Towards A Framework for Forecasting Pollutant Concentration

Jana Faganeli Pucer[1], **Rahela Žabkar**[2], **Simon Oberžan**[1], **Greta Gašparac**[1], **Erik Štrumbelj**[1,⋆]

[1] *University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana, Slovenia*
[2] *Slovenian Environment Agency, Vojkova 1b, 1000 Ljubljana, Slovenia*
⋆ *E-mail (corresponding author): erik.strumbelj@fri.uni-lj.si*

Abstract

We compare forecasting models (linear regression, regularized linear regression, and random forests) for ozone forecasting using meteorological and air pollutant measurements, forecasts of meteorological parameters, and air mass back trajectories or, alternatively, only the previous day's pollutant concentrations from other stations. Results show that random forests improve on linear methods on most, but not all sites. Results of forecasting using the previous day's concentrations from stations across Europe show substantial predictive power.

1 Introduction

Ground level ozone is created photochemically when nitrogen oxides (NO_x) and volatile organic compounds react with UV-radiation from sunlight [1]. During warmer parts of the year, when concentrations rise significantly, ozone can affect overall mortality and morbidity [2]. To minimize the negative health effect, most countries accurately monitor ozone levels and alert the public. The European parliament passed Directive $2008/50/\text{EC}$ which states that the inhabitants of European countries have to be warned about expected exceedances of the information and alert threshold (180 $\mu g/m^3$ and 240 $\mu g/m^3$, respectively), for the current and following day. This legislature created the need for ozone forecasting systems.

Ozone forecasts are usually made using dispersion models [3]. They provide a suitable spatial and temporal display, but are highly dependent on meteorological models and on emission inventories and usually do not have a good spatial resolution. For forecasting ozone levels at specific locations, statistical or machine learning models are frequently used. The most popular method used for ozone modelling are artificial neural networks [4, 5, 6]. Non-linear regression [7], support vector machines [4, 8] and Bayesian networks [9] were also used. In comparative studies, non-linear models outperformed linear models [4], non-linear regression and neural networks performed comparably to each other [7]. Most models used meteorological parameters, ozone concentrations, and emission measurements as inputs [4, 5, 6].

Forecasting pollutant concentrations is a standard forecasting task. However, there are several issues that make the task more difficult in practice. These issues can be broken down according to their corresponding steps in the forecasting pipeline: *Data transformation* - some data relevant to the forecast are structured and have to be transformed before they can be used as inputs. In particular, air mass movement, which is most often provided as a path of geographical locations over time (that is, trajectories). *Missing values* - due to faults in meteorological equipment, missing values are frequent and have to be dealt with in ways other than simply discarding the entire input. *Feature subset and forecasting model selection* - the optimal choice of features and a forecasting model might vary from forecasting site to forecasting site.

Therefore, a practical approach towards forecasting pollutant concentrations would have to be a modelling framework with different modelling options at each modelling step and an automated model selection and parameter tuning procedure that is robust in terms of over-fitting the data. This paper is a step in that direction, while also providing some practical results relevant to forecasting ozone at Slovenian sites and in general.

2 Methods

2.1 Transforming air mass trajectories

Each backward trajectory is described as a path of spatial coordinates $(x, y, z)_i$, $i = 0 \dots m$, calculated at regular intervals going back in time with increasing i. The coordinate $(x, y, z)_0$ is the station's location and can be ignored. In practice, we can assume that all trajectories will be of the same length m and that index i always denotes the same time backwards from arrival at time 0. The used air-mass trajectories are 2D on constant pressure surfaces. The following options were implemented for transforming air mass trajectories:

Trajectories are not used (NONE). Self-explanatory.

Clustering-based grouping (CLUST). This is based on a very common technique for converting trajectories into categorical variables before statistical analysis [10]. We use `kmeans` clustering from the R core packages [11] to group trajectories. The group assignment is then used as a categorical input variable. The number of clusters is treated as a hyperparameter (in all experiments it was set to 5, based on prior experience). Note that the dissimilarity of trajectories is defined as the Euclidean distance between trajectories if they are represented as vectors in R^{3m} by stacking all the x, y, and z coordinates. Z-axis

values are scaled so that their scale matches the x and y coordinates' scales (that is, equal importance is assigned to all coordinates).

Using raw coordinates (XYZ). For each trajectory path point, we add input variables, one for each coordinate x, y, and z. Therefore, $3m$ new input variables are added.

2.2 Dealing with missing values

We used single imputation (SINGLE), where missing values are replaced with the mean (numeric) or mode (categorical) of the input feature on the learning data, or multiple imputation (MULTI) by fitting models to missing data (see `aregImpute` from the Hmisc package [12]).

