CHEMICAL ENGINEERING TRANSACTIONS
VOL. 83, 2020
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Bruno Fabiano, Valerio Cozzani, Genserik Reniers

ISBN 978-88-95608-81-5; **ISSN** 2283-9216

A data driven model for ozone concentration prediction in a coastal urban area

Tomaso Vairo [a,c] *, Andrea Rapuzzi [b], Mario Lecca [c], Bruno Fabiano [a]

[a] DICCA - Civil, Chemical and Environmental Engineering Dept. – Genoa University, via Opera Pia 15 - 16145 Genoa, Italy
[b] A-SIGN S.r.l - via XXV Aprile 10/3a - 16121 Genoa, Italy
[c] ARPAL, via Bombrini 8 - 16149 Genoa Italy
* tomaso.vairo@edu.unige.it

This study is focused on the development of a Data Driven model, based on LightGBM, an algorithm for gradient boosting on decision trees, which is able to predict the ozone concentration in the urban area of Genova. The system represents a pragmatic and scientifically credible approach to data driven modelling applied to complex and uncertain situations. In particular, the present work concerns the application of data analytic standard methodologies to air quality analysis, which includes the pre-treatment of data, the choice of a suitable configuration of the learning algorithm, the identification of the fitting parameters, and how to minimize the errors. The data are significant statistical time-series of the past years from the air quality monitoring network in the urban area of Genova. From that data series, the learning, test and validation dataset are taken. The system used for data assimilation, construction and network learning, testing and validation, is completely based on an open source statistical processing software.

Keywords: data driven model, machine learning, air quality, ozone.

1. Introduction

The protection of air quality from atmospheric pollution and the reduction of greenhouse gas emissions are essential goals and are increasingly important in international, national and regional strategies and policies. Air pollution increases the risk of respiratory and heart disease, and it is recognized as a major environmental and health risk.
Tropospheric ozone (O_3) is a secondary pollutant, formed as a result of chemical reactions that take place in the atmosphere starting from the precursors (in particular nitrogen oxides and volatile organic compounds). These reactions take place by strong solar radiation and high temperatures.
Ozone pollution is a characteristic phenomenon of the summer period and the highest concentrations are usually found in the afternoon and in suburban areas placed leeward with respect to the main urban areas.
The reference values for health protection in the Italian legislation are the following Table 1:

Table 1: Ozone reference values in Italian legislation

Reference	Concentration
Information threshold on the hourly average	180 µg / m^3
Alarm threshold on the hourly average	240 µg / m^3 for 3 consecutive hours
Target value	120 µg / m^3 as daily avg. over 8 hours, not to be exceeded more than 25 times/y
Long-term target value	20 µg / m^3 as daily average over 8 hours

For the O_3 parameter the following table 2 shows the exceedances of the Information threshold and the alarm threshold in 2018. The number of days of exceeding the target value and the long-term target value in the year 2018 are shown (Regione Liguria 2018):

Table 2: days in 2018 exceeding the target values for O_3

Urban station	nr. of days exceeding the target value	nr. of days exceeding the long-term target value
Quarto	69	6
Corso Firenze	52	9
Parco Acquasola	108	89

The main operational tools for air quality planning are monitoring systems and the regional inventory of emissions with indications on the regulatory framework. In order to plan useful actions for achieving environmental objectives, it is important to have reliable forecasting tools. Two main modelling approaches can be outlined: the former is based on numerical modelling by simulating atmospheric dispersion and transport starting from emission inventory. The latter is based on advanced statistical models based on machine learning algorithms.
The focus of this work is to evaluate the results that proper data analysis techniques (i.e. proper regularization, data pre-treatments ...), and the selected learning algorithm, could achieve to perform a reliable forecasting of critical ozone concentrations.

1. Methodology

The development of this ozone forecasting model includes:

1.1 Data collection and preprocessing

Meteorological and pollutant data for the time period starting on May 2015 until the end of 2018 has been obtained for the 3 above mentioned metropolitan zones (Quarto, Corso Firenze, Parco Acquasola) respectively. Data have been summarized on a daily basis (Garcìa et al. 2011). The following input variables has been considered:

1. Time variables: day of the year (doy), day of the week (dow), month.
2. Meteorological variables (daily aggregate): mean sea level pressure (MSLP), solar radiation (SLHR, SSHR), temperature (TEMP), wind direction and speed (UWIND, VWIND, MOD) humidity (HUM) and rain (RAIN).
3. Pollutants (daily aggregate): ozone (O3) daily mean.
4. Bank holiday information for each day (true or false) to consider the influence of holidays on O_3 concentration.

