

# Actionable Data Visualization for Air Quality Data in the Istanbul Location


Damla Mengüş and Bihter Daş

**Abstract**— Air pollution is increasing day by day due to the increasing population, urbanization, and industrial development. In our country, the amounts of pollutants in the air are recorded every day at different points. These recorded data continue to be collected in an increasing amount day by day. Information overload, which renders the data meaningless, complicates the interpretation of these data. One of the ways to solve this problem is to visualize curves and trends in measured pollution concentrations over time. In this study, using the data provided by the continuous monitoring center of the Turkey Ministry of Environment, Urbanization and Climate Change, visualization of different pollutants in the air was provided. Scatter plots, line scatter plots, and bar plots were used as data visualization. Data visualization makes it easy for non-experts to estimate air quality information from the concentration profiles displayed.

**Index Terms**— Data visualization, air pollution, air quality assessment, visual analytics

## I. INTRODUCTION

**AIR POLLUTION** is defined as the presence of foreign substances above normal, which should not be present in the air, which adversely affect human health and environmental balance. There are many factors to air pollution, especially the increasing population, urbanization, and industrial developments [1-3]. Carbon monoxide (CO), nitrogen dioxide (NO<sub>2</sub>), sulfur dioxide (SO<sub>2</sub>), ozone (O<sub>3</sub>), particulate matter (PM), and NO<sub>x</sub>, which is a combination of nitric oxide and nitrogen dioxide, are the leading gases that cause air pollution[4-6]. Particulate matter PM<sub>10</sub> and PM<sub>2.5</sub> are one of the most important pollutants affecting human health. Since it is very small in size, it passes through our respiratory system very easily and penetrates our lungs. May cause cancer if inhaled for a long time. As other pollutants can easily penetrate the lungs, all of them are very dangerous for human health [7-9]. Evaluation of air quality is made according to the air quality index. This index is a measure used to express air quality.

 **Damla MENGÜŞ** is with the Department of Computer Engineering, Technology Faculty, Marmara University, İstanbul TÜRKİYE e-mail: [damla.mengus@gmail.com](mailto:damla.mengus@gmail.com)

 **Bihter DAŞ** is with the department of Software Engineering, Technology Faculty, Firat University, Elazığ TÜRKİYE e-mail: [bihterdas@firat.edu.tr](mailto:bihterdas@firat.edu.tr)

As the measured value gets larger, it is understood that the air starts to have negative effects on human health. Air quality index 0-50 range is good, 51-100 range is medium, 101-150 range is sensitive, 151-200 range is unhealthy, 201-300 range is bad, 301-500 range is dangerous. There is a high probability of experiencing health problems with a value of 151 and above, and it is not necessary to go to the open area during these times [10,11].

In this study, data visualization is carried out to observe the air quality by using the air pollutant data of the Basaksehir district of Istanbul city. There are continuous monitoring centers (SIMs) at 39 different points in Istanbul. It is very suitable for monitoring the air quality in different parts of the city and making inferences from the data. The differences in the districts of Istanbul, located in the middle or by the sea, provide better observation with visualization tools. Figure 1 shows the location of the Başakşehir district.



Fig.1. The location of the Basaksehir

### A. Contributions of the paper

The main contributions of the study are listed below.

- 1) Air quality indices (AQI) and environmental data are combined with data visualization.
- 2) A practical experiment on which visualization tools would be appropriate for what type of data is provided.
- 3) Visualization methods help non-experts to interpret big data.

The rest of the article is organized as follows. The methods used are mentioned in Chapter 2. In Chapter 3, the application steps and the data used are mentioned. In Chapter 4, the results of the study are mentioned. In section 5, inferences are made.

