Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Low-cost sensors and Machine Learning aid in identifying environmental factors affecting particulate matter emitted by household heating

Amirhossein Hassani [a,*], Sebastian Bykuć [b], Philipp Schneider [a], Paweł Zawadzki [b], Patryk Chaja [b], Núria Castell [a]

[a] NILU - The Climate and Environmental Research Institute, P.O. Box 100, Kjeller, 2027, Norway
[b] Institute of Fluid-Flow Machinery Polish Academy of Sciences, Fiszera 14, Gdańsk, 80231, Poland

HIGHLIGHTS

- PM variability is analyzed using residents' network of low-cost sensors.
- A data quality assurance scheme for PM sensors is proposed.
- Relative Humidity-induced uncertainty of PM sensors is estimated using a new approach.
- $PM_{2.5}$ and environmental settings relations are explored with an explainable ML model.
- Sensors identify fossil-fuel-induced air pollution hotspots.

ARTICLE INFO

Keywords:
Poland air pollution
Citizen engagement
Household heating
Fossil-fuel combustion
Meteorological variables
Generalized additive model
LIME analysis

ABSTRACT

Poland continues to rely heavily on coal and fossil fuels for household heating, despite efforts to reduce Particulate Matter (PM) levels. The availability of reliable air quality data is essential for policymakers, environmentalists, and citizens to advocate for cleaner energy sources. However, Polish air quality monitoring is challenging due to the limited coverage of reference stations and outdated equipment. Here, we report the results of a study on the spatio-temporal variability of Particulate Matter in Legionowo, Poland, using residents' network of low-cost sensors. Along with identifying the hotspots of household-emitted PM, (1) we propose a data quality assurance scheme for PM sensors, (2) suggest an approach for estimating the Relative Humidity-induced uncertainty in the sensors without co-location with reference instruments, and (3) develop an interpretable Machine Learning (ML) model, a Generalized Additive Model (RMSE = 6.16 μg m^{-3}, and $R^2 = 0.88$), for unveiling the underlying relations between $PM_{2.5}$ levels and other environmental parameters. The results in Legionowo suggest that as air temperature and wind speed increase by 1 °C and 1 km h^{-1}, $PM_{2.5}$ would respectively decrease by 0.26 μg m^{-3} and 0.14 μg m^{-3} while $PM_{2.5}$ increases by 0.03 μg m^{-3} as RH increases by 1%.

1. Introduction

Using solid fuels for electrical generation and heating has caused Poland to have some of Europe's highest air pollution levels (Carvalho, 2019; Junninen et al., 2009; Kerimray et al., 2017). Emissions of particles in Poland are primarily caused by "low-altitude" emitters, i.e., the ones with a height of 40 m or less (Wielgosiński and Czerwińska, 2020). During the winter months, heating activities are one of the primary sources of Particulate Matter (PM) in Poland's ambient air (Reizer and Juda-Rezler, 2016). Inhabitants often use coal and other fossil fuels as primary heating sources (Holnicki et al., 2022; Mrozowska et al., 2021; Sokołowski and Bouzarovski, 2022). Approximately 83–86% of heat in Poland is produced by solid fuels and 8–12% by natural gas (Nyga-Łukaszewska et al., 2020). Stoves and boilers that burn solid fuels in inefficient ways can release high levels of particulate matter, such as $PM_{2.5}$ (PM smaller than 2.5 μm) and PM_{10} (PM smaller than 10 μm) (Xu et al., 2011).

In 2018, around 47% of the total PM_{10} dust emissions and 52% of the total $PM_{2.5}$ emissions in Poland came from solid-fuel burning sources (Holnicki et al., 2022). Among other pollutants, coal combustion

[*] Corresponding author.
E-mail address: ahas@nilu.no (A. Hassani).

https://doi.org/10.1016/j.atmosenv.2023.120108
Received 21 March 2023; Received in revised form 7 August 2023; Accepted 21 September 2023
Available online 23 September 2023

produces nitrogen oxides, sulfur dioxide, and mercury (Xu et al., 2011; Zhao et al., 2019). In addition to respiratory illnesses and cardiovascular disease (Kuźma et al., 2021), there is also a risk of lung cancer and premature death caused by these pollutants (Badyda et al., 2017; Finkelman, 2007; Munawer, 2018). Using green technologies instead of coal and eliminating fossil fuel-fired boilers will improve air quality, reducing particulate air pollution-related premature mortality (Mrozowska et al., 2021).

The Polish government has implemented several policies and regulations to reduce PM levels in the ambient air (Brauers and Oei, 2020). A few of them are to encourage greater use of cleaner heating technologies, such as natural gas, and to encourage homes to replace old heating systems with newer, more efficient ones (Jagiełło et al., 2022; Sokołowski and Bouzarovski, 2022). The Polish government has instituted measures like restricting fossil-fuel burning during high PM episodes, air quality monitoring, and public awareness campaigns (Jagiełło et al., 2022; Piwowar and Dzikuć, 2019). However, Polish households continue to have the highest coal consumption rate across the European Union (Sokołowski and Bouzarovski, 2022). Air quality improvement efforts often face challenges due to a lack of public awareness about the health risks of air pollution (Sokołowski and Bouzarovski, 2022).

Providing air quality data to support policies for reducing air pollution is crucial to eliminating fossil fuel-fired boilers in Poland (Attia et al., 2022). Data can be used by engaged citizens, environmental organizations, and administrative agencies to advocate for cleaner energy sources and policies that encourage the transition from fossil fuels to cleaner energy (Attia et al., 2022). Systematic air quality monitoring can contribute to identifying pollution hotspots, measuring the impact of policy interventions, and raising public awareness about pollution caused by fossil fuel-burning boilers (Nadarajah, 2008; Snyder et al., 2013).

However, monitoring air quality in Poland is challenging for several reasons (Parascandola, 2018). Pollution levels can be highly local in some episodes and vary significantly between neighborhoods, making it challenging to monitor air quality accurately. Polish air quality monitoring stations are distributed across the country, but their coverage is not comprehensive, and some regions do not have any monitoring stations. Additionally, some stations' air quality monitoring equipment needs to be upgraded, which can be costly and time-consuming.

Air quality sensors are cost-effective solutions to monitor air pollution at higher geospatial resolutions in real-time (Alfano et al., 2020; Castell et al., 2017; Snyder et al., 2013). The relationship between air pollution levels, heating activities, and air pollution hotspots can be detected/investigated using official monitoring data complemented with data from low-cost sensor networks (Giordano et al., 2021; Morawska et al., 2018). Due to their small size, portability, and ease of installation, low-cost sensor networks allow community members to collect data about pollution levels in their neighborhoods, thus raising their awareness about air quality (Kumar et al., 2015; Watne et al., 2021). Ambient parameters such as air temperature, humidity, and cross-sensitivity with other pollutants, however, can affect the accuracy of the measurements, and calibration and maintenance may be necessary to maintain consistency and reliability (Kang et al., 2022; Karagulian et al., 2019; Stavroulas et al., 2020). Although the accuracy and reliability of these devices may not be as high as those of traditional air quality monitoring systems (Bulot et al., 2019), they can provide valuable information about air quality when official monitoring stations are unavailable or standard monitoring equipment is prohibitively expensive (Castell et al., 2017; Lim et al., 2019).

Low-cost PM sensors have gained popularity among community groups, researchers, and individuals interested in monitoring air pollution due to some general advantages of PM low-cost sensors as compared to low-cost gaseous sensors (Kang et al., 2022; Mahajan et al., 2020; Wesseling et al., 2019). Generally, a PM sensor is less sensitive to temperature changes and cross-sensitivity with other pollutants than a gaseous sensor (Kang et al., 2022; Rai et al., 2017). The long-term maintenance requirements of PM sensors are also typically lower than those of gaseous sensors, as PM sensors are more stable over time and less prone to drift and degradation (Lee et al., 2020; Liu et al., 2020; Malings et al., 2020; Sayahi et al., 2019). Sensors and calibration methods for low-cost PM sensors are continually improving, and new sensors are being developed (Giordano et al., 2021). A PM sensor typically measures the amount of light scattered by air particles using light sources and detectors. In the presence of PM in the air, light is scattered or absorbed in proportion to its concentration.

Our goal in this study was to identify air pollution hotspots and to investigate the relationship between air pollution levels ($PM_{2.5}$ in particular) and heating activities in the town of Legionowo, Poland, by combining historical data from a network of air-quality sensor systems (Airly) operated by citizens, data from the official air quality stations, and meteorological data. Monitoring air quality data using the low-cost sensor network allowed us to identify periods of high pollution levels and correlate them with heating activities. Additionally, by developing a statistical Machine Learning (ML) model for predicting $PM_{2.5}$ levels, we quantified the effects of environmental factors, such as Relative Humidity (RH), air temperature, and wind intensity, on the dispersion of PM pollutants from heating activities. We also identified periods of the day or week when air pollution levels are especially high due to household heating activities.

We focused here on Legionowo, the pilot town in the GREEN HEAT – "Towards Collaborative Local Decarbonization" project (https://greenheat.kezo.pl/en/, accessed in Feb 2023), and the sensor data are primarily gathered by a selected Pilot Case in a chosen community in the town of Legionowo in Poland. The methodology developed here based on the procedures established during the Pilot Case analysis has the potential for future implementation in other cities in Poland. PM air quality data provided here may aid in developing effective strategies for reducing emissions and improving air quality, for example, by identifying specific areas where heating activities contribute more to air pollution and prioritizing strategies to change to less polluting heating sources than currently still prevalent coal-fired boilers.

The focus was not solely on developing and comparing an ML model. Instead, the ML model serves as a tool to understand complex relationships between environmental parameters and household-related $PM_{2.5}$. Our main emphasis is interpreting the model's outcomes and revealing underlying connections rather than just comparing it to other ML models or predicting $PM_{2.5}$. Furthermore, our research introduces a method for estimating the uncertainty in sensor measurements due to RH variations while the sensors were in the field. The technique has practical implications and can benefit other sensor networks in comparable towns or cities, where RH fluctuations can also impact sensor readings. Additionally, we propose and apply a data assurance scheme outlined in the "Data Quality Assurance" section. The proposed three-stage pre-processing (filtering) scheme addresses data reliability concerns without needing sensor co-location. The method can potentially be applied to similar sensor networks in various locations, extending its usefulness beyond the context of Legionowo.

2. Materials and methods

2.1. Study location

Legionowo is situated in the Masovian province in central Poland, about 23 km north of Warsaw, Poland's capital. Legionowo is approximately located at 52.4012° N and 20.9369° E. The town has about 55,000 residents (2019) and covers about 15 square kilometers (https://bdl.stat.gov.pl/bdl/dane/teryt/jednostka). Legionowo has a relatively flat terrain with no significant hills or mountains nearby, where 100 m above sea level is the highest point in the town. The climate in Legionowo is temperate continental, with cold winters and mild summers (Beck et al., 2018; Kundzewicz and Matczak, 2012), and an average temperature of −3 °C in January, the coldest month, and

18 °C in July, the hottest month. The town receives the most rainfall in June, July, and August.

