sciendo INTERNATIONAL JOURNAL ON SMART SENSING AND INTELLIGENT SYSTEMS

Research article | DOI: 10.2478/ijssis-2024-0004 Issue 1 | Vol. 17 (2024)

Comparison of outlier detection approaches in a Smart Cities sensor data context

Sofia Zafeirelli*, Dimitris Kavroudakis*

Department of Geography, University of the Aegean, 81100 Mytilene, Greece

*Emails: s.zafeirelli@aegean.gr; dimitrisk@aegean.gr

Received for publication: September 06, 2023

Abstract

This study examines outlier detection in time-series sensor data from PurpleAir low-cost sensors in Athens, Greece. Focusing on key environmental parameters such as temperature, humidity, and particulate matter (PM) levels, the study utilizes the Interquartile Range (IQR) and Generalized Extreme Studentized Deviate (GESD) methods on hourly and daily basis. GESD detected more outliers than IQR, most of them in PM, while temperature and humidity data had fewer outliers; applying filters before outlier detection and adjusting alpha values based on time scales were crucial, and outliers significantly affected spatial interpolation, emphasizing the need for spatial statistics in smart city air quality management.

Keywords

Outlier detection, smart cities, IoT, sensors, environmental data

I. Introduction

"Smart cities" is a relatively novel approach to city planning and management that aims to improve their residents' life quality, reduce cities' environmental footprint, and support decision-making using advanced technologies, transforming all these into data-driven processes (Loo et al., 2019; Yigitcanlar et al., 2022). Such technologies consist of sensors and smart devices connected to the Internet that create networks called the Internet of Things (IoT) or gather data individually. They record large volumes of heterogeneous data (typically called "Big Data"), requiring modeling and analytics to understand city activity (Lovén et al., 2019), as there is an exigent need for such analysis methods to effectively handle the operations of smart cities (Sayeed et al., 2023; Tahmasseby, 2022; Amini et al., 2019). IoT can be considered as the extension and expansion of the Internet, affecting people's daily lives, but the quality (and therefore the use value) of data can be degraded by missing and abnormal values (Ding et al., 2020).

Smart city sensor data is often captured in time-series form. A time series is a sequence of measurements recorded over time with properties such as Trend, Seasonality, Cycles, Level, Stationarity, and White noise (Braei & Wagner, 2020). Raw IoT sensor data often presents noise, missing values, and outliers—anomalies. There is no consensus on outlier and anomaly definitions (Braei & Wagner, 2020). Blázquez-García et al. (2021) demarcate outliers from anomalies, defining outliers as unwanted data and anomalies as events of interest. Ayadi et al. (2017) define outliers as sensor readings that significantly deviate from ground truth data and don't follow the normal pattern of data. In Samara et al. (2022), outliers are defined as observations that deviate significantly from normal data.

Outlier sources in IoT data usually occur due to (1) noise and errors, (2) events, and (3) malicious attacks (Ayadi et al., 2017). According to Elmenshawy & Helmy (2018), Sharma et al. (2019), and Samara et al. (2022), outliers can be categorized into three groups: (1) Point outliers, (2) Collective outliers, and (3) Contextual outliers, and in the first two it is referred as anomalies and in the third as outliers.

Open Access. Published by Sciendo. CC BY-NC-ND
© 2024 Zafeirelli and Kavroudakis. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License https://creativecommons.org/licenses/by-nc-nd/4.0/

The detection of outliers is important in data preprocessing for data analysis, decision-making (Krishnamurthi et al., 2020), time-series analysis (Ogasawara et al., 2010), and IoT systems (Alvear-Puertas et al., 2022). Outlier detection plays a crucial role in spatial interpolation as outliers can significantly affect the accuracy and reliability of the interpolation results. Spatial interpolation techniques aim to estimate values at unmeasured locations based on the values observed at nearby sampled locations. Outliers can distort the interpolation process and lead to erroneous predictions. Detecting them leads to improving accuracy, preserving spatial continuity, assessing data quality, and facilitating informed decision-making. By identifying and appropriately handling outliers, more reliable and meaningful interpolated surfaces can be generated for various applications.

There is a plethora of literature regarding time-series outlier detection techniques. In Samara et al. (2022) seven of them are analyzed: statistical, clustering, nearest neighbors, classification, artificial intelligence, spectral decomposition, and hybrid based. Braei & Wagner (2020) classify outlier-anomaly detection on time-series data into three classes namely statistical, classical machine learning, and deep learning methods. Cieplak et al. (2019) categorized the techniques into four groups: (1) statistical approaches, (2) distance and density-based, (3) profiling based, and (4) model-based.

Pastorio et al. (2022) evaluated and calibrated five time-series data particulate matter (PM) sensors, in order to use them in the smart cities' context. In data processing, they removed outliers with the Interquartile Range (IQR) method. Also in the context of air quality monitoring, Schilt et al. (2023) performed outlier detection techniques and applied filters on low-cost sensor environmental data. In 2017, Csaji et al. applied statistical data analysis on smart city sensor data, and during data pre-processing, they filtered outliers using the Hampel filter (Liu et al., 2004). An unsupervised anomaly detection method was developed by Liu et al. (2020) on IoT real-time temperature data. A deep-learning method was proposed by Elbaz et al. (2023) for spatiotemporal air quality forecasting in a smart city context. More specifically, the method was conducted on hourly and daily data, including nine pollutants and meteorological factors, and during data preprocessing, they identified outliers.

Aslan et al. (2022) use the DBSCAN algorithm to find outliers and characterize them as noise and extreme values, capitalizing on its independence from statistical assumptions and limited prior use. They detect irregularities in hourly PM time series and examine factors behind increased PM levels, acknowledging context-specific results due to dataset structure, thus enhancing the field. Chen et al. (2018) propose an anomaly detection framework (ADF) for environmental sensing systems, identifying outliers and perceptible anomalies. Their framework informs data visualization services and assists Environmental Protection Agency (EPA) authorities in policy formulation, using real data from AirBox, a PM2.5 monitoring system. In Liang and Yu (2021), Air quality from monitoring stations (AQMS) spatial interpolation data is used to assess two sensor systems' performances by estimating their linearity, sensitivity, offset, precision, accuracy, and bias. The proposed enhancement platform involves monitoring and automatic correction loops, including outlier detection, temporal and spatial anomaly analysis, and a trajectory analysis module, aiming to enhance low-cost sensor system performance, particularly for alerting reportable events.

Aix et al. (2023) introduced a method that combines a gas-pollutant approach with dust event preprocessing to calibrate urban PM low-cost sensors (PMS7003). The developed protocol encompasses outlier selection, model tuning, and error estimation, facilitating comprehensive data analysis and calibration through Multiple Linear Regression (MLR) and Random Forest Regression (RFR) regressions. Ma et al. (2017) introduced an outlier detection algorithm for temperature time series in meteorological sensor networks. It employs sliding window prediction to forecast data and identifies outliers by comparing predicted and actual values. Experimental results affirm its efficacy in improving temperature time-series analysis. Another approach is presented by Esnaola-Gonzalez et al. (2017) who introduced the semantic outlier detection (SemOD) Framework which uses Semantic Technologies to contribute to Outlier Detection in Wireless Sensor Network (WSNs). It employs ontology-based reasoning to identify factors impacting sensor measurements and offers methods to detect context-based outliers. The approach reveals potential outlier sources, enhancing decision-making even in complex contexts where traditional methods are insufficient.

Stavroulas et al. (2020) conducted a field evaluation of PurpleAir sensors in Athens and Ioannina, two Greek cities, on different seasons. They tested their performance on hourly PM 2.5 concentrations in relation to reference PM 2.5 instruments, high particle concentration events, and meteorological conditions. Also, inter-site correlations and intra-urban homogeneity were assessed. Finally, a nearest

neighbor interpolation was performed during an event of a forest fire, but they did not go deep into it.

In this paper, we investigate the issue of outlier management with three research questions: (a) how can we identify outlier values in a real-time series dataset such as the PurpleAir dataset, (b) what management approaches should be considered when detecting outliers on environmental data in a smart city, and (c) how can outliers affect spatial interpolation. The overall purpose of this work is to detect outliers efficiently and effectively in real-time-series datasets that present challenges as malfunctions and pauses occur. Ayadi et al. (2017) outlier definition is followed and two outlier detection statistical techniques on environmental spatial time-series sensor data are performed to answer these questions.

In this work, two temporal scales of data are used, which are based on hourly and daily averages. It is the first long-term sensor data analysis for this area (Athens, Greece). The two approaches are compared and discussed for their implementation in a smart city context, especially when there is a need to identify problematic or misleading data values due to malfunctions. They were chosen due to their fast and simple methodology, and they can easily detect outliers and can be adapted to real-time data outlier detection.