## B. Related works

There are many areas in the literature where data visualization is used. In most places with big data, visualization is used to understand the data [12,13]. Bachechi et al. used information visualization techniques to analyze urban traffic data and the effect of traffic emissions on urban air quality. Traffic data statistics for months or years provided a clear understanding of the similarities and differences between days [14]. Von Bromssen et al. studied the acidification of the river by using Swedish riverbank data during 1988–2017. They have tried to summarize complex information over nearly thirty years of data and have used data visualization while doing this [15]. Huang et al. Shenzhen, a mega-city in China, has made efforts to promote the transition to green transport by enforcing license plate restrictions. However, it is unclear whether the restrictions improve urban air quality. They have studied the effect of diesel vehicles on air pollution. Thanks to the data visualization tools used in the study, the result was easily achieved [16]. Pérez-Campuzano et al. took 30 years of data from 18 different US passenger airlines and made visualization on these data in their study. As a result of the study, it was revealed how much the passenger airline of the USA was affected during the Covid-19 period [17]. In the study of Prasad et al., existing data visualization methods have been enhanced by spectral modeling to overcome the problem of cluster bias on non-CS datasets, which efficiently recognizes the spectral features of non-CS datasets and cluster patterns [18]. New visualization techniques are used not only for environmental data, but also for other types of data such as satellite information, X-ray spectra processing or big data [19–22]. For this purpose, a software platform was developed with an integration in the form of two measurement stations and satellite information in a unified view to process publicly available data from various sources. In [23], authors superposed health risk information from 9 different Air Quality Indices (AQIs) on different kinds of graphs. They visualized the data, which is obtained from two monitoring stations located in regions, Ghent and Vielsalm in Belgian Environment Agency.

## II. BACKGROUND

In this part of the study, general information about data visualization methods and incomplete data completion techniques is given. The process of making large and complex data meaningful and understandable with certain graphics is called data visualization. Three different types of charts were used in this study:

**Point scatter plot:** Point scatter plot is a data visualization method created using two different numerical data. It allows us to directly see the connection between these two numerical values. We can visualize not only two values, but many values at the same time through different colors or sizes. Thus, complex data becomes more understandable thanks to visualization.

**Line scatter plot:** Line scatter plot gives similar output when used in the same way as point scatter plot. However, the line scatter plot is easier to read than the point scatter plot in some cases. Breakpoints that are not clearly visible in the point scatter plot can be observed better with the line scatter plot. For this reason, data visualization with line scatter plot may be more advantageous depending on the usage area.

**Bar plot:** It provides the opportunity to visualize data in categories with bar plot. Data can be displayed in multiple groups in a single image. This provides an advantage for the bar plot.

One of the other problems when working with data is the problem of missing data. Accurate inference or visualization cannot be made due to missing data. In such cases, there are various algorithms that can be used to complete the missing data. Two different methods were used for this study. Based on the results, it was decided that the most appropriate method for this study was to complete the missing data with the mean technique.

**Complete missing data with k-nn technique:** K-nearest neighbor (KNN) is a kind of algorithm used for classification and regression in supervised learning. Training and testing are pretty much the same. It is not an ideal algorithm to complete missing values in large datasets.

**Complete missing data with mean technique:** In this method, the missing data part is filled by taking the average of the other data in the area where the missing data is located. Data range is very important for using this method. Using this method in data sets with very high data range increases the error rate.

## III. METHODOLOGY AND IMPLEMENTATION

In this study, firstly annual data were collected and then missing data was completed with k-nn and average algorithms. Then, different visualization methods were applied to the data. Point scatter plot, line scatter plot and bar plot were used on the data, respectively. The mean technique was used as the missing data completion algorithm in the study because the data range is low, using mean in such data provides better performance. The flowchart showing the application steps is shown in Figure 2.