2.3 Prediction algorithms

The following statistical and machine learning algorithms are currently included in the framework (the first two serve as a baseline for comparison):

Previous day (PERSISTENCE). A simple model where the latest available concentration in the training set (typically the previous day's) is used as the forecast. Serves as a baseline for comparison.

Predicting the mean (MEAN). A simple model where the training set mean is used as the forecast.

Linear regression (LR). Classic least-squares linear model as implemented in `lm` from the R core packages [11]. Categorical input variables are dummy-coded (that is, replaced by $k-1$ binary indicator variables, where k is the number of categories).

L1-regularized linear regression (L1). L1-norm type regularization, also known as the lasso. We used the implementation in `lars` from the lars package [13]. The regularization parameter is selected using 10-fold cross-validation on the training set. Dummy-coding is used.

Random forests (RF). Breiman's random forests algorithm [14] as implemented in `randomForest` from the randomForests package [15]. Non-liner methods have performed best on forecasting pollutants, but, unlike artificial neural networks, RF is robust both in terms of overfitting and in terms of tuning hyperparameters.

2.3.1 Input variable subset selection

Currently, input variable subset selection is an option only with the LR model (the other algorithms are by design robust to overfitting). We use stepwise (forward) selection with the AIC criterion as implemented in `step` from the R core packages [11]. The maximum number of features to select is treated as a hyperparameter (it was set to 5 and 10, separately; not restricting the number of maximum results in overfitting).

2.4 Model evaluation process

The evaluation process is time-respecting (only data available prior to the day we are forecasting for are made available to the model) using a rolling window with step size 1 (after a forecast is made for a particular day, that day is added to the training set and we move to the next day). The first 50 days are used as the initial training set and are not used in the evaluation. Mean squared error is used to evaluate the point forecasts and standard errors of the estimated errors are provided to aid in the interpretation of the differences between models.

3 Data

We used two separate sets of data. The first set are ozone concentration data and other relevant input variables for eight Slovenian sites. These data are what the current ozone forecasts are based on. The main purpose was to measure how accurately we can forecast ozone concentrations at these sites. The second set are pollutant concentration data only, but from a large number of European sites. The main purpose was to explore if data from other stations contain some predictive power.

3.1 Eight Slovenian sites

Ozone forecast models for 1-hour daily maximum concentrations were developed for eight Slovenian monitoring sites (Koper, Nova Gorica, Otlica, Ljubljana, Krvavec, Iskrba, Hrastnik and Murska Sobota) taking into account measured air quality and meteorological data, weather forecast data and the predicted backward trajectories for the warm part of the year (Apr-Sep).

3.1.1 Measurements

Meteorological (temperature, pressure, relative humidity, direct and diffusive solar radiation, wind speed and wind gust) and air quality (O_3, NO_2, NO_x, SO_2 and PM_{10}) measurements from Aug 2011 to Sep 2015 were used. In some occasions due to missing or less accurate results, measurements from another representative nearby station were used. Measurements were collected with the time resolution of 30 minutes, where in the case of pollutant concentrations the aggregated 1h values were calculated. Furthermore, only previous day's values at 12:00, 15:00, 18:00 and 21:00 LT, daily maximum, minimum and average value between 9:00 and 19:00 LT, as well as today's early morning values (at 07:00 LT) were included. Precipitation was included as the daily cumulative between 7:00 LT of the previous day and today 7:00 LT hour.

3.1.2 Weather forecast

Forecasts from The European Centre for Medium-Range Weather Forecasts meteorological model were used. 24 h predictions at 12:00UTC of the forecast day for temperature, dew point temperature, wind speed, geopotential height, relative humidity and vertical velocity at different vertical levels (1000, 925, 850, 500, 300 hPa) were taken into account. We included predictions of ground and upper level conditions: precipitation, cloudiness, solar radiation, convection, and showalter index, temperature triggering convection, the height of zero isotherm, relative topography, dew point depression, and indicators of inversion depth rate. The locations of points with predictions were not necessarily exactly the locations of the monitoring sites, but were selected among the synop stations with archived time series of predictions. In addition, sinusoidal day of the year was also included.

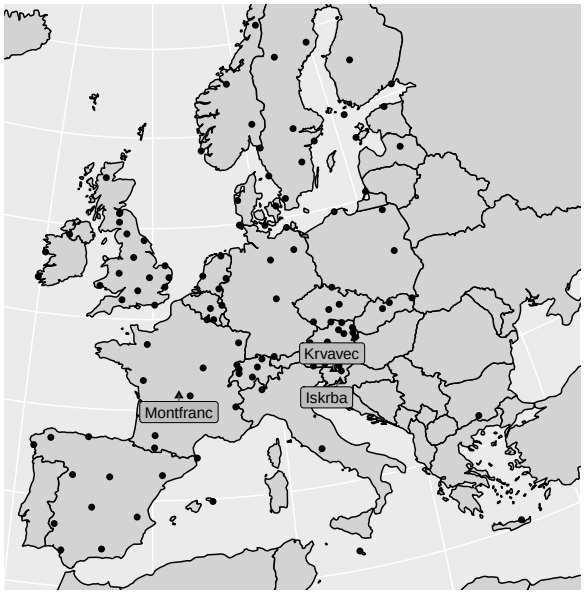


Figure 1: Locations of European sites in our data (circles), with emphasis on the three sites included in the results (triangles).

