CHEMICAL ENGINEERING TRANSACTIONS
VOL. 82, 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 978-88-95608-80-8; **ISSN** 2283-9216
DOI: 10.3303/CET2082064

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

As amply known, ozone concentration in the coastal area of study is well relevant in connection with "photochemical smog", due to high levels of solar radiation and temperature values and possible photochemical oxidation of volatile organic compounds (VOCs) in the presence of nitrogen oxides (NOx). In this paper, a framework for predicting ozone concentration in urban area is presented, relying a *LightGBM* algorithm for gradient boosting on decision trees. The system represents a pragmatic and scientifically credible approach to data driven modelling applied to complex and uncertain situations. The study 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 error minimization. Training and verification data are significant statistical time-series over the past years validated from the air quality monitoring network in the urban area of Genoa (Italy).

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

1. Introduction

The protection of air quality from pollution and the reduction of greenhouse gas emissions are essential goals gaining increasing attention in international and national strategies and policies. In this context, many attributes (such as safety, environment, reputation, policy, costs, etc.) need to be properly taken into consideration when prioritising safety plant/industry investments (Abrahamsen et al., 2020). The transition to low-carbon economy and the ambition to reach net zero emissions offers research challenges addressing pollution prevention, e.g. by advanced pyrolysis processes recovering mass and energy (Chiarioni et al., 2006). Further to emission reduction process, the enhancement of climate change resilience requires advanced pollution modelling forecasting for both emergency situations (Fabiano et al., 2017) and conventional environmental risk assessment (Sikorova et al., 2017). Two different approaches can be sorted in air pollution modelling: the former relies on atmospheric dispersion modelling of pollutants by simulating diffusive and transport mechanism (e.g. Vairo et al., 2014) and once correctly defined source terms and chemical processes involved, can be properly applied also to non-stationary sources (Vairo et al., 2017). The latter is based on advanced statistical models, such as machine learning methodologies, e.g. relying on statistical data elaboration from air monitoring networks.

Table 1: Ozone (O_3) reference values set down by Italian legislation.

Reference	Ozone 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 on 8-hour average	120 µg / m^3 as daily, not to be exceeded more than 25 times/y
Long-term target value on 8-hour average	20 µg / m^3 as daily average

Paper Received: 5 January 2020; Revised: 27 March 2020; Accepted: 20 July 2020
Please cite this article as: Vairo T., Rapuzzi A., Lecca M., Fabiano B., 2020, A Data Driven Model for Ozone Concentration Prediction in a Coastal Urban Area, Chemical Engineering Transactions, 82, 379-384 DOI:10.3303/CET2082064

As a matter of fact, analogously to the risk assessment domain, main improvement challenges are based on the application of machine learning techniques and big data exploitation (De Rademaeker et al., 2014). In this regard, predicting ability is strictly connected to spatial and time interpolation schemes, such as Multiple Linear Regression (MLR), or Artificial Neural Network (ANN) for non-linear problems (e.g. Wand & Quian, 2018). Air pollution increases the risk of respiratory and heart disease, being recognized as a major environmental and health risk. As amply known, tropospheric ozone is a secondary pollutant, formed as a result of chemical reactions occurring in the atmosphere starting from the precursors (nitrogen oxides and volatile organic compounds), under high solar radiation level and elevated temperature conditions. Ozone pollution is a well relevant and characteristic phenomenon of the summer period, with the highest concentrations usually recorded in the afternoon, in suburban areas placed leeward with respect to the main urban areas. The forecasting ability of well-developed data driven model can outperform the predictions attained by mechanistic models due to inherent approximations and uncertainties in the emission source estimation (Cobourn et al., 2010).

Table 2: Exceedances of the target and long-term target values set out by legislation in the year 2018.