The time variables in (1) are assimilated via trigonometric functions (sin and cos), in order to account for the cyclic nature of their impact (Eapi et al. 2013).
The meteorological variables in (2) have been summarized on a daily frequency with min, max, average and standard deviation functions.
In addition, to provide the model with easier to correlate information, we have added redundant values to each row related to previous days values (e.g. meteorological values from one day before or for one year before).

1.2 Validation strategy

In accordance to the best practices for time series validation, we have cross validated our results with a Walk-Forward approach (Cao et al. 2003).
This is an approach that allow to have a robust estimation of the model performances without leaking information from the training to the validation set (see Figure 1).
Each fold has the following sub-sets:

- Training: contains data points belonging to the time interval from T_0 to T_t (included), where T_0 is the oldest available data point and $T_t - T_0$ is a sufficient time interval to train the model on the problem
- Validation: contains data points belonging to T_{t+1}
- Test: contains data points belonging to T_{t+2}
- Unused: contains data points belonging to a time more recent than T_{t+2}

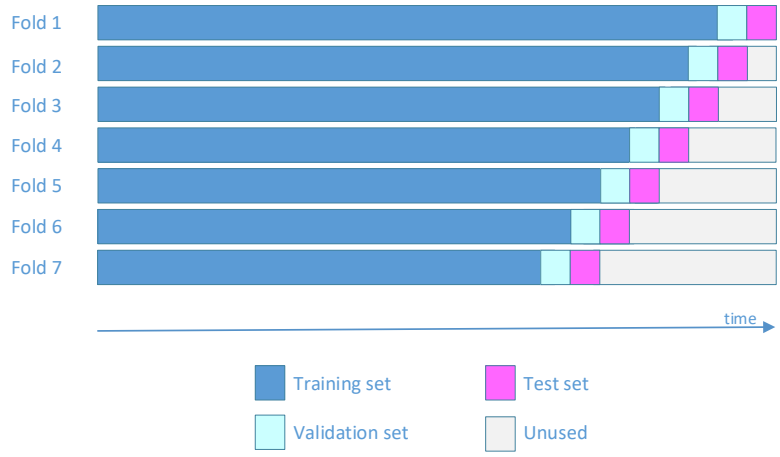

Figure 1: Walk-Forward Validation

Validation data is used to perform early stopping of the learning process to identify when the model starts to overfit. When we use a single day to select the validation interval we have an increased variance due to, among other factors, a premature stopping of the training for an initial (random) fitness of the model to the small validation data.
In order to limit noise in the early stopping process (introduced by random good initial fit on such a small validation set) we tested two variations of the strategy:

- A. Using a 7 days interval for the validation set (see Figure 2). Since the model final performances are measured on the test set (whose interval is kept one-day long), we can introduce a small data leakage between the training and validation to stabilize the validation score and the early stopping strategy.
- B. Running a small number of training epochs without early stopping before the full training process (the same as in Figure 1, with a “warmup” step).