3.1.3 Air mass trajectories

24 h backward trajectories approaching at 15:00 UTC of the forecast day were calculated at Slovenian Environment Agency on ALADIN/SI [16] wind field predictions. For Krvavec station 850 hPa winds were used, while other trajectories were calculated on the 925 hPa vertical level. Trajectory end points were located directly above the station for Koper, Nova Gorica, Otlica, Krvavec, and Iskrba. For the rest of the stations trajectories calculated for Krvavec were used. Trajectories were saved as pairs of geographical latitudes and longitudes (every 3 hours) that approximate 1 day of air mass travel.

3.2 European sites

We compiled ozone concentration measurements data for 121 European sites (see Figure 1). The data run from 2011 to 2013, Apr to Sep. As with the first data set, the goal is to predict maximum concentrations today at a particular site, but using only yesterday's concentrations from all sites, including the site we are forecasting for.

4 Results

Figure 2 shows the results for the Slovenian sites. Note that out of the many possible combinations of model and preprocessing steps we selected only a few. First, each model type was run with single imputation and xyz trajectory transformation. The RF model, which was the best performing model on all sites, was then run with different preprocessing or different handling of missing values to explore how these changes affect performance.

Figure 3 shows the results for three European sites: Iskrba, Krvavec (both SLO) and Montfranc (FRA). Ozone was forecast only with pollutant concentrations (including ozone) from other European monitoring sites (EMEP).

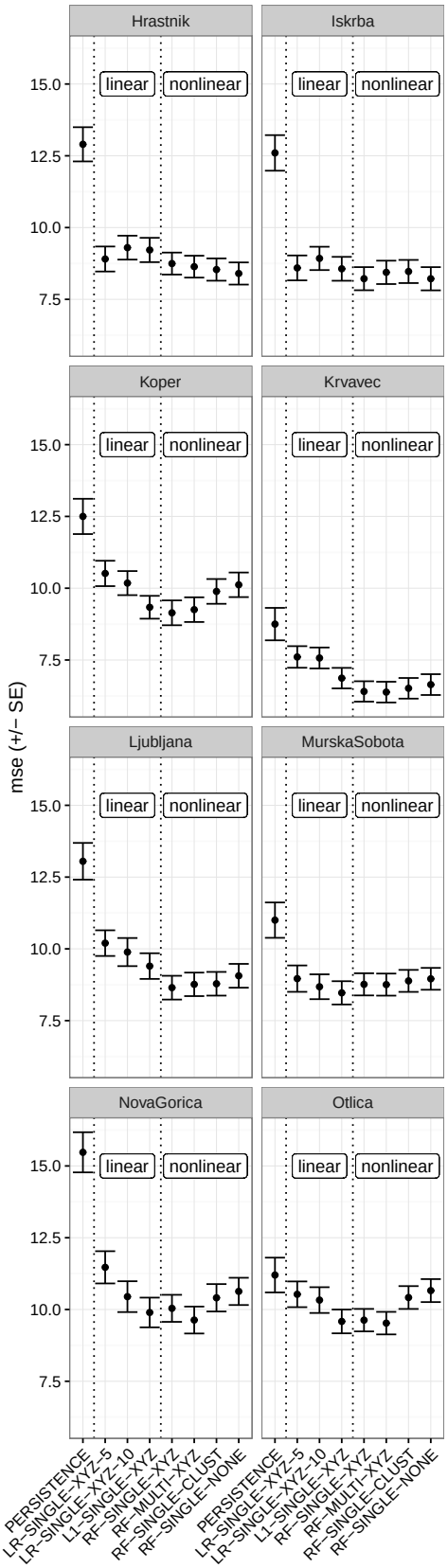
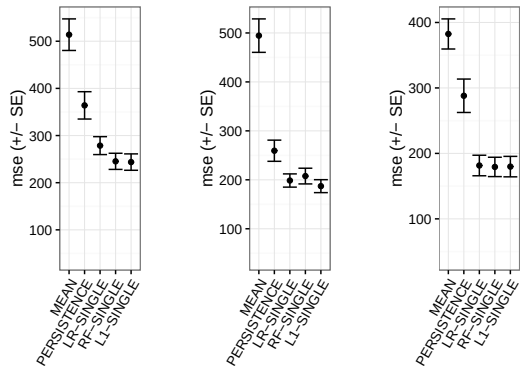


Figure 2: Estimated forecasting accuracy for 8 Slovenian sites.



