Environmental Modelling and Software 110 (2018) 52–61

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A novel approach to forecast urban surface-level ozone considering heterogeneous locations and limited information

Álvaro Gómez-Losada[a,*], G. Asencio–Cortés[b], F. Martínez–Álvarez[b], J.C. Riquelme[c]

[a] European Commission, Joint Research Centre (JRC) Edificio Expo, C/ Inca Garcilaso 3, 41092 Seville, Spain
[b] Division of Computer Science, Pablo de Olavide University, ES-41013 Seville, Spain
[c] Department of Computer Science, University of Seville, Spain

ARTICLE INFO

Keywords:
Time series
Forecasting
Data science
Ozone concentration

ABSTRACT

Surface ozone (O_3) is considered an hazard to human health, affecting vegetation crops and ecosystems. Accurate time and location O_3 forecasting can help to protect citizens to unhealthy exposures when high levels are expected. Usually, forecasting models use numerous O_3 precursors as predictors, limiting the reproducibility of these models to the availability of such information from data providers. This study introduces a 24 h-ahead hourly O_3 concentrations forecasting methodology based on bagging and ensemble learning, using just two predictors with lagged O_3 concentrations. This methodology was applied on ten-year time series (2006–2015) from three major urban areas of Andalusia (Spain). Its forecasting performance was contrasted with an algorithm especially designed to forecast time series exhibiting temporal patterns. The proposed methodology outperforms the contrast algorithm and yields comparable results to others existing in literature. Its use is encouraged due to its forecasting performance and wide applicability, but also as benchmark methodology.

1. Introduction

Ozone (O_3) is an ubiquitous, secondary photochemical air pollutant that is formed when volatile organic compounds, nitrogen oxides and carbon monoxide –the three ozone precursors– react in the presence of short wavelength solar radiation. To date, surface O_3 is considered as the most damaging air pollutant in terms of adverse effects on human health, vegetation crops and material (Paoletti, 2006; Sicard et al., 2016).

Concentrations of surface O_3 can shift rapidly over hours and days, sometimes reaching levels that can exceed prescribed thresholds considered to be safe for health, particularly for the most vulnerable segments of the population. Predicting the temporal evolution of O_3 concentration in specific urban locations emerges as a priority for guaranteeing quality of life, providing the population in affected areas with accurate information and alerting them when exceptionally high levels are present.

Any threshold value exceedance, accurately forecasted in advance, allows environmental authorities to apply short-term pollution control measures and abatement strategies to protect the population. Traditionally, environmental modelers have relied on multiple information to perform predictions (Corani and Scanagatta, 2016), incorporating numerous predictors related to O_3 formation to the general formulation of the forecasting models.

However, on many occasions, the observation data available from monitoring sites do not ensure quality requirements or may often be limited to few parameters. Hence, relevant O_3 chemical precursors or originators traditionally used as input parameters which strongly contribute to perform better forecasts, cannot be considered.

Traditional time series (TS) techniques fail to forecast O_3 accurately (Chattopadhyay and Bandyopadhyay, 2007). As a replacement, machine learning (ML) techniques have emerged and proved to be more effective for O_3 prediction (Gong and Ordieres-Meré, 2016; Martínez-Ballesteros et al., 2010; Martínez-Ballesteros et al., 2011). Since 2006, ensemble forecasting has begun to receive more attention, as ensemble algorithms can improve forecasting accuracy and enhance the generalization capability (Zhang et al., 2012). However, they hold the inherent limitations associated with the single ML model's accuracy.

This study introduces a methodology based on bagging and ensemble learning to forecast 24 h-ahead hourly O_3 concentrations, which was evaluated using ten-year (2006–2015) hourly O_3 TS obtained from three major urban areas of Andalusia (Spain). The different air quality monitoring site locations permitted to evaluate the forecasting results in a wide range of pollution levels and urban scenarios. The main

* Corresponding author.
E-mail addresses: alvaro.gomez-losada@ec.europa.eu (Á. Gómez-Losada), guaasecor@upo.es (G. Asencio–Cortés), fmaralv@upo.es (F. Martínez–Álvarez), riquelme@us.es (J.C. Riquelme).

https://doi.org/10.1016/j.envsoft.2018.08.013
Received 3 March 2017; Received in revised form 5 August 2018; Accepted 16 August 2018
Available online 22 August 2018

contribution of this methodology is to use just lagged O_3 concentrations as predictors, without requiring the participation of any other variable related to the O_3 formation in urban environments. This methodology was compared with the Pattern Sequence-based Forecasting PSF (Martínez-Álvarez et al., 2011; Bokde et al., 2017) algorithm, which is especially conceived to forecast on TS exhibiting regular patterns, as in O_3 TS. A detailed revision of the existing literature was also performed to compare the forecasting performance of the proposed methodology.

The rest of the paper is structured as follows. Relevant and related works are reviewed in Section 2. Section 3 introduces the methodology proposed to forecast O_3 when limited historical data is available. Results from its application to several urban environments in Spain are reported and discussed in Section 4. Finally, the conclusions drawn from this study are summarized in Section 5.

2. Related works

Ozone TS forecasting problem has been addressed using a wide variety of techniques, from statistical approaches to deterministic models. The most recurrent techniques in literature are based on ML algorithms, specially artificial neuronal networks (ANN) and ensemble methods. In this section, the most relevant approaches for O_3 TS forecasting are presented.

ANNs were used in (Pires et al., 2012), where both the activation function and the number of hidden neurons are tuned using a genetic algorithm which also optimizes a threshold value that helps to differentiate between regimes of O_3 behaviour. Different correction techniques were applied to ANN to improve their performance based on the average O_3 profile and training errors (Pires and Martins, 2011). ANNs were compared to a deterministic model named WRF-Chem in (Hoshyaripour et al., 2016) resulting that the latter performs better in predicting mean and extreme O_3 concentrations, while ANN achieved better results in predicting daily O_3 values.

The combination of support vector regression algorithms and numerical models were studied in (Carro-Calvo et al., 2017). In the work of Lu et al. (Lu and Wang, 2014), the authors explained the limitations of both ANN and support vector machines (SVM) in the field of the ground-level O_3 prediction. They claim that ANN-based techniques can easily incur in overfitting, local minima problems and they not provide interpretable models (ANN are black-box schemes).

Gaussian processes (GP) are statistical models for regression problems with an infinite-dimensional generalization of multivariate normal distributions. GP has been applied to O_3 TS prediction in (Kocijan et al., 2016; Petelin et al., 2013). Specifically, in these works an on-line learning-based variant of GP named evolving Gaussian processes was used, enabling the possibility of considering a mobile air-quality measurement station. Such methodology is able to predict O_3 concentrations for a specific geographical location without a large quantity of historical of measurements.