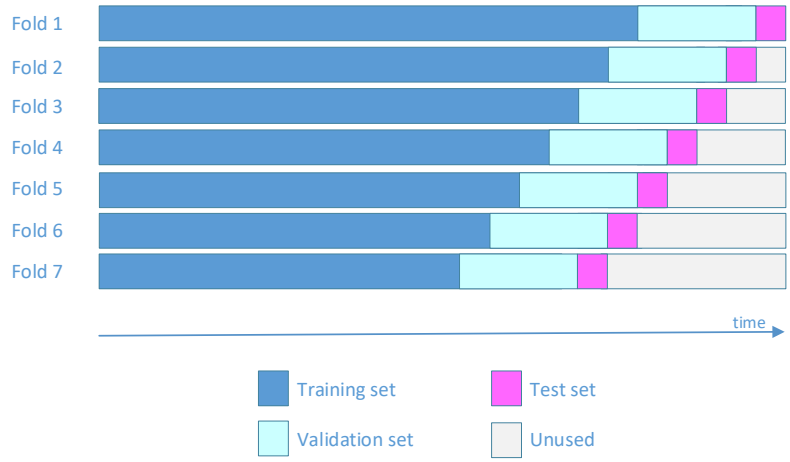

Figure 2: Walk-Forward Validation – Type A

1.3 Learning algorithm

We used a *LightGBM* model for the regression task (Zhang et al. 2019). *LightGBM* is a gradient boosting library based on decision trees algorithms.
It is a highly optimized library that performs very well in structured/tabular data problems, capable of managing gracefully a mix of scalar and categorical variables (Ke et al. 2017).

2. Results and discussion

Several run tests were run, with very wide Walk-Forward window to test the model convergence and its dependence on the training data dimension.
As depicted in Figure 3, higher folds use a progressively smaller training set, having bigger unused data.

Figure 3 shows an example of the performances (Mean Absolute Error) on the validation and test sets across 300 folds. Validation and test data are quite noisy, but their average (in the interval Fold 0, Fold i) tends to converge. After around 130 folds the model performances degrade slowly.

This is even more evident considering the Validation and Test moving averages (across the 20 Folds) for the same experiment depicted in Figure 4. This means the model is needs to be trained with most of the available data to reach top performances.

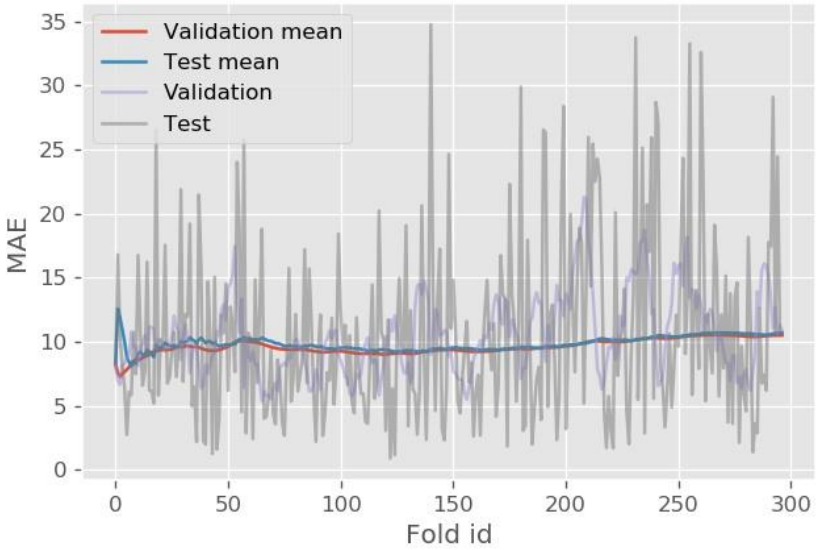

Figure 3: Model Performance by Fold Id

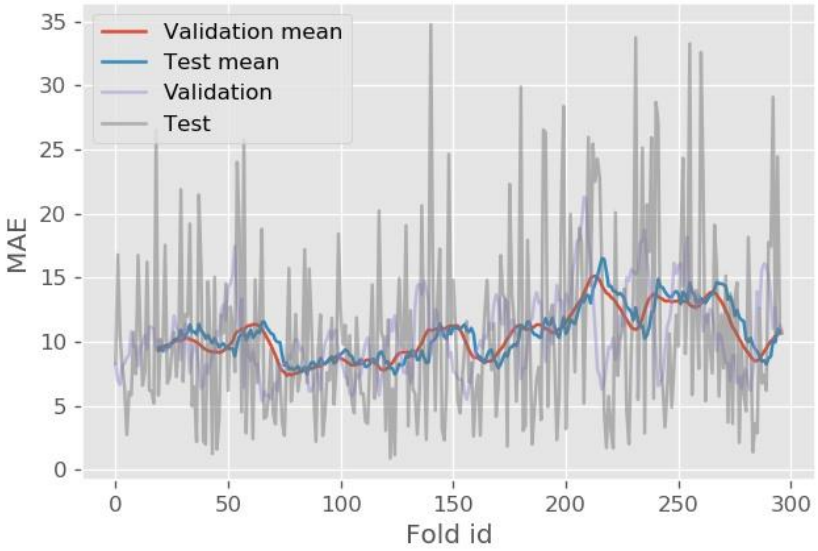

Figure 4: Model Performance by Fold Id - Moving Averages

The variation A to the validation strategy (7 days for the validation interval), described in par. 2.2, has both helped in stabilizing the validation score (see Figure 5) and increasing the test performance.

