Regressor Selection for Ozone Prediction

Juš Kocijan[a,b], Marko Hančič[a], Dejan Petelin[a], Marija Zlata Božnar[c], Primož Mlakar[c]

[a] *Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia*
[b] *University of Nova Gorica, Vipavska 13, SI-5000 Nova Gorica, Slovenia*
[c] *MEIS d.o.o., Mali Vrh pri Šmarju 78, SI-1293 Šmarje-Sap, Slovenia*

Abstract

Being able to predict high concentrations of tropospheric ozone is important because of its negative impact on human health. In this paper eight regressor-selection methods are utilised in a case study for ozone prediction in the city of Nova Gorica, Slovenia. The comparison of the selected methods proved to be useful for building models that successfully predict the ozone concentrations for the treated case. Different regressors are selected for different models, with different methods based on the validation procedure's cost functions. Namely, for the model to predict the maximum daily ozone concentration, ten regressors are selected; for the average concentration of ozone between 8.00 and 20.00 hours, fifteen regressors are selected; and for the average daily concentration, ten regressors are selected. The result of the study is a regressor selection that is specific for a particular geographical location. Moreover, the study reveals that regressor selection, as well as the obtained models, differ depending on the kind of averaging interval of the ozone concentration.

Keywords: Regressor selection, regression modelling, black-box modelling, prediction of ozone concentration, dynamical systems

[☆]This work was supported by the Slovenian Research Agency with Grant Development and Implementation of a Method for On-Line Modelling and Forecasting of Air Pollution, L2-5475 and Grant Systems and Control, P2-0001.

1. Introduction

The generation of ozone at ground level depends on many factors, but primarily on meteorological variables and pollution. Even though ozone modelling is a matter of intensive research, the physical and chemical mechanisms of ozone generation are not understood in detail. This means that experimental modelling methods can be very useful. The ozone concentration can be modelled and predicted, or forecasted, using a variety of methods, and the methods that describe the nonlinear dynamics from the available data are particularly useful. The accuracy of these models depends crucially on the set of regressors as well as on the input variables or the features that are used when modelling with regression methods.

The United States Environmental Protection Agency (EPA) issued guidelines for ozone prediction [1] in which it lists nitrogen oxides (NOx), volatile organic compounds (VOCs) of various origins and various meteorological variables as being influential. The final selection of the regressors, however, is left to the model developers and depends on the modelling method being used, the regressor-selection method, the geographical site and the professionals' judgement.

An overview of recent literature reveals that there are a variety of models for the ozone prediction in cities and regions, e.g., Kuwait city, Kuwait [2], Delhi, India [3], Hsinchu, Taiwan [4], Malaga, Spain [5], Beijing, China [6], Lisbon and Tagus valey, Portugal [7], Baltimore, Maryland, USA [8], 6 regions in the state of Kentucky, USA [9], Athens, Greece [10], Mexico City, Mexico [11], Bourgas, Bulgaria [12], the Hamilton region, Ontario, Canada [13], and the Dallas-Fort Worth region, Texas, USA [14]. Various black-box models obtained with a range of regression methods from Principal Component Regression to Takagi-Sugeno fuzzy models are used, as are different regressor selection methods. The objectives of ozone prediction also differ and one can find models for the prediction of hourly ozone values, e.g., [2], [5], [6], [9], [12], [13], maximum ozone values, e.g., [3], [4], [7], [10], [11], [14], or different average ozone values, e.g., [7], [8], [14]. These models use various selections of pollutants and various meteorological variables and their lagged values as the regressors.

From this overview it can be inferred that regressor selection differs depending on the sort of averaging interval of the ozone, the geographical region and most likely also from the availability of the measurements. No generalisation whatsoever can be made based on the research results regarding

regressor selection for the various sorts of averaging interval of the ozone concentrations in other places, different from the particular place of interest.

The objective of this case study is to systematically select a method for regressor selection that will later be used for the development of a regression model for the short-term prediction of the ozone concentration in the city of Nova Gorica, Slovenia.

The paper is structured as follows. The problem description is given in Section 2. In Section 3, an overview of the methods for regressor selection is briefly reviewed together with the criteria for regressor selection. The results are discussed in Section 4, and the concluding remarks end the paper.

2. Problem description

At ground level, ozone (O_3)[15] is an air pollutant that damages human health and the equilibrium of the ecosystem [16]. Overexposure to ozone can cause serious health problems in plants and people. Ozone levels tend to increase during periods of high temperatures and sunny skies. The ozone content changes in the troposphere, and the complexity of the processes defining these changes is the reason why atmospheric ozone dynamics is the subject of intensive research.

Fixed measurements of the hourly ozone concentrations, in compliance with the European Directive on ambient air quality and cleaner air for Europe [17], give continuous information about the evolution of the surface ozone pollution at a large number of sites across Europe. The European standards that guarantee human-health protection are as follows: 'health protection level', 120 μg/m^3 eight hours mean concentration; 'informing the public level', 180 μg/m^3 one hour mean concentration; and 'warning the public level', 240 μg/m^3 one hour mean concentration. Therefore, predicting the ozone concentration and informing the population when the air-quality standards are not being met are important tasks.

As was stated in the previous section the selection of regressors for modelling differs for various geographical locations. Our problem is to find methods for regressor selection and, consequently, sets of regressors for three different models of ozone in the city of Nova Gorica, Slovenia: for the prediction of the maximum daily ozone concentration, for the prediction of the average concentration of ozone between 8.00 and 20.00 hours, and for the prediction of the average daily concentration.