Sequential aggregation (Kolesárová et al., 2015) is a type of ensemble techniques where a linear sequential aggregation rule produces a weight vector based on the past observations and the past predictions. The final prediction is then obtained by linearly combining the predictions of the models according to the weight vector. Sequential aggregation was applied to O_3 prediction in (Debry and Mallet, 2014; Mallet et al., 2009). In the work of Debry et al. (Debry and Mallet, 2014), the predictions of the French platform Prev'Air were ensembled via sequential aggregation improving original predictions.

Bagging, boosting and stacking are well-known ensemble approaches that intend to improve the accuracy of a set of predictors by reducing their bias and variance. Bagging is designed to reduce the variance, whereas boosting and stacking can help to reduce both the bias and variance. Such three approaches were applied to predicting the exceedances of daily maximum O_3 in (Gong and Ordieres-Meré, 2016). Specifically, bagging technique was used in combination with classification/regression trees and random forests (Breiman, 2001). Boosting technique was applied using stochastic gradient boosting machines (Friedman, 2002) and AdaBoost (Freund and Schapire, 1996). Stacking technique was implemented using a multiple linear regressor as the metalearner and support vector machines, ANNs, classification/regression trees, random forests, AdaBoost and gradient boosting machines as ensembled methods.

Fuzzy logic in combination with ANNs were applied in (Taylan, 2017) using the adaptive neuro-fuzzy inference system (Jang, 1993) to predict ground-level O_3 concentrations. Different feature selection techniques were applied to O_3 prediction in (Kocijan et al., 2015) using different methods based on cost functions through a validation procedure. Resulting regressor selections are specific for particular geographical locations and O_3 concentration intervals.

It must be considered that after an exhaustive search and to the best of the authors' knowledge, no other similar approaches as the introduced in this study have been found in literature. Therefore, the novelty of our methodology led us to expose in this Section, (i) the more avant-garde proposals to forecast O_3 from a Data science approach, or (ii) the methodologies which share with ours some of the applied procedures, namely, bagging and ensembles. The main contribution of our proposal is the ability to forecast O_3 when no information about its precursors is available and when historical data are scarce. It seems no other forecasting approaches allow coping with these limitations.

3. Methodology

This section describes the methodology proposed to forecast O_3 when limited information is available.

A linear method forecasts hourly O_3 concentrations and is based on an ensemble from a set of two models, $\mathscr{M}_1$ and $\mathscr{M}_2$, that use three linear regression models (LR_1, LR_2 and LR_3) to estimate the forecasted O_3, using simply actual and lagged O_3 concentrations as regressors. If it is denoted $y_{h,d+1}$ as the O_3 concentration at hour h and day $d+1$, the observations $y_{h,d}$ and $y_{h,d-1}$, for $d = 1, \ldots, 30$, represent the actual and 24-h lagged hourly O_3 concentrations prior to $y_{h,d}$, respectively. LR models are defined as follow:

$$LR_1: \hat{y}_{h,d+1} = \beta_0 + \beta_1 y_{h,d} + \varepsilon_{h,d+1} \quad (1)$$

$$LR_2: \hat{y}_{h,d+1} = \beta_0 + \beta_1 y_{h,d-1} + \varepsilon_{h,d+1} \quad (2)$$

$$LR_3: \hat{y}_{h,d+1} = \beta_0 + \beta_1 y_{h,d} + \beta_2 y_{h,d-1} + \varepsilon_{h,d+1} \quad (3)$$

where β_0 is the intercept, β_1 and β_2 are regression coefficients, $\varepsilon_{h,d+1}$ is an error term and $d = 1, \ldots, 30$. To verify the validity of regression models (LR_1, LR_2 and LR_3) during the model building processes, an study of non-linearity of the data was performed using residual plots, and no indication of discernible patterns or trends in the residuals were detected. Complementarily, for the simple linear (LR_1 and LR_2) and multiple (LR_3) regression models, hypothesis tests were carried out to confirm the association between predictor and response variables, using the t and F statistics, respectively, which yielded statistically significant values. Using a linear classical TS modelling approach, LR_1 and LR_2 are equivalent to an autoregressive model of first order -AR (1)- in which the autoregressive term is shifted back 24 h and 48 h with respect the hourly observation at time t, respectively, with the length of TS, $T = 30$. Similarly, LR_3 is equivalent to an AR (2) model, with the first autoregressive term shifted back 24 h, and the second one, 48 h. The error term $\varepsilon_{h,d+1}$ in equations (1)–(3) is equivalent to white noise with mean zero and variance one, and β_0 a constant.

The hourly O_3 forecasted value from $\mathscr{M}_1$ model ($\hat{y}_{\mathscr{M}_1}$) is obtained after averaging the forecasting results from three LRs. $\mathscr{M}_2$ model uses a bagged averaging using LR_1, LR_2 and LR_3 as base regression models. The $\hat{y}_{\mathscr{M}_2}$ value is obtained following the next algorithm, with $t = 10$ iterations, divided into two different phases:

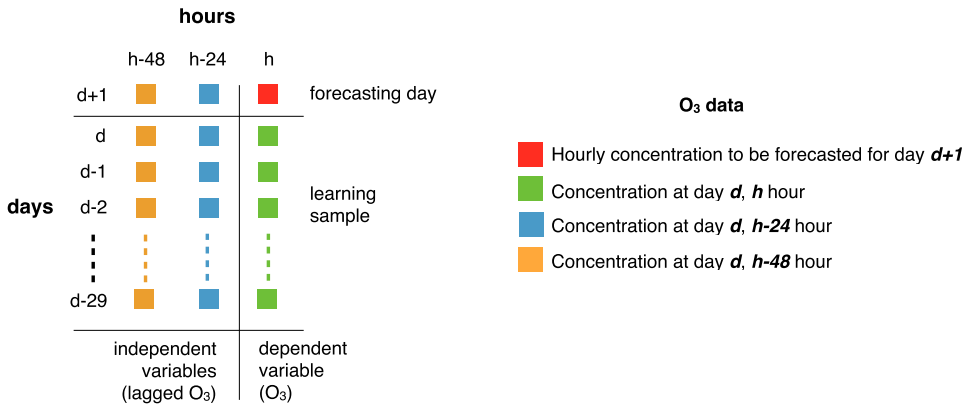

Fig. 1. Strategy for forecasting O_{33} concentration at h hour, using last 30 days as learning period with 24 and 48 h lagged concentrations. Blue and orange squares indicated hourly O_3 with different lagging (24 h and 48 h, respetively), with respect the concentration at hour h. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

1. Model generation phase. For each of t iterations:
   (a) Sample with replacement from observations $(y_{h,d}, y_{h,d-1}, y_{h,d-2})$, $d = 1, ..., 30$.
   (b) Build LR_1, LR_2 and LR_3 from the sample.
   (c) Store the resulting models.
2. Forecasting phase. For each of the t models:
   (a) Forecast value of O_3 by averaging LR_1, LR_2 and LR_3 forecasting.
   (b) Return the average value $(\hat{y}_{\mathscr{M}_2})$ of the forecasted values.

