ALFA BK UNIVERSITY
FACULTY OF INFORMATIONTECHNOLOGY
FACULTY O F MATHEMATICS AND COMPUTER SCIENCE
International Scientific Conference

UDK: 351.777.6
DOI: 10.46793/ALFATECHproc25.016S
Original scientific paper

AIR POLUTTION – CASE STUDY Podgorica UB

Biljana Stamatovic[1], Sinisa Stamatovic[2]

Abstract: In December 2024, the EU defined new rules of air quality standards by cutting allowed limit values. Five years' data about air pollutants, PM2.5, PM10, and SO2, are extracted from the measuring station in Podgorica – Podgorica UB. Analysis of the data is presented. The study used classical statistical analysis and machine learning K-mean and K-medoid methods to classify data and to make conclusions about air quality and air pollutants in Podgorica.

Keywords: Statistics, data analysis, K-mean, K-monoid.

1 INTRODUCTION

The European Union (EU) is striving to deliver cleaner air by reducing pollution levels for both short-term and long-term exposure. On October 23rd, 2024, the European Parliament adopted a new directive defining limits and models for calculating values for air pollutants [1]. Concrete, the revised Ambient Air Quality Directive updates air quality standards, lowering the allowable levels for twelve air pollutants: particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2) and nitrogen oxides (NOx), sulphur dioxide (SO2), ozone (O3), carbon monoxide, benzene, benzo(a)pyrene, arsenic, cadmium, nickel, and lead [2]. Lowering these values improves both the lives of people and the environment since it forces polluters to modify their operations to these levels.

Additionally, the World Health Organization (WHO) has guidelines for measuring the limits of certain air pollutants, including carbon monoxide, nitrogen dioxide, sulphur dioxide, ozone, and particle matter (PM2.5 and PM10) [3]. It is estimated that exposure to polluted air causes millions of deaths and leads to loss of health. The ecosystem of the world and our lives are negatively impacted by air pollution. [4].

Data on pollutants produced over the previous five years by the Podgorica UB monitoring station of the Environmental Protection Agency of Montenegro were used in the study. Data on the pollutants SO2, PM2.5, and PM10 are gathered by this station. [5]

In order to identify the source of pollution, we perform exploratory analyses using statistical analysis, data visualization, and machine learning clustering approaches. We make use of the RStudio IDE and the R programming language.

1.1 Data

Cleaning the extracted data is the first step in the process. The rows that have a NA (not available) are also removed, and all nonnumeric data is altered ("<2" is altered by 1.9, and "<5" is altered by 4.9). We discovered that much of the data for 2021 was missing after this data modification. Data from the years 2020–2024 were chosen. Figure 1 shows the results of transforming data by adding columns for the year and month. The month and year are taken from the first column of the row data (the first column represents the data collection time).

Date <chr>	SO2 <dbl>	PM10 <dbl>	PM2.5 <dbl>	year <chr>	month <chr>	day <chr>
15.01.2020 13:00	26.6	59.1	46.4	2020	01	15.01.2020
15.01.2020 14:00	23.5	40.5	31.0	2020	01	15.01.2020
15.01.2020 15:00	27.2	36.5	27.7	2020	01	15.01.2020
15.01.2020 16:00	29.8	45.3	35.4	2020	01	15.01.2020
15.01.2020 17:00	15.0	99.4	90.6	2020	01	15.01.2020
15.01.2020 19:00	9.3	186.9	151.0	2020	01	15.01.2020
15.01.2020 20:00	11.2	237.9	194.5	2020	01	15.01.2020
15.01.2020 21:00	10.1	391.5	317.2	2020	01	15.01.2020

Figure 1 Data **-** 30,618 entries, 6 total columns

Visualisation of data are presented on Figure 2.

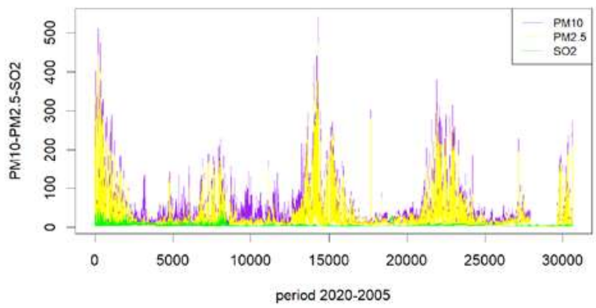

Figure 2 Visualization of data.