Houses in Poland are heated using a variety of energy sources, with coal (40.28%), district heating (31.1%), wood (12.04%), and gas (11.76%) as the most dominant ones (Karpinska et al., 2021). According to local experts, the most popular option for multi-family buildings in Legionowo is to have central heating provided by a district heating system. Multi-family and public buildings not connected to the heating network are supplied with heat using individual boilers and heating systems. Natural gas is most often used for heating purposes, but also, to a lesser extent, hard coal, wood, electricity, and heat pumps. Single-family buildings in Legionowo are supplied with heat from individual heating systems powered mainly by natural gas. Also, coal, wood, heating oil, and wood pellets are burned. Heat pumps are becoming more popular.

2.2. Network of citizen-operated sensors

Data from 14 low-cost Airly PM sensor systems (https://airly.org/en/, accessed in February 2023), installed in Legionowo formed this analysis's basis. Legionowo residents operated all sensors. Five Airly sensor systems (IDs: 6436, 6437, 31103, 86548, and 86697) were provided by the Airly company. Sensors 86548 and 86697 were serviced in the summer of 2021, resulting in data gaps. The rest of the sensor systems were installed (1.5–8 m above ground) gradually during 2022 in the pilot (Legionowo) as part of the GREEN HEAT project, particularly in areas with a high density of fossil-fueled boilers. The locations of the low-cost sensors were chosen based on a bottom-up citizen science approach. We engaged with the local community through a series of workshops and online resources, where citizens had the opportunity to express their interest in participating in the project and installing sensors. As a result, the sensor deployment locations were determined by the eagerness and voluntary participation of the citizens. Citizens were provided guidance on proper sensor placement, maintenance, power supply monitoring, and the use of weather-resistant enclosures to ensure accurate measurements.

According to the manufacturer, in the Airly sensor system (kit), PM mass concentrations are measured in the fractions of PM_1 (particle effective range: 0–500 μg m^{-3}), $PM_{2.5}$ (0–1000 μg m^{-3}), and PM_{10} (0–1000 μg m^{-3}) as outputs, all with claimed accuracies of ±10 μg m^{-3} (https://airly.org/en/EN_AIRLY_PRODUCTCARD_SENSOR_GEN2_2022.pdf). The kit is water-resistant and weighs 440 g with dimensions of 74 × 77 × 83.5 mm. A Plantower PMS5003 sensor (https://www.plantower.com/en/products_33/74.html) is integrated into the Airly sensor system, which uses laser-based light scattering technology for measuring PM concentration in the air (aerodynamic diameters of 0.3–10 μm).

Other environmental parameters such as temperature (DHT22-Thermistor, measurement range: −40 °C–85 °C, accuracy: ±0.5 °C, resolution: 0.1 °C), RH (DHT22-Capacitive, measurement range: 0–100%, accuracy: ±3% RH, resolution: 0.1%), and air pressure (BMP280, measurement range: 700–1200 hPa, accuracy: ±1 hPa) are also measured by the sensor kits. Sensor data is collected and processed in real-time by the Airly Cloud. The historical data at hourly resolution is available through the Airly Data Platform (https://app.airly.org/). The platform additionally provides wind speed data from the Dark Sky's forecast technology for individual sensors.

The study's sensors were strategically placed in residential areas with less traffic, where solid/fossil fuel burning for space heating is more prevalent. By locating the sensors near emission sources of solid fuel burning, such as in households' backyards, our goal was to capture $PM_{2.5}$ levels more heavily influenced by solid/fossil fuel burning. While it may not be feasible to entirely distinguish the impact of solid/fossil fuel burning emissions from other $PM_{2.5}$ pollution sources like background levels, we believe that residential solid fuel burning likely contributes significantly to the measured $PM_{2.5}$ levels due to the proximity of the low-cost sensors to these emission sources.

Hofman et al. (2022) evaluated the performance of three Airly PM sensor systems against a reference grade instrument (FIDAS 200, Palas) on the roof of a regulatory air quality monitoring station in Antwerp, Belgium, over two 14-day co-location periods. The statistical analysis of the first co-location campaign showed that measurements from the sensor are correlated with those from the reference measurements (Table 1). Additionally, calculations of Min-Max correlations and MAEs between the considered sensor units confirmed good intra-sensor performance (referring to Table 3 in their paper).

Similarly, Vogt et al. (2021) evaluated the performance of three Airly PM sensor systems against the FIDAS optical reference-equivalent instrument, co-located at the Kirkeveien air quality station located in Oslo, Norway (28 August to 19 October 2020). The RMSE was between 4.39 and 10.39 μg m^{-3} following the implementation of sensor-specific multi-linear regression models to the measured $PM_{2.5}$ values. The Sensor-to-sensor intercomparison of the factory-calibrated sensors showed correlations between 0.89 and 0.96.

We undertook an inter-comparison test among sensors to assess the sensor-to-sensor consistency and correlation between the measurements obtained from different sensors. The test involved 16 sensors strategically placed at one site. Among the sensors tested, only one sensor displayed relatively lower performance, with a correlation coefficient less than 0.98 with other sensor measurements, although even for that sensor, the sensor-to-sensor correlations remained above 0.78. The inter-comparison test was conducted in Kirkeveien, Oslo, spanning almost 20 days. Despite using different sensors in Legionowo and Oslo, the results indicate that the Airly brand low-cost PM sensors exhibit low sensor-to-sensor variability.

2.3. Data quality assurance

There is only one air quality monitoring station in Legionowo, *Zegrzynska*, with an elevation of 91 m. and coordinates of 52.4075° N, 20.9559° E (International code: PL0129A). The station is operated by the Chief Inspectorate of Environmental Protection of Poland and monitors $PM_{2.5}$ and PM_{10} components using a Grimm EDM180 optical dust reference-grade instrument at hourly resolution. Sensors were, on average, 1,848 m from the *Zegrzynska* reference station. Despite its proximity to a national road (*Ulica Warszawska*), the station is classified as a sub-urban background station due to its sharp elevation from the roadside (https://powietrze.gios.gov.pl/pjp/current/station_details/info/471?lang=en).

PM ($PM_{2.5}$ and PM_{10}) and environmental data (air temperature, RH, wind speed) from the 14 sensors installed in different neighborhoods of Legionowo were retrieved from January 2020 to January 2023. The quality assurance of the network data was performed through a three-stage pre-processing (filtering) scheme. In the first step, the sensor's data for a specific month were removed if its $PM_{2.5}$ (and PM_{10}) data coverage during that particular month was less than 75%. The second step was to calculate the Pearson Correlation Coefficient (r) between each sensor's hourly $PM_{2.5}$ and PM_{10} measurements during the noon/afternoon hours — 10:00 until 15:59 — and the average corresponding hourly $PM_{2.5}$ and PM_{10} measurements of all sensors in a month (including the reference station), followed by the removal of the sensor-month data with $r \leq 0.75$. The reasoning behind the second step is that

Table 1
Evaluating the performance of three Airly PM sensor systems against a reference-grade instrument (FIDAS 200, Palas).

Component	Root Mean Square Error (μg m^{-3})	Mean Absolute Error (μg m^{-3})	Pearson's Liner Correlation Coefficient
PM_1	3.14–5.31	2.50–3.95	0.91–0.94
$PM_{2.5}$	11.72–16.45	9.63–13.58	0.89–0.92
PM_{10}	13.71–20.37	11.28–16.28	0.72–0.75

we assumed these hours' PM levels are less affected by anthropogenic activities, referring to the diurnal difference of the sensor measurements from the official *Zegrzynska* station data (Supplementary Fig. 1). Moreover, the assumption was that the *Zegrzynska* official data represents the background air pollution in the study region, and the sensors measure the PM levels added to the background pollution. Thus, they are highly correlated with the official measurements.

Other studies on environmental sensor network data analysis have utilized the correlation coefficient of individual sensors with the sensor network. Fu et al. (2023) removed sites with substantially lower correlations with any other site, setting a threshold of the Pearson Correlation Coefficient to be lower than μ - 3σ (μ: mean, σ: standard deviation) of all the correlations between sites.

Applying steps one and two reduced the number of hourly measurements from 176,983 to 168,955, with the majority of lost data related to only one sensor with limited data coverage. The third stage was sensor drift diagnosis. We fitted a linear model to the absolute deviation of the monthly mean $PM_{2.5}$/PM_{10} measured by each sensor from the monthly mean of $PM_{2.5}$/PM_{10} *Zegrzynska* official data. If the fitted model (slope in particular) was statistically significant (*p*-value for the F-test on the model $\leq$0.05), that sensor was flagged as susceptible to sensor drift. For both $PM_{2.5}$ and PM_{10}, only sensor ID 31103 was flagged following this step (Supplementary Fig. 2). Further analysis of sensor ID 31103 signal, as compared to the rest of the sensors and official data time series did not persuade us to diagnose it as a degraded sensor; accordingly, its measured data remained in the analysis, and all the sensors' data passed the third filter. Data coverage of the 13 remaining sensors is represented in Supplementary Fig. 3.

PM can be highly variable spatially, so the second step may filter the real PM measurements. To ensure the robustness of the second step of the proposed data-filtering approach, we divided the whole official station data available from the AirBase dataset — provided by the European Environment Agency (https://www.eea.europa.eu/data-and-maps/data/aqereporting-9) — into grid cells of 4 km size (Fig. 1). All stations within a grid cell were assumed to be an imaginary network with a unique ID. Respectively, only 1.71% and 4.39% of the calculated station-month-network correlation values were less than 0.75 (station-month-network correlation: for each month, with at least 75% data coverage, we calculated the monthly correlation (*r*) between hourly $PM_{2.5}$/PM_{10} data of each reference station and the mean hourly $PM_{2.5}$/PM_{10} measurements of all its nearby stations, i.e., reference stations located in the same imaginary network). As we did not distinguish the type of reference stations (traffic, industrial, and background/urban and suburban), this could be a convincing case that the nearby $PM_{2.5}$ and PM_{10} measurements are typically correlated, and the sensor-month data of low correlation with the network mean can be removed from the analysis.

In summary, firstly, data with less than 75% coverage for a specific month were removed for $PM_{2.5}$. Secondly, the *r* was calculated between each sensor's hourly $PM_{2.5}$ measurements during particular hours (10:00–15:59) and the average of all sensors' measurements in a month. Sensor-month data with $r \leq 0.75$ were removed. The third stage addressed sensor drift by fitting a linear line to the deviation of monthly mean $PM_{2.5}$ measured by each sensor from the official station data.

We did similar pre-processing/data filtering on meteorological measurements (temperature and RH); however, the monthly-sensor correlation threshold was 0.9, as we assumed those parameters were more correlated in the geographical scales of Legionowo. Additionally, data from the nearest official meteorological station — Modlin station (52.4511° N, 20.6517° E) —were retrieved in the weather station's FM-15 Surface Meteorological Airways Format from the Integrated Surface Dataset (Global) of the National Centers for Environmental Information during the analysis period (https://www.ncei.noaa.gov/access/search/data-search/global-hourly, accessed in Feb 2023). The days with a data coverage of less than 75% were removed. The descriptive statistics of the wind, air temperature, and RH variation recorded at Modlin station are represented in Supplementary Fig. 4.