The final O_3 hourly forecasted value is obtained after averaging the $\hat{y}_{\mathscr{M}_1}$ and $\hat{y}_{\mathscr{M}_2}$ values.

Figure 1 depicts how the proposed methodology makes a prediction. First, the historical data for the last 30 days are only considered for learning the algorithm. Let us suppose that, for instance, hour h at day $d + 1$ is going to be predicted. In that case, only hour h at day d and hour h at day $d - 1$ are considered $(d = 1, ..., 30)$. Other possible window lengths were studied (15, 45, 60 and 90 days), and the selected one (30 days), was chosen according to its forecasting performance using the quality measures described in Section 4.2. Apart from showing a better accuracy behaviour, it seems that a window length of 30 days conveniently balances robust prediction accuracy and model training with enough recent observations, and therefore, allows to capture the seasonal temperature conditions governing the O_3 formation in the different study periods along the year. From a modelling perspective, the daily pattern of hourly O_3 concentrations could be assumed to be influenced by an underlying seasonal cycle which varies through the year. Since the above regression models (LR_1, LR_2 and LR_3) do not consider this latter seasonal component, its possible influence during the time span covered by the TS (30 days) was studied using the *auto. arima* function from the *forecast* package (Hyndman, 2017) in R (R Core Team, 2017). This function allows for conducting a search over possible seasonal ARIMA models within the order constraints provided, and then returning the best model according to a bayesian criterion. To that end, hourly TS from the 30 days prior to the hour to be forecasted were modeled using the seasonal ARIMA $(p, d, q)(P, D, Q)_{24}$. The values of q and Q were set to 0, and those of p, d, P and D were restrained to a maximum of 2 orders. The best models were selected according to the BIC criterion.

Finally, Fig. 2 illustrates the methodology aforementioned. It can be seen that three linear regressions, LR_1, LR_2 and LR_3, are created and their average is calculated $(\hat{y}_{\mathscr{M}_1})$. Alternatively, ten models are built from 10 bootstrap samples generating 10 averaged bagging models $(\hat{y}_{\mathscr{M}_2})$. The average of ensembles and linear regressions results in the final forecasting.

4. Results

This section reports the results obtained by the application of the proposed methodology to the datasets described in Section 4.1. The used metrics to evaluate its performance is introduced in 4.2. Finally, errors and comparison to other well-established methods are discussed in Section 4.3.

4.1. Data description

O_3 data were collected at five urban sites from three air quality monitoring networks in Andalusia, Spain (Cordova, Jaen and Seville) from 2006 to 2015, following the reference monitoring method established in Directive (2008)/50/EC on ambient air quality and cleaner air for Europe. Table 1 presents the type (suburban, urban) and predominant emission sources (background, traffic) of each monitoring site selected in this study. O_3 data were provided by the Regional Ministry of Environment and Land Planning of Andalusia (Seville, Spain) after validation. The cities of Cordova, Jaen and Seville are located in southern Spain, and during 2014, had a total population of 328,041, 115,837 and 696,676 (data collected from the Institute of Statistics and Cartography from Andalusia, last accessed 2017), respectively.

4.2. Quality parameters

Many error measures can be used to assess a prediction performance (Hyndman and Koehler, 2006). However, in the context of this study, the most common are root mean square error (RMSE) and mean absolute error (MAE), and for this reason, they were the ones selected. Their formulas are:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \text{ and } MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|,$$

where n is the number of evaluated samples, y_i the actual value and $\hat{y}_i$ the predicted value.

4.3. Discussion

This section includes the results obtained by the proposed methodology and the comparison to the successful PSF algorithm, especially designed to forecast TS with temporal patterns. Briefly, this algorithm forecasts the behaviour of TS based on similarity of pattern sequences. The prediction of a data point is provided as follows: first, the pattern sequence prior to the day to be predicted is extracted. Then, this sequence is searched for within the historical data and the prediction is calculated by averaging all the samples immediately after the matched sequence.

Tables 2 and 3 show the results, in terms of RMSE and MAE, respectively, for all the five stations and years 2006–2015. As it can be seen, the proposed methodology clearly outperforms results of PSF for every year. On average, the RMSE achieved is 18.16 μ g/m^{33}, whereas

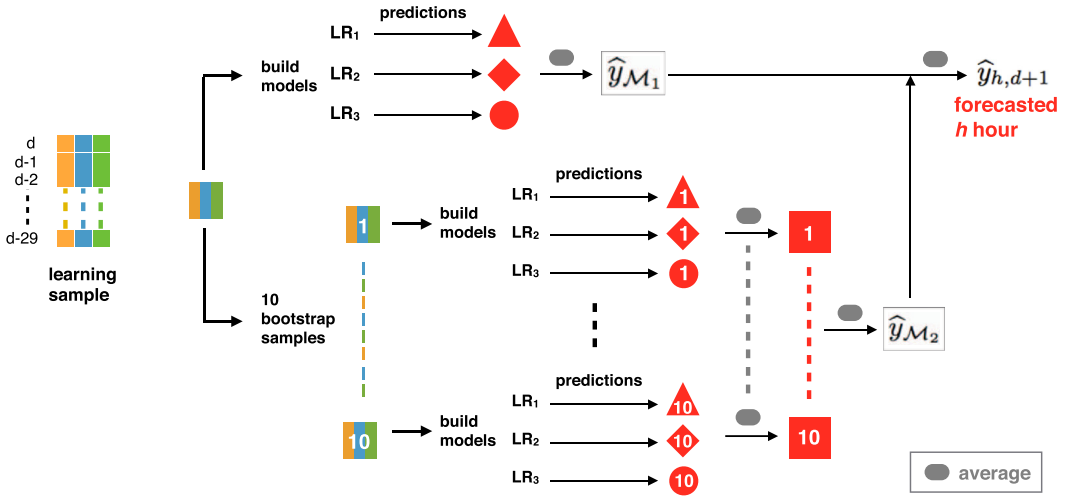

Fig. 2. Linear regression and ensemble models generation for forecasting one hourly O_3 concentration ($\hat{y}_{h,d+1}$). $\hat{y}_{\mathcal{M}_1}$ estimation is obtained after averaging the LR_1, LR_2 and LR_3 estimations, represented with red figures. $\hat{y}_{\mathcal{M}_2}$ is obtained after averaging the estimations obtained in each bootstrap sample with LR_1, LR_2 and LR_3. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Classification of monitoring sites where O_3 data were obtained (locations are given in X,Y ETRS89-UTM coordinates, zone 30).