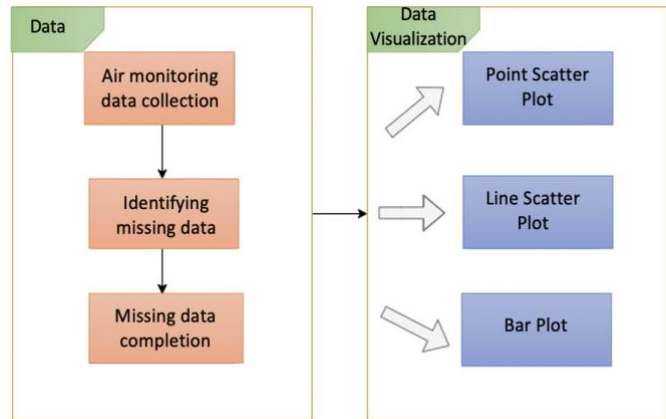


Fig. 2. The flowchart of the study

### A. Data description

In this study, the data provided by the continuous monitoring center of the Turkey Ministry of Environment, Urbanization and Climate Change were used [24]. It consists of data on different pollutants recorded daily or hourly. In our study, visualization was made on the data of six pollutants for one year. These inhibitors are NO<sub>x</sub>, which is a combination of carbon monoxide (CO), nitrogen dioxide (NO<sub>2</sub>), sulfur dioxide (SO<sub>2</sub>), ozone (O<sub>3</sub>), particulate matter (PM), nitric oxide and nitrogen dioxide. Table 1 shows the first six rows of the data set.

Table 1. First six rows of data set

Date	İstanbul- Basaksehir					
	PM10 (µg/m3)	SO2 (µg/m3)	NO2 (µg/m3)	O3 (µg/m3)	CO (µg/m3)	NOX (µg/m3)
05.09.2022 00:00:56	11,48	2,66	26,35	45,15	751,43	5,22
06.09.2022 00:00:56	16,21	2,73	26,37	44,00	757,58	5,58
07.09.2022 00:00:56	18,33	2,86	26,77	45,61	744,61	5,68
08.09.2022 00:00:56	18,28	2,31	27,05	42,35	789,02	5,73
09.09.2022 00:00:56	19,40	2,55	28,71	41,63	746,33	6,11
10.09.2022 00:00:56	21,30	2,08	22,28	41,63	739,98	6,23

### B. Data preprocessing

In this study, the average method was used while filling the missing values of the data set consisting of one-year data. The reason for choosing the incomplete data completion algorithm with the mean is that the data set is large and the data range is small.

C. Actionable data visualization

The most important reason for choosing point scatter, line scatter and bar plot as visualization methods in this study is that the data set is independent of each other. While simpler visualization tools are used in independent and unrelated data sets, different visualization methods are used in linked related data.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The visualizations that emerged as a result of the study are given below, respectively. In Figure 3, it is seen that the PM10 value is not very high throughout the year and is high on very rare dates. In particular, it is clearly seen that it reached the highest point in April 2022.

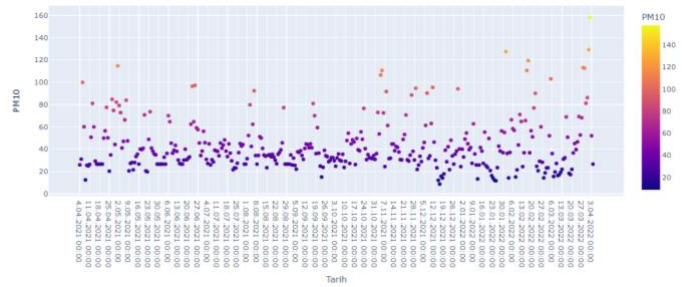


Fig. 3. Point Scatter for PM10

In Figure 4, it is seen that the SO2 value is much higher, especially in April and May, compared to the rest of the year. It can be said that the annual sulfur dioxide value is higher than normal.

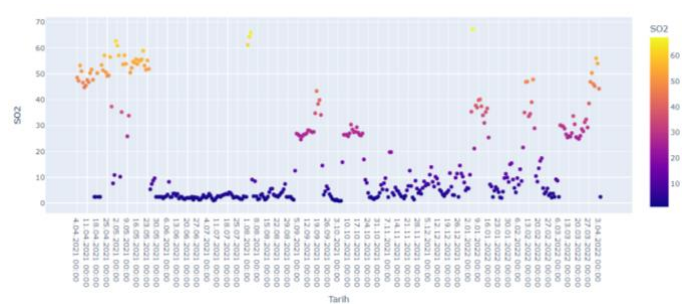


Fig. 4. Point Scatter for SO2

In Figure 5, annual NO2 values are more evenly distributed. High values can be easily seen.

