Journal of Modeling and Optimization 2019;11(2):63-68
https://doi.org/10.32732/jmo.2019.11.2.63

# Forecasting Model Validation of Particulate Air Pollution by Low Cost Sensors Data

Nicoletta Lotrecchiano[1,2], Filomena Gioiella[1], Aristide Giuliano[3], Daniele Sofia[1]

*1. Sense Square srl, Piazza Vittorio Emanuele 11, Fisciano (SA) 84084, Italy*
*2. University of Salerno, Via Giovanni Paolo II 132, Fisciano (SA) 84084, Italy*
*3. ENEA, Italian National Agency for New Technologies, Energy and Sustainable Economic Development, Rotondella (MT) 75026, Italy*
*E-mail:info@sensesquare.eu*

Received: 24 May 2019; Accepted: 24 June 2019; Available online: 25 July 2019

**Abstract:** Environmental pollution in urban areas may be mainly attributed to the rapid industrialization and increased growth of vehicular traffic. As a consequence of air quality deterioration, the health and welfare of human beings are compromised. Air quality monitoring networks usually are used not only to assess the pollutant trend but also in the effective set-up of preventive measures of atmospheric pollution. In this context, monitoring can be a valid action to evaluate different emission control scenarios; however, installing a high space-time resolution monitoring network is still expensive. Merge of observations data from low-cost air quality monitoring networks with forecasting models can contribute to improving significantly emission control scenarios. In this work, a validation algorithm of the forecasting model for the concentration of small particulates (PM10 and PM2.5) is proposed. Results showed a satisfactory agreement between the PM concentration forecast values and the measured data from 3 air quality monitoring stations. Final average RMSE values for all monitoring stations are equal to about 4.5 µg/m$^3$.

**Keywords:** Air quality; Forecast models; Monitoring; Pollution; Particular matter.

## 1. Introduction

Particulate Matter (PM) causes acute and chronic effects, particularly at the respiratory level since they can penetrate deep into the lungs. Primary PM sources are industrial production, transport and residential. Therefore, it is crucial to make many efforts to monitor and control air pollution in an urban context [1]. The monitoring stations are an efficient solution in collecting vast amounts of pollutant concentrations in real-time in urban and extra-urban areas and they are a support for citizens who can know the pollution status of their city [2]. Increasing the numbers of sampling points is possible to obtain a more detailed view of the environmental situation [3]. In recent years, many approaches are developed to predict air quality based on existing historical air quality and meteorological data [4]. A model is a simplified representation of the reality and it gives an approximate description of the modeled phenomenon. One of the main purposes of modeling is the phenomenon explanation, sometimes it can be used to describe the mechanism behind the reality we are investigating. Considerable relative humidity usually causes increases in PM concentrations due to the hygroscopic effect of aerosols, but not for PM10 in spring and summer, mainly due to the suppression of dust emissions under wet air conditions in spring and the impact of wet scavenging under high summer rainfall [5].

The relationships between meteorological factors, traffic flow, topographical factors and particulate concentration have been analyzed using models with inferential statistics, such as linear regression or correlation analysis [6]. In this work, a new methodology is proposed able to forecast the urban quality air using the concentration levels of particulate matter and meteorological conditions. The air quality data recorded by monitoring stations installed in a southern city of Italy during the whole month of August 2018 were considered. From 1st September to 10th November, a comparison was made between the predicted concentrations of PM2.5 and PM10 using the model with the measured concentrations by monitoring stations to fine-tune the model. Finally, a correlation between air pollution and some meteorological factors (wind speed and direction, humidity) was investigated.

## 2. Air quality monitoring area

Battipaglia is a city in the south of Italy. Recently, the city has been characterized by a lot of environmental issues regarding the air quality from its extend industrial area and its three dumps sites which are monitored continuously. An air quality-monitoring network composed of three stations has been located with the purpose to monitor the particular matter concentration (PM 2.5 and PM 10). As shown in Figure 1, the stations are localized between the industrial area and the city center.