II. Methods

The study was designed to assess and improve the accuracy of air quality data obtained from PurpleAir low-cost sensors in Athens, Greece. It addressed accuracy and calibration challenges through a comprehensive analysis that included outlier detection, data decomposition, and geostatistical assessment. PurpleAir low-cost sensors are used by individuals, organizations, and communities to monitor air quality in real-time, especially PM. The sensors have gained popularity for their affordability and ease of use, but one of their primary drawbacks is their accuracy and calibration. While they provide valuable real-time air quality data, their precision may not be as reliable as professional-grade monitoring equipment. Calibration can drift over time, leading to potential inaccuracies in the readings. Factors such as dust, humidity, and temperature can also affect the sensor's performance, resulting in fluctuations or inconsistencies in the measurements. The study was carried out using a systematic approach that involved data collection, data preprocessing, outlier detection, data decomposition, geostatistical analysis, and reporting of results.

The data were analyzed for outlier detection of the PurpleAir time-series datasets using two methods: the IQR and the Generalized Extreme Studentized Deviate (GESD). Their results were compared for both hourly and daily data. All analysis was conducted on R statistical software and the two methods used were included in the "*anomalize*" R package, which helps toward detecting outliers in numerical data.

The IQR method of extreme values initially divides the values of a dataset into four quartiles (Q1: 0%–25%, Q2: 25%–50%, Q3: 50%–75%, Q4:75%–100%), using the median value and then calculates the difference between the first and the third quartile, IQR = Q3 − Q1. Q1 is the first quartile, Q2 is the median, and Q3 is the third quartile, which means Q1 is the median of lower half values and Q3 is the median of upper half values. Additionally, there is an extra calculation of lower and upper bounds to define the outliers. Observations less than Q1 − 3 × IQR and greater than Q3 + 3 × IQR are considered outliers, which means that observations three or more times above and three or more times below the IQR are considered outliers. This definition was proposed by Tukey (1977, p. 44), who introduced it as a boxplot with a factor of 1.5 (1.5 times around IQR instead of 3). For alpha = 0.05, the factor is adjusted as 3X and alpha = 0.10 results in a 1.5X factor. This is the first method of defining outlier observations of the environmental data.

Another method is GESD which was first introduced by Rosner (1983). GESD can detect outliers in a univariate dataset that follows an approximately normal distribution. First, the null and the alternative hypotheses are tested as follows:

H_0 = Data has no outliers

H_1 = Data has up to r outliers.

There are performed r tests for 1 to r outliers. Tests compute

$$R_i = \frac{\max_i |x_i - \underline{x}|}{s} \tag{1}$$

where $\underline{x}$ is the mean and s is the standard deviation of the dataset.

R_i is calculated by removing the value maximizing the difference $|x_i - \underline{x}|$ and then repeating the operation for n –1 observations, recalculating mean and standard deviation. The process is repeated until the r observations are removed. Next, λ_i critical values are computed for a specified significance level α:

$$\lambda_i \frac{(n-i)t_{p,n-i-1}}{\sqrt{\left(n-i-1+t^2_{p,n-i-1}\right)(n-i+1)}},\ i = 1,2,\ldots,r \tag{2}$$

where $t_{p,v}$ is the 100p percentage point from t distribution with v degrees of freedom

$$\text{and } p = 1 - \frac{\alpha}{2(n-i+1)}.$$

The outlier total is known when ***i*** is great enough such that $R_i > \lambda_i$. The Alpha parameter adjusts the critical value width. H_0 is rejected once $R_i > \lambda_i$, so when the test statistic drops below the critical value the outliers are removed. GESD method calculations are somehow relatively slower than the IQR calculations as GESD requires more calculations.

The analysis was performed in the open software and open source RStudio platform, which uses R programming language. Outliers were detected using "*anomalize*" R package (Dancho & Vaughan, 2022). This package works with the two aforementioned outlier detection methods, IQR and GESD, allowing users to change some parameters e.g., the alpha value. It is identifying outliers, especially for time series data, and works with tibble objects.

Anomalize package decomposes time-series data before performing the outlier detection. There are many statistical decomposition models (Error-Trend-Seasonality (ETS), Autoregressive Integrated Moving Average (ARIMA), Prophet, Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components (TBATS), etc.). The decomposition of the data series results in remainder values in which outlier detection is performed.

Seasonal and trend decomposition using loess (STL) procedure (Cleveland et al., 1990) decomposes seasonal time series data into seasonal, trend, and remainder components, and is based on locally estimated scatterplot smoothing (LOESS) estimate, introduced by Cleveland et al. (1979). LOESS in STL extracts smoothed component estimates. The STL decomposition algorithm applies a series of filtering operations to extract the trend and seasonal components. It typically involves a combination of moving averages, weighted least squares, and locally weighted regression. The remainder component is obtained by subtracting the trend and seasonal components from the original time series. Trend describes the general data movement along time via a line pattern, showing potential increases or decreases in the data. Time-series seasonal component is similar to the trend component. The main difference is that the seasonal component is repeated periodically over time. The Remainder is what remains after removing trend and seasonal components from the time series data.

The components can be described as follows:

$$Yi = Si + Ti + Ri$$

where:
Y_i = Time series at data point $_i$
S_i = Seasonal component at $_i$
T_i = Trend component at $_i$ point
R_i = Remainder component at $_i$

Twitter decomposition instead of the "trend" component creates and removes median spans, which is a series of numerical medians.

The *anomalize* package allows users to control frequency and trend parameters. Frequency is adjusting the seasonal component and trend is adjusting the trend window. *Auto* mode adjusts both parameters according to the time series scale. After data series decomposition, outlier detection is performed for IQR or GESD method. The final step is time recomposition, where limits separating normal data from anomalies are calculated. The package returns three columns: the lower limit of the remainder, the higher limit of the remainder, and whether the measurement is an anomaly or not.

In this study, STL was selected as the decomposition method, with automatic frequency and trend settings. Outlier detection was performed for each sensor using both methods for variables: temperature, humidity, and PM. Temperature was converted from °F to °C. Outliers were detected for daily and hourly data separately. The results of all sensors were summarized for IQR and GESD individually for hourly and daily data. Furthermore, the same values considered as outliers in both methods were also summarized. These calculations were converted to percentages (relative frequencies) of the outlier number of total measurements for each category.

We repeated the same steps, but filters were applied before outlier detection to each measurement type, except for humidity. Humidity didn't need a filter due to the sensor's default value limit (0%–100%). For temperature, an upper filter of 50°C and a lower filter of −20°C were applied considering the case study climate. The PM sensor can't record values below 0 but can record extremely high values. The upper PM value limit was set to 1,000 μg/m^3. These filters were applied to exclude certain sensors' faults. Also, in all sensor datasets, recordings appear to be duplicated. Hence, the whole methodology was applied with duplicates removed.

We performed the Ordinary Kriging (OK) geostatistical technique to see how outliers affect root mean square error (RMSE) on interpolated surfaces. RMSE was calculated according to the formula:

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{(P_i - O_i)^2}{N}},$$

i = variable i,
N = total number of observations
P_i = predicted values
O_i = observed values

OK was applied to hourly temperature data without a filter. The chosen date was 2019-05-24 00:00:00, when an extreme value was present out of 13 total sensor observations. First, the temperature variogram was computed. Data was divided into test and train data. Test data was a selection of three random nonoutlier values and train data was the rest values. Next, OK was executed for 10 times where each repetition contained different test data and RMSE estimated for each repetition. The same steps were followed, but this time train data consisted only of nonoutlier values. Through this process, OK models were assessed to examine the influence of outliers on spatial interpolation and gain a deeper understanding of how they can impact the accuracy of interpolated data.

II. a. Data

Initial data comes from PurpleAir low-cost sensors provided by EPA. The sensors are located in Athens, Greece and some of them have records from 2018 to 2022. Each sensor consists of two PMS*003 series laser counters (labeled channel A and channel B), alternating every 5 s and averaged in 120 s. For each channel, there are two data sets “Primary” and “Secondary” data, and their fields are described in Table 1.

Sensors collect environmental data measuring temperature (°F), humidity (%), and air quality in PM concentration (1.0 μm/m^3, 2.5 μm/m^3, 10.0 μm/m^3). These measurements are uploaded in the ThingSpeak data cloud via Wi-Fi and can be downloaded in averages of 10 min, 15 min, 20 min, 30 min, 60 min, 240 min, 720 min, and 1.440 min.

PM is calculated by the number of particles sized 0.3 μm/m^3, 0.5 μm/m^3, 1 μm/m^3, 2.5 μm/m^3, 5 μm/m^3 and 10 μm/m^3. PM 1.0 μm/m^3, 2.5 μm/m^3, and 10.0 μm/m^3 have two different outputs, cf_1 and cf_atm. Cf_atm is performing atmospheric correction on raw data, for which additional information is not provided. The correction tends to reduce PM values on high concentrations and keeps low concentrations about the same as cf_1 (Barkjohn et al., 2021). Literature shows sensor calibration and outlier detection techniques for both cf_1 and cf_atm (Kelly et. al, 2017; Stavroulas et al., 2020; Barkjohn et al., 2021; Nilson et al., 2022). Here both outputs are used.