(a) Iskrba (b) Krvavec (c) Montfranc

Figure 3: Estimated forecasting accuracy for 3 European sites.

Results for these three sites are representative of the results for the remaining sites.

5 Discussion and conclusion

Random forests improve on linear methods on most, but not all sites. However, the differences are relatively small and linear methods are, in practical terms, equally accurate. Overfitting was an issue, even with linear methods, because the number of input variables is in the same order of magnitude as the number of days.

Trajectories improve results when raw coordinates are used, but not if the clustering-based approach is used. More complex handling of missing values does not provide further improvement, possibly due to a small learning set which does not support more complex models for missing values. Somewhat surprisingly, and despite our relatively simple approach of using only yesterday's concentrations to predict today's concentrations, there is substantial predictive power in other sites' concentrations. Note that we've limited ourselves to using only previous day's data (in order to get results comparable to forecasting systems currently in production), however, the approach can easily be extended to using data from an arbitrary number of previous days'.

Our final goal is to simplify the process of developing and deploying a model for forecasting air pollutant concentrations for the non-(data analysis)expert. Two key features are delegated to future work. First, being able to fit all (hyper)parameter in a robust way, for example, using internal cross-validation. And second, being able to put less emphasis on older data points, with a moving window or a more elaborate decay/smoothing approach.

Acknowledgements

The authors would like to thank ARSO (Slovenian Environment Agency) for providing data and guidance.

References

[1] W. B. Group and U. N. I. D. Organization, *Pollution prevention and abatement handbook, 1998: toward cleaner production*. World Bank Publications, 1999.

[2] A. Gryparis, B. Forsberg, K. Katsouyanni, A. Analitis, G. Touloumi, J. Schwartz, E. Samoli, S. Medina, H. R. Anderson, E. M. Niciu, *et al.*, "Acute effects of ozone on mortality from the "air pollution and health: a European approach" project," *American journal of respiratory and critical care medicine*, vol. 170, no. 10, pp. 1080–1087, 2004.

[3] U. Nopmongcol, B. Koo, E. Tai, J. Jung, P. Piyachaturawat, C. Emery, G. Yarwood, G. Pirovano, C. Mitsakou, and G. Kallos, "Modeling Europe with CAMx for the air quality model evaluation international initiative (AQMEII)," *Atmospheric environment*, vol. 53, pp. 177–185, 2012.

[4] E. Salazar-Ruiz, J. Ordieres, E. Vergara, and S. Capuz-Rizo, "Development and comparative analysis of tropospheric ozone prediction models using linear and artificial intelligence-based models in Mexicali, Baja California (Mexico) and Calexico, California (US)," *Environmental Modelling & Software*, vol. 23, no. 8, pp. 1056–1069, 2008.

[5] D. Wang and W.-Z. Lu, "Ground-level ozone prediction using multilayer perceptron trained with an innovative hybrid approach," *Ecological modelling*, vol. 198, no. 3, pp. 332–340, 2006.

[6] U. Brunelli, V. Piazza, L. Pignato, F. Sorbello, and S. Vitabile, "Two-days ahead prediction of daily maximum concentrations of SO_2, O_3, PM_{10}, NO_2, CO in the urban area of Palermo, Italy," *Atmospheric Environment*, vol. 41, no. 14, pp. 2967–2995, 2007.

[7] W. G. Cobourn, L. Dolcine, M. French, and M. C. Hubbard, "A comparison of nonlinear regression and neural network models for ground-level ozone forecasting," *Journal of the Air & Waste Management Association*, vol. 50, no. 11, pp. 1999–2009, 2000.

[8] W.-Z. Lu and D. Wang, "Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme," *Science of the total environment*, vol. 395, no. 2, pp. 109–116, 2008.

[9] L. E. Sucar, J. Pérez-Brito, J. C. Ruiz-Suárez, and E. Morales, "Learning structure from data and its application to ozone prediction," *Applied Intelligence*, vol. 7, no. 4, pp. 327–338, 1997.

[10] Z. L. Fleming, P. S. Monks, and A. J. Manning, "Review: Untangling the influence of air-mass history in interpreting observed atmospheric composition," *Atmospheric Research*, vol. 104, pp. 1–39, 2012.

[11] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[12] F. E. Harrell, *Hmisc*, 2016. R package version 3.17-4.

[13] T. Hastie and B. Efron, *lars: Least Angle Regression, Lasso and Forward Stagewise*, 2013. R package version 1.2.

[14] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[16] A. Horányi, I. Ihász, and G. Radnóti, "ARPEGE/ALADIN: A numerical weather prediction model for Central-Europe with the participation of the Hungarian Meteorological Service," *Idojárás*, vol. 100, no. 4, pp. 277–301, 1996.