Problem statement & Method

Particulate matter (PM) pollution poses a major health concern worldwide, with PM2.5 (particulate matter less than 2.5 μm in diameter) being one of the most detrimental ones to human health. Within urban agglomerations, the landscape induces extra complexity, such as roads between high buildings (the so-called street canyons), where thermal phenomena also come into play [1]. As the human-generated emissions vary significantly both in time and space and influential meteorological conditions such as wind speed and relative humidity are extremely time-dependent, modeling and forecasting PM pollution at urban environments is a rather challenging problem. Various methods that utilize algorithms belonging in the broader field of Machine Learning (ML) have been developed in the literature, e.g. [2, 3].

The novelty of our approach lies in using a modified Long Short-Term Memory (LSTM) Neural Network that employs an attention-based mechanism to provide a level of interpretability for the final results, without compromising on forecasting accuracy [4]. The latter allows the main problematic regarding ML algorithms in Physics, namely their black-box approach which does not allow for insight in their output. Our method is quite general and is applicable to any set of measuring sensors.

Dataset compilation, Cleaning & Feature Selection

Our dataset consists of publicly available PM2.5 measurements from 21 Purplish Air PA-2 sensors (10 min resolution) and meteorological data from local stations (see Fig. 1). Each PM2.5 measurement, i, is taken equal to the mean of the sensor’s two channels [5] A and B, if
\[
\text{PM2.5}_i = \frac{A_i + B_i}{2} < 10\%,
\]
otherwise the data point is dropped.

We use the following features: UV High, Solar Radiation High, Wind Gust Average, Average Wind Direction, Average Wind Speed, Average Temperature, Mean Floor Area Ratio, Dew Point Average, Mean Population Density, Pressure, Precipitation Rate and Average Relative Humidity. In total, there are N = 253,546 data points, spanning the period from 2018-12-01 to 2022-06-19.

The prediction window used is 1h, taking into account the values of PM2.5 from the previous 24h, however we can make predictions for arbitrary time ranges by considering each predicted point as a new data point. The PM2.5 measurements utilized for the prediction of a new PM2.5 value are labelled as ‘Auto-recursive’ and correspond to our last feature.

Scaling, Training and Evaluation

The logarithm of PM2.5 values is taken in order to reduce the influence of possible outliers and then min-max scaling is applied for all features. The modified LSTM of [6] is trained on our dataset using an early stopping mechanism. We split the dataset into 50% training data, 20% validation data and 30% test data.

Feature importance: Insight into the Physics

The most important features in terms of their contribution to the predictions are Auto-regressive (14.39%), Mean Floor Area Ratio (14.39%), Wind Gust Average (14.39%), Temperature Average (14.39%) and Dew Point Average (14.39%). The importance of ‘Auto-regressive’ reflects the past values of PM2.5 measurements highly impact present ones. The importance of ‘Wind Gust Average’, stems from the fact that pollutants can efficiently cross the urban canopy layer and then be removed via diffusion, mostly through coherent structures of gust wind [7]. Wind Direction Average is connected to the generation and advection of turbulent structures, as well as being transported pollution. Mean Floor Ratio is also related to emergent turbulent structures, as a measure of building volume. Temperature is known to be closely correlated to PM2.5 concentrations [8]. Importantly, from training epoch 30 onwards, each feature’s importance practically converges to constant values, thus indicating a robustness in our interpretations (see Fig. 4).

Summary

Particulate matter pollution is a major concern, due to its adverse effects on human health worldwide. Towards undertaking mitigation measures, an accurate and efficient forecasting service is imperative. We construct a general framework for PM forecasting, employing publicly available PM measurements from low-cost sensors and open meteorological data with the aid of state-of-the-art machine learning algorithms. Specifically, we use a LSTM Neural Network that provides a level of interpretability. The spatial dependence of the phenomenon due to the complex urban agglomeration is taken into account using features such as population density and mean floor area ratio, for the first time in this field. The method is applicable to any type of sensors. As a case study, we apply our method to Patras, a previously unstudied Greek port-city, for the particular case of PM2.5 concentrations. It is found that the model is forecasting a accuracy that is comparable to the sensors’ resolution, combined with meaningful interpretations of its results.

Acknowledgments

The work is supported in part by project ENRIS+ under grant agreement No. MIS 507451 (from the General Secretariat for ERDF & CF, under Operational Programme Competitiveness, Entrepreneurship and Innovation 2014-2020 (ENI-CCO) of the Greek Ministry of Economy and Development (co-financed by Greece and the EU) through the European Regional Development Fund).

References


*Corresponding author: Fotios K. Anagnostopoulos, fotianag@uop.gr

The evaluation results and the test data are shown in Fig. 3. For a set of 50 runs, the Root Mean Square Error on the test set is 15.42 ± 0.30, while the Mean Absolute Error is 4.63 ± 0.27. This value corresponds to a maximum error similar to 25 μg/m^3, which is comparable to the sensor’s accuracy.

Figure 1. Locations of meteorological (blue) and particulate matter (red) measurement stations. Patras City, Greece.

Figure 2. Comparison of the two channels for these PA-2 sensors.

Figure 3. Testing of the trained model for a time period of 10 days ± 240 hours.

Figure 4. Feature importance variability over training epochs.

Figure 5. Temporal variability of each time-series point’s feature importance. The horizontal line sum to 1, with each element a_i corresponding to the percentage of importance for feature i (on time step i).