2.4. Estimating uncertainty resulting from Relative Humidity (RH) effects

Many low-cost sensors experience biases when RH exceeds certain values, as mentioned in the introduction (Brattich et al., 2020; Jayaratne et al., 2018; Kang et al., 2022). In the Airly sensor co-location conducted by Vogt et al. (2021) in Oslo, absolute Bias from reference grade instruments increased with RH values over 70%. Because we had no co-location periods, we could not directly evaluate the RH-induced uncertainty. Instead, we focused on the May–Sep period when we assumed the anthropogenic heating activities are at the lowest and the $PM_{2.5}$/PM_{10} values measured by the sensors might be very close to the *Zegrzynska* measurements (Supplementary Fig. 5 to Supplementary Fig. 7). We further limited the analysis to hours of the day between 10:00 and 15:59 when the sensors show the lowest deviation from the official measurement Supplementary Fig. 1. For the remaining sensor hour measurements, we calculated the sensor deviation from the official data of the *Zegrzynska* station (sensor measurement minus official data). We removed the outliers defined as data more than three scaled MAD (Mean Absolute Deviation) from the median. The scaled MAD was $c \times median(abs(data - median(data)))$, where $c = -1/(sqrt(2) \times erfcinv(3/2))$.

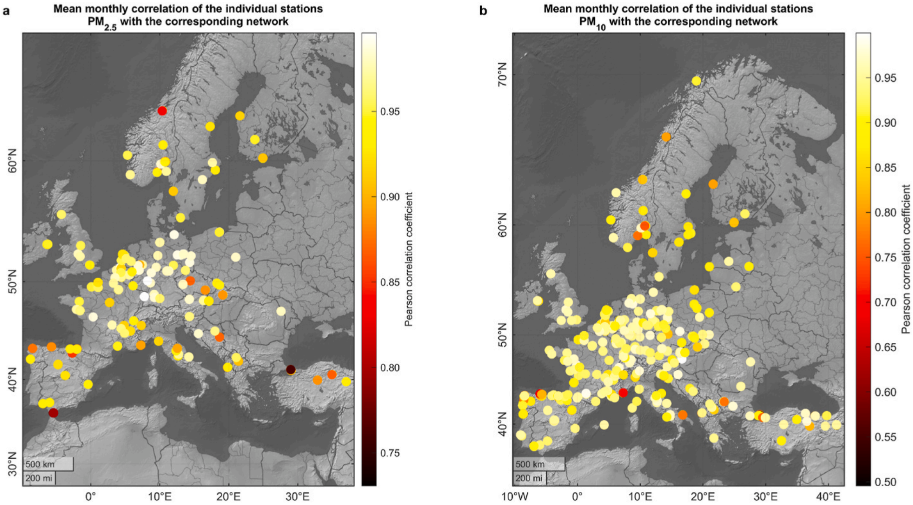
a
Mean monthly correlation of the individual stations
$PM_{2.5}$ with the corresponding network
60°N
50°N
40°N
30°N
0°
10°E
20°E
30°E
500 km
200 mi
0.95
0.90
0.85
0.80
0.75
Pearson correlation coefficient
b
Mean monthly correlation of the individual stations
PM_{10} with the corresponding network
70°N
60°N
50°N
40°N
10°W
0°
10°E
20°E
30°E
40°E
500 km
200 mi
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
Pearson correlation coefficient

Fig. 1. Mean monthly correlation of the $PM_{2.5}$ and PM_{10} data measured at air pollution monitoring stations with corresponding network mean measurements across Europe. A network here indicates all reference stations within a grid cell with a 4 km spatial resolution.

The corresponding RH values measured by sensors were divided into five bins (50:10:100%, left bin edge included), and for each bin, we fitted a normal distribution to deviation values. The Cumulative Distribution Function (CDF) of deviation for each RH bin is plotted in Fig. 2 for both $PM_{2.5}$ and PM_{10} components. As expected, the CDFs show the probability of deviation from official data increases by RH. The dependency of PM_{10} measurements on RH is more than $PM_{2.5}$ and PM_{10} measurements, showing a higher deviation from the reference measurements at higher RH values.

2.5. *Predicting the $PM_{2.5}$ using statistical modeling*

To further investigate the relationship between the PM levels and the economic-environmental parameters in Legionowo, we developed a Generalized Additive Model (GAM) for Regression (T. Hastie and Tibshirani, 1987; T. J. Hastie, 2017; Lou et al., 2012). To effectively capture the complex interactions affecting $PM_{2.5}$ and handle missing values, we chose GAM over more interpretable models like Generalized Mixed Linear Models (Lou et al., 2013). We avoided using more complex models like boosted ensemble trees or Neural Networks to prioritize interpretability, gaining insights into local predictor contributions. This approach provides more transparent and more interpretable results.

Using the GAM, we could predict PM values while balancing speed, interpretability, and flexibility. Simply put, we aimed to train an interpretable ML model — GAM — using a set of predictors (detailed below) to predict PM as an air quality indicator. Local and global interpretations of the trained model unveil the relationship between the predictors and air quality levels. Several studies have shown that PMS5003 — integrated into Airly sensor systems — has a poor performance for measuring PM_{10}, as compared to $PM_{2.5}$ (Cavaliere et al., 2018; Kuula et al., 2020; Sayahi et al., 2019; Tagle et al., 2020; Vogt et al., 2021). The RH uncertainty analysis presented earlier shows similar behavior for Airly sensors. Thus, here we only focused on $PM_{2.5}$ as the target variable for prediction by GAM.

A GAM allows a non-linear relationship between the response variable and the predictors by adding the predictors' shape (smooth) functions to a linear model. The general form of a GAM can be written as (Lou et al., 2012):

$$y = \beta 0 + f1(x1) + f2(x2) + \ldots + fp(xp) + \varepsilon, (1)$$

Where y is the response variable, β0 is the intercept, f1, f2, …, fp are univariate shape functions of the predictors x1, x2, …, xp, and ε is the error term. Bivariate shape functions of important interaction terms can be used in a model to incorporate interactions between predictors (Lou et al., 2013). We used MATLAB's built-in function "fitrgam" to fit a GAM, which uses a gradient-boosting algorithm for building the shape functions (see https://uk.mathworks.com/help/stats/fitrgam.html).

Sensor ID, hourly RH measured by sensor (%), the temperature measured by sensor (°C), wind speed from Airly data platform (km h^{-1}), $PM_{2.5}$ measured at *Zegrzynska* reference station (μg m^{-3}), day-ahead energy price for Poland's bidding zone (Euro $MW^{-1}h^{-1}$) retrieved from ENTSO-E Transparency Platform (Hirth et al., 2018), the sample hour of measurement (0–23), and the sample month of measurement (1–12) were used as the predictors while fitting the GAM. Sensor ID was the only categorical predictor. We divided the time series of $PM_{2.5}$ measured by each sensor into the train (85%) and test (15%) sets.

The GAM was initially fitted to the train set. Later, the predictive fitted model quality was evaluated by comparing the model predictions against the sensor measurements during the test period. Additionally, the final trained model was cross-validated by a 10-fold cross-validation scheme. Accuracy metrics were calculated for the cross-validated model and individual sensors' test sets, including Coefficient of Determination (R^2), RMSE, MAE, and Bias. The model included the interaction terms and the maximum p-value for detecting interaction terms was set to 0.1. We used the "*bayesopt*" optimizer with the "*expected-improvement-per-second-plus*" Acquisition Function to optimize the model hyperparameters. The hyperparameters were optimized using a hold-out cross-validation scheme, with 25% of the data being held out. The

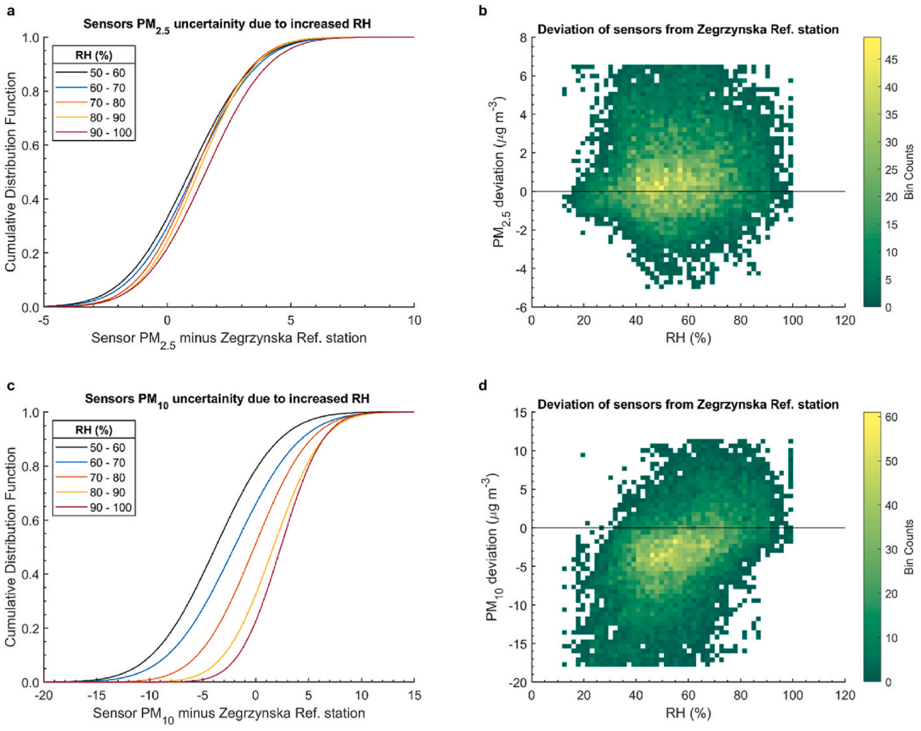

Fig. 2. Quantification of the Airly low-cost PM sensors uncertainty due to increased Relative Humidity, Legionowo, Poland. a and b, measured $PM_{2.5}$ components. c and d, measured PM_{10} uncertainty quantification. Deviation represents the factory-calibrated sensor output minus the reference instrument measurement. Cumulative Distribution Functions are estimated based on a normal distribution fitted to the sensors' deviation from the reference instrument during the year's warm months (May–Sep), at 10:00–15:59.

minimum objective function of hyperparameter optimization was not significantly different after 50 evaluations. It was assumed that the weight of observations in model training was constant, equal to one.

Accumulated Local Effect (ALE) plots (Apley and Zhu, 2016) were used for the global and Shapley value of the predictors (Lundberg and Lee, 2017), and LIME analysis (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016) were used for the local interpretation of the final fitted model — the relation between predictors and $PM_{2.5}$ concentrations.