For each sensor location, the PurpleAir platform provides real-time and averaged measurements on an interactive map. Sensors can be found worldwide working individually or creating networks, in some cities for example.

For this study, hourly and daily averaged data (in UTC) were downloaded containing temperature, relative humidity, and PM in csv format. Thereupon, Channel A was selected for the analysis. Datasets come in spatial time-series format, where each.csv represents a sensor having the coordinates and the measurements over time. Various time-scale data of the same sensors can help understand similarities and differences among the outlier detection techniques as we can identify outliers in both temporal scales.

II. b. Case study

The study was conducted in Athens, the Greek capital, which is the largest city in Greece with about 3 million inhabitants. PurpleAir sensors in Athens are located in random spots, as they have different owners. Figure 1 shows their locations, which cover a major proportion of the city. There were 58 sensors, some of them active since January 2018. The dataset includes measurements between January 2018 and May 2022. Sensors didn’t record unceasingly all the time, but there were some pauses or they would have even stopped completely. No other environmental data from external sources in the spatial scale required for the case study area are available.

III. Results

This section illustrates the results of our analysis, focusing on the identification and handling of extreme values in various environmental measurements, such as Temperature, Humidity, and PM. First, we present graphical representations of outlier scenarios, encompassing sensor malfunctions and false records, offering a visual understanding of their impact. Additionally, we highlight the effect of atmospheric correction examined, particularly as provided by PurpleAir, on PM data. Also, we show the issues that are addressed related to false data and the importance of duplicate data removal in reducing outlier occurrences. Furthermore, we represent the influence of outliers on spatial interpolation through the OK of hourly temperature data, emphasizing the necessity of robust outlier detection for accurate environmental analysis and modeling. Overall, this section provides

Table 1: PurpleAir sensor data, Primary and Secondary data sets of Channels A and B, gray cells represent the selected parameters of the study (PurpleAir, 2022)

	PRIMARY	
	CHANNEL A	CHANNEL B
Field 1	PM1.0 (CF = 1) μg/m^3	PM1.0 (CF = 1) μg/m^3
Field 2	PM2.5 (CF = 1) μg/m^3	PM2.5 (CF = 1) μg/m^3
Field 3	PM10.0 (CF = 1) μg/m^3	PM10.0 (CF = 1) μg/m^3
Field 4	Uptime (min)	Free HEAP memory
Field 5	RSSI (WiFi signal strength)	ADC0 (analog input) voltage
Field 6	Temperature (F)	FIRMWARE 2.5 and up: atmospheric pressure
Field 7	Humidity (%)	FIRMWARE 4.10 and up: Bosch BSEC IAQ when BME680 gas sensor is present
Field 8	PM2.5 (CF = ATM) μg/m^3	PM2.5 (CF = ATM) μg/m^3
	SECONDARY	
Field 1	0.3 μm particles/dL	0.3 μm particles/dL
Field 2	0.5 μm particles/dL	0.5 μm particles/dL
Field 3	1.0 μm particles/dL	1.0 μm particles/dL
Field 4	2.5 μm particles/dL	2.5 μm particles/dL
Field 5	5.0 μm particles/dL	5.0 μm particles/dL
Field 6	10.0 μm particles/dL	10.0 μm particles/dL
Field 7	PM1.0 (CF = ATM) μg/m^3	PM1.0 (CF = ATM) μg/m^3
Field 8	PM10 (CF = ATM) μg/m^3	PM10 (CF = ATM) μg/m^3

BSEC, Bosch Sensortec Environmental Cluster; IAQ, Indoor Air Quality; PM, particulate matter; RSSI, Received Signal Strength Indicator.

insights into the presence and handling of outliers in environmental measurements, shedding light on the significance of accurate outlier detection techniques in scientific analyses and data processing.

III. a. Outlies in the Data Series

As shown in Figure 2a there are cases where extremely high values can occur due to sensor malfunction. This sensor recorded temperatures up to 1,130,876°C (2,035,609°F).

Another case is when false records repeat for a long time, enough for the outlier detection methods not to consider them as outliers. The second graph (Figure 2b) is an example of that. In this sensor, extreme low temperature values are recorded but for a long time. The length of the "long time" period depends on the time window the algorithm has. This results in the algorithm declaring outlier values as normal after the window time period, where the outlier detection algorithm is calculating the nonoutlier range by only extreme high or low values. Also in Figure 2b, the values just before and after the malfunction event (window time period) are considered as outliers. That occurred due to the sudden extreme difference between the normal and the malfunction values. Therefore, the nonoutlier range was between these values.

Also, the same issue appears on daily PM 10.0 μm/m^3 data (Figure 2c). Observations seem to have high values, above 1,000. The reason here can be different than the temperature sensor. For example, the existence of dust inside the sensor.

Another issue faced here despite outliers, is false data. For example, the sensor in Figure 2d presents about the same value for a long time period.

The sensor in Figure 3a has a similar malfunction similar to the sensor in Figure 2c. But this time, outliers were detected correctly. There are some possible reasons for that: (a) The presence of some records with values between normal and extreme ones, before and after the extreme event. That can make the change not as sudden as in Figure 2c. Non-outlier range calculation isn't affected by the extreme values; (b) Trend window may be different. This means the range calculation considers both before and after extreme event values in one window; and (c) In

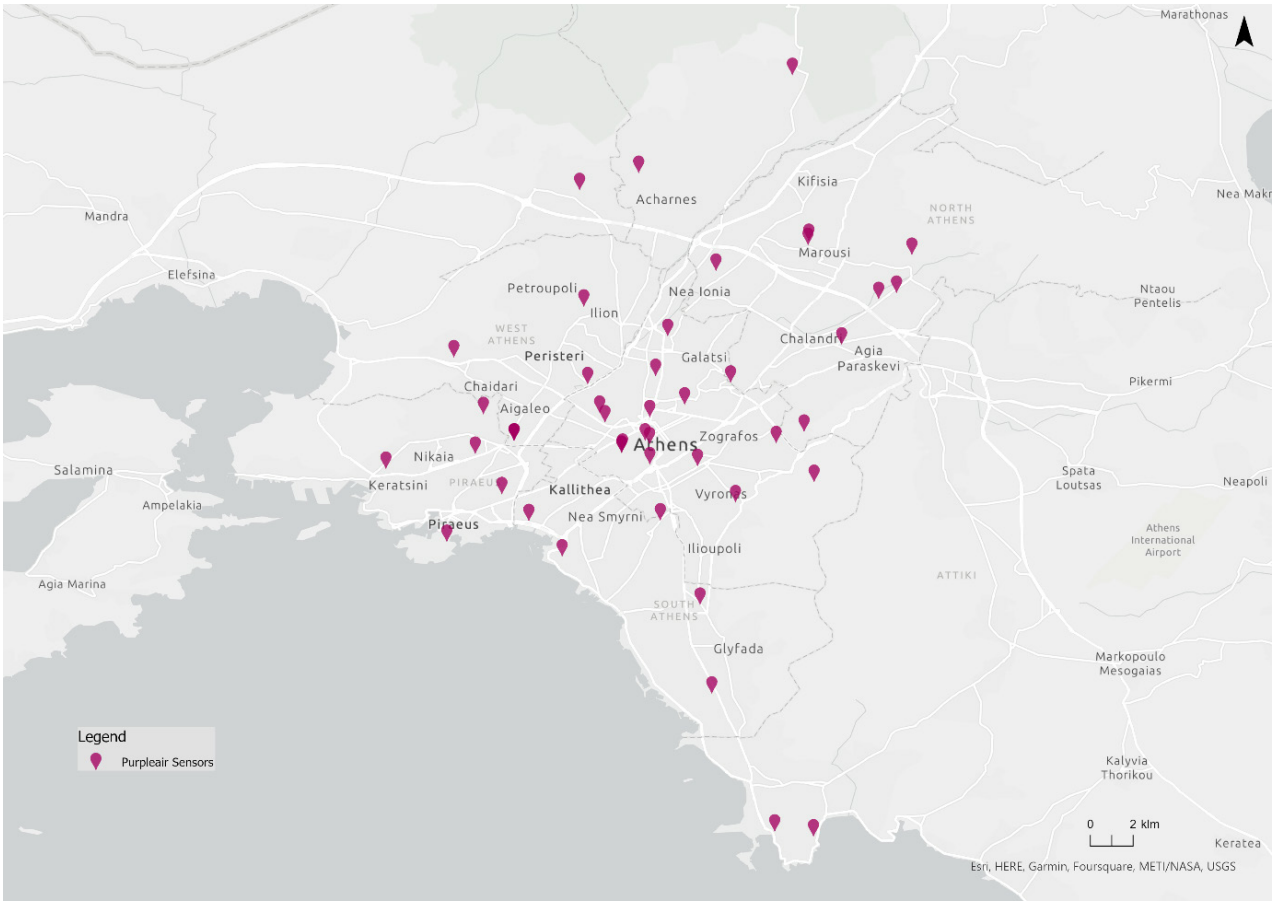


Figure 1: Map of the PurpleAir network of sensors in Athens Greece.

this sensor, before and after the extreme event, the normal records have been continuous for longer than the sensors in Figures 2b,c.