Urban station	Target value exceedances [day]	Long-term target value exceedances [day]
Quarto	69	6
Corso Firenze	52	9
Parco Acquasola	108	89

The legislative reference values for health protection in the Italian legislation, in terms of non-compliance limits are summarized in Table 1. Table 2 summarizes the number of days of the year 2018 exceeding the target value and the long-term target value, experimentally obtained by the monitoring network (Regione Liguria, 2018). 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. The focus of this work is to evaluate the results that advanced data analysis techniques (i.e. proper regularization, data pre-treatments) coupled with a learning algorithm framework can achieve in reliable forecasting ozone concentrations. Focusing on trend forecasting, the remainder of this paper is as follows. Section 2 describes the methodology including modelling dataset and learning model, Section 3 presents the data statistics and the forecasting results with descriptions of contributions made, while in Section 4 conclusions are drawn, with the strengths of the proposed technique and future work.

2. Methodology

2.1 Data collection and preprocessing

In this paper, we consider air quality and meteorological data measured in the urban area of the town of Genoa (Italy) over the time span May 2015- December 2018. Raw data were obtained for the three metropolitan zones of Genoa (Italy), i.e.: Quarto, Corso Firenze, Parco Acquasola, respectively. Upon validation, data have been statistical elaborated 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 (O_3) daily mean.
4. Bank holiday information for each day (true or false) to consider the influence of holidays on ozone concentration.

As suggested by Eapi et al. (2013,) time variables under heading 1 were assimilated via trigonometric functions, in order to account for the cyclic nature of their impact. The meteorological variables under heading 2 have been summarized on a daily frequency utilizing minimum, maximum, average and standard deviation functions. Additionally, in order to improve the prediction accuracy when forecasting ozone concentration by information correlation, we have added redundant values to each row related to previous time span (e.g. meteorological values from one day before or for one year before).

2.2 Validation strategy

In accordance to the best practices for time series validation, we have cross validated the obtained results, according to a customized and accurate Walk-Forward approach (Cao et al. 2003) as detailed in the following.

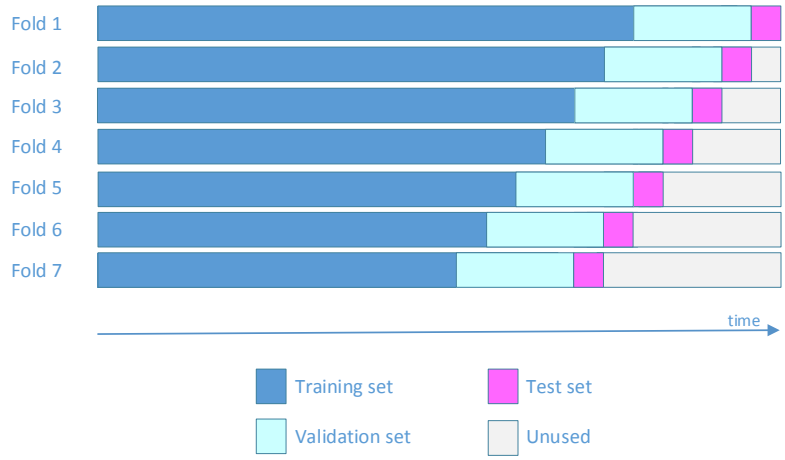

Figure 1: Validation results based on a Walk-Forward approach A.