3. Results and discussion

3.1. Spatio-temporal variability in PM

The time series of factory-calibrated $PM_{2.5}$ and PM_{10} measurements of the sensors and the *Zegrzynska* official data are presented in Fig. 3. The mean daily PM recorded by sensors between Jul 2022 and Dec 2022 is additionally represented in Supplementary Fig. 8 (the period with all sensors' data coverage). The air temperature measured at Modlin meteorological station is also plotted in Fig. 3. The role of air temperature and winter heating activities on high levels of PM pollution is evident. From Jan 2020 until Jan 2023, the average air temperature, official $PM_{2.5}$, and sensor $PM_{2.5}$ were 9.76 °C (daily min = 4.99 °C, daily max = 13.29 °C), 18.99 $\mu g\ m^{-3}$, and 20.04 $\mu g\ m^{-3}$, respectively. The respective *Zegrzynska* and sensors' average PM_{10}s were 25.38 $\mu g\ m^{-3}$ and 27.95 $\mu g\ m^{-3}$. The mean air temperature in each quarter of the year was Q1 = 2.03 °C (min = −20 °C, max = 20 °C), Q2 = 13.15 °C (min = −4 °C, max = 33 °C), Q3 = 17.8 °C (min = 1 °C, max = 35 °C), and Q4 = 5.2 °C (min = −16 °C, max = 23 °C).

The highest deviation of the sensors from the reference measurements occurs during the cold months (Oct–Apr). Similar trends are observed for Zgierz in central Poland by analyzing four years (2008–2011) of meteorological, atmospheric radon, and air quality observations (Chambers and Podstawczyńska, 2019). Air quality monitoring showed that high heating season emissions caused mean annual respective PM_{10} and $PM_{2.5}$ values of 33.6 and 21.13 $\mu g\ m^{-3}$ in Zgierz. The authors concluded that during the heating season (Oct–Mar), domestic emissions of SO_2 and PM could be substantial since centralized heating systems do not mainly heat houses in that region.

Sensor deviations from reference stations are the highest in Feb 2021 and Mar 2022. According to the Copernicus Atmospheric Monitoring Service, in addition to household heating activities, this high deviation can be attributed to the continent-scale high PM episodes during Feb 2021 and Mar 2022. A major inflow of Saharan air and significant dust caused daily mean PM_{10} concentrations of 50–100 $\mu g\ m^{-3}$ in a vast region of south and middle Europe between 19 and 27 February 2021 (Schulz et al., 2021). An extensive anticyclone with dry and stagnant conditions under a high-pressure system led to even higher PM levels in Northern Europe between 20 and 27 March 2022 (Tsyro et al., 2022). A higher concentration of ambient $PM_{2.5}$ increases uncertainties in PMS5003 measurements, according to the sensor reference manual (https://www.plantower.com/en/products_33/74.html, retrieved in Feb 2023; the error is ±10% @ 100–500 $\mu g\ m^{-3}$ while @ 0–100 $\mu g\ m^{-3}$, it is ±10 $\mu g\ m^{-3}$) and some previous studies, such as Hong et al. (2021) and Kang et al. (2022).

As the period with the best data coverage was Nov 2022–Jan 2023, we calculated the descriptive statistics of measured $PM_{2.5}$ and PM_{10}, represented in Fig. 4 and Supplementary Fig. 9. During that period, sensors 90531 and 97546, located respectively in the middle and on the edge of the town, showed the highest PM level, while the average $PM_{2.5}$ measured at the reference station was 24.7 $\mu g\ m^{-3}$. The spatial patterns observed in the intensity of measured $PM_{2.5}$ values during the three months cannot be easily generalized. We noticed that low and high concentrations can coexist at very close distances within the study area. This spatial variability suggests that $PM_{2.5}$ levels are influenced by local factors and sources, leading to distinct pockets of pollution close to each other.

However, the observed $PM_{2.5}$ averages can, to some extent, be explained by the building and landscape properties of Legionowo. High pollution levels were observed north of the town, including sensor 90531 (the highest PM levels), a neighborhood with old buildings, and single domestic house heating sources. Additionally, frequent

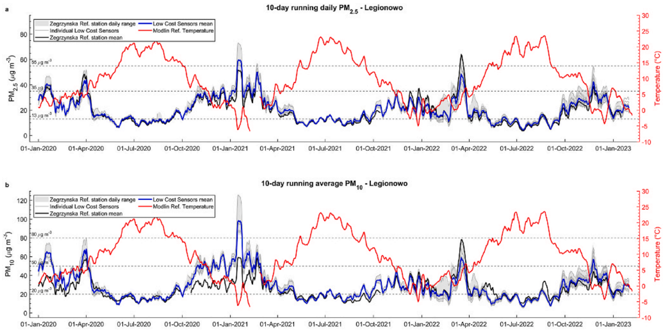

Fig. 3. 10-day running average of $PM_{2.5}$ and PM_{10} measured by Airly PM sensors, Legionowo, Poland. The air temperature data are retrieved from the nearby reference meteorological station, Modlin. Limits (horizontal lines) are daily acceptable levels of PM concentrations, according to Polish Chief Inspectorate for Environmental Protection..

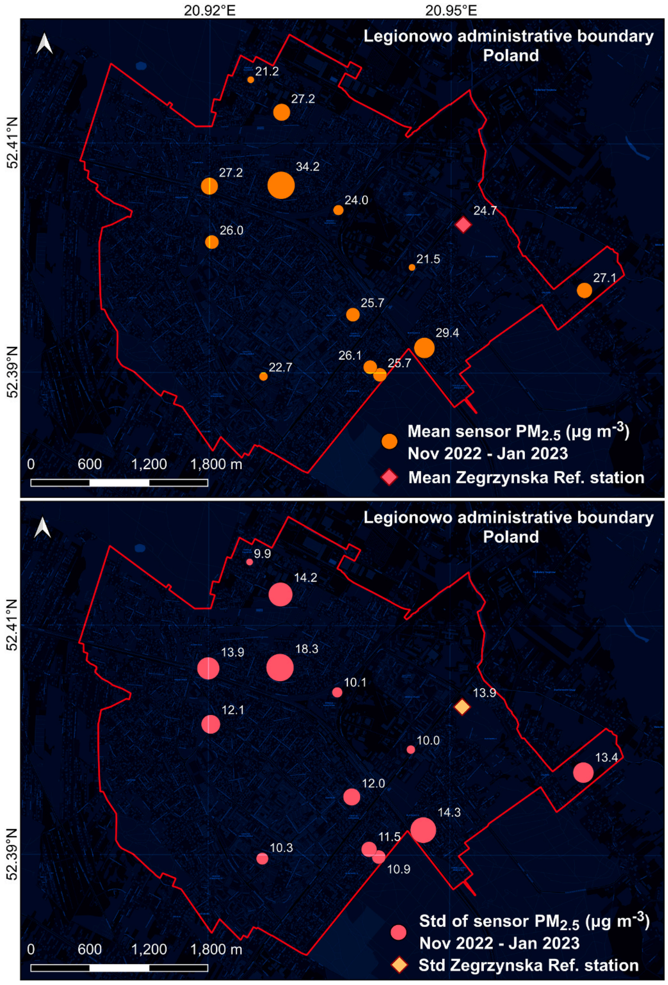

Fig. 4. Descriptive statistics of the measured $PM_{2.5}$ by low-cost Airly PM sensors, Legionowo, Poland. a, mean $PM_{2.5}$ concentration between Nov 2022 and Jan 2023. b, Standard deviation of the measured $PM_{2.5}$ concentrations between Nov 2022 and Jan 2023.

complaints of waste material/trash burning have been reported by the citizens in the region. In contrast, the neighborhoods where sensors 86548, 98115, and 6437 are located are regions with district heating and more detached building blocks — primarily central and south-western parts of the city (Supplementary Fig. 9). Sensors 96155 and 101796 were deliberately installed very close to each other as a measure to ensure the measurements' quality. The mean $PM_{2.5}$ values for those two sensors are very close (25.7 and 26.1 μg m^{-3}), confirming the measurements' reliability.

The time series of Polish day-ahead electricity prices (the final price for the electricity producers) against the $PM_{2.5}$ levels in Legionowo are presented in Supplementary Fig. 10. There is no statistically significant correlation between PM levels and electricity prices. The increase in day-ahead electricity prices, starting in mid-2021, does not impact the pollution levels. The lack of a relationship between the rise in day-ahead electricity prices and the $PM_{2.5}$ levels measured by sensors supports the idea that heating activities and PM levels primarily depend on fossil-fuel combustion in Legionowo. However, it should be considered that we used day-ahead prices (the price the producers requested), and consumers may have been tariffed differently, such as by receiving subsidies for electricity.

The diurnal and weekly cycle of $PM_{2.5}$ calculated using the low-cost sensors and the reference station is shown in Fig. 5. Two peaks in $PM_{2.5}$ concentrations (bimodal distribution) are observed for all sensors between 6:00 and 10:00 and 16:00 and 23:00 (local time). The enhancement factors in the diurnal cycle of air pollution for $PM_{2.5}$, referring to the ratio of the concentration of $PM_{2.5}$ during rush hours (16:00–23:00) to the concentration of $PM_{2.5}$ during other hours (associated with reduced human activities or background pollution levels), calculated by sensors and the reference station measurements were 1.22 and 1.30, respectively (Fig. 5a). There may be a lower boundary layer, a fumigation effect during rush-hour traffic early in the morning (American Meteorological Society, 2020), as well as house heating responsible for the peak in the morning (Kompalli et al., 2014; Singh et al., 2020). By creating stronger thermals after sunrise, the nighttime inversion is broken due to the fumigation effect, and aerosols stabilized in the residual layer are mixed downward. Additionally, $PM_{2.5}$ concentrations are lower in the afternoon (12:00–15:00) because of a higher boundary layer height (Barlow, 2014) and fewer fossil-fuel burnings and household heating activities (Zhang and Cao, 2015). Previous studies have also shown similar bimodal distributions, for example, Schnell et al. (2018), Zhang and Cao (2015), and Yadav et al. (2017).

The study finds higher $PM_{2.5}$ levels during weekends and early weekdays (Fig. 5b), likely due to increased anthropogenic activities, including higher solid/fossil fuel burning for household heating. In contrast, Thursdays exhibit lower $PM_{2.5}$ concentrations. However, these findings are specific to the winter of 2022–2023 and may be influenced by natural variability and other external factors, such as meteorological conditions and changes in human behavior.

3.2. *Predicting $PM_{2.5}$ using GAM*

3.2.1. *Model performance*

The performance of fitted GAM in predicting $PM_{2.5}$ is visualized in Fig. 6 ($R^2 = 0.88$). In Fig. 6, we present the outcomes of 10-fold cross-validation for the fitted GAM utilized to predict $PM_{2.5}$ levels in Legionowo, Poland, spanning from January 2020 to January 2023. This figure illustrates the validation plot, depicting the relationship between true and predicted values, and the residual plot, showcasing the differences between predicted and true values against the predicted values. The model was trained on a dataset comprising 135,778 rows. For the training set, 10-fold cross-validation normalized RMSEs — respectively normalized by interquartile and the data range — were 33.15% and 1.96%. We compared the predicted $PM_{2.5}$ with the sensors' measured $PM_{2.5}$ during the test periods to assess our models' performance (Supplementary Table 1). Fig. 7 represents the model performance for six out of the 13 sensors. The quality of predictions for the rest of the sensors is illustrated in Supplementary Fig. 11 and Supplementary Fig. 12. On average, the RMSE and R^2 values for the 13 sensors during their test periods were 5.39 μg m^{-3} and 0.80, respectively.