Hourly PM of all sizes (1.0 μm/m^3, 2.5 μm/m^3, and 10.0 μm/m^3) tend to have many outliers, especially those of the GESD method. Figure 3b is an example of PM10.0 μm/m^3 outliers in hourly data using the GESD method.

It can be assumed that most high PM concentrations happen during winter months. A reason for this can be smog formation by heating sources like wood-burning fireplaces and fossil fuel burning for building heating. Winter smog tends to occur in big cities and can be unhealthy for residents.

III. b. Outlier Detection Methods Comparison

The results of identifying extreme values with two different methods (IQR and GESD) (Table 2) show the outliers on daily data for Temperature, Humidity, and PM. There are two sections, one for the results before using filters on data and another showing the results after filters are applied. Observations (*n*) are the total observations of each measurement type. The third and fourth columns refer to the number of outliers with IQR and GESD methods respectively. Outlier observations in both methods are found identical in both methods. In Table 3, the results of hourly data are presented the same way as in Table 2.

As no filter is applied to humidity data, there is no change in the after-filter tables. Humidity outlier percentages for both methods and both datasets were somehow low. Most outliers found in the PM 10.0 μm/m^3 dataset in every case were in contrast to the rest of the measurement types. In general, GESD outliers include almost every IQR outlier. The GESD method always finds more outliers than IQR and hourly PM data tend to have more outliers than daily data (Figure 4).

PM data have fewer outliers on both time scales and techniques with atmospheric correction (cf_atm) PurpleAir provides (Tables 2 and 3).

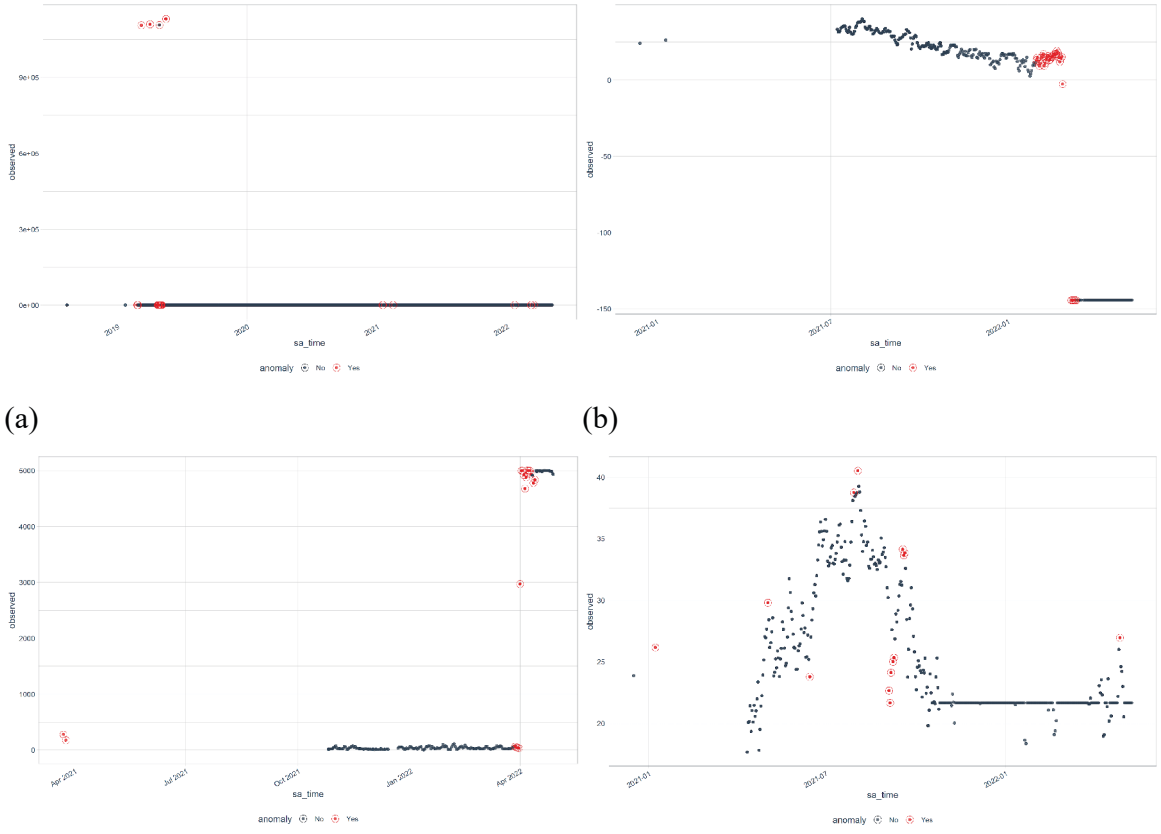


Figure 2: Outliers by IQR method on daily data with (a) extreme high-temperature values on sensor, (b) continuous malfunction of temperature sensor, (c) continuous malfunction of PM10.00µm/m^3 on sensor, (d) continuous malfunction of temperature sensor. IQR, interquartile range; PM, particulate matter.

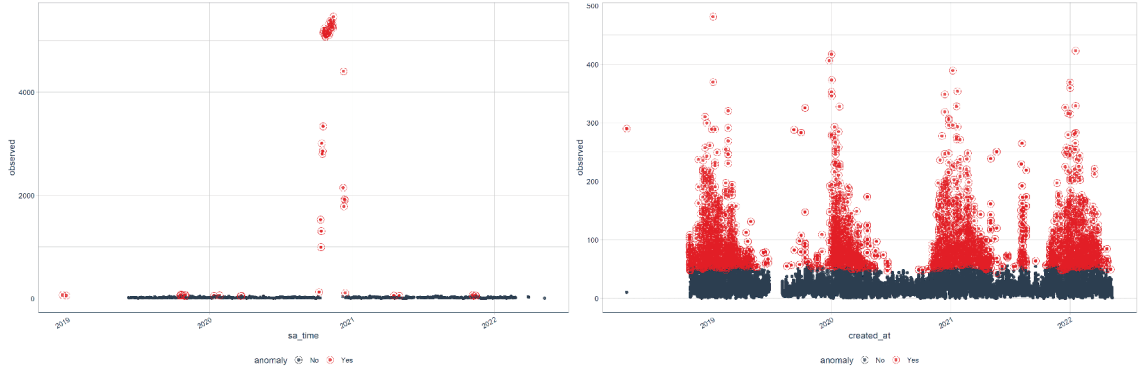


Figure 3: (a) Daily PM 10.0µm/m^3 outliers by IQR method with extremely high values due to sensor malfunction, (b) Hourly PM 10.0µm/m^3 outliers with GESD method. GESD, generalized extreme studentized deviate; IQR, interquartile range; PM, particulate matter.

Table 2: Outliers of IQR and GESD methods on daily data for temperature (°C), humidity (%), and PM (1.0 μm/m³, 2.5 μm/m³, 10.0 μm/m³) before and after filter application

	Observations (*n*)	IQR outliers (*n*)	GESD outliers (*n*)	Outlier observations in both methods (*n*)	IQR outliers/ observations (%)	GESD outliers/ observations (%)	Both methods/ observations (%)
BEFORE FILTER							
Temperature (°C)	45,740	1,094	1,932	1,034	2.4	4.2	2.3
Humidity (%)	45,740	260	556	260	0.6	1.2	0.6
PM1.0 μm/m³ cf_1	46,305	1,655	2,745	1,655	3.6	5.9	3.6
PM2.5 μm/m³ cf_1	46,305	1,822	3,042	1,815	3.9	6.6	3.9
PM10.0 μm/m³ cf_1	46,305	1,869	3,146	1,862	4.0	6.8	4.0
PM1.0 μm/m³ cf_atm	46,299	1,498	2,019	1,488	3.2	4.4	3.2
PM2.5 μm/m³ cf_atm	46,305	1,632	2,193	1,537	3.5	4.7	3.3
PM10.0 μm/m³ cf_atm	46,299	1,762	2,558	1,626	3.8	5.5	3.5
AFTER FILTER							
Temperature (°C)	44,928	624	1,470	624	1.4	3.3	1.4
Humidity (%)	45,740	260	556	260	0.6	1.2	0.6
PM1.0 μm/m³ cf_1	46,091	1,386	2,449	1,378	3.0	5.3	3.0
PM2.5 μm/m³ cf_1	46,091	1,549	2,738	1,545	3.4	5.9	3.4
PM10.0 μm/m³ cf_1	46,091	1,598	2,854	1,593	3.5	6.2	3.5
PM1.0 μm/m³ cf_atm	46,089	1,232	1,741	1,225	2.7	3.8	2.7
PM2.5 μm/m³ cf_atm	46,095	1,376	1,897	1,282	3.0	4.1	2.8
PM10.0 μm/m³ cf_atm	46,089	1,475	2,231	1,340	3.2	4.8	2.9

GESD, generalized extreme studentized deviate; IQR, interquartile range; PM, particulate matter.