The data used are obtained from the database of the measurement station in Nova Gorica and Bilje in the close vicinity of Nova Gorica. The data are available to the public via the web page of the Slovenian Environment Agency. Ground-based measurements of the air quality are in the form of a series of simultaneous observations of the time evolution of the surface ozone concentrations. In addition, ground-level meteorological measurements and other air-pollutant concentration measurements are available. As the ozone concentration depends on the present, and not only on the past, conditions, the forecasts of variables were added, as is common in this type of investigations. To avoid the forecasts' uncertainty we applied the measurements of these variables, which, in our opinion, provides a more accurate picture of the regressors' relevance.

The utilised data contain hourly and half-hourly concentration measurements of various pollutants and meteorological variables for the years 2012 and 2013. Since the ozone changes dynamically, lagged regressors also need to be incorporated for the modelling of the system's dynamics.

3. Methods used for the regressor selection and its validation

Our goal is to select only as many regressors for each of the models as are really necessary. Every additional regressor increases the complexity of the regression model and makes the optimisation of the model more demanding. While the input dimension increases linearly, the complexity of the model increases exponentially [18] and we end up with the so-called curse of dimensionality.

A quick look at the literature reveals lots of methods and algorithms for regressor selection. However, the various authors divide the methods up differently. We adopt the division of the regressors' selection into three major groups [19],[20], [21],[18]: wrappers or wrapper methods, embedded methods and filter methods.

Wrapper methods are the so-called brute-force methods for regressor selection. The basic idea behind these methods is that they form a kind of wrapper around the system model, which is considered as a black-box. The search for the optimal vector of regressors is initiated from some basic set of regressors. After the model's optimisation and cross-validation, the regressors are added to, or removed from, the model. The successful models, according to the selected performance criteria, are kept, while the poorly performing models are rejected. Some of these methods or groups of methods

are [18]: forward selection, backward elimination, nested subset, exhaustive global search, heuristic global search, single-variable ranking and other ranking methods. The wrapper methods are also known by the names Validation-based regressor selection or Exhaustive search for the best regressors [22].

Embedded methods have the regressor selection built into the model's optimisation procedure. For example, if a certain sort of model has a property that the values of the model's parameters correspond to the importance of the used regressors, then properly selected, lower-valued regressors can be eliminated. This property assumes that the global minimum of the parameter-optimisation cost function is achieved. Some other embedded methods are coupled with model optimisation, e.g., the direct optimisation method, or are weight-based, e.g., stepwise regression, recursive regressor elimination.

Filter methods do not rely on the model structure that we identify, like the other two groups of methods. The measure of relevance for the regressors or combinations of regressors is extracted directly from the identification data. The relevant regressors are then selected on the basis of this measure. The relevance measure is usually computed based on the statistical properties of identification data, or based on measures from information theory, or based on other properties. These methods are attractive because they are computationally efficient in comparison with the wrapper and embedded methods. This computational efficiency comes from the fact that multiple optimisation runs are not necessary and from the relatively straightforward computation of the filter-method measures.

The other possible division is on the methods that focus on subsets of regressors and others that focus on regressor ranking [20].

The methods that focus on *subsets of regressors* construct models with different subsets of regressors to build a good model for a specific purpose, i.e., frequently to be a good predictor. The majority of wrapper methods can be classified as these kinds of methods.

The methods that focus on *regressor ranking* score regressors in terms of their information content. The majority of filter methods can be classified as regressor-ranking methods. Nevertheless, wrapper methods can also be used as regressor-ranking methods, when the aside-developed models are not used as the final predictors.

In this paper we focus on some of the possible regressor-selection methods from the literature that can be employed to rank the candidate regressors. The drawback of such an approach is that the most relevant regressors do not necessarily yield an optimal model. Therefore, we further enhance the

use of regressor-selection methods with model-based selection. The set of regressor-selection methods is, in our case, limited to the following methods that are contained in the ProOpter-IVS programme package utilised in [23]: Pearson's correlation coefficient - CCorr [18], distance correlation - dCorr [24], partial correlation - PCorr [18], mutual information - MI [25], partial mutual information - PMI [26], the method of model linear in the parameters - LIP [27], analysis of variance - ANOVA [22] and method with sensitivity analysis and regularisation of a neural network model - NNSA [28]. While LIP and NNSA are embedded methods, the others are filter methods. This list of used methods for regressor ranking is not exhaustive and can be expanded further.

The complete procedure is carried out in two stages. The first stage is pursuing the regressor-ranking and the selection of most informative regressors for each of the used methods. The second stage is to use previously selected regressors and the attached methods and validate them with regression modelling for the prediction. In this stage we reduce the number of used regressors and determine the method by which the best set of regressors is obtained.

For the validation of the prediction results, five criteria are used. The first three are standard criteria used in computational intelligence applications, while the last two are specific to the validation of ozone-concentration models.

- The mean standardised log loss - MSLL[29] is obtained by subtracting the loss of the model that predicts using a Gaussian with the mean $E(\mathbf{y})$ and the variance σ_y^2 of the measured data from the model log predictive density.

$$\text{MSLL} = \frac{1}{2N}\sum_{i=1}^{N}\left[\log(\sigma_i^2) + \frac{(E(\hat{y}_i) - y_i)^2}{\sigma_i^2}\right] - \frac{1}{2N}\sum_{i=1}^{N}\left[\log(\sigma_y^2) + \frac{(y_i - E(\mathbf{y}))^2}{\sigma_y^2}\right]. \tag{1}$$

  where y_i and $\hat{y}_i$ are the system's output measurement, i.e., observation, and the model output in the i-th step, respectively, σ_i^2 is the model output variance in the i-th step, and N is the number of used measurements. The MSLL is approximately zero for the simple models and negative for the better ones.

- The standardised mean-squared error - SMSE:

$$\text{SMSE} = \frac{1}{N}\frac{\sum_{i=1}^{N}(E(\hat{y}_i) - y_i)^2}{\sigma_y^2}. \tag{2}$$

This measure normalises the mean-squared error between the mean of the model output and the measured output of the process by the variance of the outputs of the validation data.