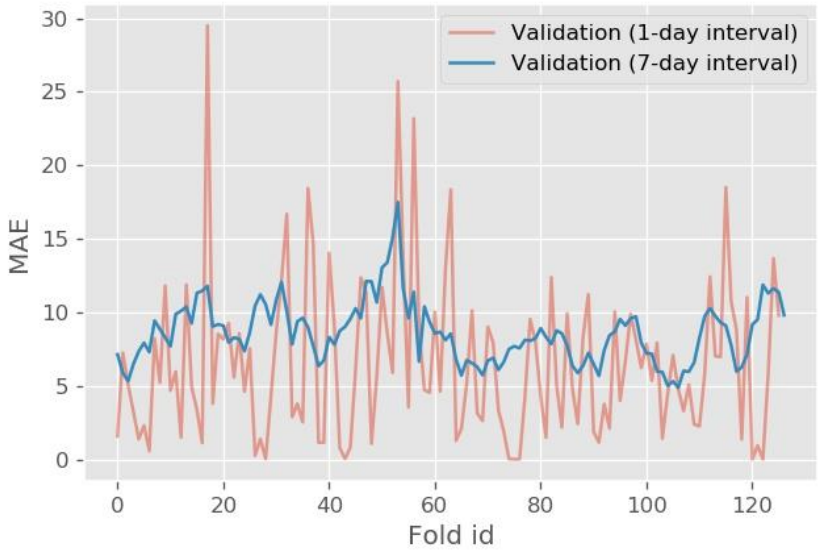

Figure 5: Validation Performance - Variation A

The variation B to the validation strategy (small training pre-run), described in par. 2.2, has not helped in stabilizing the validation score (as visible in Figure 6) but has provided the best overall test performance.

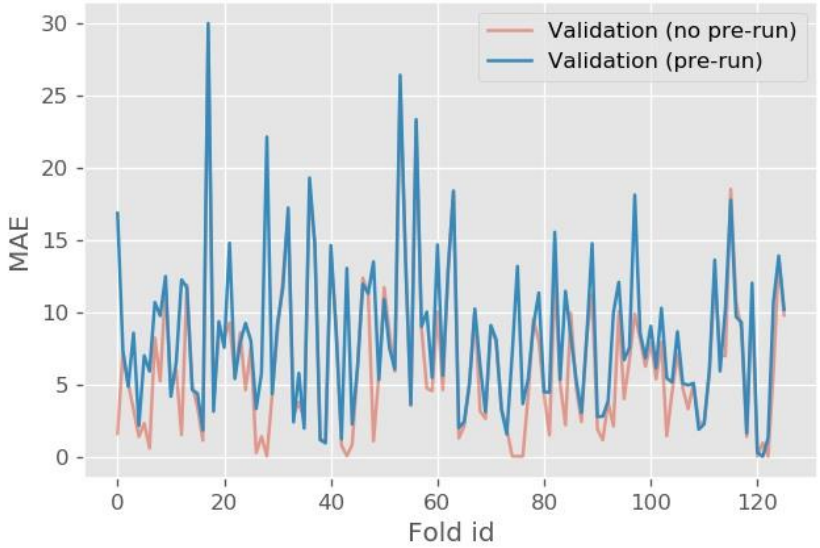

Figure 6: Validation Performance - Variation B

To provide a reference point a naïve prediction has been performed using previous-day value as a prediction. In the following Table 3, the scores in different configurations are shown.

Table 3: Model scores

	Validation mean score	Test mean score
Naïve prediction	NA	14.011
Walk-Forward	6.509	9.456
A - Walk-Forward 7-day validation	8.593	9.066
B - Walk-Forward pre run	8.320	**8.875**

The relation between the ground truth and the prediction (ozone ppm), along with the identity line representing the perfect prediction space, is depicted in Figure 7.