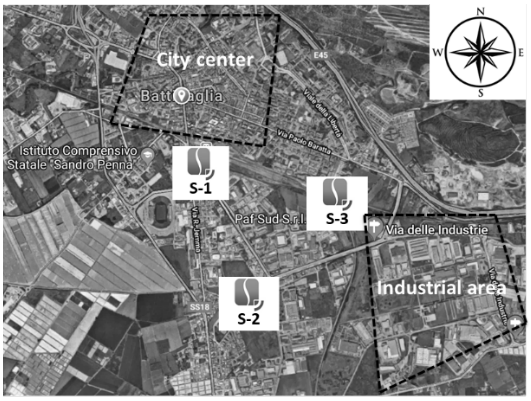


Figure 1. Graphical representation of the installed monitoring network between the Industrial area of Battipaglia (Italy) and City center.

## 3. Methodology

### 3.1 Data collection

The concentration of small particulates (PM10 and PM2.5) and meteorological parameters (wind direction and velocity, temperature, humidity) have been collected for all days of August 2018 by three monitoring stations installed in Battipaglia city (Italy) in the position located between the industrial area and the urban center (Figure 1).

### 3.2 August-day data analysis

Provided the data at every hour, the data processing returns average values, $\bar{c}_{i,j,k}$ in Eq. 1. Data were collected by value intervals for each different parameter (relative humidity, wind intensity and intensity, temperature, daily hour). The raw data were examined and some adjustments were made to purify data by outlier values. The time variability of data was considered studying its behavior during 24 hours of a day.

### 3.3 PM concentration values by the forecasting model

These data have been used to find the best fitting parameters of the model. Initial weight was considered based on the magnitude of the difference between 75° and 25° percentiles (box in Figure 2). So, for each parameter, the variability of data was evaluated considering the PM2.5/PM10 mean concentration in the parameter interval. A deterministic approach was used in this work. In Eq. 1, $c_{j,h}$ is the forecasted concentration for the day *j* and hour *h*, $p_{i,j-1}$ is the model coefficient for the parameter *i* optimized by day *j-1*, $\bar{c}_{i,j-1,k_{i,j,h}}$ is the average concentration (diamond in Figure 2) considering only parameter *i*, until day *j-1*, in the parameter interval (by Figure 2) $k_{i,j,h}$, $k_{i,j,h}$ is the interval related to parameter *i* from forecasting database for the day *j* for the daily hour *h*.

$$c_{i,h} = \sum_{i=1}^{n_p} p_{i,j-1} c_{i,j-1,ki,j,h} \quad (1)$$

### 3.4 Objective function description

To find the best value of the coefficient associated with each parameter, an objective function was described as the sum of the quadratic deviations of the residuals between the estimated and the measured value. In Eq. 2, *n* is the day considered (from 1st to 71st), *h* in the daily hour, $\dot{c}_{j,h}$ is measured PM concentration by the monitoring stations. The coefficient estimation starts the first day using all the background of the previous 31 days of August. Going forward in the estimation of the coefficients, the number of days used as background for the estimation will progressively increase. In this way, the minimization takes into account not only what happened in the last 24 hours, but also in the previous *n* days considered.

$$f_{OB,n} = \sum_{j=1}^{n} \sum_{h=1}^{24} (\dot{c_{j,h}} - c_{j,h})^2 \quad (2)$$

### 3.5 Model tuning

From the 1st September to 10th November 2018 (71 days) a comparison was made between the predicted concentrations of PM2.5 and PM10 using the dispersion model with the measured concentrations by monitoring stations in order to fine-tune the model. It was found that after $n_{TOT}$ (71) days, a reliable model was obtained and no further adjustment was required to adjust prediction with experiments.