- The mean-relative-square error - MRSE:

$$\text{MRSE} = \sqrt{\frac{\sum_{i=1}^{N}(E(\hat{y}_i) - y_i)^2}{\sum_{i=1}^{N} y_i^2}}. \tag{3}$$

Some authors call this performance measure the relative-root-mean-square error (RRMSE).

- The success index - si, proposed by the European Environment Agency[30] with a threshold 140 μg/m^3:

$$\text{si} = \left(\frac{a}{m} + \frac{N + a - m - f}{N - m} - 1\right) \cdot 100\%, \tag{4}$$

where a is the number of correctly predicted values of the ozone concentration above the threshold, m is the number of measured values of the ozone concentration above the threshold and f is the number of all the predicted values of the ozone concentrations above the threshold. A larger value of the index means a better model prediction.

- The performance index - p^6[31]:

$$\text{p}^6 = \frac{1}{N}\sum_{i=1}^{N} J_i^6, \tag{5}$$

where J_i^6 is the cost function, which equals 1, meaning 'correctly classified', if there is no case of a false alarm and the measured concentration is high and if, at the same time, the absolute error is less than 20 μg/m^3 or if the relative error is less than 20%, and equals 0 otherwise.

4. Results and discussion

In general, we are interested in predicting the ozone concentration in the air for the following day, so the population can be informed about a possible high ozone concentration one day in advance. For that purpose, we are modelling the maximum ozone concentration, the average ozone concentration and the average ozone concentration between 8.00 and 20.00 hours, when the daily cycle of the ozone concentration has its highest values. The prediction of the ozone concentration for the following day can be made right after the last measurement for the current day is available.

From the available databases the following measurements are selected: ozone concentration (O_3), solid particles (PM_{10}), nitrogen oxides concentration (NO_x), nitrogen dioxide concentration (NO_2), air temperature, relative humidity, global solar radiation, diffuse solar radiation, wind speed, wind direction and precipitation. The measurements in the database are hourly based, but were averaged for our purpose. Depending on the model's objective the maximum or average values on a daily basis were built. As ozone forming is a dynamic process the preprocessed measurements were taken for four consecutive days. A total of 43 regressors are obtained this way to form the initial set for the regressor ranking.

The problems of predicting the maximum ozone concentration for the next day, the average ozone concentration between 8.00 and 20.00 hours for the next day and the average daily ozone concentration for the next day will be described next.

4.1. The first stage

The first stage of our investigation is the regressor ranking with selected regressor-selection methods. The relevance scores from the used methods are scaled between 0 and 1 for a visual comparison, though the absolute values of the relevance scores of the different methods do not have exactly the same meaning.

Maximum daily value of the ozone concentration

The ranking of the regressor selection for the selected methods is given in Table A.5 in Appendix A. From the results of the regressor selection for the maximum daily ozone concentrations in Table A.5 it is clear that the methods MI, CCorr and dCorr favor lagged ozone-concentration, air-temperature and global-solar-radiation regressors. The nnSA method puts an emphasis on the air temperature. The methods PCorr, PMI and LIP put excessive weight on

the regressors that they assess to be more informative, which in our case is the lagged ozone concentration. The results of the ANOVA method, on the other hand, do not show such large differences in the ranking weights of the regressors.

Figure 1 shows the cumulative results of all the methods for ranking the available 43 regressors for the prediction of the maximum daily value of the ozone concentration.

