Forecasting PM2.5 concentration in port-cities

Fotios K. Anagnostopoulos*[1], Spyros Rigas [2], Michalis Papachristou [3], Christos Tryfonopoulos [1], Ioannis Chaniotis [3]

[1]Dep. of Informatics & Telecommunications, Univ. of Peloponnese [2]School of Electr. & Comp. Engineering, Nat. & Tech. Univ. of Athens [3]Dep. of Physics, Nat. and Kapodistrian Univ. of Athens

Problem statement & Method

Particulate matter (PM) pollution poses a major health concern worldwide, with PM2.5 (aerodynamic diameter less than 2.5 μm) 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, modelling 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 alleviates 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

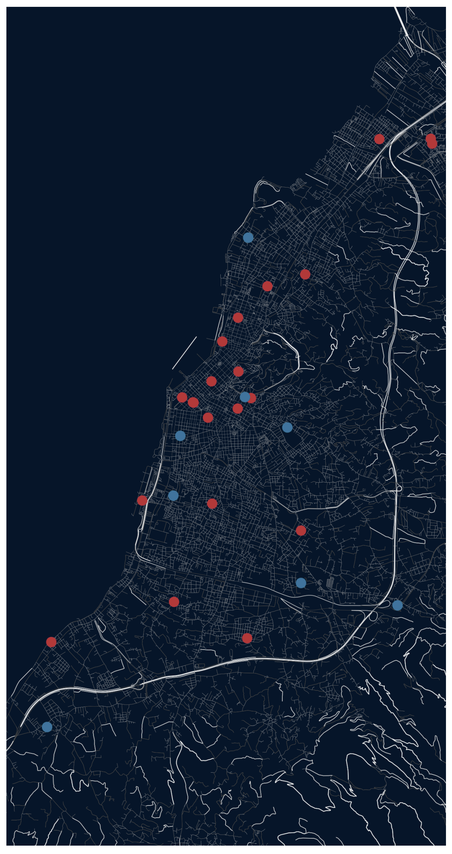

Figure 1. Locations of meteorological (blue) and particulate matter (red) measurement stations in Patras city, Greece.

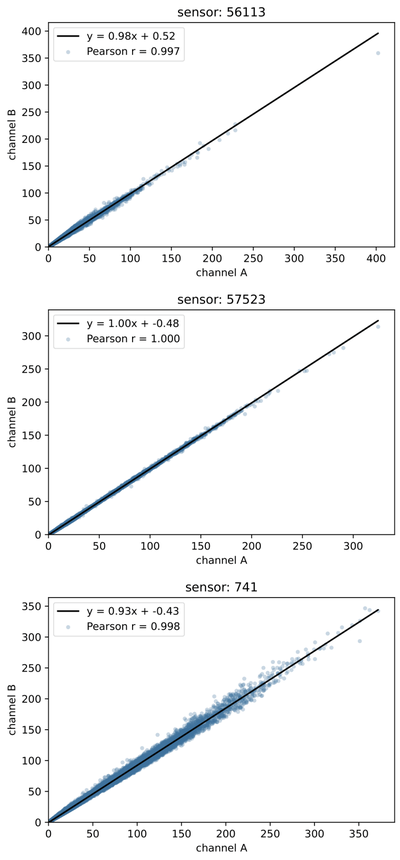

Figure 2. Comparison of the two channels for three PA - II sensors.

Our dataset consists of publicly available PM2.5 measurements from 21 PurpleAir PA - II 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

$$\frac{|\mathrm{PM2.5}_{\mathrm{A},i} - \mathrm{PM2.5}_{\mathrm{B},i}|}{\mathrm{PM2.5}_{\mathrm{A},i} + \mathrm{PM2.5}_{\mathrm{B},i}} < 10\%,$$

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 [6], Dew point Average, Mean Population Density, Pressure, Precipitation Rate and Average Relative Humidity. In total, there are $N = 253546$ 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-regressive" 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 [4] 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.

