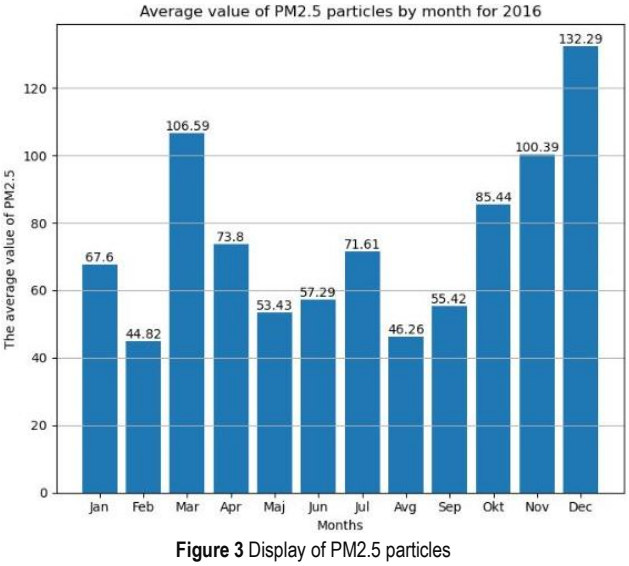

Figure 3 Display of PM2.5 particles

On the Figure 3 shows the mean value of PM2.5 particle concentration by month in 2016. Based on the diagram, it can be seen that the values according to the months are approximate compared to the months with PM10 particles, with the difference that the month of December with PM2.5 particles has a higher average value than the month of March with PM10 particles.

PM10 and PM2.5 particles can lead to serious consequences for the health of residents, including respiratory and cardiovascular diseases. Therefore, understanding these indicators can be key to making decisions about reducing pollution and improving air quality [15].

The second part of the data analysis also about to the analysis of PM10 and PM2.5 particles. The main goal of the second part of the analysis was to obtain the total number of days according to the categories, that is, the level of the concentration of PM10 and PM2.5 particles. In this analysis, the Chinese air quality control standard was used to determine the range of categories values according to which days belong to a certain category [16]. Based on this standard, there are six air quality categories, which can be seen in the following diagrams (Figure 4 and Figure 5).

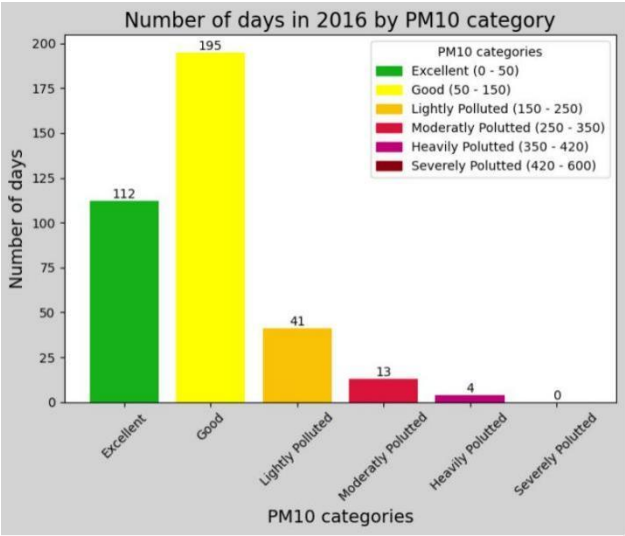

Figure 4. Number of days in 2016 by PM10 category

By analyzing the diagram (Figure 4), it can be concluded that according to PM10 particles, the largest number of days in 2016 is in the category of Good air quality with 195 days, followed by Excellent air quality with 112 days, Lightly Polluted with 41 days, Moderately Polluted with 13 days, Heavily Polluted with 4 days, while no days belong to the Severely Polluted air quality category.

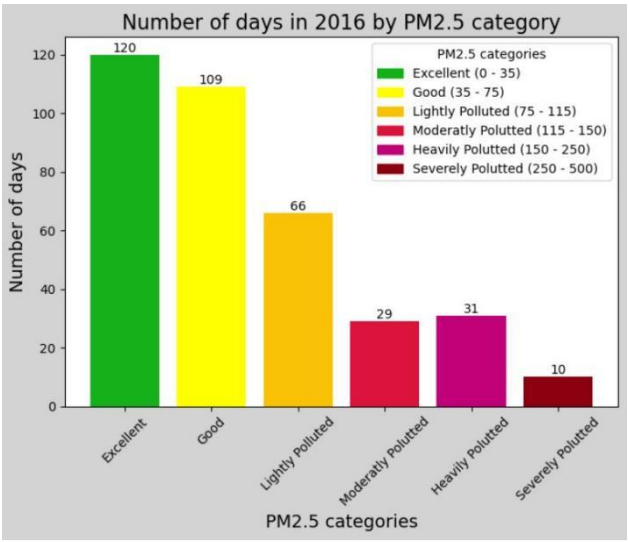

Figure 5 Number of days in 2016 by PM2.5 category

By analyzing the next diagram (Figure 5), it can be seen that the range of values of PM2.5 particle categories is different compared to PM10 particles, because different ranges are used according to the standard. Based on this, according to the concentration of PM2.5 particles in the air, it can be observed that the largest number of days in 2016 are in the first category called Excellent air quality with 120 days. After that, the Good air quality category follows with 109 days, while the Severely Polluted air quality category has 10 days.