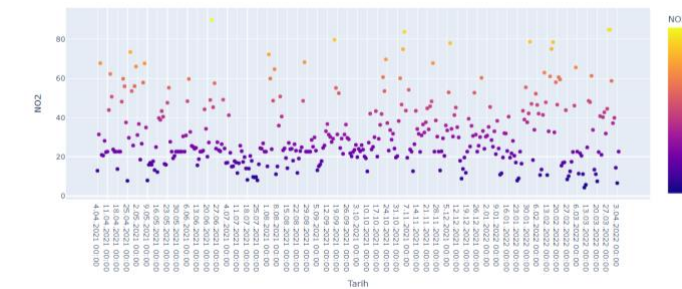


Fig. 5. Point Scatter for NO2

Figure 6 shows a more balanced distribution for the NOX value. It is observed that the annual values are normal and very slightly increase to high values.

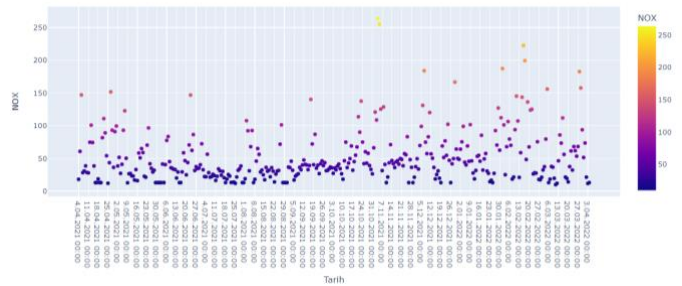


Fig. 6. Point Scatter for NOX

In Figure 7, very high O3 values were observed in May, July and August. Even looking at the annual value in general, it is mostly seen to be high.

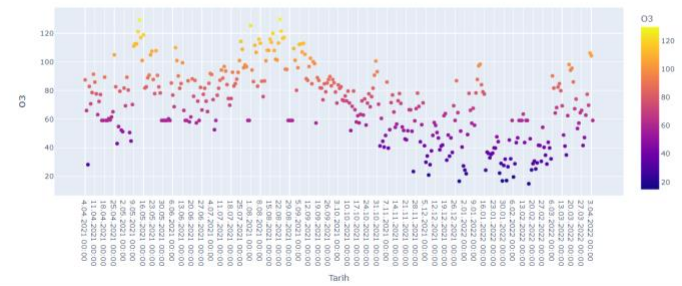


Fig. 7. Point Scatter for O3

The visualizations for the line scatter on the same data set are given below, respectively. In the Figure 8-12, the peaks are more prominent. The highest and lowest values can be easily observed in these figures. In Figure 8, unlike the point scatter, the highest PM10 value was shown in March. The upper values are more prominent in the image.

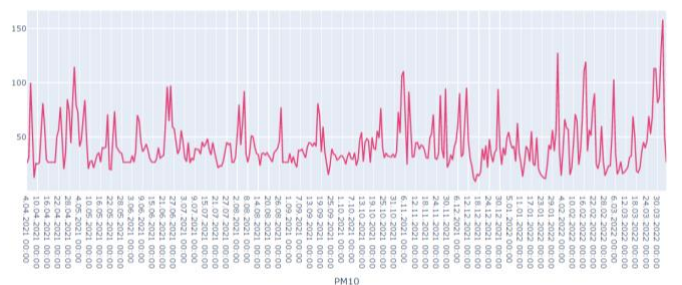


Fig. 8. Line Scatter for PM10

Figure 9 shows that the values are distributed unevenly. High and low values are easily visible.

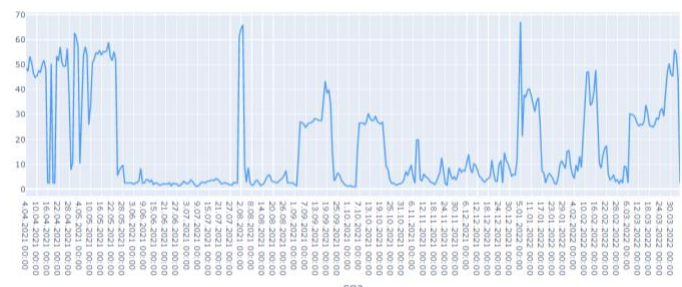


Fig. 9. Line Scatter for SO2

Figure 10 shows that the NO2 value is evenly distributed. High and low values are evident.