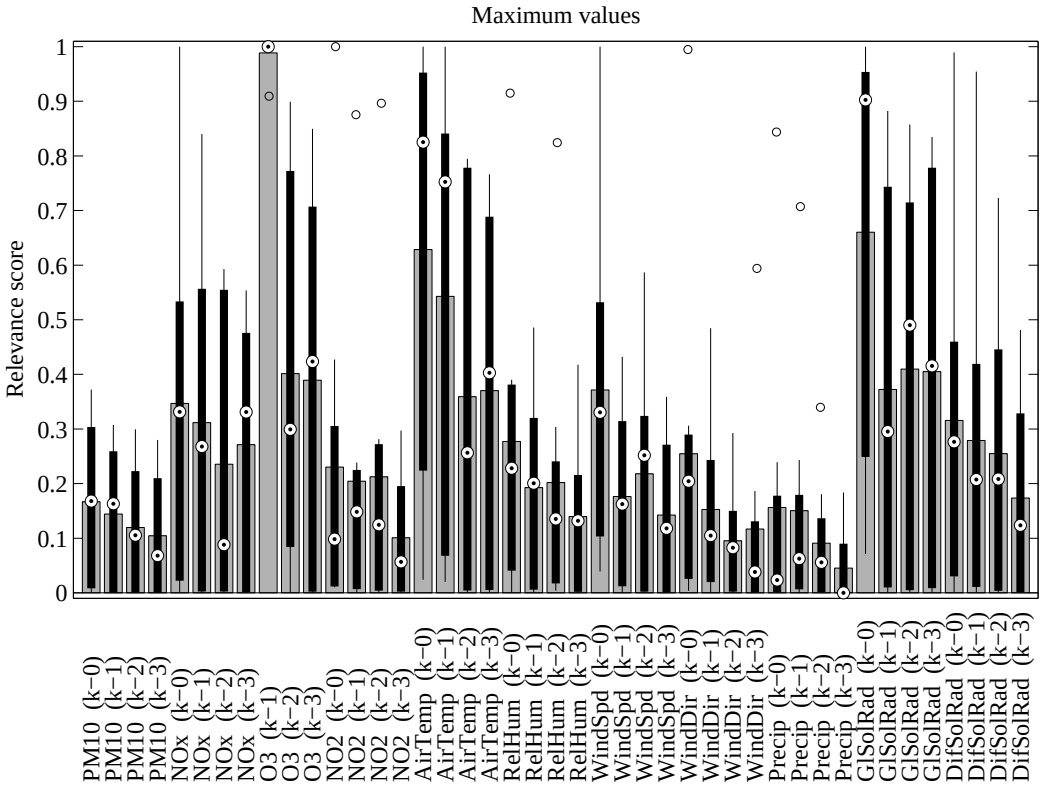

Figure 1: Box-and-whisker diagram of the cumulative regressor-ranking results for all used regressor-selection methods for the prediction of the maximum daily value of the ozone concentration. Regressors: ozone concentration (O_3), solid particles (PM10), nitrogen oxides concentration (NO_x), nitrogen dioxide concentration (NO_2), air temperature (AirTemp), relative humidity (RelHum), global solar radiation (GlSolRad), diffuse solar radiation (DifSolRad), wind speed (WindSpd), wind direction (WindDir) and precipitation (Precip). $k-i, i=0,\ldots,3$ denotes consecutive time instants.

Average value of the ozone concentration between 8.00 and 20.00 hours
The ranking of the regressor selection for the selected methods is given in

Table A.6 in Appendix A. From the results of the regressor selection for the prediction of the average value of the ozone concentrations between 8.00 and 20.00 hours in Table A.6 it is clear that the methods CCorr and dCorr favor a lagged ozone-concentration, air-temperature and global-solar-radiation regressors, as well as NO_x and NO_2 concentrations. The MI method provides similar results, but is not so clear. The nnSA method puts the most emphasis among the listed regressors on the NO_x concentration. The methods PCorr, PMI and LIP put excessive weight on the regressors that they evaluate as being more informative, which in our case are the lagged ozone concentration, followed by the global solar radiation, the relative humidity, the NO_2 concentration and the wind speed. Again, the results of the ANOVA method, on the other hand, do not show such large differences in the ranking weights of the regressors.

Figure 2 shows the cumulative results of all the methods for the ranking of the available 43 regressors for the prediction of the average value of the ozone concentration between 8.00 and 20.00 hours.

Average daily value of the ozone concentrations

The ranking of the regressor selection for the selected methods is given in Table A.7 in Appendix A. The results of the regressor selection for the prediction of the average daily value of the ozone concentrations in Table A.7 is similar to those for the prediction of the average value of the ozone concentration between 8.00 and 20.00 hours. The methods CCorr and dCorr also favor a diffuse solar radiation and wind speed with their lagged values. The nnSA method this time puts an emphasis on the past ozone concentration, the air temperature, the NO_2 and the global solar radiation. The methods PCorr, PMI, LIP and ANOVA provide comparable results as for the prediction of the average value of the ozone concentration between 8.00 and 20.00 hours.

Figure 3 shows the cumulative results of all the methods for the ranking of the available 43 regressors for the prediction of the average daily value of the ozone concentrations.

It was concluded from the ranking results that the 20 most relevant regressors for each of the 8 methods out of all 43 regressors are selected for each of the three models to enter the second stage. As the absolute values of the different relevance indices cannot be directly compared, the number of relevant regressors is selected so that it contains the most relevant regressors for each of the used relevance-selection methods.

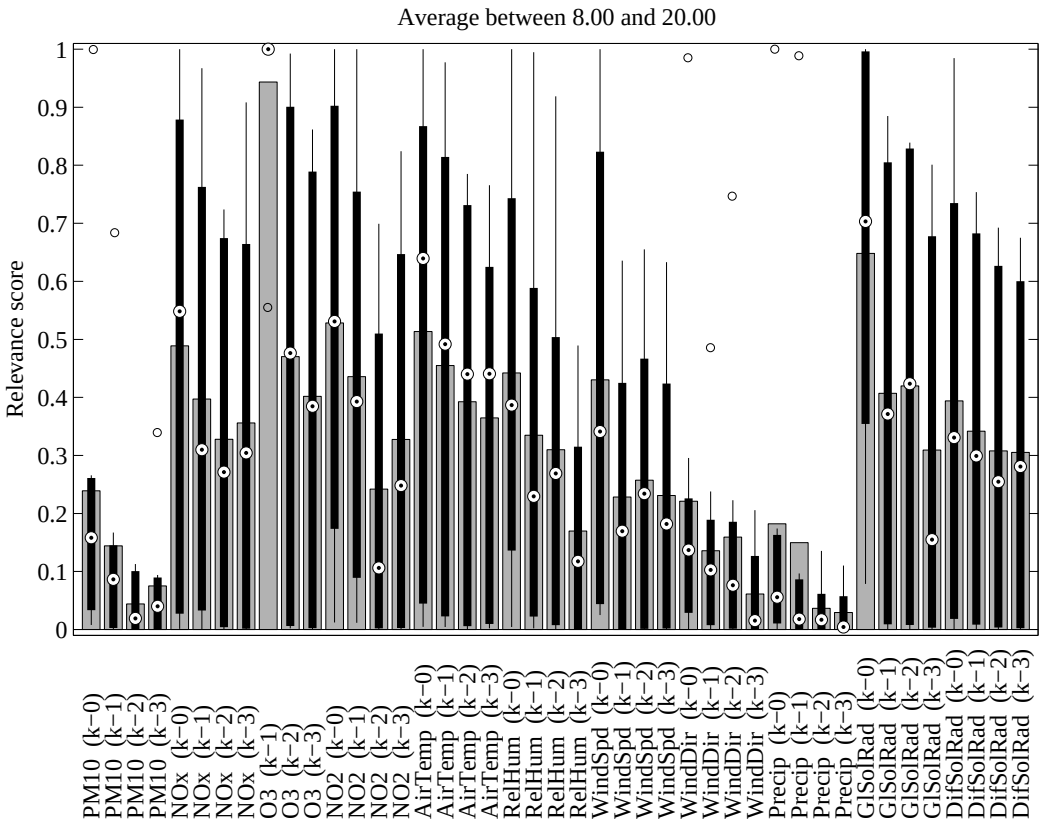

Figure 2: Box-and-whisker diagram of the cumulative regressor-ranking results for all the used regressor-selection methods for the prediction of the average value of the ozone concentration between 8.00 and 20.00 hours. Regressors: ozone concentration (O_3), solid particles (PM10), nitrogen oxides concentration (NO_x), nitrogen dioxide concentration (NO_2), air temperature (AirTemp), relative humidity (RelHum), global solar radiation (GlSolRad), diffuse solar radiation (DifSolRad), wind speed (WindSpd), wind direction (WindDir) and precipitation (Precip). $k-i, i = 0, \ldots, 3$ denotes consecutive time instants.