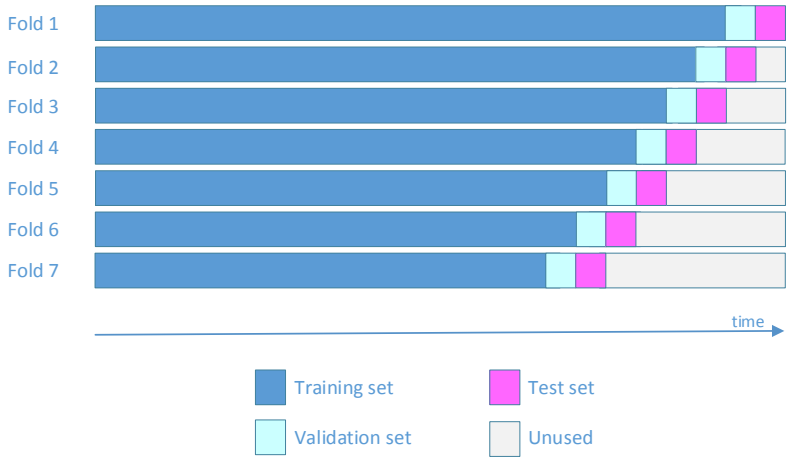

Figure 2: Validation results based on a Walk-Forward approach B.

This is an approach that allows achieving a robust estimation of the model performances, without leaking information from the training to the validation set shown as a general example in Figure 1. Each fold consists of 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}

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 strategy implementations, as follows. A validation scheme based on the selection of 7 days interval for the validation set. As evidenced in Fig. 1, 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, in order to stabilize the validation score and the early stopping strategy. B. validation scheme based on running a small number of training epochs without early stopping before the full training process. In this case, the resulting trend will evidence a “warm-up” step (see Figure 2).

2.3 Learning algorithm

We selected a *Light GBM* model learning method mainly based on the Decision Tree algorithm, and frequently used in classification tasks. (Zhang et al., 2019). *Light GBM* is based on a highly optimized library that performs very well in structured/tabular data problems, capable of gracefully managing a mix of scalar and categorical variables (Ke et al., 2017). However, research on *Light GBM* application for spatial forecasting ability in the field of air quality is limited. It is noteworthy, noting that the system used for data assimilation, construction and network learning, testing and validation, is completely based on an open source statistical processing software. In the next chapter, the application of the validation schemes is thoroughly discussed, in order to evidence how the newly built model reflects better fitting effect and predictive data feature.

3. Results and discussion

Several run tests were performed according to a wide Walk-Forward window in order to test the model convergence and its dependence on the training data dimension. As clearly depicted in Figure 3 a, higher folds use a progressively smaller training set: it provides an example of the performances in terms of 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 nearly 130 folds the model performances degrade slowly. The trend is even more evident considering the Validation and Test moving averages (across the 20 Folds) for the same experiment depicted in Figure 3 b, evidencing that the model needs to be adequately trained with nearly 80% of the available data to reach top performances.

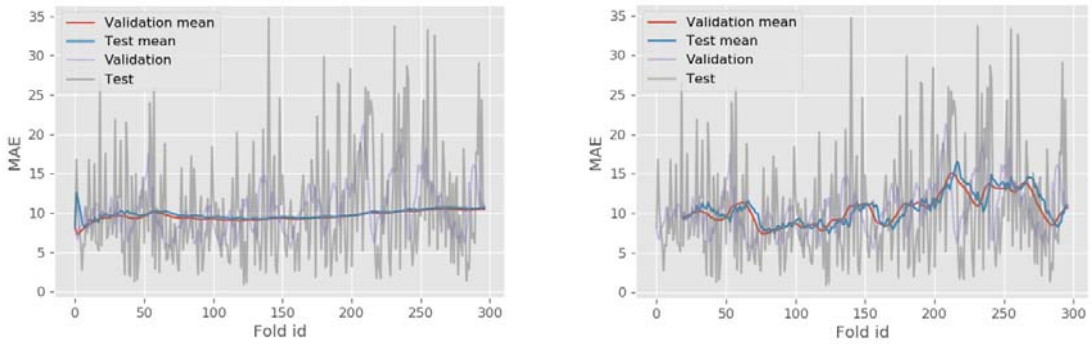

Figure 1: (a) Model Performance by Fold Id; (b) Model Performance by Fold Id - Moving Averages

The variation a in relation with the 7 days-validation strategy above described allowed stabilizing the validation score and increasing the test performance, as shown in Figure 4 a. Conversely, the variation B according to the validation strategy based on a small training pre-run approach previously outlined, was un-effective in stabilizing the validation score but has provided the best overall test performance (see Fig. 4 b). The model is sensitive to modelling data size and its performance degrades when data are too few. In order to provide a reference point a naïve prediction has been performed using previous-day value as a prediction.