City	Site	Type	Main pollution source	Location (X, Y)
Cordova	Asomadilla	Suburban	Background	(343546, 4196519)
Jaen	Ronda del Valle	Urban	Background	(431177, 4181976)
Seville	Aljarafe	Suburban	Background	(230473, 4137017)
Seville	Bermejales	Urban	Background	(236063, 4137554)
Seville	Torneo	Urban	Traffic	(234151, 4142873)

Table 2
Performance of proposed and PSF algorithms to forecast hourly O_3 concentrations at five studied sites from 2006 to 2015, using RMSE (in $\mu g/m^3$).

Year	Aljarafe		Asomadilla		Bermejales		Ronda del Valle		Torneo	
	Proposed	PSF	Proposed	PSF	Proposed	PSF	Proposed	PSF	Proposed	PSF
2006	18.6	20.5	18.6	40.7	20.7	22.3	22.3	25.3	14.6	15.7
2007	17.7	20.0	17.8	20.0	20.3	22.1	21.5	24.0	14.7	15.7
2008	18.8	20.9	18.9	21.1	20.1	22.3	22.6	24.3	14.5	16.1
2009	19.0	21.2	18.0	19.6	20.7	22.6	21.6	23.3	15.6	17.2
2010	18.7	21.3	17.6	19.7	20.5	22.8	22.8	26.1	–	17.8
2011	17.7	19.2	17.3	19.8	19.2	21.2	20.4	22.5	15.4	17.0
2012	18.7	20.4	–	17.7	19.6	21.8	16.7	19.3	16.1	17.5
2013	16.9	19.0	16.4	18.9	19.1	21.2	19.2	21.6	10.7	17.1
2014	17.3	18.3	17.7	19.7	18.8	21.1	19.7	22.6	15.7	18.1
2015	15.7	20.1	17.0	18.9	18.3	20.7	19.7	22.3	16.0	17.2
Average	17.9	20.1	17.7	21.6	19.7	21.8	20.7	23.1	14.8	16.9

Table 3
: Performance of both algorithms, as in Table 2, using MAE (in $\mu g/m^3$).

Year	Aljarafe		Asomadilla		Bermejales		Ronda del Valle		Torneo	
	Proposed	PSF	Proposed	PSF	Proposed	PSF	Proposed	PSF	Proposed	PSF
2006	14.8	15.8	14.5	31.4	15.9	17.0	17.8	19.9	11.2	11.8
2007	14.0	15.5	14.0	15.4	15.8	17.1	16.9	18.5	11.2	11.8
2008	15.1	16.4	15.1	16.4	16.1	17.4	18.0	19.1	11.3	12.5
2009	15.3	16.6	14.1	15.2	16.5	17.6	17.1	18.3	12.2	13.3
2010	14.7	16.8	14.0	15.6	16.4	17.8	18.4	20.7	–	14.0
2011	14.1	15.2	13.8	15.5	15.3	16.7	16.2	17.7	11.9	13.2
2012	14.9	16.2	–	13.8	15.7	17.0	13.2	15.2	12.6	13.4
2013	13.4	14.9	12.8	14.8	15.1	16.8	15.0	16.9	8.1	13.1
2014	13.9	14.6	14.2	15.4	14.9	16.6	15.8	17.8	12.3	13.9
2015	11.6	16.0	13.4	14.8	14.6	16.2	15.5	17.4	12.5	13.3
Average	14.2	15.8	14.0	16.8	15.6	17.0	16.4	18.2	11.5	13.0