The third part of the data analysis also refers to the analysis of PM10 and PM2.5 particles. The goal of the third part of the analysis was to obtain the average values of PM10 and PM2.5 particles for four years, namely 2013, 2014, 2015 and 2016. The year 2017 was not taken into consideration because it contains data for only the first two months, namely January and February.

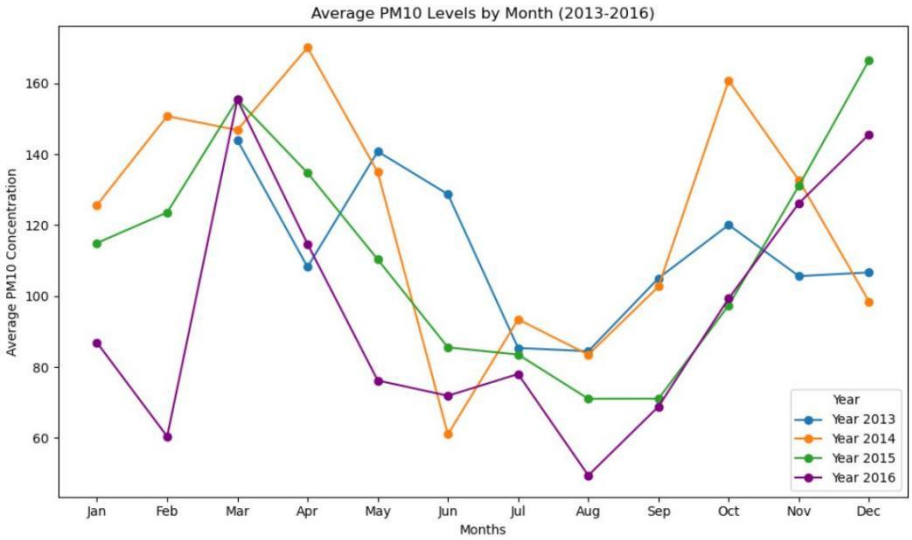

Figure 6 Display of average monthly values of PM10 particles for a period of four years (2013-2016)

The multi-line diagram (Figure 6) shows the average monthly values of the concentration of PM10 particles for four years. The horizontal axis shows the months of the year, while the vertical axis shows the average concentration of PM10 particles in micrograms per cubic meter (μg/m³). Each line on the diagram represents one year, so average PM10 values for different months over four years can be compared.

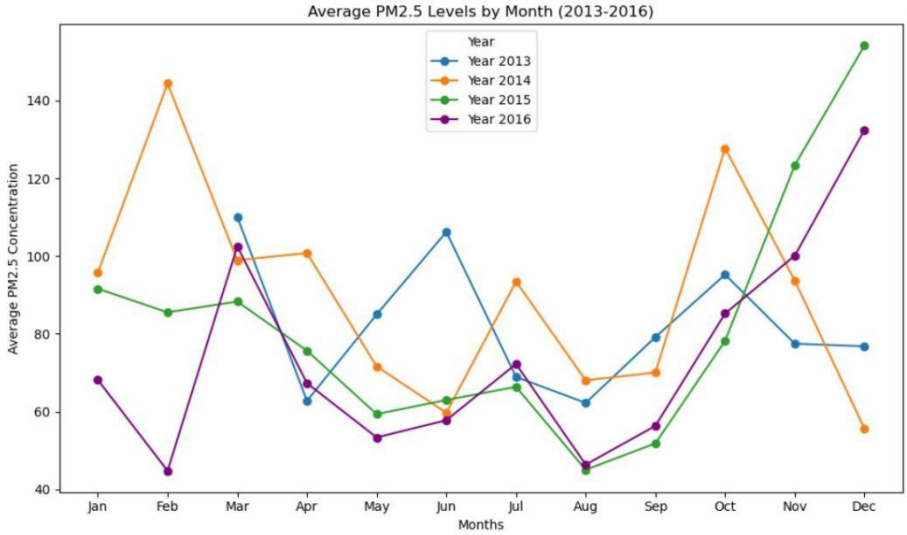

Figure 7 Display of average monthly values of PM2.5 particles for a period of four years (2013-2016)

The following, also a multi-line diagram (Figure 7), shows the average monthly values of the concentration of PM2.5 particles for four years. The diagram enables visual analysis of seasonal changes in the concentration of PM10 and PM2.5 particles and comparison of trends between different years. For example, it is possible to note which months have the highest concentrations of PM10 and PM2.5 particles, as well as how these levels change from year to year. This diagram can indicate possible changes in pollution that occur due to weather factors, industrial activities and long-term similar factors, as well as whether measures taken to reduce pollution have had an impact over time.