Table 3: Outliers of IQR and GESD methods on hourly data for temperature (°C), humidity (%), PM (1.0 µm/m³, 2.5 µm/m³, 10.0 µm/m³) before and after filter application

	BEFORE FILTER						
	Observations (*n*)	IQR outliers (*n*)	GESD outliers (*n*)	Outlier observations in both methods (*n*)	IQR outliers/ observations (%)	GESD outliers/ observations (%)	Both methods/ observations (%)
Temperature (°C)	1,074,342	5,643	7,471	4,272	0.4	0.7	0.4
Humidity (%)	1,074,342	6,373	7,196	6,026	0.6	0.7	0.6
PM1.0 µm/m³ cf_1	1,087,434	49,742	70,944	48,046	4.4	6.5	4.4
PM2.5 µm/m³ cf_1	1,087,434	52,848	73,647	51,091	4.7	6.8	4.7
PM10.0 µm/m³ cf_1	1,087,434	54,936	75,768	53,141	4.9	7.0	4.9
PM1.0 µm/m³ cf_atm	1,087,362	37,216	46,946	34,170	3.4	4.3	3.1
PM2.5 µm/m³ cf_atm	1,087,434	38,954	46,011	34,936	3.6	4.2	3.2
PM10.0 µm/m³ cf_atm	1,087,362	49,344	67,686	45,595	4.5	6.2	4.2
	AFTER FILTER						
Temperature (°C)	1,056,463	2,984	4,682	2,812	0.3	0.4	0.3
Humidity (%)	1,074,342	6,373	7,196	6,026	0.6	0.7	0.6
PM1.0 µm/m³ cf_1	1,082,638	46,121	67,057	44,444	4.3	6.2	4.1
PM2.5 µm/m³ cf_1	1,082,631	49,387	69,968	47,650	4.6	6.5	4.4
PM10.0 µm/m³ cf_1	1,082,619	50,824	71,449	49,052	4.7	6.6	4.5
PM1.0 µm/m³ cf_atm	1,082,576	33,896	43,637	30,887	3.1	4.0	2.9
PM2.5 µm/m³ cf_atm	1,082,646	35,257	42,116	31,249	3.3	3.9	2.9
PM10.0 µm/m³ cf_atm	1,082,573	45,929	63,842	42,188	4.2	5.9	3.9

GESD, generalized extreme studentized deviate; IQR, interquartile range; PM, particulate matter.

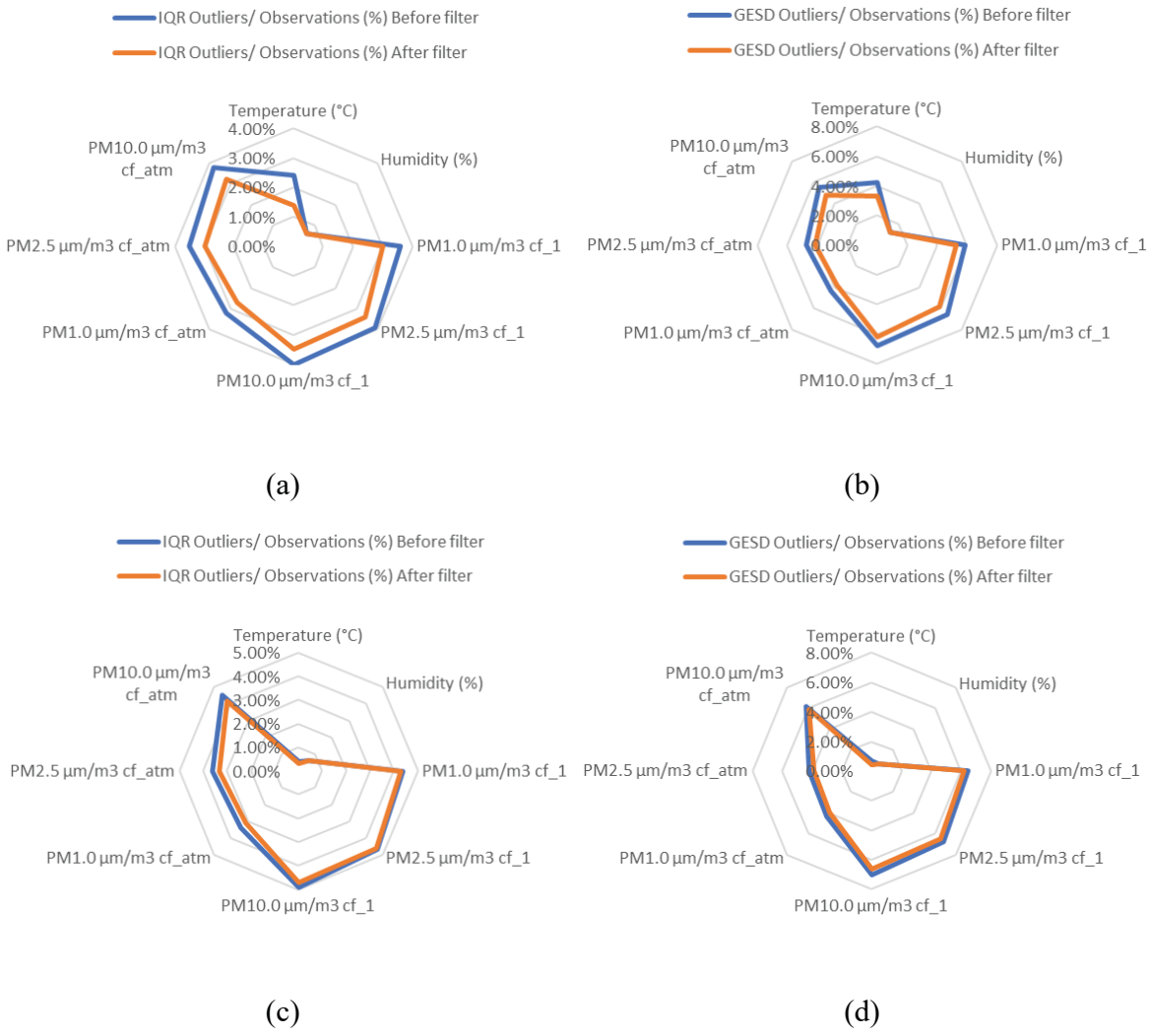


Figure 4: Outliers/observations (%) before and after filter for (a) IQR method on daily data, (b) GESD method on daily data, (c) IQR method on hourly data, and (d) GESD method on hourly data. GESD, generalized extreme studentized deviate; IQR, interquartile range; PM, particulate matter.

Additionally, duplicate data removal is a determining factor of outlier reduction in every case. Table 4 shows the outliers detected on hourly data for all the environmental factors we studied before and after the filters and Table 5 presents the corresponding results for hourly data. These two tables have the least outliers achieved by every method.

III. c. Spatial Interpolation of Data

Figure 5 depicts OK with hourly temperature data captured at 2019-05-24 00:00:00 Coordinated Universal Time (UTC). OK was first applied to raw data with one extreme value (Figure 5a) and then to the same data without the extreme value (Figure 5b). This is an example of how extreme outliers can affect spatial interpolation. The results are quantified in Table 6, where RMSE is calculated for both cases (before and after filtering the extreme value).

This analysis has shed light on the identification and management of extreme values in various environmental measurements, including Temperature, Humidity, and PM. Through graphical representations and tables, the impact of sensor malfunctions, false records, and dust presence has been highlighted. Additionally, the significance of atmospheric correction from PurpleAir and the role of duplicate data removal in reducing outliers have been emphasized.

Table 4: IQR and GESD outliers on hourly data without duplicates, for Temperature (°C), Humidity (%), and PM (1.0 µm/m³, 2.5 µm/m³, 10.0 µm/m³) before and after filter application

	Observations (n)	IQR outliers (n)	GESD outliers (n)	Outlier observations in both methods (n)	IQR outliers/ observations (%)	GESD outliers/ observations (%)	Both methods/ observations (%)
BEFORE FILTER							
Temperature (°C)	682,028	3,533	4,763	3,531	0.5	0.7	0.5
Humidity (%)	682,028	2,685	3,817	2,685	0.4	0.6	0.4
PM1.0 µm/m³ cf_1	691,210	28,161	40,473	28,161	4.1	5.9	4.1
PM2.5 µm/m³ cf_1	691,210	29,515	42,624	29,515	4.3	6.2	4.3
PM10.0 µm/m³ cf_1	691,210	30,364	43,831	30,364	4.4	6.3	4.4
PM1.0 µm/m³ cf_atm	691,159	18,076	22,099	18,074	2.6	3.2	2.6
PM2.5 µm/m³ cf_atm	691,210	18,874	23,095	18,866	2.7	3.3	2.7
PM10.0 µm/m³ cf_atm	691,159	22,396	33,156	22,020	3.2	4.8	3.2
AFTER FILTER							
Temperature (°C)	671,277	2,068	3,486	2,066	0.3	0.5	0.3
Humidity (%)	682,028	2,685	3,817	2,685	0.4	0.6	0.4
PM1.0 µm/m³ cf_1	688,544	26,450	38,706	26,450	3.8	5.6	3.8
PM2.5 µm/m³ cf_1	688,537	27,692	40,662	27,692	4.0	5.9	4.0
PM10.0 µm/m³ cf_1	688,527	28,398	41,690	28,398	4.1	6.1	4.1
PM1.0 µm/m³ cf_atm	688,500	16,234	20,235	16,232	2.4	2.9	2.4
PM2.5 µm/m³ cf_atm	688,549	17,050	21,219	17,042	2.5	3.1	2.5
PM10.0 µm/m³ cf_atm	688,497	20,508	31,115	20,132	3.0	4.5	2.9

GESD, generalized extreme studentized deviate; IQR, interquartile range; PM, particulate matter.