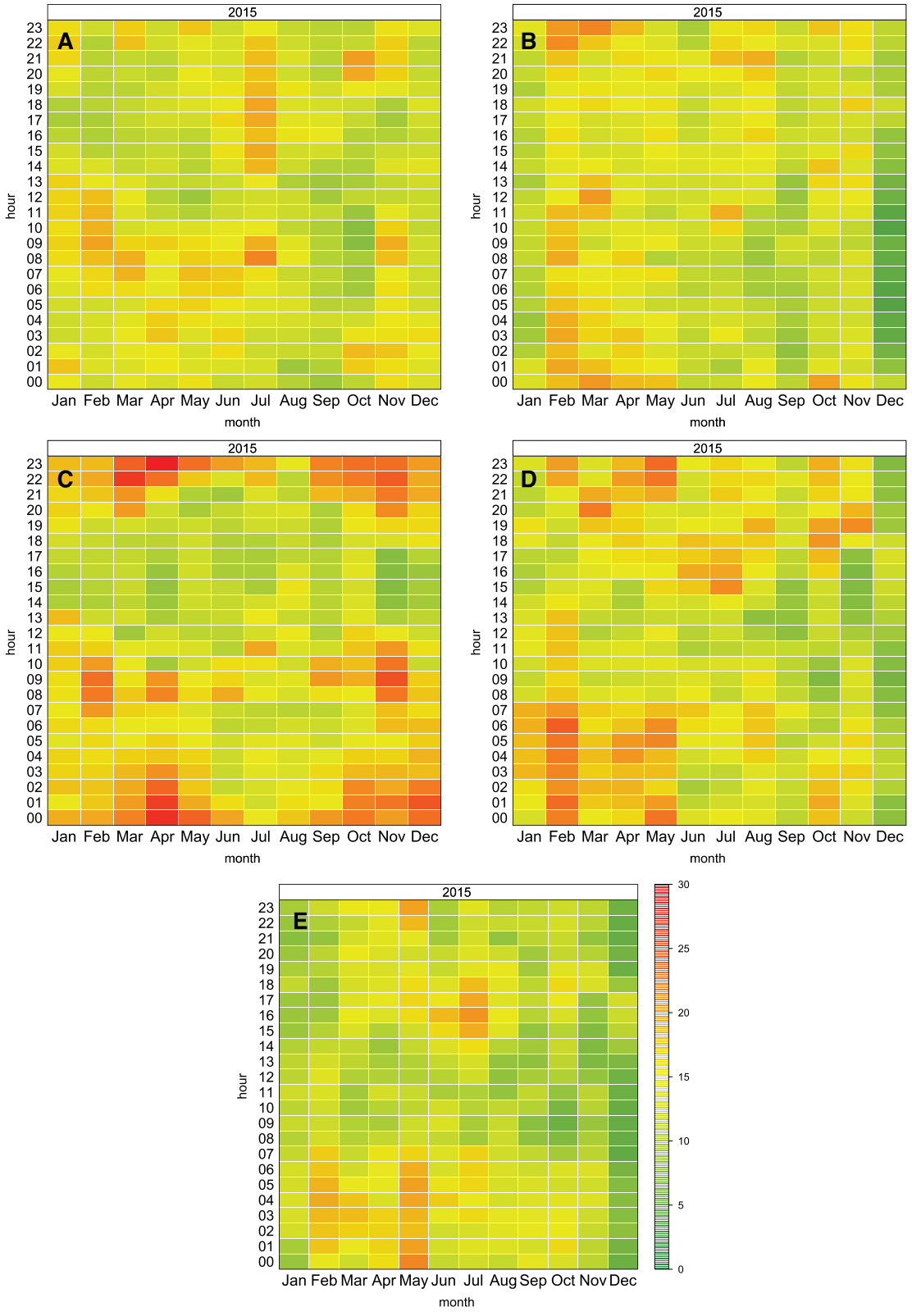

Fig. 3. Average RMSE values for all the five stations for 2015, distributed by hour and month, (in $\mu g/m^3$).

PSF achieved 20.72 μ g/m^3, being the best obtained values 10.7 μ g/m^3 and 15.7 μ g/m^3, respectively. With respect to MAE, the average results were 14.33 μ g/m^3 and 16.17 μ g/m^3 for the proposed and PSF algorithms, an the best values, 8.1 μ g/m^3 and 11.8 μ g/m^3, respectively. The seasonal ARIMA models described previously yielded forecasting performances lower than the proposed and PSF algorithms (results not shown). The reason the latter algorithms forecast more accurately than seasonal ARIMA models in the context of this study remains open for

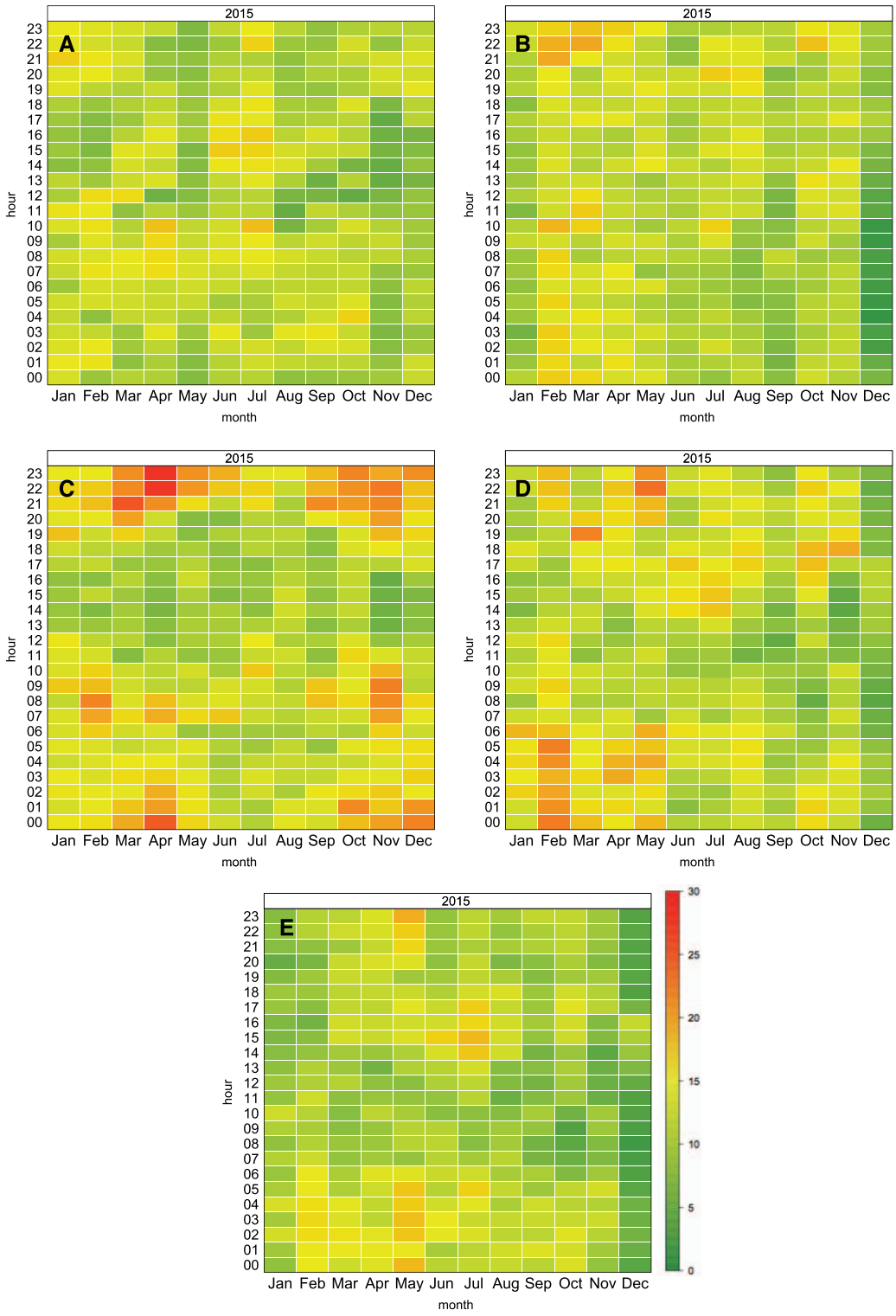

Fig. 4. Average MAE values for all the five stations for 2015, as in Fig. 3.

further investigation.

For illustrative purposes, only graphical results from 2015 for RMSE and MAE are depicted in Figs. 3–6. Letters A, B, C, D and E identify Aljarafe, Asomadilla, Bermejales, Ronda del Valle, and Torneo stations, respectively. In particular, a temporal distribution of the RMSE and MAE per month and hour of the day can be seen in Figs. 3 and 4, respectively, being more consistently obtained higher values of both statistics at night during the first six months of the year in Asomadilla, Bermejales and Ronda del Valle stations. This observation, however, cannot be extended to the rest of stations because smoother RMSE and MAE values are obtained in Aljarafe and Torneo stations during the whole year.

Broadly speaking, Figs. 3 and 4 show how the forecasting performance of our proposed methodology behaves along the day and year.