The predictive performance of the statistical and ML supervised models is case-sensitive. Several parameters affect the fitted models' final performance, such as the choice of predictors and models, data cleaning steps, hyper-parameter optimization schemes, and accuracy evaluation schemes (Molnar, 2020). Comparing the model performance of this study with those of similar studies, however, allows a better understanding of the model's place in the literature. Li et al. (2017) proposed a GAM combined with a Principal Component Analysis (PCA) to estimate the $PM_{2.5}$ concentrations in the Beijing-Tianjin-Hebei region over one year. The cross-validation of the proposed model showed an adjusted $R^2 = 0.94$ and a RMSE = 4.08 μg m^{-3}. Using a generalized additive model, bagging method, and variogram simulation, Li et al. (2017) proposed an approach for predicting $PM_{2.5}$ concentrations. As predictors, they used PM_{10} data, meteorological parameters, remote sensing data, and land use data from 96 monitoring stations in Shandong Province, China. With or without PM_{10} as a predictor, they reached

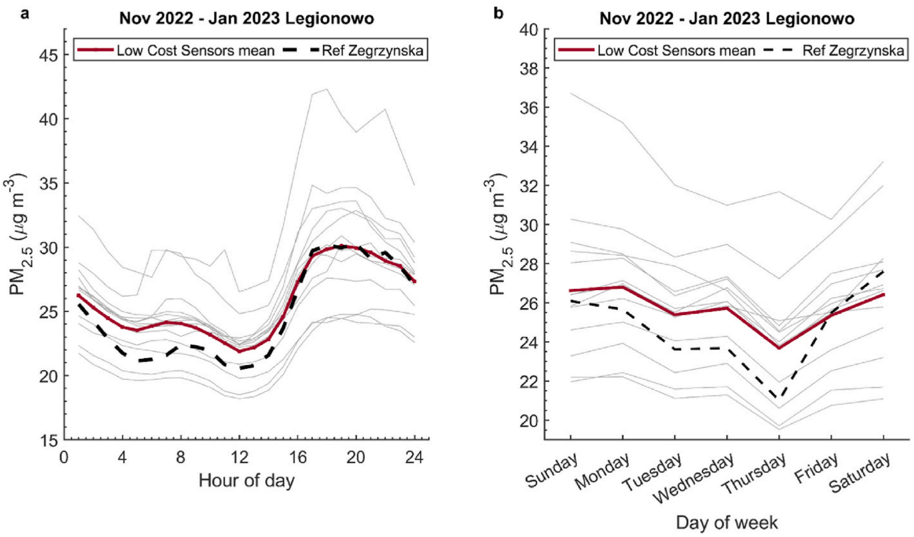

Fig. 5. Diurnal (a) and weekly (b) variations of $PM_{2.5}$ measured by Airly low-cost PM sensors during the Nov 2022–Jan 2023, Legionowo, Poland. The gray lines represent the measurements of different sensors.

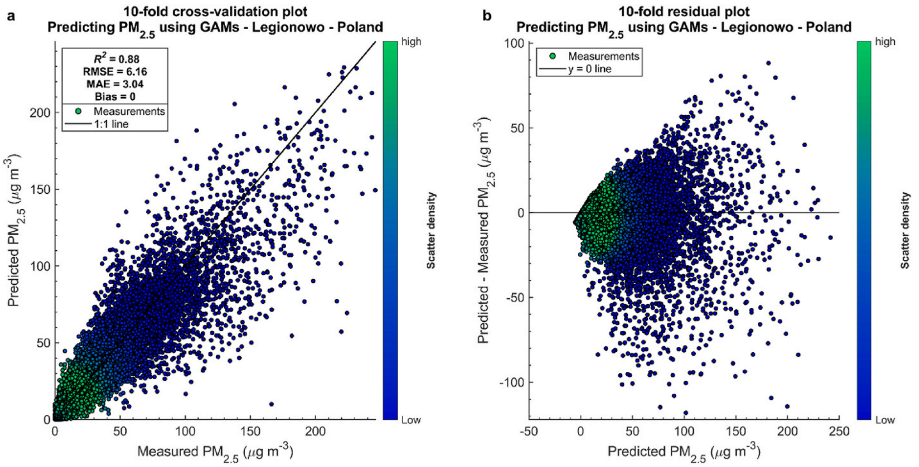

Fig. 6. 10-fold cross-validation of the fitted Generalized Additive Model (GAM) for prediction of $PM_{2.5}$ in Legionowo, Poland, Jan 2020–Jan 2023. a, Validation plot. b, Residual plot. The training set size was 135,778 rows..

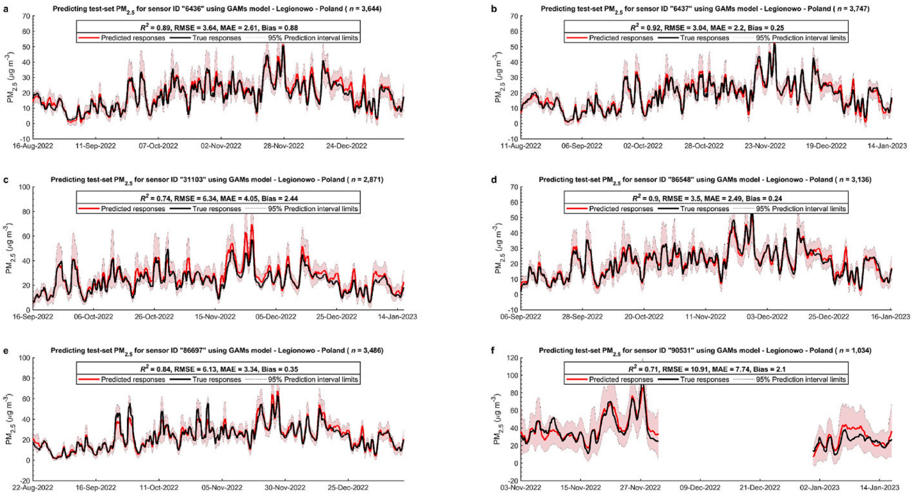

Fig. 7. Predictions of the fitted Generalized Additive Model (GAM) for estimating $PM_{2.5}$ measured by Airly low-cost PM sensors during the test period, Legionowo, Poland. 15% of the data measured by each sensor was held out as a test set. Time series represent the 24-h running average.

cross-validated R^2 values of 0.89 and 0.86, respectively. For an urban area encompassing seven counties, Brokamp et al. (2018) used a Random Forest model based on satellite, meteorological, atmospheric, and land use data to estimate daily $PM_{2.5}$ concentrations at a resolution of 1 × 1 km with an overall cross-validated RMSE of 2.22 μg m^{-3} and a cross-validated R^2 of 0.91. Cross-validated R^2 and RMSEs in a similar range (0.8–0.95 and 2–10 μg m^{-3}) have been reported by several other studies, predicting $PM_{2.5}$ using ML models such as Yu et al. (2022), Shtein et al. (2019), and Reid et al. (2021). Overall, the performance of the trained GAM here is acceptable and in line with the accuracies observed in the previous studies.

3.2.2. Accumulative local effects

The first-order ALE plots, describing the average influence of predictors on GAM predictions, are represented in Fig. 8. Due to the high correlation between the predictors (e.g., month and air temperature), we avoided the partial dependency plots (Molnar, 2020). The ALE (Y-axis) quantifies the deviation of the prediction from the average prediction at the predictor value (X-axis). Some physical concepts can explain the relationships between predictors and predictions. High wind speeds can disperse and dilute PM in the air, resulting in lower $PM_{2.5}$, assuming the significant portion of the predicted $PM_{2.5}$ (measured by sensors) is due to heating activities, not dispersion processes. According to the output of the CALMET model (2014–2015) of air pollution simulation in Krakow, Southern Poland (Oleniacz et al., 2016), there was a negative correlation between $PM_{2.5}/PM_{10}$ concentrations and wind speed as well as mixing layer height, indicating that in the winter months, the concentrations of these pollutants in the air may be primarily attributed to low wind speeds and low mixing layer heights, in line with the calculated ALE results for wind speed in this analysis.

The study of winter weather conditions against PM_{10} levels in Tricity, Northern Poland, showed the highest concentrations of PM occur when the air temperature is low, wind speed is low, pressure is high, and relative air humidity is lower than the average (i.e., anticyclonic weather conditions) (Nidzgorska-Lencewicz and Czarnecka, 2015). However, the results here suggest that the predicted $PM_{2.5}$ is lower than the average at low RHs (≤60%) and higher than the average at high (≥80%) RHs. This can be attributed to high levels of $PM_{2.5}$ during cold months when RH is also on average higher; however, the increased Bias of the low-cost PM sensors due to moisture should not be neglected.

As compared to reference instruments that measure dry particle concentration, the low-cost PM sensor measures in ambient conditions, resulting in a positive bias in measured PM values (Jayaratne et al., 2018). The humidity in the sampling system affects light intensity, absorbing infrared radiation (Zieger et al., 2013). As a result, the phototransistor receives less light, which can lead to an overestimation of