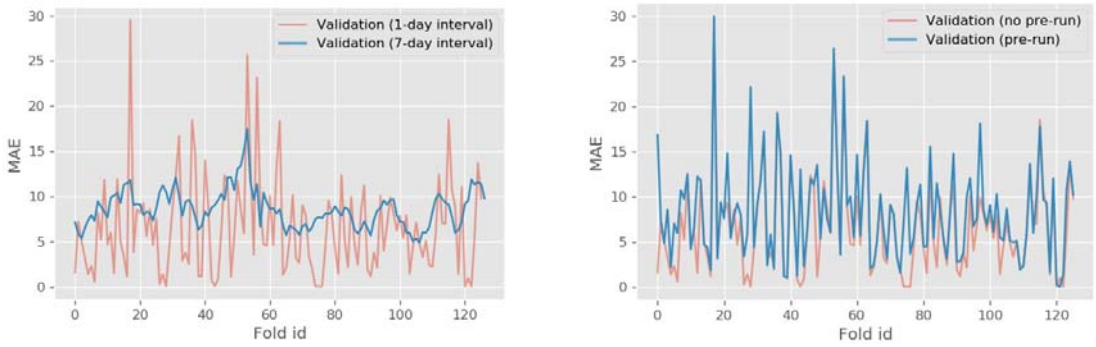

Figure 4: (a) Validation Performance - Variation A; (b) Validation Performance - Variation B.

Table 3: Model scores on validation and testing.

	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

In Table 3, the model performance on validation and testing scores are summarized, by considering the different configurations previously outlined. The comparison of Ozone concentrations [ppm] experimentally observed (ground truth) and the model prediction is depicted in Figure 5.

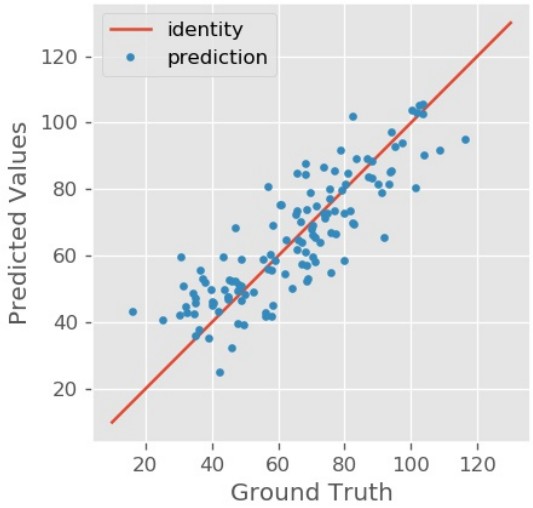

Figure 5: Predicted Ozone concentration [ppm] vs experimental values [ppm] (Ground Truth).

Figure 6 shows the comparison of Ozone concentrations [ppm] experimentally observed (ground truth) to the naïve prediction: results reveal that the model yields again satisfactory predictions evidently less clustered around the identity line. Because of its predicting ability, this method can not only be used to forecast surface ozone concentrations, but also be used to make predictions of other air pollutants, upon proper refinement

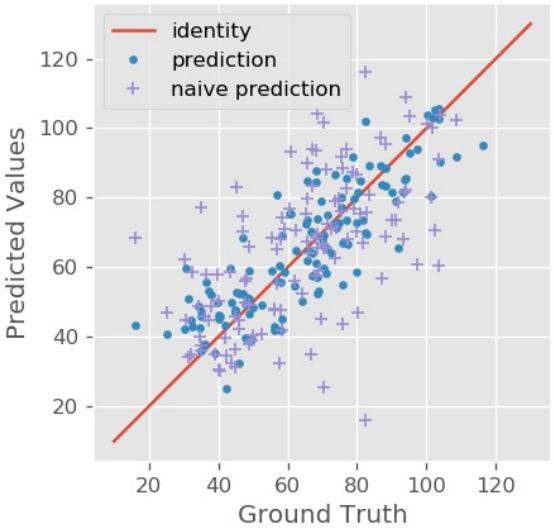

Figure 6: Predicted Ozone concentration [ppm] vs experimental values [ppm] (Ground Truth) and Naive Prediction.