7 CONCLUSION

As urban areas develop and grow, the importance of innovative solutions becomes more and more necessary. This paper provides a comprehensive framework to approach the application of Big Data analytics in solving one of the most pressing challenges in urban environments today, air quality management, and therefore air quality analysis in a smart city. By combining IoT devices and big data analytics in this paper, precise monitoring and analysis of air quality is enabled. The application of these technologies contributes to improving urban living conditions by offering data-driven insights for better policy making, public health interventions and sustainable urban planning. Integrating Big Data analytics into air quality management systems is a critical step in building smarter and

healthier cities. Big Data analytics can be applied to any task in any area in a smart city, which is characterized by the generation of huge amounts of data. In particular, this paper uses air quality data collected with the help of IoT devices. After data collection, Big Data analytics was applied in order to analyze the collected data. Data on the parameters of PM10 and PM2.5 particles were analyzed. Considering the increasing focus on the protection of the environment and human health, it is important to analyze the causes of these polluting particles and work to reduce their values. Based on all of the above, it can be concluded that the application of big data analysis in a smart city is necessary for the normal functioning of the city and that it drastically contributes to preserving and improving the quality of life, public life and health. Also, this paper highlights the potential of smart cities to use data-driven approaches to create healthy, sustainable living spaces.

8 REFERENCES

[1] Picioroagă, I., Ermia, M., & Sănduleac, M. (2018). SMART CITY: Definition and Evaluation of Key Performance Indicators. *2018 International Conference and Exposition on Electrical And Power Engineering (EPE)*, pp. 217-222. https://doi.org/10.1109/ICEPE.2018.8559763

[2] Alshawish, R. A., Alfagih, S. A. M., & Musbah, M. S. (2016). Big data applications in smart cities. *2016 International Conference on Engineering & MIS (ICEMIS)*, pp. 1-7. https://doi.org/10.1109/ICEMIS.2016.7745338

[3] Hashem, I. A. T., Chang, V., Anuar, N. B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E., & Chiroma, H. (2016). The Role of Big Data in Smart City. *International Journal of Information Management*, 36(5), pp. 748-758. https://doi.org/10.1016/j.ijinfomgt.2016.05.002

[4] Madyatmadja, E. D., Munassar, A. H., Sumarlin & Purnomo, A. (2021). Big Data For Smart City: An Advance Analytical Review. *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, pp. 307-312. https://doi.org/10.1109/ICCSAI53272.2021.9609728

[5] Talebkhah, M., Sali, A., Marjani, M., Gordan, M., Hashim S. J. & Rokhani, F. Z. (2021). IoT and Big Data Applications in Smart Cities: Recent Advances, Challenges, and Critical Issues. *IEEE Access*, vol. 9, pp. 55465-55484, https://doi.org/10.1109/ACCESS.2021.3070905

[6] Kitchin, R., & McArdle, G. (2016), What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Kitchin, Rob and Gavin Mcardle. "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1). https://doi.org/10.1177/2053951716631130

[7] Buyya, R., Calheiros, R. N. & Dastjerdi, A. V. (2016). *Big Data: Principles and Paradigms*. Morgan Kaufmann.

[8] See https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-big-data-analytics

[9] See https://www.tableau.com/analytics/what-is-big-data-analytics

[10] Savira, P., Marrinan, T. & Papka, M. E. (2021). Writing, Running, and Analyzing Large-scale Scientific Simulations with Jupyter Notebooks. *2021 IEEE 11th Symposium on Large Data Analysis and Visualization (LDAV)*, pp. 90-91. https://doi.org/10.1109/LDAV53230.2021.00020

[11] See https://www.python.org/downloads/release/python-3109

[12] See https://pandas.pydata.org/docs

[13] See https://matplotlib.org/stable/index.html

[14] See https://www.kaggle.com/datasets/kurniatilaelimunifah/prsa-data-aotizhongxin

[15] See https://www.compactweathersensor.com/info/the-hazards-of-pm2-5-and-pm-50447990.html

[16] See http://asianfootprint.blogspot.com/2013/12/the-2013-mep-readings-of-new-aqi-in.html

Contact information:

Lazar Radovanović[1], Master of Science in National Security
Year of birth: 1999
E-Mail: lradovanovic800@gmail.com
https://orcid.org/0009-0000-9625-693X

Vojkan Nikolić[2], Associate Professor
(Corresponding author)
Year of birth: 1971
University of Criminal Investigation and Police Studies
Postal address: Cara Dusana 196, 11080 Beograd
E-Mail: vojkan.nikolic@kpu.edu.rs
https://orcid.org/0000-0002-9230-7549