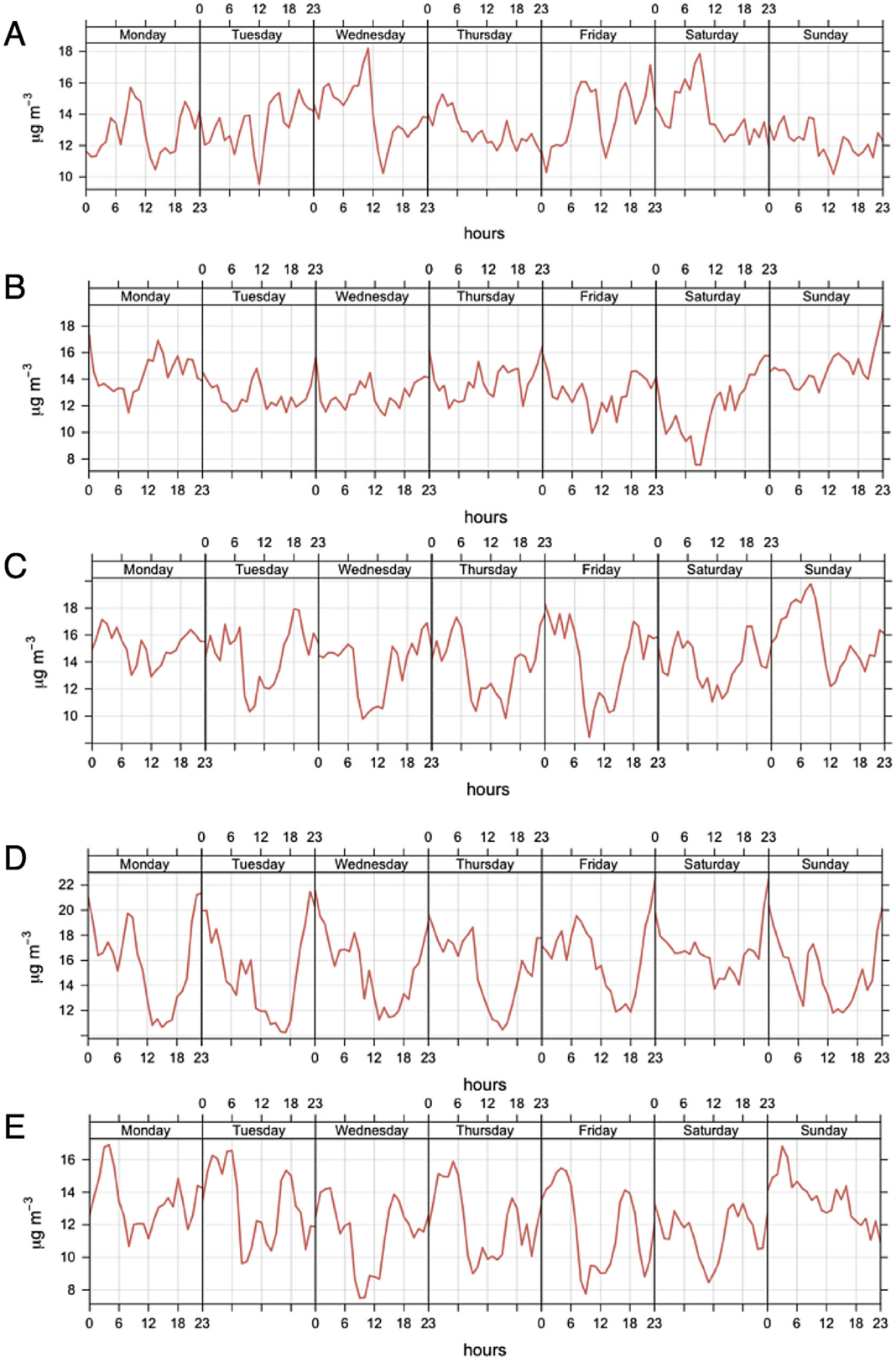

Fig. 5. Behaviour of the RMSE values during the week, for all the five stations and year 2015.

As said before, Aljarafe (A) and Torneo (E) sites seem to experience a less acute declining of the RMSE and MAE performance. Aljarafe site is placed at the Seville's outskirts, and could receive O_3 from a transported origin (Huelva city, with an intense industrial activity). The transported O_3 during night is not coupled with the O_3 genesis photochemical reactions, which are (light) ultra-violet dependent. With respect the Torneo site, in this area it is produced a high concentration of NO_2, a typical marker of traffic origin, which in case of adequate presence of sunlight intensity, can also produce high O_3 concentrations during the day. If this O_3 is not conveniently washed-out by wind conditions, it