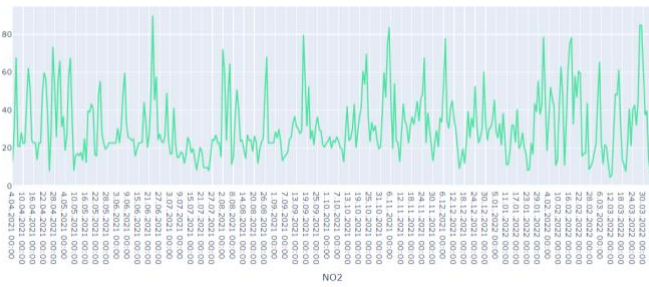


Fig. 10. Line Scatter for NO2

Figure 11 shows a balanced distribution in annual values. The highest value belongs to November.

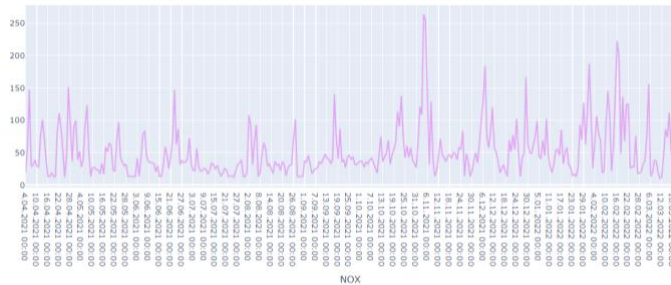


Fig. 11. Line Scatter for NOX

The annual data for Figure 12 are shown below. Annually the values are mostly high.

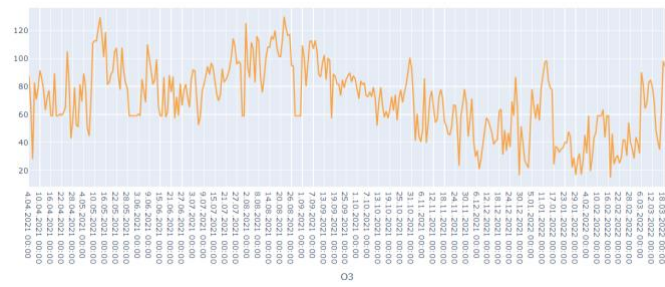


Fig. 12. Line Scatter for O3

The coloring process used for the bar plot is important in terms of both understanding the value of the data and seeing the peak values. Again, the images created on the same data set are shown in Figures 13-17 respectively. Figure 13 shows the distribution of annual PM10 values. High and low values are clearly visible and its coloring helps a lot.



Fig. 13. Bar Plot for PM10

An uneven distribution is seen in Figure 14. High and low values are clearly visible thanks to both colors and the use of bar plots.

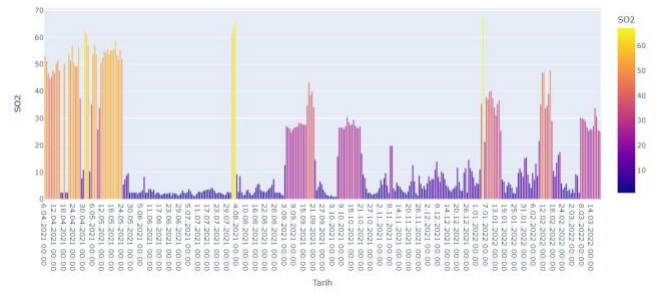


Fig. 14. Bar Plot for SO2

Annual NO2 values are shown in Figure 15. It is seen that there is a balanced distribution.