2 STATISTIC DATA ANALYSIS

Summary of the pollutants' data values are presented in Figure 3 (minimum, 25% quartile, median, average value, 75% quartile and maximum).

```
      SO2               PM10              PM2.5
Min.   : 1.90    Min.   :  4.90    Min.   :  4.90
1st Qu.: 2.30    1st Qu.:  9.10    1st Qu.:  6.40
Median : 3.70    Median : 15.50    Median : 10.40
Mean   : 4.36    Mean   : 28.82    Mean   : 21.06
3rd Qu.: 5.40    3rd Qu.: 30.30    3rd Qu.: 17.90
Max.   :47.60    Max.   :541.00    Max.   :471.70
```

Figure 3 Summary of data.

2.1 SO2

Sulphur dioxide (SO2) is a gaseous air pollutant that is mainly produced by the combustion of fuels (power plants, internal combustion engines, manufacturing, oil refining, diesel engines, etc.). SO2 causes harmful effects on the lungs (difficulty breathing, chest tightness, inflammation of the airways, etc.) [6]. According to the new EU directive [1], it is recommended that the level of SO2 higher than 50 μg/m^3 should not be exceeded more than 18 times during the year. Our data shows a maximum value of 47.6 μg/m^3, and no hours over the 2020–2024 timeframe had a greater value. Podgorica is therefore in very good shape with relation to the SO2 pollution.

2.2 PM2.1 and PM10

A variety of chemical species are mixed together to form particulate matter in the atmosphere. These particles vary greatly in size, shape, and chemical composition and can contain inorganic ions, metal compounds, elemental carbon, organic compounds, and compounds from the earth's crust. For air quality, their diameter is important. The diameter of PM2.5 is 2.5 microns or less, and PM10 is 10 microns or less. PM2.5 contains a fraction of PM10. [7].

Their correlation in our data is positive and it is greater than 0.97 with 95 percent confidence interval [0.9732772, 0.9744332].

On the Figure 4 are presented mean values of PM10 and PM2.5 per months for the 2020-2024 period.

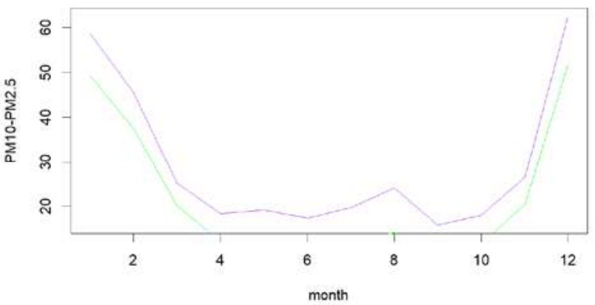

Figure 4 Average values of PM10 and PM2.5 per months

Premature mortality, acute and chronic bronchitis, asthma episodes, emergency room visits, respiratory symptoms, limited activity days, and other negative health effects have been linked to both short-term and long-term exposure to PM2.5 and PM10.

According to the new EU directive [1], we find the number of days per year with PM10 values greater then 45 μg/m^3 (reccomendation is not to be exceeded more than 18 times per calendar year) and annual mean values (reccomendation is 20 μg/m^3). The data are presented in the Table 1:

Table 1 Number of days and annual mean PM10 values per year

PM10	2020	2021	2022	2023	2024
days >45 μg/m^3	55	23	53	58	46
Annual average	32.57	32.82	32.98	28.59	22.38

Also, according to the new EU directive [1], we find the number of days with PM2.5 values greater then 25 μg/m^3 (not to be exceeded more than 18 times per calendar year) and annual mean values (reccomendation is 10 μg/m^3). The data are presented in the Table 2:

Table 2 Number of days and annual mean PM2.5 values per year

PM2.5	2020	2021	2022	2023	2024
days >25 μg/m^3	65	31	42	88	69
Annual average	23.08	27.06	19.18	23.02	17.66

We can observe from Tables 1 and 2 that there are signific antly more particulate matter particles in the air in Podgorica tha n what the new EU directive recommends.

The most likely source of the air pollution is what we are searching for.

3 CLUSTERING DATA

Clustering is an unsupervised machine learning technique that discovers similar data across rows, grouping such data into the same group. Less similar data are grouped into different groups.