Table 5: IQR and GESD outliers on daily data without duplicates, for Temperature (°C), Humidity (%), and PM (1.0 µm/m³, 2.5 µm/m³, 10.0 µm/m³) before and after filter application

	Observations (n)	IQR outliers (n)	GESD outliers (n)	Outlier observations in both methods (n)	IQR outliers/ observations (%)	GESD outliers/ observations (%)	Both methods/ observations (%)
BEFORE FILTER							
Temperature (°C)	29,040	380	735	380	1.3	2.5	1.3
Humidity (%)	29,040	94	234	94	0.3	0.8	0.3
PM1.0 µm/m3 cf_1	29,437	665	1,302	665	2.3	4.4	2.3
PM2.5 µm/m3 cf_1	29,437	716	1,494	708	2.4	5.1	2.4
PM10.0 µm/m3 cf_1	29,437	751	1,552	751	2.6	5.3	2.6
PM1.0 µm/m3 cf_atm	29,435	560	835	553	1.9	2.8	1.9
PM2.5 µm/m3 cf_atm	29,437	596	926	596	2.0	3.1	2.0
PM10.0 µm/m3 cf_atm	29,435	608	1,051	606	2.1	3.6	2.1
AFTER FILTER							
Temperature (°C)	28,552	221	579	221	0.8	2.0	0.8
Humidity (%)	29,040	94	234	94	0.3	0.8	0.3
PM1.0 µm/m3 cf_1	29,316	554	1,188	554	1.9	4.1	1.9
PM2.5 µm/m3 cf_1	29,316	592	1,360	584	2.0	4.6	2.0
PM10.0 µm/m3 cf_1	29,316	624	1,417	624	2.1	4.8	2.1
PM1.0 µm/m3 cf_atm	29,316	443	713	436	1.5	2.4	1.5
PM2.5 µm/m3 cf_atm	29,318	485	807	485	1.7	2.8	1.7
PM10.0 µm/m3 cf_atm	29,316	495	930	493	1.7	3.2	1.7

GESD, generalized extreme studentized deviate; IQR, interquartile range; PM, particulate matter.

Table 6: OK RMSE of hourly temperature data on 2019-05-24 00:00:00 UTC, before and after outlier filter for 10 repetitions

Before filter	4,083.351	8,997.641	4,102.043	4,080.238	544.752	4,141.303	426.213	4,087.449	8,030.859	3,272.878
After filter	0.209	0.540	0.501	0.204	0.155	0.507	0.285	0.312	0.503	0.245

OK, ordinary kriging; RMSE, root mean square error.

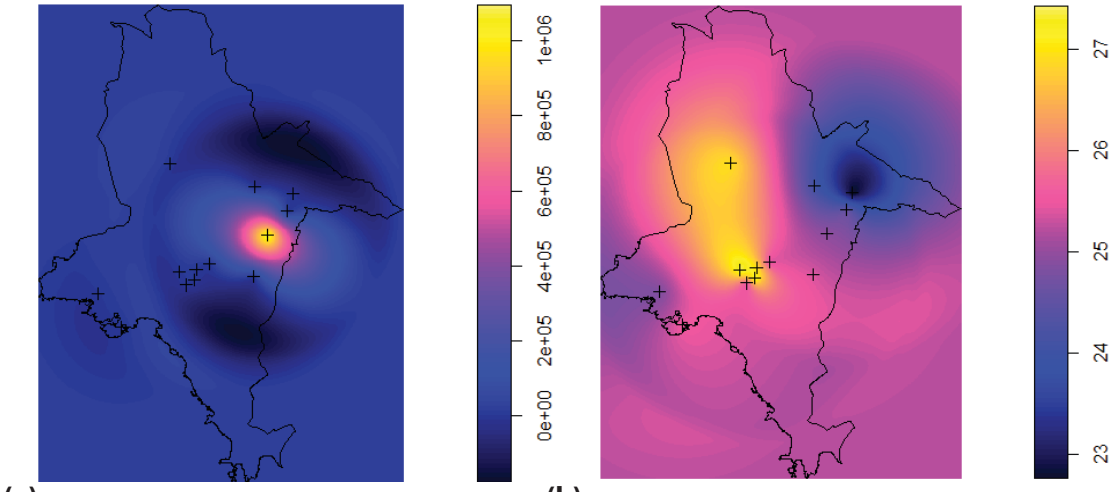


Figure 5: OK of hourly temperature data on 2019-05-24 00:00:00 UTC (a) with an extreme value, (b) without outliers. OK, ordinary kriging.

IV. Discussion

iv. a. Main findings and comparison to other studies

In this paper, we investigated outlier management in time series datasets, in line with many papers in different stands of the literature on data management (Fan et al., 2020; Stavroulas et al., 2020; Bi et al., 2020). IoT devices' data in a smart city context, ask for advanced data warehousing and data filtering before being used for decision-making. We have used alternative methods to identify outlier values in a time series dataset concerning environmental data in a so-called smart city environment and tested how outliers affect spatial interpolation techniques in line with Pinder et al. (2019) that recognize the need for using spatial statistics in sensor network data analysis for air quality management.

Our findings demonstrate that in all cases where the application of big data analysis is required, filtering and managing outliers is very important as the presence of unmanaged outliers can result in distorting the results of any types of models used. This has been recognized by Feenstra et al. (2020) among others, while Bi et al. (2020) calibrated PurpleAir PM 2.5 sensor measurements at a large spatial scale by performing geographically weighted regression and associating these with external sources (land use, meteorological data, etc.) have made clear that outlier management is critical in such approaches. Moreover, our findings point toward a need for using spatial statistics in sensor network data analysis for air quality management (for a similar approach see Pinder et al., 2019).

Our findings show that temperature and humidity variables present fewer outliers for both time scales than PM data. This is expected considering the range of the values of the above environmental factors. PM data (of all sizes) should have a larger nonoutlier range on outlier detection particularly on hourly data. This is in line with what Aslan et al. (2022) reported.

In comparison with other studies, the outliers occurrence here is similar to that that has been reported in other studies, both in "smart-cities" contexts (Csaji et al., 2017; Elbaz et al., 2023; Pastorio et al., 2022) and in other types of approaches (Stavroulas et al., 2020; Liu et al., 2022; Cieplak et al., 2019; Liang and Yu, 2021). It is important to consider the proportionality of the nonoutlier range in relation to the time scale when working with PM data. Smaller time scales necessitate larger ranges for nonoutlier values. Interestingly, PM data that has undergone atmospheric correction tend to exhibit fewer outliers across all cases.

A notable point of our work is the identification of the need for data filtering on temperature and PM data before outlier detection to omit extreme data. On temperature data, the filter should be adapted to the city's climate conditions (see also Pereira et al., 2023; Ma et al., 2017). PM should be eliminated at a range of 0–1,000 μm/m^3, so extreme values can be excluded (see also Aix et al. 2023 that suggest other values).

One notable example is the presence of environmental phenomena that can be overshadowed by outliers and the methods used for their estimation and management. In our case, such an event is winter smog formation, a type of air pollution characterized by a mixture of smoke and fog, resulting from the interaction of pollutants like sulfur dioxide and PM in classical smog, or the complex chemical reactions involving sunlight, nitrogen oxides, and volatile organic compounds in photochemical smog, both of which can have harmful effects on human health and the environment (Fenger, 1999). Winter smog seems to be detected every winter in almost every sensor of our network. The solution here could be an extra parameter checking neighboring sensors' data. If the neighboring sensor's data show event existence, then data should be perceived as an event and not as outliers. Although data among these sensors could deviate depending on the event spatial scale, event management could have different time and space scales.

iv. b. Strengths and limitations

The calibration of sensors has emerged as an important issue in our approach. Becnel et al. (2019) have proposed an algorithm for a low-cost sensor network, using as a reference a single station. Their algorithm calibrates according to the sensors' distance from the reference station. However, the use of a single reference station to calibrate all sensors seems a nonoptimal choice in our research, as spatial scale emerges as a determining factor. Some events tend to happen close to a single sensor and are not recorded by nearby sensors, something reasonable at a city scale. When nearby sensors also record the events, the event can be determined safely.