4. Conclusions

The work presented in this study aims to examine the feasibility of applying a machine learning algorithm based on gradient boosting techniques to predict the concentration of O_3 in the metropolitan area of the city of Genoa. The model is based on a relative novel algorithm used in many different kinds of data mining tasks, such as classification, regression and ordering, while its application in the given urban context is still rather limited. 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 was developed by a Bayesian inference approach. As a further refinement and extension of the study, it seems interesting to extend the framework to cover nitrogen oxides concentration too, in order to develop an overall predictive system of the main pollutants relevant for photochemical pollution and their environmental synergistic impact.

References

Abrahamsen E.B., Milazzo M.F., Selvika J.T., Asche F., Abrahamsen H.B., 2020, Prioritising investments in safety measures in the chemical industry by using the Analytic Hierarchy Process, Reliability Engineering & System Safety, 198, article 106811.

Cao L.J., Tay F., 2003. Support vector machine with adaptive parameters in financial time series forecasting. Neural Networks, IEEE Transactions on Neural Networks 14(6), 1506-1518.

Cobourn W.G., Dolcine L., French,M., Hubbard M.C., 2000. A comparison of nonlinear regression and neural network models for ground-level ozone forecasting. J. Air Waste Manage. Assoc. 50, 1999-2009.

Chiarioni A., Reverberi A.P., Fabiano B., Dovì V.G., 2006, An improved model of an ASR pyrolysis reactor for energy recovery, Energy 31, 2460-2468.

De Rademaeker, E., Suter, G., Pasman, H.J., Fabiano, B. 2014. A review of the past, present and future of the European Loss Prevention and Safety Promotion in the Process Industries. Process Safety and Environmental Protection 92, 280-291.

Eapi G.R., Sattler M., Manry M.T., 2013, Comprehensive ozone forecasting model using neural networks, Conference paper. https://www.researchgate.net/publication/280026476.

Fabiano B., Vianello C., Reverberi A.P., Lunghi E., Maschio G., 2018, A perspective on Seveso accident based on cause-consequences analysis by three different methods, Journal of Loss Prevention in the Process Industries, 49, 18-35.

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.

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, 3146-3154.

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

Sikorova K., Bernatik A., Lunghi E., Fabiano, B., 2017, Lessons learned from environmental risk assessment within the framework of Seveso Directive in Czech Republic and Italy, Journal of Loss Prevention in the Process Industries, 49, 47-60.

Vairo T., Currò, F., Scarselli, S., Fabiano, B., 2014. Atmospheric emissions from a fossil fuel power station: dispersion modelling and experimental comparison. Chemical Engineering Transactions 36, 295-300, DOI:10.3303/CET1436050

Vairo T., Del Giudice T., Quagliati M., Barbucci A., Fabiano B., 2017, From land- to water-use-planning: A consequence-based case-study related to cruise ship risk, Safety Science 97, 120-133.

Vairo T., Lecca M., Trovatore E., Reverberi A., Fabiano B., 2019, A Bayesian Belief Network for local air quality forecasting, Chemical Engineering Transactions, 74, 271-276 DOI:10.3303/CET1974046

Wang B., Quian F. 2018. Three-dimensional gas dispersion modeling using cellular automata and artificial neural network in urban environment. Process Safety and Environmental Protection, 120, 286-30.

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 7, 30732-30743.