Here, we employ the K-mean and K-medoid clustering techniques. Using a distance metric, they both separate the data into distinct groups. Reducing the sum of squared distances between data points and their respective cluster centroids (mean of the cluster) is the goal of the K-means algorithm. In order to minimize the total absolute distance between data points and their respective cluster medoids (the cluster median), the K-medoid algorithm is used. K-medoid is less sensitive to outliers compared to K-means. [8].

Only PM10 and PM2.5 data will be used after the SO2 data has been analyzed. We also exclude the *year*, *month*, and *day* columns because these clustering techniques are sensitive to categorical data. Column *Date* is olso excluded. Thus, we have two columns in the data frame at the end: PM10 and PM2.5.

For both methods, the number of clusters must be predetermined. An explicit approach for determining the ideal number of clusters does not exist. Here, we use within cluster sums of squares (wss) method [9]. Figure 5 shows the results of utilizing the R fviz_nbclust function for determining and visualizing the optimal number of clusters. We therefore selected three clusters.

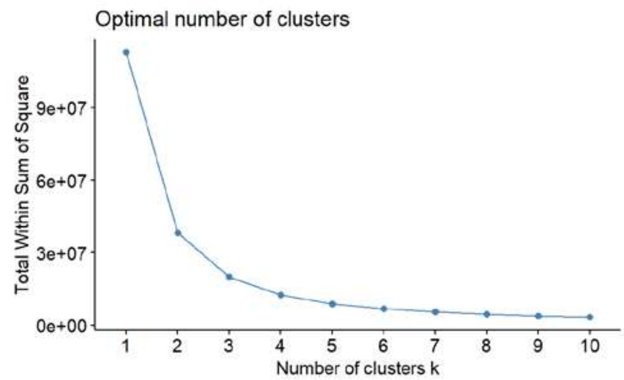

Figure 5 Number of clusters using wss method.

Figure 6 and Figure 7 show K-mean and K-monoid clustering with 3 clusters.

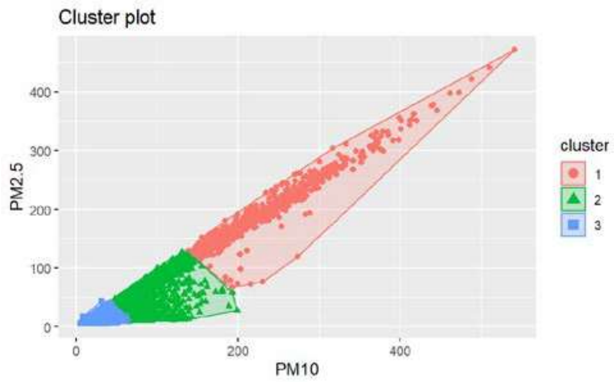

Figure 6 K-mean clusters

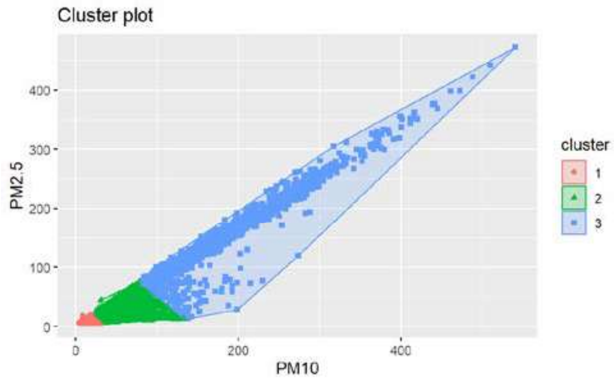

Figure 7 K-medoid clusters

The disparity in clusters is predicted because the median is smaller than the mean, for both PM10 and PM2.5 data. We may infer that there are fewer large values, but those values have a significant influence on air pollution, since the third quartiles of the summary data (Figure 3) are significantly smaller than the highest values.

The six-level yearly PM2.5 and PM10 pollution scales are developed by the WHO. We incorporate two columns that contain categorical data (values ranging from 1 to 6) according to PM2.5 or PM10 readings and the relevant EU regulatory thresholds. The barplots look like what's in Figure 8.