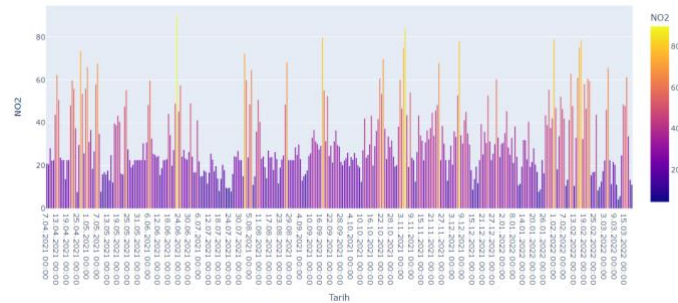


Fig. 15. Bar Plot for NO2

Figure 16 shows NOX values. High values are clearly visible. In general, it is seen to be at normal annual levels. The highest value belongs to November.

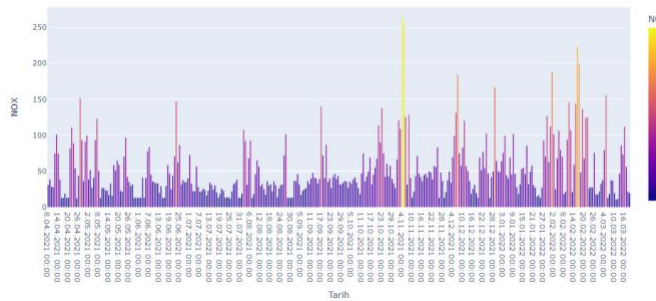


Fig. 16. Bar Plot for NOX

In Figure 17, it is seen that the annual O3 values are generally high.

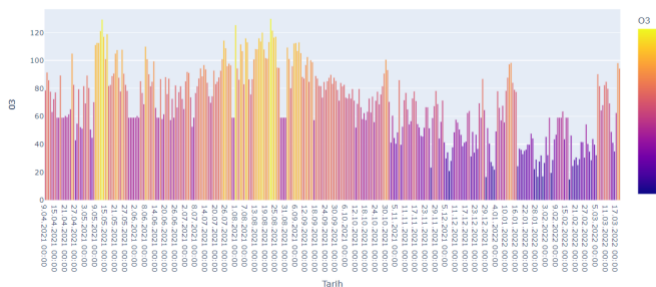


Fig. 17. Bar Plot for O3

Point scatter helped to show the distribution of air pollutants more clearly than other graphs. However, the line scatter allowed us to see the peak values more clearly. Like the point scatter plot, the bar plot allowed us to see the distribution more clearly as a result of coloring. In addition, the peaks can be easily seen in the bar plot.

## V. CONCLUSION

In this study, trends in measured air pollution concentrations are visualized to make sense of large amounts of air data. One-year air pollutants data provided by the continuous monitoring center of the Turkish Ministry of Environment, Urbanization and Climate Change were obtained, and missing data were completed with the mean algorithm. Then, using data visualization methods such as point scatter plot, line scatter plot and bar plot, it was ensured that large amounts of data on environmental parameters were understandable by different stakeholders. Here, the excess and confused data has become more understandable with visualization. In future studies, it is planned to work on longer-term data, to use related data, to use different data visualizations suitable for these data types, to conduct a study that reveals the values of pollutants and what environmental effects are. Thus, researchers will be shown which visualization methods will have more appropriate use on which data types. Although we can see the distribution and minimum-maximum values with these three visualization methods used in our study, we could not get much detail about the data. This is a limiting feature of our study.