## 4. Results

For the sake of brevity, only for one monitoring station (S1) graphs are shown. On each box, the central mark indicates the median, the diamond indicates the mean value and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Points are considered as outliers, indicated by a cross. Figure 2 shows clearly that the PM concentration varies in different ways on each parameter. The most significant variation of concentration depends on relative humidity (*(a)* in Figure 2), wind intensity (*(b)* in Figure 2) and daily hour (*(e)* in Figure 2).

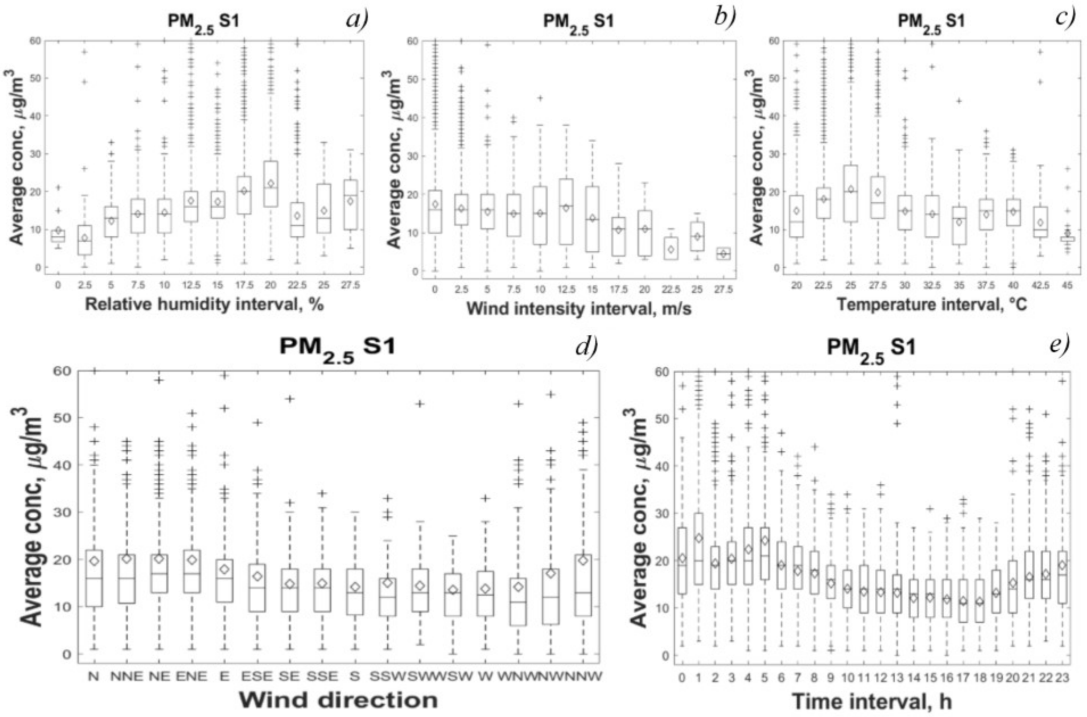


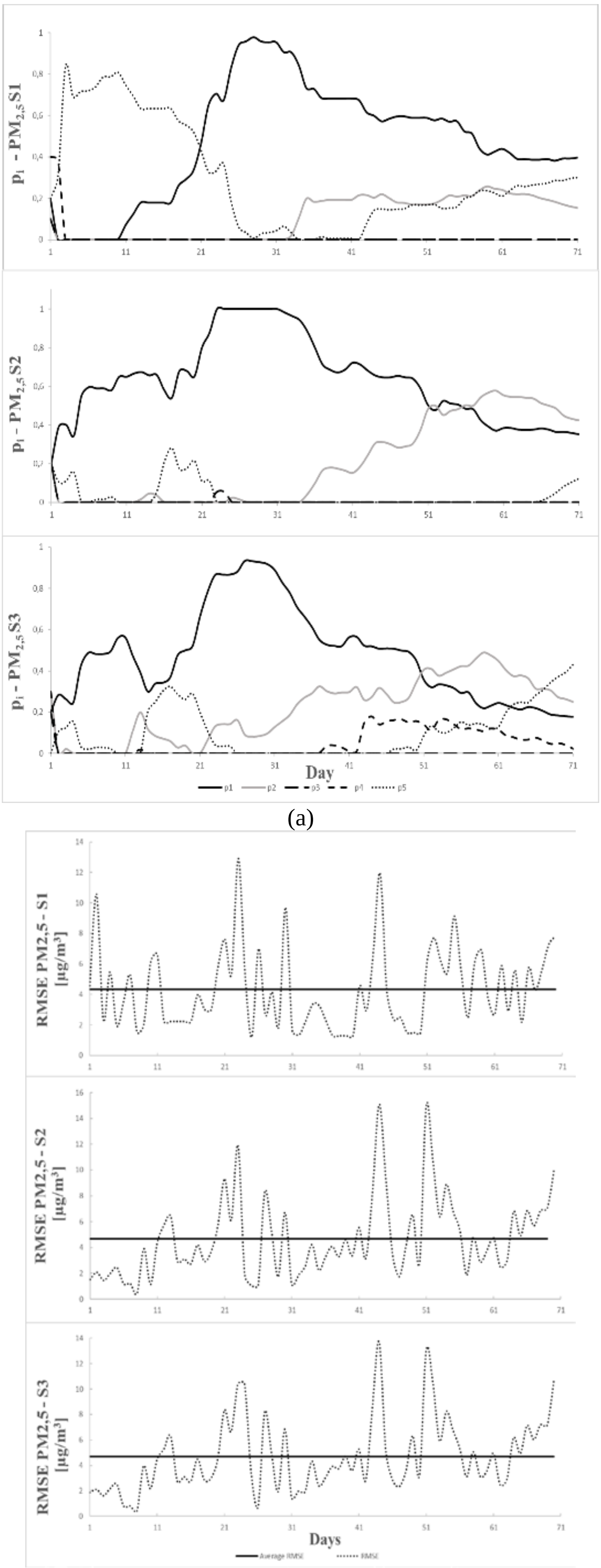
Figure 2. Variation of the average concentration of PM2.5 for S1 respect to intervals of a) relative humidity; b) wind intensity; c) temperature; d) wind direction; e) daily time. Data processing was carried out between 1st-31st August. Central mark indicates the median. Diamonds indicate the mean value.

In order to estimate the robustness of the model, the RMSE (Root-Mean-Square Error) values were evaluated every day as:

$$RMSE_j = \sqrt{\frac{1}{24} * \sum_{h=1}^{24} (c_{j,\dot{h}} - c_{j,h})^2} \tag{3}$$

Finally, to summarize the performance of the model for each station, the average RMSE was calculated by:

$$\overline{RMSE} = \sum_{j=1}^{n_{TOT}} RMSE_j \tag{4}$$



(a)

(b)

Figure 3. *(a)* Model coefficients $p_i$ from 1st to 71st day for PM2.5 concentrations for stations S1, S2, S3. *(b)* RMSE values from 1st to 71st day for PM2.5 concentrations for S1, S2, S3.

Figure 3 shows the behavior of the model coefficients for each day used for the minimization. The coefficients are more varying in the first days, while they are taking around a constant value after 40-50 days of observation (Table 1). The coefficient estimation is also essential to understand what parameter influences the phenomenon studied. As a result, relative humidity, wind intensity and time are the parameters that influence mainly the PM concentrations. This evidence is also underlined in Figure 2, where PM concentrations change during the day above all, according to the anthropic activities. The PM concentrations decrease by increasing the wind intensity because of air origin from the sea (SE direction) or mountains (other directions).