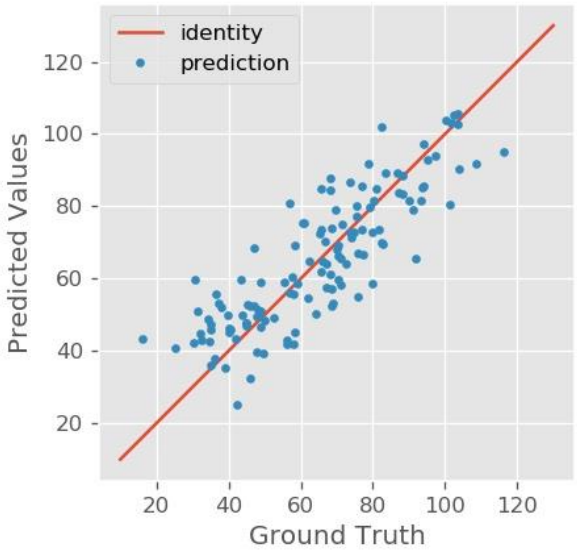

Figure 7: Prediction vs Ground Truth (ppm)

Figure 8 shows the same relation in comparison to the naïve prediction which is evidently less clustered around the identity line:

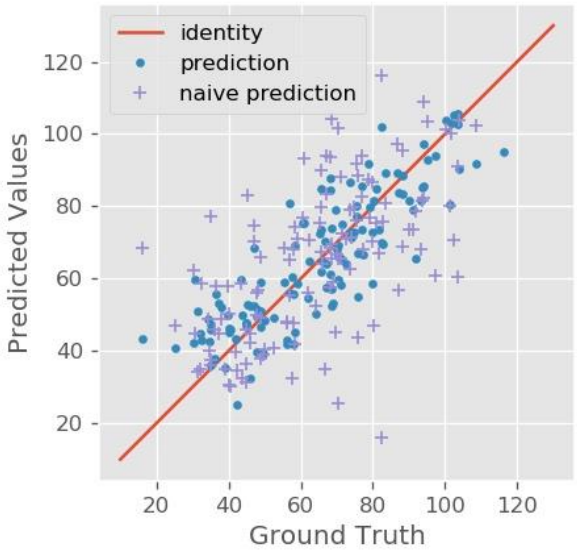

Figure 8: Prediction vs Ground Truth vs Naive Prediction (ppm)

3. Conclusions

A machine learning algorithm based on gradient boosting techniques has been built to predict the concentration of O_3 in the metropolitan area of the city of Genoa. The predictive model was trained with meteorological data, ozone measurements in three urban areas, and time variables, all suitably pretreated as described above.

The best cross-validation strategy was therefore selected, in order to balance bias and variance in the prediction results, and thus avoid situations of under-specification and over-specification. The model thus built showed excellent results.

This work complements and improves the previous predictive model developed for PM10 prediction (Vairo et al. 2019), which is based on Bayesian inference.

As a further development it seems interesting to study further models to predict the NO_X concentration, in order to have a predictive system for all the main pollutants that have an effect on air quality.

References

Vairo, T., Lecca, M., Trovatore, E., Fabiano, B., Reverberi, A.P., 2019, A Bayesian Belief Network for Local Air Quality Forecasting, Chemical Engineering Transactions, 76.

Eapi, G.R., Sattler, M., Manry, M.T., 2013, Comprehensive Ozone Forecasting Model using Neural Networks, Conference paper. https://www.researchgate.net/publication/280026476.

García, I., Rodríguez, J.G.,Tenorio, Y.M., 2011, Artificial Neural Network Models for prediction of ozone concentrations in Guadalajara, Mexico, Air Quality-Models and Applications, InTechOpen 2011.

Regione Liguria, ARPAL, Valutazione annual di qualità dell'aria, 2018, Regional annual air quality report, http://www.ambienteinliguria.it/eco3/DTS_GENERALE/20191014/ValutazioneAnnuale_2018.pdf

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T., 2017, LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems, 30.

Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R., Wang, Q., Huang, L,. 2019. A Predictive Data Feature Exploration-Based Air Quality Prediction Approach. IEEE Access. 1-1.

Cao, L.J., Tay, F., 2003. Support vector machine with adaptive parameters in financial time series forecasting. Neural Networks, IEEE Transactions on. 14.