## REFERENCES

- [1] B. Li, Z. Qiu, J. Zheng. "Impacts of noise barriers on near-viaduct air quality in a city: a case study in Xi'an". *Build. Environ.*, 196 (2021), Article 107751, [10.1016/j.buildenv.2021.107751](https://doi.org/10.1016/j.buildenv.2021.107751)
- [2] K.F. Lu, H.D. He, H.W. Wang, X.B. Li, Z.R. Peng Characterizing temporal and vertical distribution patterns of traffic-emitted pollutants near an elevated expressway in urban residential areas *Build. Environ.*, 172 (2020), Article 106678, [10.1016/j.buildenv.2020.106678](https://doi.org/10.1016/j.buildenv.2020.106678)
- [3] H.D. He, H.O. Gao Particulate matter exposure at a densely populated urban traffic intersection and crosswalk *Environ. Pollut.*, 268 (2021), Article 115931, [10.1016/j.envpol.2020.115931](https://doi.org/10.1016/j.envpol.2020.115931)
- [4] A. Lak, M. Ramezani, R. Aghamolae Reviving the lost spaces under urban highways and bridges: an empirical study *J. Place Manag. Dev.*, 12 (2019), pp. 469-484, [10.1108/JPM-12-2018-0101](https://doi.org/10.1108/JPM-12-2018-0101)
- [5] A. Sharma, D.D. Massey, A. Taneja A study of horizontal distribution pattern of particulate and gaseous pollutants based on ambient monitoring near a busy highway *Urban Clim.*, 24 (2018), pp. 643-656, [10.1016/j.uclim.2017.08.003](https://doi.org/10.1016/j.uclim.2017.08.003)
- [6] K.F. Lu, H.D. He, H.W. Wang, X.B. Li, Z.R. Peng Characterizing temporal and vertical distribution patterns of traffic-emitted pollutants near an elevated expressway in urban residential areas *Build. Environ.*, 172 (2020), [10.1016/j.buildenv.2020.106678](https://doi.org/10.1016/j.buildenv.2020.106678)
- [7] B. Das, Ö. O. Dursun, and S. Toraman, "Prediction of air pollutants for air quality using deep learning methods in a metropolitan city," *Urban Climate*, (2022), vol. 46, p. 101291, , doi: [10.1016/j.uclim.2022.101291](https://doi.org/10.1016/j.uclim.2022.101291).
- [8] C. Wu, H. He, R. Song, Z. Peng Prediction of air pollutants on roadside of the elevated roads with combination of pollutants periodicity and deep learning method *Build. Environ.*, 207 (2022), Article 107436, [10.1016/j.buildenv.2021.108436](https://doi.org/10.1016/j.buildenv.2021.108436)
- [9] G. Kurnaz, A.S. Demir Prediction of SO<sub>2</sub> and PM<sub>10</sub> air pollutants using a deep learning-based recurrent neural network: case of industrial city Sakarya *Urban Clim* 41 (2021), Article 101051, [10.1016/j.uclim.2021.101051](https://doi.org/10.1016/j.uclim.2021.101051)
- [10] P. Perez, C. Menares, C. Ramirez PM<sub>2.5</sub> forecasting in Coyhaique, the most polluted city in the Americas *Urban Clim.*, 32 (2020), p. 100608, [10.1016/j.uclim.2020.100608](https://doi.org/10.1016/j.uclim.2020.100608)
- [11] A. Aggarwal, D. Toshniwal A hybrid deep learning framework for urban air quality forecasting, *J. Clean. Prod.*, 329 (2021), Article 129660, [10.1016/j.jclepro.2021.129660](https://doi.org/10.1016/j.jclepro.2021.129660)
- [12] M.Sülü, R. Daş, "Graph visualization of cyber threat intelligence data for analysis of cyber attacks", *Balkan Journal of Electrical and Computer Engineering (BAJECE)*, (2022),10(3), 300-306.
- [13] M.Sülü, R. Daş, "QR Algoritması Kullanarak Spektral Çizge Bölümleme", *Firat Üniversitesi, Fen Bilimleri Dergisi*, (2022), vol. 34, no. 2, pp. 207-218.
- [14] C. Bachechi, L. Po, and F. Rollo, "Big Data Analytics and Visualization in Traffic Monitoring," *Big Data Research*, (2022), vol. 27, p. 100292, doi: [10.1016/j.bdr.2021.100292](https://doi.org/10.1016/j.bdr.2021.100292).
- [15] C. von Brömssen, S. Betnér, J. Fölster, and K. Eklöf, "A toolbox for visualizing trends in large-scale environmental data," *Environmental*

*Modelling & Software*, 2021, vol. 136, p. 104949, doi: [10.1016/j.envsoft.2020.104949](https://doi.org/10.1016/j.envsoft.2020.104949).