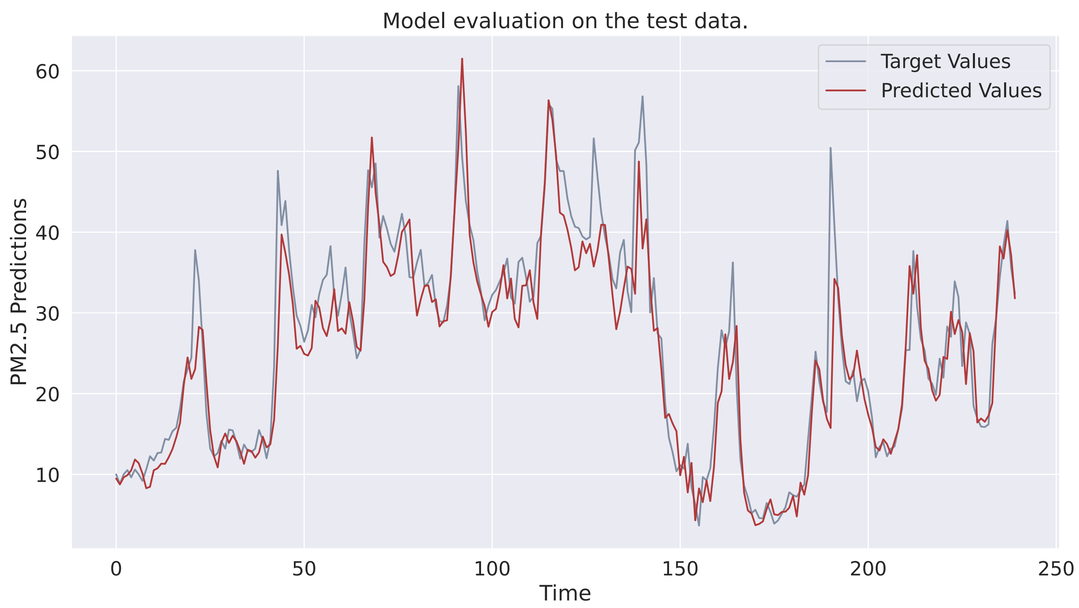

Figure 3. Testing of the trained model for a time period of 10 days ≡ 240 hours.

The evaluation results on (part of) 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 4.53 ± 0.27. This value corresponds to a maximum error similar to 15 $\mu g/m^3$, which is comparable to the sensor's accuracy.

Feature Importance: Insight into the Physics

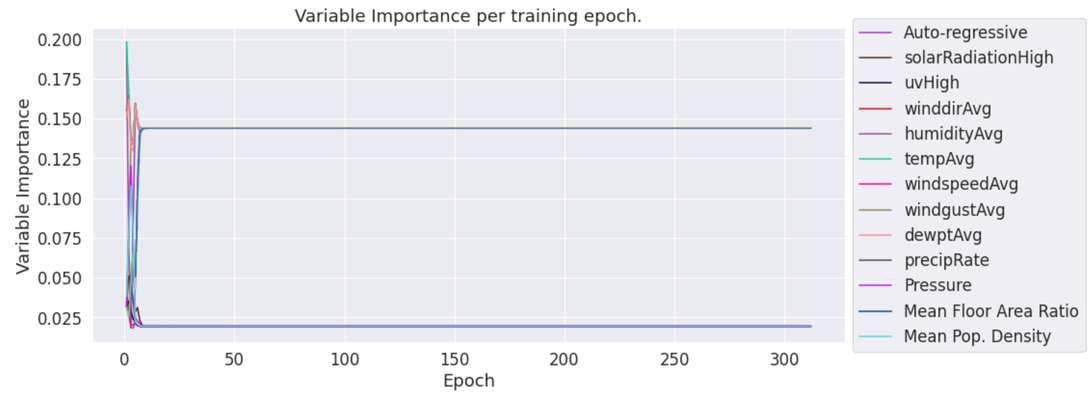

Figure 4. Feature importance variability over training epochs.

The most important features in terms of their contribution to the predictions are **Auto-regressive** (14.39%), **Mean Floor Area Ratio** (14.39%), **Wind Direction Average** (14.39%), **Wind Gust Average** (14.39%), **Temperature Average** (14.39%) **and Dew Point Average** (14.39%). The importance of "Auto-regressive" is obvious: 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 disruption of turbulent structures, as well as 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).