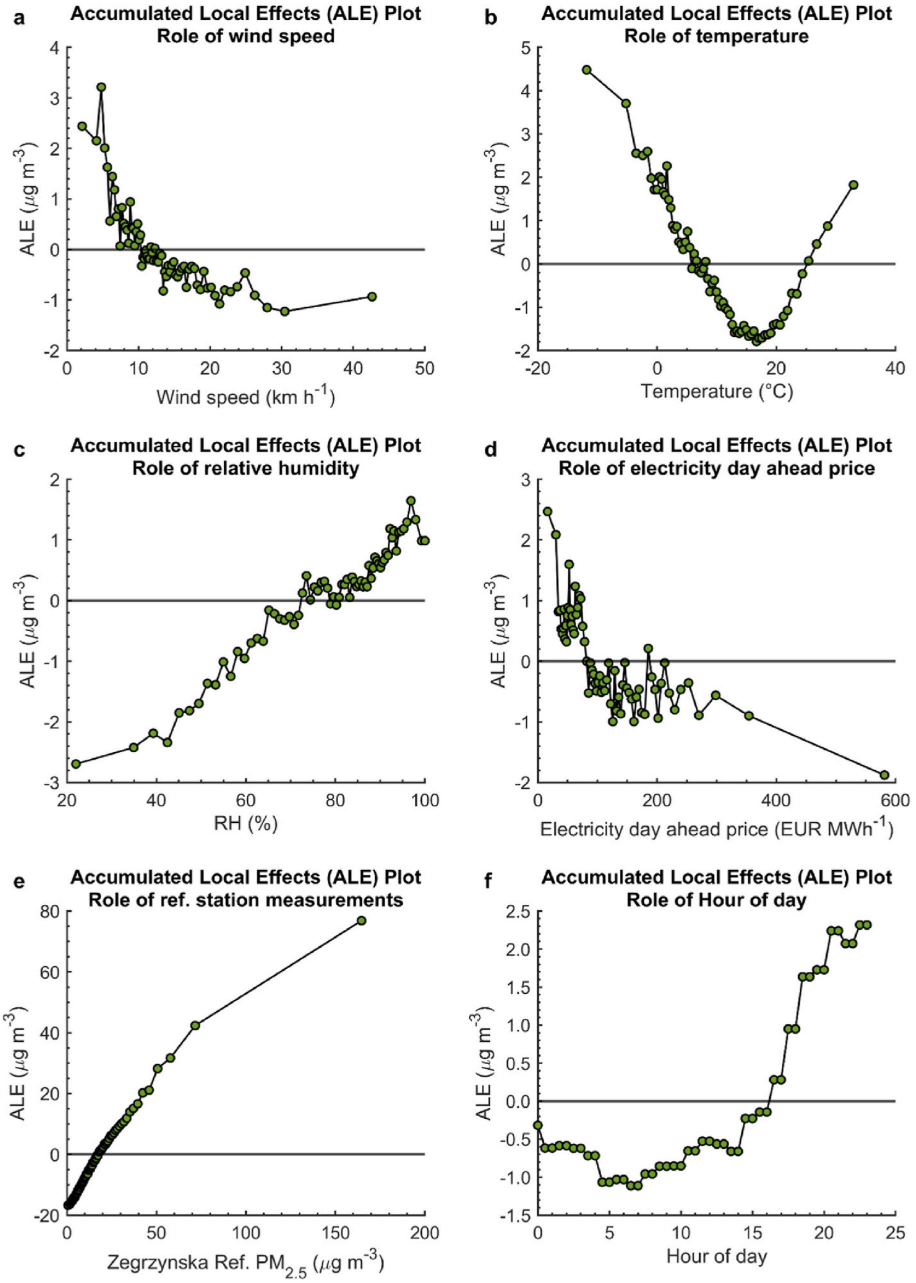

Fig. 8. Accumulated Local Effect (ALE) plots for predictors used for estimating $PM_{2.5}$ measured by Airly low-cost PM sensors in Legionowo, Poland, using a Generalized Additive Model (GAM).

particle mass concentrations.

The peak in $PM_{2.5}$ measurements in evening hours (16:00–23:00) is also evident from ALE plots. The ALE plot for temperature also shows that at high air temperatures ($\geq$25 °C), the predicted $PM_{2.5}$ is higher than the average prediction. The presence of dry and stagnant air conditions may accelerate the release of dust and other particles, as well as the formation of secondary particles, which include $PM_{2.5}$.

Unlike our analysis, a developed GAM by Li et al. (2017) for analyzing the relationship between air temperature and $PM_{2.5}$ in Shandong Province, China, showed that an increase in air temperature would increase $PM_{2.5}$ levels at all temperature ranges. Li et al. (2017) covered a much larger geographical area (216,957 km^2) over one year, while we focused on a smaller urban area or small city. Moreover, our analysis specifically focuses on $PM_{2.5}$ induced by solid/fossil fuel burning during winter, with sensors placed close to emission sources. The emphasis on solid fuel burning might overshadow the impact of other pollution sources in the small city. The primary source of $PM_{2.5}$ in Li et al. (2017) study could have been more influenced by windblown and background pollutants, potentially resulting in higher concentrations during warmer months when temperature-driven dispersion mechanisms are more pronounced.

3.2.3. *Shapley values and interpretable model-agnostic explanations*

The fitted GAM is evaluated according to predictor Shapley values to find the most important predictors. For each query point, the Shapley value represents the predictor's contribution to the prediction's deviation from the average prediction Sundararajan and Najmi (2020). Due to the high correlation between the predictors, we used the *conditional-kernel* method to calculate the Shapley values (Aas et al., 2021). We used sub-samples of the data to reduce the computational load. We randomly extracted 500 samples from the training set and calculated the Shapley values for ten randomly selected query points. This procedure was repeated 1,000 times to generate the Shapley values plot in Fig. 9a. The categorical variable Shapley values (including 12 dummy variables) are not shown. Neglecting the sensor ID, the most important predictors were $PM_{2.5}$ measured at the *Zegrzynska* reference station (mean Shapley value = 0.58), wind speed (0.12), and month of the year (−0.16).

The impact of the different predictors on predicted $PM_{2.5}$ estimated using LIME at other sensor locations is also presented in Fig. 9b, c, and d. LIME is a method that offers explanations for individual predictions made by ML models. It fits a simple, interpretable model (like linear regression) to approximate the local behavior around the query point, providing insights into the complex ML model's predictions in a transparent manner (Molnar, 2020). Here, an interpretable linear model approximates the GAM in the LIME technique (Lozano et al., 2011; Swirszcz et al., 2009), using all predictor values. For each sensor data in the training set, we applied the LIME technique to fit a linear model at 500 randomly selected query points and estimated the model's coefficients. This procedure was repeated 200 times, and the mean of the calculated coefficients for each sensor is represented in Fig. 9 and Supplementary Table 2. The mean of predictions for query points made by GAM was 20.57 μg m^{-3} while the surrogate linear models' predictions for the query points fitted by the LIME technique had an average of 21.46 μg m^{-3}. Surrogate model predictions deviation from the GAM predictions was the highest for sensor ID 86548 (mean = 2.53 μg m^{-3}), and the lowest was calculated for sensor ID 97546 (mean = 0.03 μg m^{-3}).

The calculated coefficients range for temperature was between −0.2754 (sensor ID = 3417) and −0.2530 (sensor ID 96337). This represents, for example, by a one °C increase in air temperature for sensor ID 3417, the $PM_{2.5}$ at the sensor location would decrease by 0.2754 μg m^{-3} (The negative sign indicates reverse relation). The calculated minimum and maximum calculated coefficients for wind speed and RH were −0.1534 (sensor ID = 6437)/-0.1298 (sensor ID 96337) and 0.0189 (sensor ID 3417)/0.0328 (sensor ID 96337), respectively. The calculated minimum, maximum, and mean coefficients for $PM_{2.5}$ measured at the *Zegrzynska* reference station were 0.6366 (sensor ID 3542), 0.6418 (sensor ID 12441), and 0.6389, respectively.

3.3. Limitations and constraints

- The study results are influenced by lockdown measures during Covid-19 outbreaks, impacting air quality data during those periods.
- Citizens' eagerness determined sensor locations, possibly leading to underrepresentation in certain areas with higher PM pollution levels, affecting overall air quality representation.
- Airly sensors used in this analysis are not calibrated in Legionowo. Individual calibration is recommended but may not be feasible in some studies with multiple sensors.
- Sens-to-sensor variability may affect results, which can introduce some level of uncertainty.
- Data processing steps (in particular step 2) may exclude accurate sensor measurements, as distinguishing incorrect and correct data can be challenging.
- The type of aerosol present in the air can significantly impact the measurements made by $PM_{2.5}$ sensors. Different aerosols may have

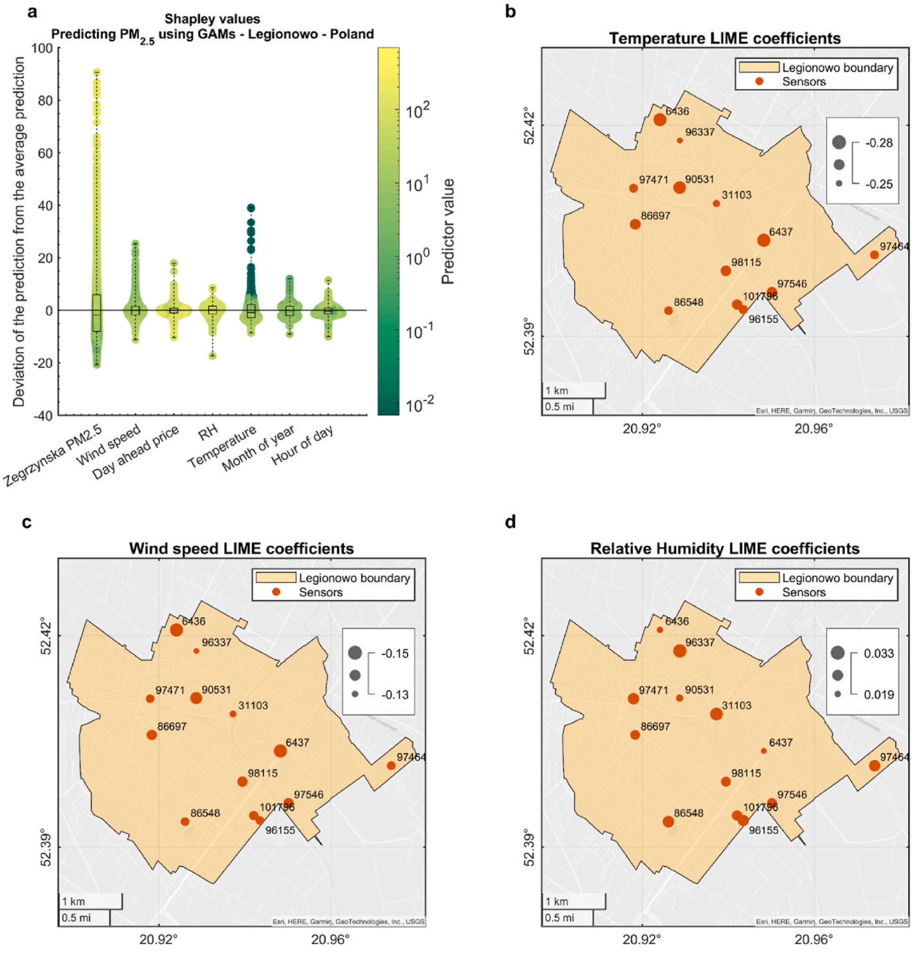

Fig. 9. Interpreting the fitted Generalized Additive Model (GAM) for prediction of $PM_{2.5}$ measured by Airly low-cost PM sensors in Legionowo, Poland. a, Shapley values for individual predictors. b, c, and d, calculated Local Interpretable Model-agnostic Explanations (LIME) coefficients for environmental predictors.

varying sizes, compositions, and physical properties, which can affect how they interact with the sensor's detection mechanism.
- The sensor data processing did not explicitly consider RH influence. However, advanced calibration techniques like ML could address this aspect effectively.
- Without detailed predictor data, we couldn't create a spatial map. Access to such data could generate high-resolution gridded data, improving the representation of spatiotemporal variability and enhancing air quality analysis.
- Correlating measured $PM_{2.5}$ values with the energy price paid by the users (user price) would provide better insights into household heating patterns, but hourly data wasn't available. However, here, we used day-ahead energy prices (producer price) data as a proxy, which may have limitations due to hidden subsidies and other factors.

4. Conclusions

This paper discusses a study conducted in Legionowo, Poland, to identify air pollution hotspots and investigate the temporal and spatial air pollution distribution and its relationship with household heating activities. The general conclusions include:

- Proposed a data quality assurance scheme for PM sensors to ensure the accuracy and reliability of the collected data. This scheme involved three stages of pre-processing, including data coverage filtering, sensor-to-sensor correlation analysis, and sensor drift diagnosis.
- Introduced an approach to estimate the uncertainty induced by RH on PM sensor measurements. By considering the influence of RH, we aimed to enhance the accuracy of PM data interpretation and better understand the impact of environmental factors on measured PM levels.
- Employed an explainable ML model, specifically Generalized Additive Models (GAMs), to explore the relationships between $PM_{2.5}$ and various environmental settings.