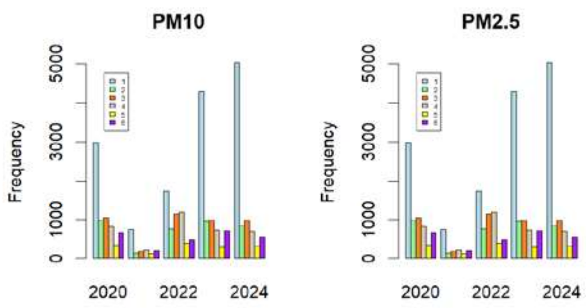

Figure 8 Frequancies of data per year

The most likely cause of the air pollution in Podgorica is what we are now searching for. A significant periodicity may be seen in the higher numbers, according to the visualization of our data in Figure 2. We created Figures 9 and 10 by clustering the PM2.5 and PM10 data independently, then joining the clusters with the monthly data. The K-mean method is used.

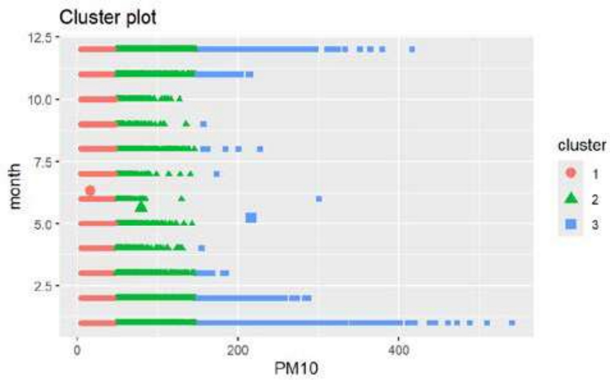

Figure 9 Clustering PM10 data and months

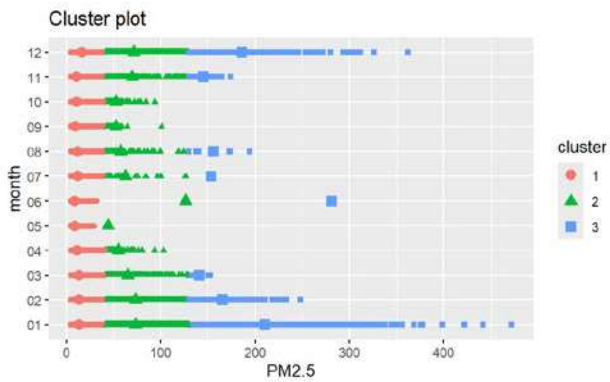

Figure 10 Clustering PM2.5 data and months

From these figures we can see that months with critical, high level, PM10 and PM2.5 data is extreme in the November, December, January and February.

3 CONCLUSION

According to the new EU air pollution directive, two conclusions can be drawn from historical air quality data generated by the Environmental Protection Agency of Montenegro's Podgorica UB monitoring station:

- In Podgorica, the measured SO2 air pollutant is within the advised range for the 2020–2024 timeframe.
- In Podgorica, measured levels of PM2.5 and PM10 air pollutants are higher than the advised threshold for the 2020–2024 timeframe.

Podgorica residents are exposed to PM2.5 and PM10 particulate matter for longer than is advised by a recent EU directive. The contaminants' yearly average levels are also higher than what the directive suggests.

The months of November, December, January, and February have particularly high measured levels of PM10 and PM2.5 particulate matter. This city's heating season is represented by these months. All year long, the measured SO2 levels fall within the allowed ranges. Thus, the primary cause of air pollution in Podgorica can be attributed to heating fuels.

4 COPYRIGHT TRANSFER AGREEMENT

5 REFERENCES

[1] Directive (EU) 2024/2881 of the European Parliament and of the Council of 23 October 2024 on ambient air quality and cleaner air for Europe (recast)

[2] See https://environment.ec.europa.eu/news/new-pollution-rules-come-effect-cleaner-air-2030-2024-12-10_en

[3] See https://www.who.int/publications/i/item/9789240034228

[4] See https://www.who.int/teams/environment-climate-change-and-health/air-quality-energy-and-health/health-impacts

[5] See https://www.epa.org.me/vazduh/

[6] See https://www.epa.gov/so2-pollution/sulfur-dioxide-basics.

[7] See https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health

[8] See https://rpubs.com/folwalsh/UL

[9] See https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm

Contact information:

Biljana Stamatovic, full professor
(Corresponding author)
F-Mail: biljana.stamatovic@gmail.com
https://orcid.org/0000-0002-3934-9686

Sinisa Stamatovic, full professor
G-Mail: sstamatovic@ucg.ac.me
https://orcid.org/0000-0002-7013-3154