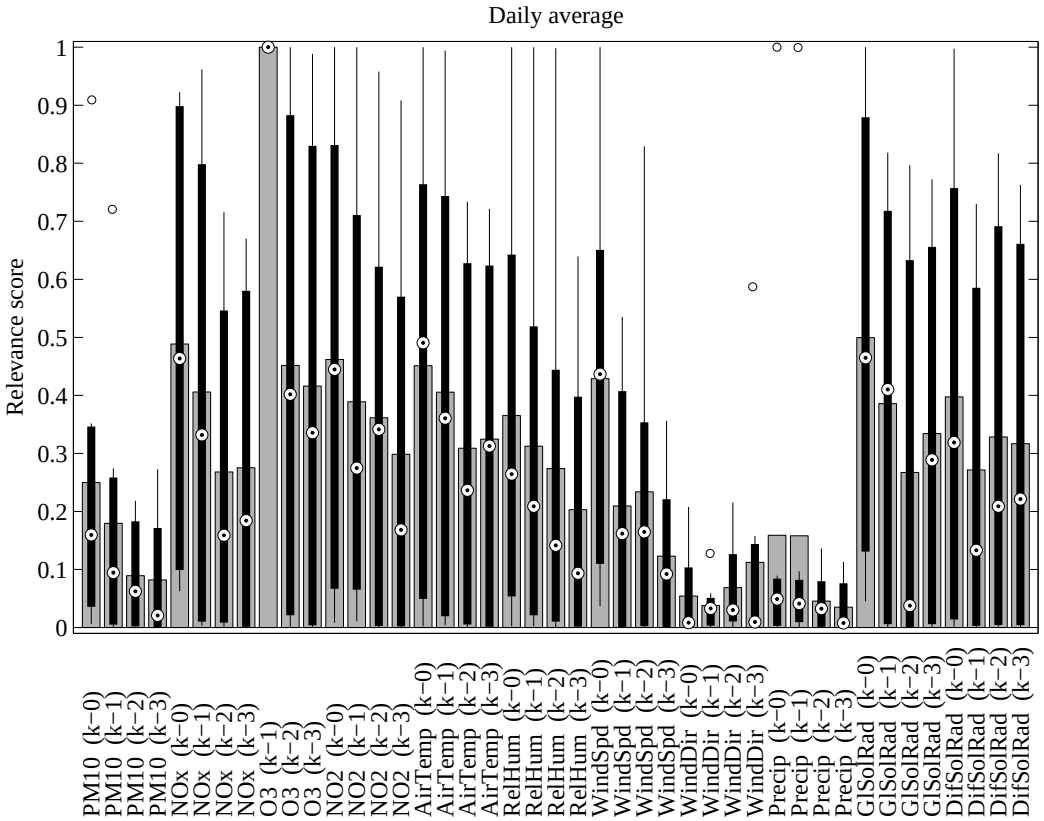

Figure 3: Box-and-whisker diagram of the cumulative regressor-ranking results for all the used regressor-selection methods for the prediction of the average daily value of the ozone concentration. Regressors: ozone concentration (O_3), solid particles (PM10), nitrogen oxides concentration (NO_x), nitrogen dioxide concentration (NO_2), air temperature (AirTemp), relative humidity (RelHum), global solar radiation (GlSolRad), diffuse solar radiation (DifSolRad), wind speed (WindSpd), wind direction (WindDir) and precipitation (Precip). $k-i, i=0,\ldots,3$ denotes consecutive time instants.

4.2. The second stage

The second stage of our investigation is the selection of a single method and the set of the most informative regressors based on a test with a model prediction.

A Gaussian-process (GP) model [32],[29] is used for the second stage. Any other suitable regression method can be used for the modelling and the prediction test. This method is selected arbitrarily, mainly because of the method's properties described in the continuation, and because it contains a relatively small number of elements to be decided by the modeler, and because it is one of the candidates for the final model. However, the study for the selection of the regression method is not the focus of this paper.

A GP model is a probabilistic, nonparametric, kernel model based on the principles of Bayesian probability. In other words, it provides a Bayesian interpretation of the kernel methods. A GP model differs from most of the other black-box identification approaches in that it searches for relationships among the measured data, rather than trying to approximate the modelled system by fitting the parameters of the selected basis functions. The output of the GP model is a normal distribution, expressed in terms of the mean and the variance. Their modelling properties are reviewed in [29], [33], and [34] with applications in, e.g., [35], [12]. The idea of GP models is rather simple. A GP model assumes that the output is a realisation of a GP with a joint probability density function:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{m}, \mathbf{K}), \tag{6}$$

with the mean and covariance being functions of the inputs $\mathbf{x}$. Usually, the mean function is defined as $\mathbf{0}$, while the covariance function or kernel

$$\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

defines the characteristics of the process to be modelled, i.e. the stationarity, smoothness, etc. The most commonly used function is the squared exponential (SE) covariance function with an automatic relevance determination (ARD) [29]. This covariance function is smooth and stationary and makes it possible to determine the impact on the model for each input by an optimisation of the parameters, called hyperparameters. Once the covariance matrix $\mathbf{K}$ is calculated, the predictive (normal) distribution for the new input $\mathbf{x}^*$ is simply calculated using:

$$\begin{aligned} \mu &= \mathbf{k}(\mathbf{x}^*)^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}, \\ \sigma^2 &= \kappa(\mathbf{x}^*) - \mathbf{k}(\mathbf{x}^*)^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{k}(\mathbf{x}^*), \end{aligned} \tag{8}$$

where $\mathbf{k}(\mathbf{x}^*)$ is the vector of covariances between the new input sample and the training input samples, and $\kappa(\mathbf{x}^*)$ is the covariance between the new input sample itself. As can be seen from Equations (8) the GP model, in addition to the mean value, also provides information about the confidence in the prediction by the variance. Usually, the confidence of the prediction is depicted with a 2σ interval, which is an about 95% confidence interval. It highlights the areas of the input space where the prediction quality is poor, due to the lack of data or noisy data, by indicating a wider confidence interval around the predicted mean.

It is not the purpose of this paper to explain the details of the modelling method. The reader is referred to [32],[29] for more details about this regression-modelling method. The square exponential covariance function, combined with the white-noise covariance function, is used as the kernel, and the marginal likelihood maximisation is used for the selection of the hyperparameters. To simplify the procedure, we test only the models obtained from sets of the 5, 10, 15 and 20 most informative regressors for each of methods we used in the first stage. This choice of sets is a compromise because the absolute values of the scores of different methods are not directly comparable and we wished to avoid testing each and every regressor's inclusion and coming down to one of the time-consuming wrapper methods like forward selection or backward elimination.

The available data contained data instants for two years, i.e., 2012 and 2013. These data are divided into 11 subsets: 10-fold cross-validation is used for the prediction validation on the 10 subsets, while the remaining, larger data subset is used for testing the prediction.

Each prediction result is evaluated with the criteria described in Section 3. The final table of results contains 160 values, i.e., the averages of 10-fold cross-validation. From these results the most efficient method for regressor selection used in the first stage and with the best selection of regressors can be selected.

Maximum daily value of the ozone concentration

One of the important issues is which validation criterion is the most suitable for a specific modelling objective. In the case of modelling the maximum daily value of the ozone concentration, the success index si, described with Eq. (4), is considered as the most suitable for the task at hand. This is due to the purpose of the model to be developed, which is intended for informing the public about the maximum daily value of the one-hour mean concentration in line with environmental regulations [30].

Table 1 shows the results of the prediction validation with criteria from Eqs. (1)–(5). The best performance is shown by the model with the most

Table 1: Results of the model validation for the maximum daily value of the ozone concentration for 5, 10, 15 and 20 regressors as listed in Table A.5.

	$CCorr_5$	$PCorr_5$	MI_5	PMI_5	LIP_5	$nnSA_5$	$ANOVA_5$	$dCorr_5$	$Cummul_5$
MSLL	0.66	0.66	0.60	0.71	0.63	0.63	0.63	0.65	0.52
MRSE	0.16	0.14	0.16	0.17	0.15	0.16	0.14	0.18	0.16
SMSE	0.18	0.14	0.15	0.18	0.14	0.16	0.12	0.19	0.13
si	72.46	73.16	57.55	65.11	68.34	74.34	81.52	61.31	65.08
p^6	0.90	0.88	0.92	0.83	0.94	0.95	0.92	0.87	0.90
	$CCorr_{10}$	$PCorr_{10}$	MI_{10}	PMI_{10}	LIP_{10}	$nnSA_{10}$	$ANOVA_{10}$	$dCorr_{10}$	$Cummul_{10}$
MSLL	0.65	0.63	0.66	0.60	0.65	0.57	0.60	0.60	0.56
MRSE	0.16	0.13	0.16	0.14	0.13	0.16	0.12	0.14	0.16
SMSE	0.17	0.11	0.17	0.12	0.11	0.14	0.09	0.14	0.14
si	70.40	**97.02**	69.76	75.65	69.99	81.32	82.61	73.06	63.83
p^6	0.94	0.90	0.85	0.95	0.97	0.90	1.00	0.89	0.92
	$CCorr_{15}$	$PCorr_{15}$	MI_{15}	PMI_{15}	LIP_{15}	$nnSA_{15}$	$ANOVA_{15}$	$dCorr_{15}$	$Cummul_{15}$
MSLL	0.64	0.64	0.65	0.64	0.51	0.64	0.61	0.68	0.59
MRSE	0.14	0.12	0.14	0.13	0.13	0.12	0.13	0.14	0.15
SMSE	0.13	0.10	0.14	0.11	0.09	0.10	0.10	0.13	0.13
si	57.28	73.07	68.77	85.10	70.10	84.21	75.41	54.78	73.10
p^6	0.95	0.98	0.92	0.91	0.93	0.97	0.94	0.96	0.91
	$CCorr_{20}$	$PCorr_{20}$	MI_{20}	PMI_{20}	LIP_{20}	$nnSA_{20}$	$ANOVA_{20}$	$dCorr_{20}$	$Cummul_{20}$
MSLL	0.68	0.64	0.62	0.59	0.56	0.64	0.57	0.63	0.56
MRSE	0.14	0.14	0.15	0.14	0.13	0.12	0.14	0.13	0.15
SMSE	0.14	0.11	0.13	0.13	0.11	0.10	0.11	0.11	0.11
si	79.94	92.74	63.03	80.71	87.06	62.55	74.75	86.33	75.35
p^6	0.91	1.00	0.94	0.89	0.96	0.97	0.92	0.92	0.90

informative ten regressors obtained with a partial correlation - the PCorr method. The model contains the following regressors, which are labelled as described in the caption of, e.g., Figure 1, and sorted by relevance: $O_3(k-1)$, $GlSolRad(k-0)$, $WindSpd(k-0)$, $AirTemp(k-0)$, $AirTemp(k-1)$, $DifSolRad(k-0)$, $NO_x(k-0)$, $NO_2(k-0)$, $RelHum(k-2)$, $WindDir(k-0)$.