In our study, we used two alternative methods for outlier calculation and management. The findings suggest that the IQR method is faster than the GESD. This may be important with very big quantities of data and can save significant time and resources for calculations and management. Considering the particular variables measured, in the context of analyzing hourly PM data, both methods tend to label a significant number of values as outliers, especially the GESD method which is prone to detect a high number of outliers. Notably the GESD method detects almost all the outliers identified by the IQR method. Both techniques appear to be effective in identifying outliers in datasets related to temperature and humidity. The overall comparisons show that the IQR method emerges as the optimal approach.

Another issue that emerges is the alpha value that should be used in the approach. The findings suggest that the alpha value used in the analysis should be adjusted according to the time scale. For daily data, an alpha of 0.05 proves effective for both methods. However, when dealing with hourly data, an alpha of 0.05 leads to the misclassification of nonoutlier values as outliers. Consequently, a smaller alpha of 0.009 is recommended for hourly data. Merello et al. (2014) have performed a similar study on the use of the alpha metric.

This study pointed out the outliers using two time-series methods for each environmental sensor in a city. Except for the type of outliers considered here, the methods applied can be used for events such as fires, fireworks (Feenstra et al., 2020), road gas emissions, and building heating sources (wood-burning fireplaces, fossil fuel burning heating sources, etc.). Events like these have a variety of durations (e.g., 10 min, an hour, a week, a month, etc.) and areas (e.g., 1 m^2, 1 km^2, 10 km^2 around the sensor or even the whole city area). They mostly affect PM measurements, the bigger the PM size, the bigger the effect. The limitations of this study refer to the lack of specific event detection algorithms or calibration methods that could reduce the errors. Also, it primarily relies on a single dataset, potentially limiting its applicability to widely varying contexts.

The most important finding that emerges refers to the need for implementing spatial statistics in sensor network data analysis for air quality management (Pinder et al., 2019). Issues such as comparing data from nearby sensors for outliers (Feenstra et al., 2020), combining reference stations with the use of algorithms (Becnel et al., 2019), and calibrating them according to spatial proximity in large-scale environments such as cities and reducing spatial sensor noise, are all key factors for the fast and efficient identification and management of outliers in long and spatially extended time series. Our recommended solution is to employ the IQR method while ensuring that the alpha value is proportional to the time scale being analyzed (an alpha of 0.009 should be used for hourly data, while an alpha of 0.05 suffices for daily data).

V. Conclusions

In the context of the so-called smart cities, large numbers of measurements accumulate over extended spatial and time spans. Understanding these data, calculating indicators, and determining events or patterns in the data, it is vital to incorporate smart decision-making tools that include Anomaly Detection modules as well as Outlier detection modules to

detect irregularities and extreme events as soon as possible and adapt decisions accordingly.

Incorporating outlier detection systems in smart cities offers several benefits and is essential for efficient and effective urban management. Outlier detection systems help identify abnormal or anomalous events, behaviors, or patterns in real time. These anomalies could be unusual traffic patterns, abnormal energy consumption, sudden spikes in crime rates, or any other unexpected events. By detecting and flagging outliers, city authorities can respond promptly and take necessary actions to mitigate risks or address emerging issues.

Outlier detection systems act as early warning systems by identifying anomalies that may indicate potential risks or crises. For instance, sudden changes in air quality, abnormal weather patterns, or atypical disease outbreaks can be detected through outlier analysis. Early detection enables authorities to respond quickly, implement preventive measures, and minimize the impact of adverse events.

In this paper, we compared two outlier detection techniques on environmental sensor data, in a smart city context. The study was conducted in Athens, the capital of Greece, where PurpleAir sensors record temperature, humidity, and PM. Outliers were identified for both daily and hourly data of the same sensors with two techniques: IQR and GESD, adapted on time series, meaning that before their application, the time series were decomposed into their three components, seasonal, trend, and remainder. Outlier detection was performed on the remainder component. The results depict the diversity of issues encountered with low-cost sensor data. In this paper, the performance of the two outlier detection techniques was evaluated and the long-term data behavior for the first time for this area was examined as well.

Our approach suggests that it is possible to detect outliers rapidly and with a level of confidence that is related to the time scale used, the variable, and the spatial scale. Space appears to be an important spatial factor for outlier detection in such sensor networks that can work autonomously and detect spatial patterns and/or analyze and calibrate spatial events, irregularities, and outliers that emerges as an area for further analysis. Issues of spatial visualization and interpolation of sensor data should also be studied more extensively.

Acknowledgments

The research of this paper has received funding from the ISLANDS ERASMUS MUNDUS Masters' Programme "Islands and Sustainability", Grant Agreement 101050349.

References

[1] Aix, M. L., Schmitz, S., & Bicout, D. J. (2023). Calibration methodology of low-cost sensors for high-quality monitoring of fine particulate matter. *Science of the Total Environment*, *889*(February). https://doi.org/10.1016/j.scitotenv.2023.164063

[2] Al Samara, M., Bennis, I., Abouaissa, A., & Lorenz, P. (2022). A Survey of Outlier Detection Techniques in IoT: Review and Classification. *Journal of Sensor and Actuator Networks*, *11*(1). https://doi.org/10.3390/jsan11010004

[3] Alvear-Puertas, V. E., Burbano-Prado, Y. A., Rosero-Montalvo, P. D., Tözün, P., Marcillo, F., & Hernandez, W. (2022). Smart and Portable Air-Quality Monitoring IoT Low-Cost Devices in Ibarra City, Ecuador. *Sensors*, *22*(18), 1–17. https://doi.org/10.3390/s22187015

[4] Amini, M. H., Arasteh, H., & Siano, P. (2019). *Sustainable Smart Cities Through the Lens of Complex Interdependent Infrastructures: Panorama and State-of-the-art* (pp. 45–68). https://doi.org/10.1007/978-3-319-98923-5_3

[5] Aslan, M. E., & Onut, S. (2022). Detection of Outliers and Extreme Events of Ground Level Particulate Matter Using DBSCAN Algorithm with Local Parameters. *Water, Air, and Soil Pollution*, *233*(6). https://doi.org/10.1007/s11270-022-05679-6

[6] Ayadi, A., Ghorbel, O., Obeid, A. M., & Abid, M. (2017). Outlier detection approaches for wireless sensor networks: A survey. *Computer Networks*, *129*(2), 319–333. https://doi.org/10.1016/j.comnet.2017.10.007

[7] Barkjohn, K. K., Gantt, B., & Clements, A. L. (2021). Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor. *Atmospheric Measurement Techniques*, *14*(6), 4617–4637. https://doi.org/10.5194/amt-14-4617-2021

[8] Becnel, T., Sayahi, T., Kelly, K., & Gaillardon, P. E. (2019). A recursive approach to partially blind calibration of a pollution sensor network. *2019 IEEE International Conference on Embedded Software and Systems, ICESS 2019*. https://doi.org/10.1109/ICESS.2019.8782523

[9] Bi, J., Wildani, A., Chang, H. H., & Liu, Y. (2020). Incorporating Low-Cost Sensor Measurements into High-Resolution PM2.5 Modeling at a Large Spatial Scale. *Environmental Science and Technology*, *54*(4), 2152–2162. https://doi.org/10.1021/acs.est.9b06046

[10] Blázquez-García, A., Conde, A., Mori, U. & Lozano, J. A. (2021). A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Computing Surveys*, 54(3). https://doi.org/10.1145/3444690

[11] Braei, M., & Wagner, S. (2020). Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art. *ArXiv, abs/2004.00433*

[12] Chen, L. J., Ho, Y. H., Hsieh, H. H., Huang, S. T., Lee, H. C., & Mahajan, S. (2018). ADF: An Anomaly Detection Framework for Large-Scale PM2.5 Sensing Systems. *IEEE Internet of Things Journal*, *5*(2), 559–570. https://doi.org/10.1109/JIOT.2017.2766085

[13] Cieplak, T., Rymarczyk, T., & Tomaszewski, R. (2019). A concept of the air quality monitoring system in the city of Lublin with machine learning methods to detect data outliers. *MATEC Web of Conferences*, *252*, 03009. https://doi.org/10.1051/matecconf/201925203009

[14] Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A Seasonal-Trend decomposition Procedure Based on Loess. *Journal of Official Statistics*, *6*(1), 3–73.