- [16] W. Huang, X. Xu, M. Hu, and W. Huang, "A license plate recognition data to estimate and visualise the restriction policy for diesel vehicles on urban air quality: A case study of Shenzhen," *Journal of Cleaner Production*, 2022, vol. 338, p. 130401, doi: [10.1016/j.jclepro.2022.130401](https://doi.org/10.1016/j.jclepro.2022.130401).
- [17] G. Carro, O. Schalm, W. Jacobs, and S. Demeyer, "Exploring actionable visualizations for environmental data: Air quality assessment of two Belgian locations," *Environmental Modelling & Software*, 2022, vol. 147, p. 105230, doi: [10.1016/j.envsoft.2021.105230](https://doi.org/10.1016/j.envsoft.2021.105230).
- [18] D. Pérez-Campuzano, L. Rubio Andrada, P. Morcillo Ortega, and A. López-Lázaro, "Visualizing the historical COVID-19 shock in the US airline industry: A Data Mining approach for dynamic market surveillance," *Journal of Air Transport Management*, 2022, vol. 101, p. 102194, doi: [10.1016/j.jairtraman.2022.102194](https://doi.org/10.1016/j.jairtraman.2022.102194).
- [19] K. R. Prasad, G. R. Kamatam, M. B. Myneni, and N. R. Reddy, "A novel data visualization method for the effective assessment of cluster tendency through the dark blocks image pattern analysis," *Microprocessors and Microsystems*, 2022, vol.93, 104625, doi: [10.1016/j.micpro.2022.104625](https://doi.org/10.1016/j.micpro.2022.104625).
- [20] A. Eldawy, M. Mokbel, A. Alharthi, A. Azaidy, K. Tarek, S. Ghani SHAHED: a MapReduce-based system for querying and visualizing spatio-temporal satellite data *IEEE 31st International Conference on Data Engineering (2015)*, pp. 1585-1596, [10.1109/ICDE.2015.7113427](https://doi.org/10.1109/ICDE.2015.7113427)
- [21] G. Van Snickt, S. Legrand, J. Caen, F. Vanmeert, M. Alfeld, K. Jansses Chemical imaging of stained-glass windows by means of macro X-ray fluorescence (MA-XRF) scanning. *Microchem. J.* (2016), pp. 615-622
- [22] A. Syed, N. Gupta, G. Nayak, R. Lenka Big Data Visualization: Tools and Challenges *IEEE 2nd Int. Conference on Contemporary Computing and Informatics (2016)*, pp. 656-660, doi: [10.1109/IC3I.2016.7918044](https://doi.org/10.1109/IC3I.2016.7918044)
- [23] G. Carro, O. Schalm, W. Jacobs, and S. Demeyer, "Exploring actionable visualizations for environmental data: Air quality assessment of two Belgian locations," *Environmental Modelling & Software*, 2022, vol. 147, p. 105230, doi: [10.1016/j.envsoft.2021.105230](https://doi.org/10.1016/j.envsoft.2021.105230).
- [24] SIM Air Quality- Station Data Download Continuous Monitoring Center. [https://sim.csb.gov.tr/STN/STN\\_Report/StationDataDownloadNew](https://sim.csb.gov.tr/STN/STN_Report/StationDataDownloadNew) (202 0) (accessed 17 June 2022)

## BIOGRAPHIES



**DAMLA MENGÜŞ** is currently studying in his Bachelor's degree in the Department of Software Engineering, Technology Faculty, at the Firat University. She works at the Department of Computer Engineering, Marmara University. Her current research areas include data visualization, machine learning, data science, and artificial intelligence.



**BİHTER DAŞ** graduated B.S. and M.S. degrees from the Department of Computer Science at the Firat University in 2004 and 2007 respectively. Then she received Ph.D. degree at the Department of Software Engineering at the same university in 2018. She also worked between September 2017 and June 2018 as a visiting scholar at the Department of Computing Science at the University of Alberta, Edmonton, Canada. Her current research areas include data science, big data, data analytics, bioinformatic, digital signal processing, genome data analysis.