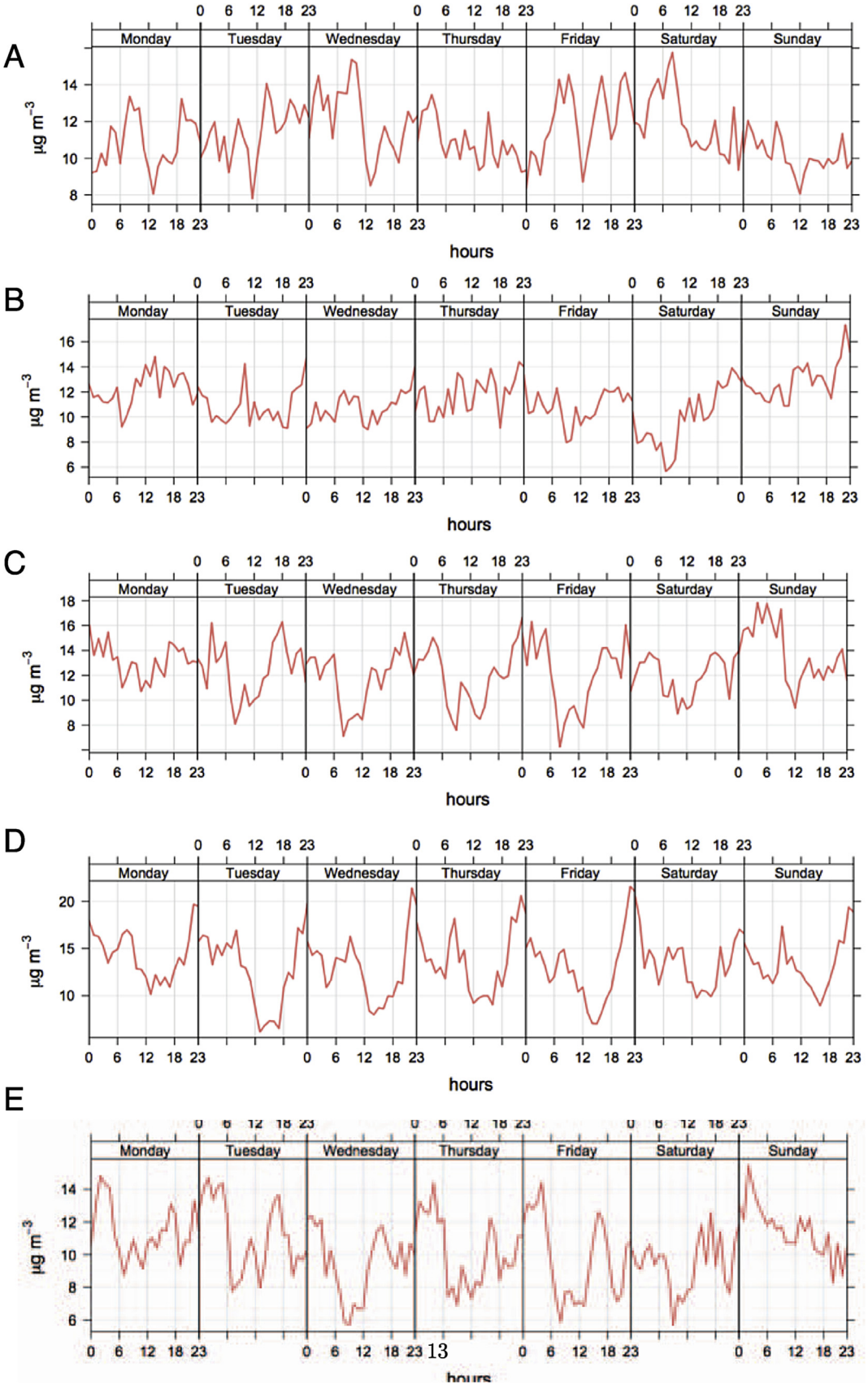

Fig. 6. Behaviour of the MAE values during the week, as in Fig. 5.

Table 4
Average RMSE (in $\mu g/m^3$) reported by other models.

Reference	Models	Average RMSE	Best RMSE
(Lu and Wang, 2014)	Four MLP, one SVM	25.28	22.50
(Sousa et al., 2009)	MLR	21.55	18.78
(Pires and Martins, 2011)	Two ANN, two MLR	23.63	23.00
(Pires et al., 2010)	Seven AR, one ANN	23.18	22.11
(Mallet et al., 2009)	Fifty-one ensemble models	19.21	17.43
(Pires et al., 2012)	Three ANN	19.11	17.35
(Debry and Mallet, 2014)	Three-model ensemble	15.57	12.80
(Kumar et al., 2017)	MLP trained with backpropagation	12.05	8.83
(Hoshyaripour et al., 2016)	One ANN	9.66	7.06

could become stagnant during night hours. Therefore, in Aljarafe and Torneo sites, the O_3 concentrations would experience less range variations than in the rest of sites. It seems that exogenous O_3 concentrations present at night, caused by transport from other origins, or the stagnated one produced during the day, could be captured by our methodology. However, the more wide O_3 concentration ranges experienced in Asomadilla, Bermejales and Ronda del Valle sites (due to high O_3 concentration values at midday and low at night hours) would not be entirely described with our approach. This aspect remains open for further investigation.

Figs. 5 and 6 illustrate the particular behaviour of the RMSE and MAE values per day of the week, which also depends on the studied station. Based on Tables 2 and 3 values for the rest of the years, it can be concluded that RMSE and MAE values are approximately constant and, in general, robust over the 10 years studied. Likewise, these values seem to be locally controlled by generating factors of pollution that have an effect on the performance of the proposed algorithm. It is worth to note the more steady behaviour of RMSE and MAE in Asomadilla site (B), a typical suburban background monitoring site, as Aljarafe site. In this case, Asomadilla site experiences higher O_3 concentrations due to its location far from Cordova's downtown, such area of the city likely being the origin of this higher O_3 pollution due to transport dynamic. This fact could support the performance explanation derived from Figs. 3 and 4. In general terms, from Figs. 5 and 6, it can be concluded that our approach seems not to properly describe the weekend effect (lower O_3 concentrations during Saturday and Sunday) since similar RMSE and MAE values are obtained with respect to working days.

Despite the short available time period for learning the proposed algorithm (30 days), these results must be considered satisfactory in absolute terms, since similar RMSE values were achieved in other consistent studies. Table 4 shows other RMSE models applied to the same kind of data, where MLP stands for multilayer perceptron, SVM for support-vector machine, MLR for multiple linear regression, and AR for autorregressive model.

As it can be noticed, all methods obtained RMSE values higher than those of the proposed method and even higher than those of PSF. The only exception are the works in (Debry and Mallet, 2014), (Kumar et al., 2017) and (Hoshyaripour et al., 2016). However, in (Debry and Mallet, 2014) authors considered exogenous variables such as NO_2 and PM_{10}, in (Kumar et al., 2017), temperature, relative humidity and NO_2, and NO_2 and wind direction, in (Hoshyaripour et al., 2016), to generate a more robust models. It is worth highlighting that the ultimate goal of this approach is to make predictions in extreme situations, where short historical data and no other correlated variables are available. Therefore, its comparison could not be considered fair.

5. Conclusions

A new methodology based on bagging and ensembles of learning models to forecast 24 h-ahead hourly surface-level O_3 is proposed. Its main novelty lies in the ability to develop these models when no information about O_3 precursors is available, which is new in literature. This methodology was only built on two O_3 variables composed by 24 h and 48 h lagged concentrations with respect the hourly concentration to be forecasted, or equivalently, just using the information from the two days prior the forecasting time were required. Modelling introduced in this work are presented ready-to-use, without requiring further intervention from the final user to reproduce it.

Hourly TS from five O_3 monitoring sites from Andalusia (Spain) were used to test the forecasting ability of the proposed methodology. The long period of study (2006–2015) and the different monitoring sites where data were obtained permitted a wide range of pollution levels, contributions and locations to be considered for assessing its robustness. Every forecasted hourly O_3 concentration was obtained after averaging the estimated O_3 concentrations from two models: the first one averages the estimates of three linear regression models, and the second one, used an bagged averaging of them. The proposed methodology could pose how simply averaging few and slightly correlated models can improve the forecasting ability of ensembles.

The accuracy of the proposed forecasting approach outperforms the results found in the literature. Related studies make use of input variables involved in the O_3 formation in urban environments. These latter variables are provided by meteorological and air quality services which make models reproducibility dependent on the availability of such information from other similar data providers. The introduced method circumvents the use of this information related to the O_3 genesis, widening their applicability. The PSF algorithm, specially designed to forecast TS with temporal patterns, as in the O_3 case, was used to compare the performance of the proposed methodology. On average, the RMSE and MAE achieved by this algorithm after studying their forecasting performance was 20.72 $\mu m/m^3$ and 16.17 $\mu m/m^3$, whereas the proposed methodology obtained 18.16 $\mu m/m^{33}$ and 14.33 $\mu m/m^3$, respectively.