We used data from a network of low-cost air quality sensors (13 Airly sensors) operated by citizens, official air quality stations, and meteorological data. More specific conclusions based on our findings:

- Overall, the low-cost PM sensors showed a responsive/stable performance, considering the official measurements during the analysis period (Jan 2020–Jan 2023). 95.46% of sensor-hour measurements passed our proposed pre-processing data filtering measures. One limitation of this study may be a lack of co-location of low-cost sensors at reference stations and regular service of the sensors due to insufficiency of staff/funding.
- Because no co-location was conducted, we estimated the uncertainty of $PM_{2.5}$ and PM_{10} measurements by fitting probability distribution functions to the deviation of the sensors from the reference station PM data for different RH bins when we assumed PM spatial variability was low. For $PM_{2.5}$, the probability of observing a negative bias (sensor measurement lower than the reference instrument measurement) at all RH bins was almost 0.3. For the first RH bin (50–60%), the probability of observing a positive bias was 0.66; for the fifth bin (90–100%), the likelihood of a positive bias was 0.78. For PM_{10}, the influence of RH bins on biases was more extreme. The probability of a positive bias was 0.21 for the first RH bin, while the likelihood for the fifth RH bin was 0.76. Overall, the results confirm the sensors' positive Bias dependency on RH levels.
- It was observed that $PM_{2.5}$ concentrations peak between 6:00 and 10:00 and 16:00 and 23:00 in all sensors. Overall, during the winter of 2022–2023, weekends and early weekdays, particularly Mondays, exhibited higher $PM_{2.5}$ levels. On the other hand, Thursdays consistently showed the lowest $PM_{2.5}$ concentrations during this winter period. The highest $PM_{2.5}$ and PM_{10} levels were observed during cold months (Oct–Apr).
- It was estimated that the average air temperature, official $PM_{2.5}$ concentration, and sensor $PM_{2.5}$ concentration in Legionowo were 9.76 °C, 18.99 μg m^{-3}, and 20.04 μg m^{-3}, respectively, from Jan 2020 to Jan 2023.
- The results suggest that high spatio-temporal monitoring of PM levels during warm months (May–Sep) is not necessary, and the reference station suffices for monitoring purposes. As a result, costs associated with sensor data handling could be reduced during the warmer months when pollution patterns show less variation. During cold months, we recommend enhancing monitoring efforts by increasing the density of monitoring stations, particularly in areas with significant solid fuel-burning activity.
- The highest concentrations of $PM_{2.5}$ are observed in the Dec–Jan period. Due to low levels of industrial activities in Legionowo, the primary source of $PM_{2.5}$ may be attributed to heating using fossil fuels burning in domestic boilers and, to less extent, stove wood-burning.
- The ALE plots estimated using the fitted statistical model showed there is no specific relation between day-ahead electricity prices and $PM_{2.5}$ levels in Legionowo. Also, the $PM_{2.5}$ decreases at higher wind speeds, suggesting there is no dependency of $PM_{2.5}$ levels on the dispersion processes. The $PM_{2.5}$ constantly decreases with an increase in the air temperature until 20 °C, after which the $PM_{2.5}$ concentrations trend changes. As less than 7% of the data are measured at temperatures above 25 °C, the dependency of $PM_{2.5}$ in Legionowo on high air temperatures needs further research.
- According to calculated Shapley values, the most important predictors for predicting $PM_{2.5}$ in Legionowo were $PM_{2.5}$ at Zegrzynska (mean value 0.58), wind speed (0.12), and month of the year (−0.16).
- The LIME analysis showed that by a one °C increase in air temperature and a one km h^{-1} increase in wind speed, the $PM_{2.5}$ in Legionowo would decrease by 0.26 μg m^{-3} and 0.14 μg m^{-3}, respectively. On the other hand, the $PM_{2.5}$ level increases by 0.03 and 0.64 μg m^{-3} as RH and the $PM_{2.5}$ measured at the Zegrzynska reference station increase by 1% and one μg m^{-3}, respectively.

Funding

The research leading to these results has received funding from the EEA Grants 2014–2021 via the National Center for Research and Development. The grant is provided under the “Applied Research” Program for the “GREEN HEAT – towards collaborative local decarbonization” project (acronym GREEN HEAT); grant number PL-Applied Research-0043.

CRediT authorship contribution statement

Amirhossein Hassani: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Sebastian Bykuć:** Resources, Writing – review & editing, Project administration, Funding acquisition. **Philipp Schneider:** Validation, Writing – review & editing. **Paweł Zawadzki:** Resources, Writing – review & editing. **Patryk Chaja:** Resources, Writing – review & editing. **Núria Castell:** Conceptualization, Methodology, Validation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition, All authors have read and agreed to the published version of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

As the sensor data requires access credentials for the Airly data platform, the low-cost PM sensor data are available upon request from the corresponding author. The reference data for reference Zegrzynska monitoring station are available at: https://powietrze.gios.gov.pl/pjp/current/station_details/info/471 (in Polish, accessed in Feb 2023). Meteorological data from the Modlin station are available at: https://www.ncei.noaa.gov/access/search/data-search/global-hourly.

Acknowledgments

The contribution of all citizens who participated in the GREEN HEAT project is acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2023.120108.

References

Aas, K., Jullum, M., Løland, A., 2021. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artif. Intell. 298, 103502.
Alfano, B., Barretta, L., Del Giudice, A., De Vito, S., Di Francia, G., Esposito, E., et al., 2020. A review of low-cost particulate matter sensors from the developers' perspectives. Sensors 20 (23), 6819.
Fumigation. Glossary of Meteorology,, 2020. http://glossary.ametsoc.org/wiki/fumigation.
Apley, D.W., Zhu, J., 2016. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *arXiv preprint arXiv:1612.08468*.
Attia, S., Kosiński, P., Wójcik, R., Węglarz, A., Koc, D., Laurent, O., 2022. Energy efficiency in the polish residential building stock: a literature review. J. Build. Eng. 45, 103461.
Badyda, A.J., Grellier, J., Dąbrowiecki, P., 2017. Ambient PM2. 5 exposure and mortality due to lung cancer and cardiopulmonary diseases in Polish cities. Respir. Treat. Prevent. 9–17.
Barlow, J.F., 2014. Progress in observing and modelling the urban boundary layer. Urban Clim. 10, 216–240.
Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci. Data 5 (1), 1–12.
Brattich, E., Bracci, A., Zappi, A., Morozzi, P., Di Sabatino, S., Porcù, F., et al., 2020. How to get the best from low-cost particulate matter sensors: guidelines and practical recommendations. Sensors 20 (11), 3073.
Brauers, H., Oei, P.-Y.J.E.P., 2020. In: The Political Economy of Coal in Poland: Drivers and Barriers for a Shift Away from Fossil Fuels, vol. 144, 111621.
Brokamp, C., Jandarov, R., Hossain, M., Ryan, P., 2018. Predicting daily urban fine particulate matter concentrations using a random forest model. Environ. Sci. Technol. 52 (7), 4173–4179.
Bulot, F.M.J., Johnston, S.J., Basford, P.J., Easton, N.H.C., Apetroaie-Cristea, M., Foster, G.L., et al., 2019. Long-term field comparison of multiple low-cost particulate matter sensors in an outdoor urban environment. Sci. Rep. 9 (1), 1–13.
Carvalho, H., 2019. Air pollution-related deaths in Europe–time for action. J. Global Health 9 (2).
Castell, N., Dauge, F.R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., et al., 2017. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environ. Int. 99, 293–302.
Cavaliere, A., Carotenuto, F., Di Gennaro, F., Gioli, B., Gualtieri, G., Martelli, F., et al., 2018. Development of low-cost air quality stations for next generation monitoring networks: calibration and validation of PM2. 5 and PM10 sensors. Sensors 18 (9), 2843.
Chambers, S.D., Podstawczyńska, A., 2019. Improved method for characterising temporal variability in urban air quality part II: particulate matter and precursors in central Poland. Atmos. Environ. 219, 117040.
Finkelman, R.B., 2007. Health impacts of coal: facts and fallacies. AMBIO A J. Hum. Environ. 36 (1), 103–106.
Fu, J., Tang, D., Grieneisen, M.L., Yang, F., Yang, J., Wu, G., et al., 2023. A machine learning-based approach for fusing measurements from standard sites, low-cost sensors, and satellite retrievals: application to NO2 pollution hotspot identification. Atmos. Environ. 302, 119756.
Giordano, M.R., Malings, C., Pandis, S.N., Presto, A.A., McNeill, V.F., Westervelt, D.M., et al., 2021. From low-cost sensors to high-quality data: a summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. J. Aerosol Sci. 158, 105833.
Hastie, T.J., 2017. Generalized additive models. In: Statistical Models in S. Routledge, pp. 249–307.
Hastie, T., Tibshirani, R., 1987. Generalized additive models: some applications. J. Am. Stat. Assoc. 82 (398), 371–386.
Hirth, L., Mühlenpfordt, J., Bulkeley, M., 2018. The ENTSO-E Transparency Platform–A review of Europe's most ambitious electricity data platform. Appl. Energy 225, 1054–1067.
Hofman, J., Peters, J., Stroobants, C., Elst, E., Baeyens, B., Van Laer, J., et al., 2022. Air quality sensor networks for evidence-based policy making: best practices for actionable insights. Atmosphere 13 (6), 944.
Holnicki, P., Kałuszko, A., Nahorski, Z., 2022. Scenario analysis of air quality improvement in Warsaw, Poland, by the end of the current decade. Atmosphere 13 (10), 1613.
Hong, G.-H., Le, T.-C., Tu, J.-W., Wang, C., Chang, S.-C., Yu, J.-Y., et al., 2021. Long-term evaluation and calibration of three types of low-cost PM2. 5 sensors at different air quality monitoring stations. J. Aerosol Sci. 157, 105829.
Jagiełło, P., Struzewska, J., Jeleniewicz, G., Kamiński, J.W., 2022. Evaluation of the effectiveness of the national clean air programme in terms of health impacts from exposure to PM2. 5 and NO2 concentrations in Poland. Int. J. Environ. Res. Publ. Health 20 (1), 530.
Jayaratne, R., Liu, X., Thai, P., Dunbabin, M., Morawska, L., 2018. The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog. Atmos. Meas. Tech. 11 (8), 4883–4890.
Junninen, H., Mønster, J., Rey, M., Cancelinha, J., Douglas, K., Duane, M., et al., 2009. Quantifying the impact of residential heating on the urban air quality in a typical European coal combustion region. Environ. Sci. Technol. 43 (20), 7964–7970.
Kang, Y., Aye, L., Ngo, T.D., Zhou, J., 2022. Performance evaluation of low-cost air quality sensors: a review. Sci. Total Environ. 818, 151769.
Karagulian, F., Barbiere, M., Kotsev, A., Spinelle, L., Gerboles, M., Lagler, F., et al., 2019. Review of the performance of low-cost sensors for air quality monitoring. Atmosphere 10 (9), 506.
Karpinska, L., Śmiech, S., Gouveia, J.P., Palma, P., 2021. Mapping regional vulnerability to energy poverty in Poland. Sustainability 13 (19), 10694.
Kerimray, A., Rojas-Solórzano, L., Torkmahalleh, M.A., Hopke, P.K., Gallachóir, B.P.Ó., 2017. Coal use for residential heating: patterns, health implications and lessons learned. Energy Sustain. Dev. 40, 19–30.
Kompalli, S.K., Moorthy, K.K., Babu, S.S., 2014. Rapid response of atmospheric BC to anthropogenic sources: observational evidence. Atmos. Sci. Lett. 15 (3), 166–171.
Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., et al., 2015. The rise of low-cost sensing for managing air pollution in cities. Environ. Int. 75, 199–205.
Kundzewicz, Z.W., Matczak, P., 2012. Climate change regional review: Poland. Wiley Interdisciplinary Reviews: Clim. Change 3 (4), 297–311.
Kuula, J., Mäkelä, T., Aurela, M., Teinilä, K., Varjonen, S., González, Ó., Timonen, H., 2020. Laboratory evaluation of particle-size selectivity of optical low-cost particulate matter sensors. Atmos. Meas. Tech. 13 (5), 2413–2423.
Kuźma, Ł., Kurasz, A., Dąbrowski, E.J., Dobrzycki, S., Bachórzewska-Gajewska, H., 2021. Short-term effects of "polish smog" on cardiovascular mortality in the green lungs of Poland: a case-crossover study with 4,500,000 person-years (PL-PARTICLES study). Atmosphere 12 (10), 1270.
Lee, H., Kang, J., Kim, S., Im, Y., Yoo, S., Lee, D.J.S., 2020. Long-term evaluation and calibration of low-cost particulate matter (PM) sensor 20 (13), 3617.
Li, L., Zhang, J., Qiu, W., Wang, J., Fang, Y., 2017a. An ensemble spatiotemporal model for predicting PM2. 5 concentrations. Int. J. Environ. Res. Publ. Health 14 (5), 549.
Li, S., Zhai, L., Zou, B., Sang, H., Fang, X., 2017b. A generalized additive model combining principal component analysis for PM2. 5 concentration estimation. ISPRS Int. J. Geo-Inf. 6 (8), 248.
Lim, C.C., Kim, H., Vilcassim, M.J.R., Thurston, G.D., Gordon, T., Chen, L.-C., et al., 2019. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. Environ. Int. 131, 105022.
Liu, X., Jayaratne, R., Thai, P., Kuhn, T., Zing, I., Christensen, B., et al., 2020. Low-cost sensors as an alternative for long-term air quality monitoring. Environ. Res. 185, 109438.
Lou, Y., Caruana, R., Gehrke, J., 2012. Intelligible models for classification and regression. Paper presented at the. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China.
Lou, Y., Caruana, R., Gehrke, J., Hooker, G., 2013. Accurate intelligible models with pairwise interactions. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13), pp. 623–631.
Lozano, A., Swirszcz, G., Abe, N., 2011. Group orthogonal matching pursuit for logistic regression. Paper presented at the. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.
Mahajan, S., Kumar, P., Pinto, J.A., Riccetti, A., Schaaf, K., Camprodon, G., et al., 2020. A citizen science approach for enhancing public understanding of air pollution. Sustain. Cities Soc. 52, 101800.
Malings, C., Tanzer, R., Hauryliuk, A., Saha, P.K., Robinson, A.L., Presto, A.A., et al., 2020. Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation. Aerosol Science and Technology 54 (2), 160–174.