Figure 4, shows the comparison between the measured and predicted maximum daily ozone concentration of the winning model on the test data. From Figure 4 it can be seen that the model predictions of the model with the selected regressors on average predict the maximum values satisfactorily. If a prediction tolerance band was added to the figure, then it would be clear that the deviations from target values would be even more acceptable.

Average value of the ozone concentration between 8.00 and 20.00 hours We use the same procedure with the prediction of the average concentration of ozone between 8.00 and 20.00 hours and for the model to predict the average daily concentration. It is important to note that for these two models the average values of the measurements are used. The SMSE and MRSE are selected as the most suitable validation criteria because of their generality in comparison with other more specific measures and because of the special

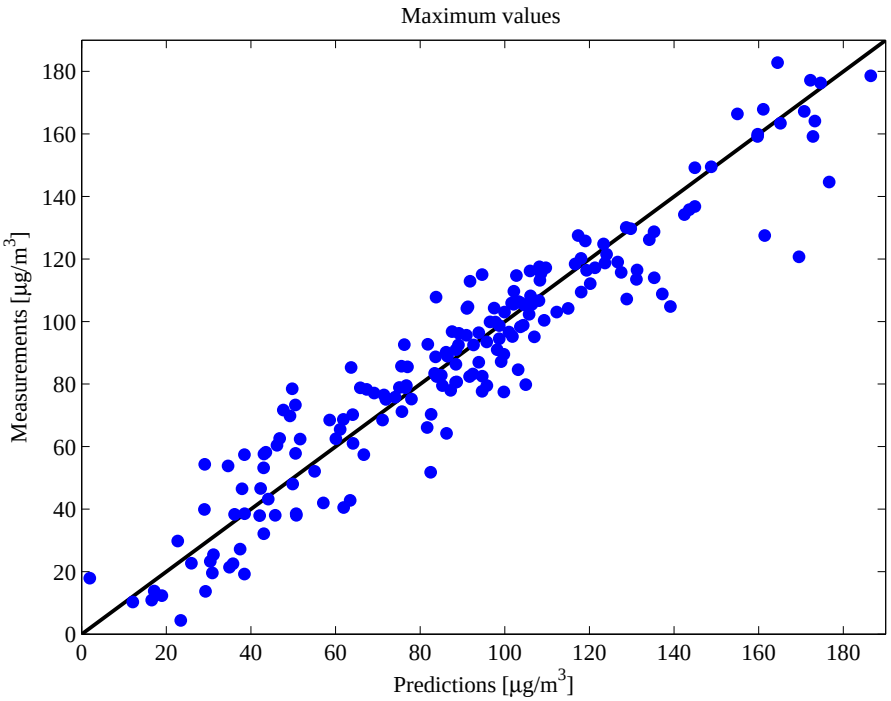

Figure 4: Measured O_3 values versus the predicted values obtained by the model with the most informative ten regressors obtained with a partial correlation.

interest in the mean value of the predictions.

Table 2 shows the results of the prediction validation with criteria from Eqs. (1)–(3). The model for the prediction of the average concentration of ozone between 8.00 and 20.00 hours of the day contains fifteen regressors obtained by the method of model linear in parameters - LIP. The model contains the following regressors sorted by relevance: $O_3(k-1)$, $GlSolRad(k-0)$, $WindSpd(k-0)$, $NO_2(k-1)$, $NO_2(k-0)$, $PM10(k-0)$, $AirTemp(k-0)$, $AirTemp(k-1)$, $RelHum(k-0)$, $NO_x(k-0)$, $Precip(k-0)$, $AirTemp(k-3)$, $O_3(k-2)$, $RelHum(k-1)$, $WindDir(k-0)$.

Figure 5 shows a comparison between the measured and predicted average ozone concentration between 8.00 and 20.00 hours of the winning model on the test data.

Average daily value of the ozone concentrations

Table 3 shows the results of the prediction validation with the criteria from Eqs. (1)–(3). The model for the prediction of the average concentration con-

Table 2: Results of the model validation for the average value of the ozone concentration between 8.00 and 20.00 hours for 5, 10, 15 and 20 regressors as listed in Table 2.

	CCorr_5	PCorr_5	MI_5	PMI_5	LIP_5	nnSA_5	ANOVA_5	dCorr_5	Cummul_5
MSLL	0.54	0.56	0.47	0.51	0.46	0.50	0.51	0.50	0.50
MRSE	0.14	0.16	0.13	0.13	0.15	0.17	0.16	0.16	0.14
SMSE	0.09	0.11	0.07	0.08	0.09	0.13	0.09	0.10	0.09
	CCorr_{10}	PCorr_{10}	MI_{10}	PMI_{10}	LIP_{10}	nnSA_{10}	ANOVA_{10}	dCorr_{10}	Cummul_{10}
MSLL	0.49	0.51	0.51	0.46	0.43	0.46	0.50	0.49	0.48
MRSE	0.13	0.12	0.12	0.13	0.13	0.14	0.13	0.16	0.13
SMSE	0.08	0.06	0.07	0.07	0.07	0.07	0.07	0.10	0.08
	CCorr_{15}	PCorr_{15}	MI_{15}	PMI_{15}	LIP_{15}	nnSA_{15}	ANOVA_{15}	dCorr_{15}	Cummul_{15}
MSLL	0.47	0.50	0.50	0.47	0.57	0.45	0.56	0.46	0.52
MRSE	0.14	0.12	0.12	0.12	**0.11**	0.13	0.12	0.12	0.12
SMSE	0.08	0.06	0.06	0.06	**0.05**	0.07	0.07	0.06	0.07
	CCorr_{20}	PCorr_{20}	MI_{20}	PMI_{20}	LIP_{20}	nnSA_{20}	ANOVA_{20}	dCorr_{20}	Cummul_{20}
MSLL	0.55	0.47	0.48	0.51	0.47	0.47	0.47	0.38	0.49
MRSE	0.12	0.12	0.12	0.11	0.12	0.13	0.11	0.11	0.13
SMSE	0.06	0.06	0.06	0.06	0.06	0.07	0.05	0.05	0.07

Table 3: Results of the model validation for the average daily value of the ozone concentration for 5, 10, 15 and 20 regressors as listed in Table A.7.