The use of the proposed methodology O_3 is encouraged to environmental modelers devoted to forecast surface-level O_3. Its use is intended when no information of the precursors involved in the formation of this air pollutant is available or when historical data are scarce. However, their forecasting accuracy can be used to provide a benchmark performance for comparative purposes with respect to other modelling approaches.

Data and software availability

The data used in this study were kindly provided by the Regional Ministry of Environment and Land Planning of Andalusia (Seville, Spain). Please contact the corresponding author for any enquiries. Models were implemented using the open-source programming environment R, version 3.2.2. This software is available for download from www.r-project.org and runs on UNIX, Windows and MacOS platforms. Source codes used in this study are available upon request.

Disclaimer

The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

Acknowledgements

The authors would like to thank the Spanish Ministry of Economy and Competitiveness and Junta de Andalucía for the support under projects TIN2014-55894-C2-R and P12-TIC-1728, respectively.

References

Bokde, N., Asencio-CortÃĺs, G., MartÃŋnez-ÃĄlvarez, F., Kulat, K., 2017. PSF: introduction to R Package for pattern sequence based forecasting algorithm. Rice J. 9 (1), 324–333.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Carro-Calvo, L., Casanova-Mateo, C., Sanz-Justo, J., Casanova-Roque, J.L., Salcedo-Sanz, S., 2017. Efficient prediction of total column ozone based on support vector regression algorithms, numerical models and Suomi-satellite data. Atmósfera 30, 1–10.

Chattopadhyay, S., Bandyopadhyay, G., 2007. Artificial neural network with back-propagation learning to predict mean monthly total ozone in Arosa, Switzerland. Int. J. Rem. Sens. 28 (20), 4471–4482.

Corani, G., Scanagatta, M., 2016. Air pollution prediction via multi-label classification. Environ. Model. Software 80, 259–264.

Debry, E., Mallet, V., 2014. Ensemble forecasting with machine learning algorithms for ozone, nitrogen dioxide and PM_{10}on the Prev'Air platform. Atmos. Environ. 91, 71–84.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. Mach. Learn.: Proc. Thirteenth Int. Conf. 96, 148–156.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38 (4), 367–378.

Gong, B., Ordieres-MerÃĺ, J., 2016. Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: case study of Hong Kong. Environ. Model. Software 84, 290–303.

Gong, B., Ordieres-Meré, J., 2016. Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: case study of Hong Kong. Environ. Model. Software 84, 290–303.

Hoshyaripour, G., Brasseur, G., Andrade, M.F., Gavidia-Calderón, M., Bouarar, I., Ynoue, R.Y., 2016. Prediction of ground-level ozone concentration in São Paulo, Brazil: deterministic versus statistic models. Atmos. Environ. 145, 365–375.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22, 679–688.

Jang, J.-S.R., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Sys., Man, Cybernet. 23 (3), 665–685.

Kocijan, J., Hančič, M., Petelin, D., Božnar, M.Z., Mlakar, P., 2015. Regressor selection for ozone prediction. Simulat. Model. Pract. Theor. 54, 101–115.

Kocijan, J., Gradišar, D., Božnar, M.Z., Grašič, B., Mlakar, P., 2016. On-line algorithm for ground-level ozone prediction with a mobile station. Atmos. Environ. 131, 326–333.

Kolesárová, A., Mesiar, R., Montero, J., 2015. Sequential aggregation of bags. Inf. Sci. 294, 305–314.

Kumar, N., Middey, A., Rao, P.S., 2017. Prediction and examination of seasonal variation of ozone with meteorological parameter through artificial neural network at NEERI, Nagpur, India. Urban Climate 20, 148–167.

Lu, W.-Z., Wang, D., 2014. Learning machines: rationale and application in ground-level ozone prediction. Appl. Soft Comput. 24, 135–141.

Mallet, V., Stoltz, G., Mauricette, B., 2009. Ozone ensemble forecast with machine learning algorithms. J. Geophys. Res.: Atmosphere 114 (D5).

Martínez-Álvarez, F., Troncoso, A., Riquelme, J.C., Aguilar, J.S., 2011. Energy time series forecasting based on pattern sequence similarity. IEEE Trans. Data Knowl. Eng. 23 (8), 1230–1243.

Martínez-Ballesteros, M., Troncoso, A., Martínez-Álvarez, F., Riquelme, J.C., 2010. Mining quantitative association rules based on evolutionary computation and its application to atmospheric pollution. Integrated Comput. Aided Eng. 17 (3), 227–242.

Martínez-Ballesteros, M., Martínez-Álvarez, F., Troncoso, A., Riquelme, J.C., 2011. An evolutionary algorithm to discover quantitative association rules in multidimensional time series. Soft Computing 15 (10), 2065–2084.

Paoletti, E., 2006. Impact of ozone on Mediterranean forest: a review. Environ. Pollut. 144, 463–474.

Petelin, D., Grancharova, A., Kocijan, J., 2013. Evolving Gaussian process models for prediction of ozone concentration in the air. Simulat. Model. Pract. Theor. 33, 68–80.

Pires, J.C.M., Martins, F.G., 2011. Correction methods for statistical models in tropospheric ozone forecasting. Atmos. Environ. 45 (14), 2413–2417.

Pires, J.C.M., Alvim-Ferraz, M.C.M., Pereira, M.C., Martins, F.G., 2010. Evolutionary procedure based model to predict ground–level ozone concentrations. Atmos. Pollut. Res. 1 (4), 215–219.

Pires, J.C.M., Gonçalves, B., Azevedo, F.G., Carneiro, A.P., Rego, N., Assembleia, A.J.B., Lima, J.F.B., A Silva, P., Alves, C., Martins, F.G., 2012. Optimization of artificial neural network models through genetic algorithms for surface ozone concentration forecasting. Environ. Sci. Pollut. Control Ser. 19 (8), 3228–3234.

Sicard, P., Serra, R., Rossello, P., 2016. Spatiotemporal trends in ground-level ozone concentrations and metrics in France over the time period 1999-2012. Environ. Res. 149, 122–144.

Sousa, S.I.V., Pires, J.C.M., Martins, F.G., Pereira, M.C., Alvim-Ferraz, M.C.M., 2009. Potentialities of quantile regression to predict ozone concentrations. Environmetrics 20 (2), 147–158.

Taylan, O., 2017. Modelling and analysis of ozone concentration by artificial intelligent techniques for estimating air quality. Atmos. Environ. 150, 356–365.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., 2012. Real-time air quality forecasting, part I: history, techniques, and current status. Environ. Res. 60, 632–655.