Table 1. Coefficients $p_i$ for the last day for PM2.5 and PM10 and stations S1, S2, S3.

| | PM2.5 | | | PM10 | | |
|---|---|---|---|---|---|---|
| | S1 | S2 | S3 | S1 | S2 | S3 |
| $p_1$ | 0.35 | 0.40 | 0.16 | 0.40 | 0.35 | 0.17 |
| $p_2$ | 0.13 | 0.43 | 0.23 | 0.15 | 0.42 | 0.25 |
| $p_3$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $p_4$ | 0 | 0 | 0.01 | 0 | 0 | 0.02 |
| $p_5$ | 0.27 | 0.1 | 0.41 | 0.3 | 0.12 | 0.43 |

RMSE values from 1st to 71st day are reported in Figure 3. The dotted line represents the RMSE of Eq. 3, the continuous black line is the average RMSE of Eq. 4. In the early days, RMSE was low because the meteorological conditions were similar to the previous days on which the model parameters were based. Between the 23rd and the 31st day and between the 43rd and the 51st day the average error is very high (up to 15 μg/m$^3$ difference) due to sudden and robust weather variations. After the 51st day and until the end of the comparison the values appeared to be close to the average RMSE values of Table 2. This demonstrates the robustness of the model. Table 2 shows how the error values are lower (about 1 μg/m$^3$) for PM2.5 compared to PM10 due to lower concentration values for PM2.5.

Table 2. Average RMSE (μg/m$^3$) values obtained after 71st days for PM2.5 and PM10 concentrations for S1, S2, S3.

| | PM 2.5 | | | PM 10 | |
|---|---|---|---|---|---|
| S1 | S2 | S3 | S1 | S2 | S3 |
| 4.33 | 4.67 | 4.7 | 4.7 | 5.02 | 5.21 |

## 5. Conclusions

This work was focused on the development of a forecast model that uses experimental data of concentration of particular matter (PM2.5 and PM10). The data was obtained through 3 outdoor positioned monitoring stations in an area of about 10 km$^2$ between an industrial zone and a densely popular urban center. The model results showed that the coefficients of the linear model are different after 71 days of optimization of the model. The coefficients changed from station to another. The more influence meteorological parameters are the relative humidity, wind speed and hour. After 71 days, the model shows a sufficient agreement model-experimental data, equal to about 4 μg/m$^3$ is for PM2.5 and PM10. This value appears reasonable considering the high variability of concentrations of pollutants and that the limit value of the law is equal to 50 and 25 μg/m$^3$ for PM10 and PM2.5, respectively. RMSE values don't improve with more days. This is due to the change of season influencing the data strongly due to the different correlations between weather and pollution. Further data (until 12 months) are necessary in order to make model able to fit very well pollution data for every season.

## 6. References

[1] Sofia D, Giuliano A, Gioiella F. Air quality monitoring network for tracking pollutants: The case study of Salerno City center. Chemical Engineering Transactions. 2018;68: 67–72.

[2] Giuliano A, Gioiella F, Sofia D, Lotrecchiano N. A novel methodology and technology to promote the social acceptance of biomass power plants avoiding Nimby Sindrome. Chemical Engineering Transactions. 2018;67: 307–312.

[3] Sofia D, Giuliano A, Gioiella F, Barletta D, Poletto M. Modeling of an air quality monitoring network with high space-time resolution. In: Computer Aided Chemical Engineering. 2018;43:193-198.

[4] Wang J, Song G. A deep spatial-temporal ensemble model for air quality prediction. Neurocomputing. 2018;314:198-206.

[5] Li X, Ma Y, Wang Y, Liu N, Hong Y. Temporal and spatial analyses of particulate matter (PM10 and PM2. 5) and its relationship with meteorological parameters over an urban city in northeast China. Atmospheric Research. 2017;198:185-193.

[6] Sahanavin N, Prueksasit T, Tantrakarnapa K. Relationship between PM10 and PM2. 5 levels in high-traffic area determined using path analysis and linear regression. Journal of Environmental Sciences. 2018;69:105-114.