[15] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*(368), 829–836. https://doi.org/10.1080/01621459.1979.10481038

[16] Csaji, B. C., Kemeny, Z., Pedone, G., Kuti, A., & Vancza, J. (2017). Wireless Multi-Sensor Networks for Smart Cities: A Prototype System with Statistical Data Analysis. *IEEE Sensors Journal*, *17*(23), 7667–7676. https://doi.org/10.1109/JSEN.2017.2736785

[17] Dancho, M. & Vaughan, D. (2022). Anomalize: Tidy Anomaly Detection. R package version 0.2.2. https://cran.r-project.org/package = anomalize

[18] Ding, Z., Mei, G., Cuomo, S., Li, Y., & Xu, N. (2020). Comparison of Estimating Missing Values in IoT Time Series Data Using Different Interpolation Algorithms. *International Journal of Parallel Programming*, 48(3), 534–548. https://doi.org/10.1007/s10766-018-0595-5

[19] Elbaz, K., Hoteit, I., Shaban, W. M., & Shen, S. L. (2023). Spatiotemporal air quality forecasting and health risk assessment over smart city of NEOM. *Chemosphere*, *313*(December 2022), 137636. https://doi.org/10.1016/j.chemosphere.2022.137636

[20] Elmenshawy, D., & Helmy, W. (2018). Detection Techniques of Data Anomalies in Iot: A Literature Survey. *International Journal of Civil Engineering and Technology (IJCIET)*, *9*(12), 794–807. http://www.iaeme.com/IJCIET/issues.asp?JType = IJCIET&VType = 9&IType = 12 1.

[21] Esnaola-Gonzalez, I., Bermúdez, J., Fernández, I., Fernández, S., & Arnaiz, A. (2017). Towards a semantic outlier detection framework in wireless sensor networks. *ACM International Conference Proceeding Series*, *2017-September*, 152–159. https://doi.org/10.1145/3132218.3132226

[22] Fan, Z., Feng, H., Jiang, J., Zhao, C., Jiang, N., Wang, W., & Zeng, F. (2020). Monte Carlo optimization for sliding window size in Dixon quality control of environmental monitoring time series data. *Applied Sciences (Switzerland)*, *10*(5). https://doi.org/10.3390/app10051876

[23] Feenstra, B., Collier-Oxandale, A., Papapostolou, V., Cocker, D., & Polidori, A. (2020). The AirSensor open-source R-package and DataViewer web application for interpreting community data collected by low-cost sensor networks. *Environmental Modelling and Software*, *134*(August), 104832. https://doi.org/10.1016/j.envsoft.2020.104832

[24] Fenger, J. (1999). Urban air quality. *Atmospheric Environment*, *33*(29), 4877–4900. https://doi.org/10.1016/S1352-2310(99)00290-3

[25] Kelly, K. E., Whitaker, J., Petty, A., Widmer, C., Dybwad, A., Sleeth, D., Martin, R., & Butterfield, A. (2017). Ambient and laboratory evaluation of a low-cost particulate matter sensor. *Environmental Pollution*, *221*, 491–500. https://doi.org/10.1016/j.envpol.2016.12.039

[26] Krishnamurthi, R., Kumar, A., Gopinathan, D., Nayyar, A., & Qureshi, B. (2020). An overview of iot sensor data processing, fusion, and analysis techniques. *Sensors (Switzerland)*, *20*(21), 1–23. https://doi.org/10.3390/s20216076

[27] Liang, C. J., & Yu, P. R. (2021). Assessment and improvement of two low-cost particulate matter sensor systems by using spatial interpolation data from air quality monitoring stations. *Atmosphere*, *12*(3). https://doi.org/10.3390/atmos12030300

[28] Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers and Chemical Engineering*, *28*(9), 1635–1647. https://doi.org/10.1016/j.compchemeng.2004.01.009

[29] Liu, W., Jiang, H., Che, D., Chen, L., & Jiang, Q. (2020). A real-time temperature anomaly detection method for iot data. *IoTBDS 2020 - Proceedings of the 5th International Conference on Internet of Things, Big Data and Security*, *IoTBDS*, 112–118. https://doi.org/10.5220/0009410001120118

[30] Loo, B. P. Y., & Tang, W. S. M. (2019). "Mapping" Smart Cities. *Journal of Urban Technology, 26*(2), 129–146. https://doi.org/10.1080/10630732.2019.1576467

[31] Lovén, L., Peltonen, E., Pandya, A., Leppanen, T., Gilman, E., Pirttikangas, S., & Riekki, J. (2019). Towards EDISON: an edge-native approach to distributed interpolation of environmental data. *Proceedings - International Conference on Computer Communications and Networks, ICCCN*, *2019-July*(May). https://doi.org/10.1109/ICCCN.2019.8847121

[32] Ma, L., Gu, X., & Wang, B. (2017). Correction of outliers in temperature time series based on sliding

window prediction in meteorological sensor network. *Information (Switzerland)*, *8*(2). https://doi.org/10.3390/info8020060

[33] Merello, P., García-Diego, F.-J., & Zarzo, M. (2014). Diagnosis of abnormal patterns in multivariate microclimate monitoring: A case study of an open-air archaeological site in Pompeii (Italy). *Science of The Total Environment*, *488–489*(1), 14–25. https://doi.org/10.1016/j.scitotenv.2014.04.068

[34] Nilson, B., Jackson, P. L., Schiller, C. L., & Parsons, M. T. (2022). Development and evaluation of correction models for a low-cost fine particulate matter monitor. *Atmospheric Measurement Techniques*, *15*(11), 3315–3328. https://doi.org/10.5194/amt-15-3315-2022

[35] Ogasawara, E., Martinez, L. C., De Oliveira, D., Zimbrão, G., Pappa, G. L., & Mattoso, M. (2010). Adaptive Normalization: A novel data normalization approach for non-stationary time series. In *Proceedings of the International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN.2010.5596746

[36] Pastorio, A. F., Spanhol, F. A., Martins, L. D., & De Camargo, E. T. (2022). A Machine Learning-Based Approach to Calibrate Low-Cost Particulate Matter Sensors. *Brazilian Symposium on Computing System Engineering, SBESC*, *2022-November*. https://doi.org/10.1109/SBESC56799.2022.9964983

[37] Pereira, F. C., Gonçalves, A. M., & Costa, M. (2023). Short-term forecast improvement of maximum temperature by state-space model approach: the study case of the TO CHAIR project. *Stochastic Environmental Research and Risk Assessment*, *37*(1), 219–231. https://doi.org/10.1007/s00477-022-02290-3

[38] Pinder, R. W., Klopp, J. M., Kleiman, G., Hagler, G. S. W., Awe, Y., & Terry, S. (2019). Opportunities and challenges for filling the air quality data gap in low- and middle-income countries. *Atmospheric Environment*, *215*(June). https://doi.org/10.1016/j.atmosenv.2019.06.032

[39] PurpleAir. (2022). *Download Data with the Sensor Data Download Tool*. https://community.purpleair.com/t/download-data-with-the-sensor-data-download-tool/316

[40] Rosner, B. (1983). Percentage points for a generalized esd many-outlier procedure. *Technometrics*, *25*(2), 165–172. https://doi.org/10.1080/00401706.1983.10487848

[41] Samara, M. Al, Bennis, I., Abouaissa, A., & Lorenz, P. (2022). A Survey of Outlier Detection Techniques in IoT: Review and Classification. *Journal of Sensor and Actuator Networks*, *11*(1), 4. https://doi.org/10.3390/jsan11010004

[42] Sayeed, M. S., Abdulrahim, H., Abdul Razak, S. F., Bukar, U. A., & Yogarayan, S. (2023). IoT Raspberry Pi Based Smart Parking System with Weighted K-Nearest Neighbours Approach. *Civil Engineering Journal*, *9*(8), 1991–2011. https://doi.org/10.28991/CEJ-2023-09-08-012

[43] Schilt, U., Barahona, B., Buck, R., Meyer, P., Kappani, P., Möckli, Y., Meyer, M., & Schuetz, P. (2023). Low-Cost Sensor Node for Air Quality Monitoring: Field Tests and Validation of Particulate Matter Measurements. *Sensors*, *23*(2), 1–29. https://doi.org/10.3390/s23020794

[44] Sharma, B., Sharma, L., & Lal, C. (2019). Anomaly Detection Techniques using Deep Learning in IoT: A Survey. *Proceedings of 2019 International Conference on Computational Intelligence and Knowledge Economy, ICCIKE 2019*, 146–149. https://doi.org/10.1109/ICCIKE47802.2019.9004362

[45] Stavroulas, I., Grivas, G., Michalopoulos, P., Liakakou, E., Bougiatioti, A., Kalkavouras, P., Fameli, K. M., Hatzianastassiou, N., Mihalopoulos, N., & Gerasopoulos, E. (2020). Field evaluation of low-cost PM sensors (Purple Air PA-II) Under variable urban air quality conditions, in Greece. *Atmosphere*, *11*(9). https://doi.org/10.3390/atmos11090926

[46] Tahmasseby, S. (2022). The Implementation of Smart Mobility for Smart Cities: A Case Study in Qatar. *Civil Engineering Journal*, *8*(10), 2154–2171. https://doi.org/10.28991/CEJ-2022-08-10-09

[47] Yigitcanlar, T., Li, R. Y. M., Inkinen, T., & Paz, A. (2022). Public Perceptions on Application Areas and Adoption Challenges of AI in Urban Services. *Emerging Science Journal*, *6*(6), 1199–1236. https://doi.org/10.28991/ESJ-2022-06-06-01