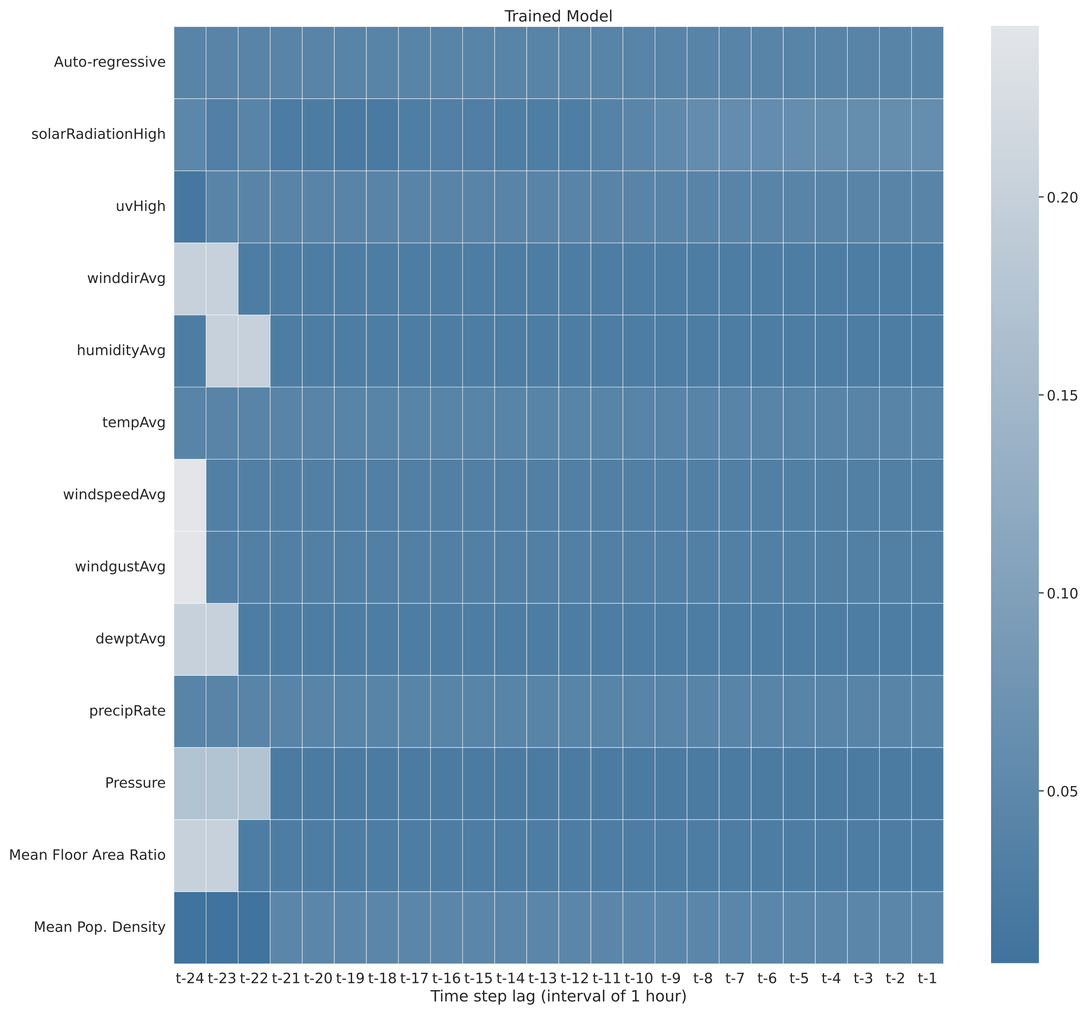

Figure 5. Temporal variability of each time-series point's feature importance. The horizontal lines sum to 1, with each element, a_{ij}, corresponding to the percentage of importance for feature i on time step j.

It is of interest to assess the contribution of each time step for a given data point (time-series) in the prediction. In Fig. 5, one observes that the most critical time-steps are from 24 to almost 22 hours before the the prediction. This could possibly be attributed to the periodicity of the phenomenon.

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 shows a forecasting accuracy that is comparable to the sensors' resolution, combined with meaningful interpretations of its results.

Acknowlegments

FA wishes to thank Ioannis Anastasiou, UoP for valuable help in obtaining meteorological data. The LaTeX theme used is based upon [9]. This work was supported in part by project ENIRISST+ under grant agreement No. MIS 5047041 from the General Secretary for ERDF & CF, under Operational Programme Competitiveness, Entrepreneurship and Innovation 2014-2020 (EPAnEK) of the Greek Ministry of Economy and Development (co-financed by Greece and the EU through the European Regional Development Fund).

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

References

[1] D. Pearlmutter, A. Bitan, and P. Berliner, "Microclimatic analysis of "compact" urban canyons in an arid zone," *Atmospheric Environment*, vol. 33, no. 24-25, pp. 4143–4150, 1999.

[2] J. Zhao, F. Deng, Y. Cai, and J. Chen, "Long short-term memory-fully connected (lstm-fc) neural network for pm2. 5 concentration prediction," *Chemosphere*, vol. 220, pp. 486–492, 2019.

[3] T. Xayasouk, H. Lee, and G. Lee, "Air pollution prediction using long short-term memory (lstm) and deep autoencoder (dae) models," *Sustainability*, vol. 12, no. 6, p. 2570, 2020.

[4] T. Guo, T. Lin, and N. Antulov-Fantulin, "Exploring interpretable lstm neural networks over multi-variable data," in *International conference on machine learning*, PMLR, 2019, pp. 2494–2504.

[5] K. K. Barkjohn, B. Gantt, and A. L. Clements, "Development and application of a united states-wide correction for $pm_{2.5}$ data collected with the purpleair sensor," *Atmospheric Measurement Techniques*, vol. 14, no. 6, pp. 4617–4637, 2021. DOI: 10.5194/amt-14-4617-2021. [Online]. Available: https://amt.copernicus.org/articles/14/4617/2021/.

[6] A. Faludi, *A reader in planning theory*. Elsevier, 2013, vol. 5.

[7] Y. Shi, Q. Zeng, L. Liu, X. Cheng, and F. Hu, "Important role of turbulent wind gust and its coherent structure in the rapid removal of urban air pollution," *Environmental Research Communications*, 2022.

[8] Y. Liu, Y. Zhou, and J. Lu, "Exploring the relationship between air pollution and meteorological conditions in china under environmental governance," *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.

[9] A. Athalye, *Github-anishathalye/gemini: Gemini is a modern latex beamerposter theme*, 2022. [Online]. Available: https://github.com/anishathalye/gemini.