Molnar, C., 2020. Interpretable Machine Learning. Lulu. com.

Morawska, L., Thai, P.K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., et al., 2018. Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: how far have they gone? Environ. Int. 116, 286–299.

Mrozowska, S., Wendt, J.A., Tomaszewski, K., 2021. The challenges of Poland's energy transition. Energies 14 (23), 8165.

Munawer, M.E., 2018. Human health and environmental impacts of coal combustion and post-combustion wastes. J. Sustain. Mining 17 (2), 87–96.

Nadarajah, S., 2008. A truncated inverted beta distribution with application to air pollution data. Stoch. Environ. Res. Risk Assess. 22, 285–289.

Nidzgorska-Lencewicz, J., Czarnecka, M., 2015. Winter weather conditions vs. air quality in Tricity, Poland. Theor. Appl. Climatol. 119, 611–627.

Nyga-Łukaszewska, H., Aruga, K., Stala-Szlugaj, K., 2020. Energy security of Poland and coal supply: price analysis. Sustainability 12 (6), 2541.

Oleniacz, R., Bogacki, M., Szulecka, A., Rzeszutek, M., Mazur, M., 2016. Assessing the impact of wind speed and mixing-layer height on air quality in Krakow (Poland) in the years 2014–2015. JCEEA 33, 315–342.

Parascandola, M., 2018. Ambient air pollution and lung cancer in Poland: research findings and gaps. J. Health Inequalities 4 (1), 3–8.

Piwowar, A., Dzikuć, M., 2019. Development of renewable energy sources in the context of threats resulting from low-altitude emissions in rural areas in Poland: a review. Energies 12 (18), 3558.

Rai, A.C., Kumar, P., Pilla, F., Skouloudis, A.N., Di Sabatino, S., Ratti, C., et al., 2017. End-user perspective of low-cost sensors for outdoor air pollution monitoring. Sci. Total Environ. 607, 691–705.

Reid, C.E., Considine, E.M., Maestas, M.M., Li, G., 2021. Daily PM2. 5 concentration estimates by county, ZIP code, and census tract in 11 western states 2008–2018. Sci. Data 8 (1), 112.

Reizer, M., Juda-Rezler, K., 2016. Explaining the high PM 10 concentrations observed in Polish urban areas. Air Quality, Atmos. & Health 9, 517–531.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should i trust you?" Explaining the predictions of any classifier. Paper presented at the. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California.

Sayahi, T., Butterfield, A., Kelly, K.J. E.p., 2019. Long-term field evaluation of the Plantower PMS low-cost particulate matter sensors. Environ. Pollut. 245, 932–940.

Schnell, J.L., Naik, V., Horowitz, L.W., Paulot, F., Mao, J., Ginoux, P., et al., 2018. Exploring the relationship between surface PM 2.5 and<? xmltex\break?> meteorology in Northern India. Atmos. Chem. Phys. 18 (14), 10157–10175.

Schulz, M., Tsyro, S., Mortier, A., Valdebenito, A., Kranenburg, R., Benedictow, A., et al., 2021. High PM Concentrations over Europe 19-27 February 2021; A Preliminary CAMS71 Analysis Episode Analysis Report N°01 in 2021. Norwegian Meteorological Institute, Copernicus Atmosphere Monitoring Service. *Reference: CAMS71_2019SC1_D3.2.1_202102_FebEpisode_v1.0*. https://policy.atmosphere.copernicus.eu/reports/CAMS71ReportFeb2021-episode.pdf.

Shtein, A., Kloog, I., Schwartz, J., Silibello, C., Michelozzi, P., Gariazzo, C., et al., 2019. Estimating daily PM2. 5 and PM10 over Italy using an ensemble model. Environ. Sci. Technol. 54 (1), 120–128.

Singh, V., Singh, S., Biswal, A., Kesarkar, A.P., Mor, S., Ravindra, K., 2020. Diurnal and temporal changes in air pollution during COVID-19 strict lockdown over different regions of India. Environ. Pollut. 266, 115368.

Snyder, E.G., Watkins, T.H., Solomon, P.A., Thoma, E.D., Williams, R.W., Hagler, G.S., et al., 2013. The changing paradigm of air pollution monitoring. Environ. Sci. Technol. 47 (20), 11369–11377.

Sokołowski, J., Bouzarovski, S., 2022. Decarbonisation of the Polish residential sector between the 1990s and 2021: a case study of policy failures. Energy Pol. 163, 112848.

Stavroulas, I., Grivas, G., Michalopoulos, P., Liakakou, E., Bougiatioti, A., Kalkavouras, P., et al., 2020. Field evaluation of low-cost PM sensors (Purple Air PA-II) under variable urban air quality conditions, in Greece. Atmosphere 11 (9), 926.

Sundararajan, M., Najmi, A., 2020. The many Shapley values for model explanation. In: International conference on machine learning, pp. 9269–9278.

Swirszcz, G., Abe, N., Lozano, A.C., 2009. Grouped orthogonal matching pursuit for variable selection and prediction. Adv. Neural Inf. Process. Syst. 22.

Tagle, M., Rojas, F., Reyes, F., Vásquez, Y., Hallgren, F., Lindén, J., et al., 2020. Field performance of a low-cost sensor in the monitoring of particulate matter in Santiago, Chile. Environ. Monit. Assess. 192 (3), 171.

Tsyro, S., Schulz, M., Mortier, A., Valdebenito, A., Benedictow, A., Timmerman, R., Kranenburg, R., 2022. High PM10 Levels: Episode of 20-27 March 2022. Norwegian Meteorological Institute, Copernicus Atmosphere Monitoring Service. Reference: CAMS71_D3.2.1-2022-2-B_20-27MarchEpisode. https://policy.atmosphere.copernicus.eu/reports/CAMS2-71_PM10_episode_20-27March2022_final.pdf.

Vogt, M., Schneider, P., Castell, N., Hamer, P., 2021. Assessment of low-cost particulate matter sensor systems against optical and gravimetric methods in a field co-location in Norway. Atmosphere 12 (8), 961.

Watne, Å.K., Linden, J., Willhelmsson, J., Fridén, H., Gustafsson, M., Castell, N., 2021. Tackling data quality when using low-cost air quality sensors in citizen science projects. Front. Environ. Sci. 9, 733634.

Wesseling, J., de Ruiter, H., Blokhuis, C., Drukker, D., Weijers, E., Volten, H., et al., 2019. Development and implementation of a platform for public information on air quality, sensor measurements, and citizen science. Atmosphere 10 (8), 445.

Wielgosiński, G., Czerwińska, J., 2020. Smog episodes in Poland. Atmosphere 11 (3), 277.

Xu, M., Yu, D., Yao, H., Liu, X., Qiao, Y., 2011. Coal combustion-generated aerosols: formation and properties. Proc. Combust. Inst. 33 (1), 1681–1697.

Yadav, R., Sahu, L.K., Beig, G., Tripathi, N., Jaaffrey, S.N.A., 2017. Ambient particulate matter and carbon monoxide at an urban site of India: influence of anthropogenic emissions and dust storms. Environ. Pollut. 225, 291–303.

Yu, W., Li, S., Ye, T., Xu, R., Song, J., Guo, Y., 2022. Deep ensemble machine learning framework for the estimation of PM 2.5 concentrations. Environ. Health Perspect. 130 (3), 037004.

Zhang, Y.-L., Cao, F., 2015. Fine particulate matter (PM2. 5) in China at a city level. Sci. Rep. 5 (1), 1–12.

Zhao, S., Pudasainee, D., Duan, Y., Gupta, R., Liu, M., Lu, J., 2019. A review on mercury in coal combustion process: content and occurrence forms in coal, transformation, sampling methods, emission and control technologies. Prog. Energy Combust. Sci. 73, 26–64.

Zieger, P., Fierz-Schmidhauser, R., Weingartner, E., Baltensperger, U., 2013. Effects of relative humidity on aerosol light scattering: results from different European sites. Atmos. Chem. Phys. 13 (21), 10609–10631.