	CCorr_5	PCorr_5	MI_5	PMI_5	LIP_5	nnSA_5	ANOVA_5	dCorr_5	Cummul_5
MSLL	0.50	0.55	0.45	0.49	0.46	0.44	0.49	0.49	0.48
MRSE	0.15	0.15	0.16	0.12	0.14	0.15	0.17	0.15	0.13
SMSE	0.11	0.12	0.10	0.08	0.09	0.10	0.13	0.12	0.09
	CCorr_{10}	PCorr_{10}	MI_{10}	PMI_{10}	LIP_{10}	nnSA_{10}	ANOVA_{10}	dCorr_{10}	Cummul_{10}
MSLL	0.46	0.44	0.47	0.46	0.42	0.44	0.43	0.42	0.46
MRSE	0.13	0.12	0.15	**0.11**	0.13	0.12	0.15	0.15	0.12
SMSE	0.07	0.06	0.09	**0.06**	0.07	0.06	0.09	0.10	0.06
	CCorr_{15}	PCorr_{15}	MI_{15}	PMI_{15}	LIP_{15}	nnSA_{15}	ANOVA_{15}	dCorr_{15}	Cummul_{15}
MSLL	0.53	0.46	0.44	0.42	0.45	0.43	0.44	0.50	0.46
MRSE	0.13	0.12	0.12	0.12	0.12	0.13	0.12	0.13	0.12
SMSE	0.10	0.06	0.05	0.06	0.07	0.08	0.07	0.08	0.07
	CCorr_{20}	PCorr_{20}	MI_{20}	PMI_{20}	LIP_{20}	nnSA_{20}	ANOVA_{20}	dCorr_{20}	Cummul_{20}
MSLL	0.47	0.47	0.47	0.44	0.39	0.44	0.43	0.48	0.45
MRSE	0.13	0.13	0.12	0.12	0.12	0.12	0.12	0.12	0.12
SMSE	0.07	0.07	0.08	0.06	0.06	0.07	0.06	0.07	0.06

tains ten regressors obtained by partial mutual information - PMI. The model contains the following regressors: $O_3(k-1)$, $WindSpd(k-0)$, $GlSolRad(k-0)$, $NO_{\text{x}}(k-0)$, $AirTemp(k-0)$, $NO_2(k-0)$, $RelHum(k-0)$, $PM10(k-0)$, $AirTemp(k-1)$, $O_3(k-2)$.

Figure 6 shows the comparison between the measured and predicted average daily ozone concentration of the winning model on the test data.

Table 4 shows the regressors for the three final models. Be aware that the values with no delay, e.g., $AirTemp(k)$, correspond to meteorological and pollution forecasts, as mentioned in Section 2. It is clear from Table 4 that the regressors, even though they are similar, are different for the different models. What is common to all three models is that they contain meteorological and pollution variables.

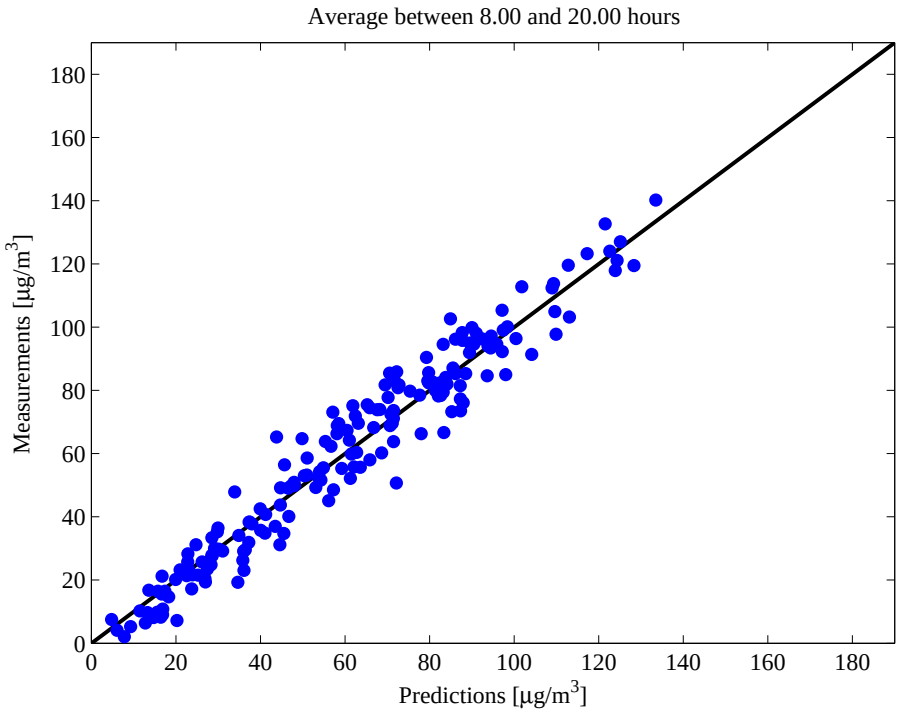

Figure 5: Measured O_3 values versus the predicted values obtained by the model with the most informative ten regressors obtained with a partial correlation.

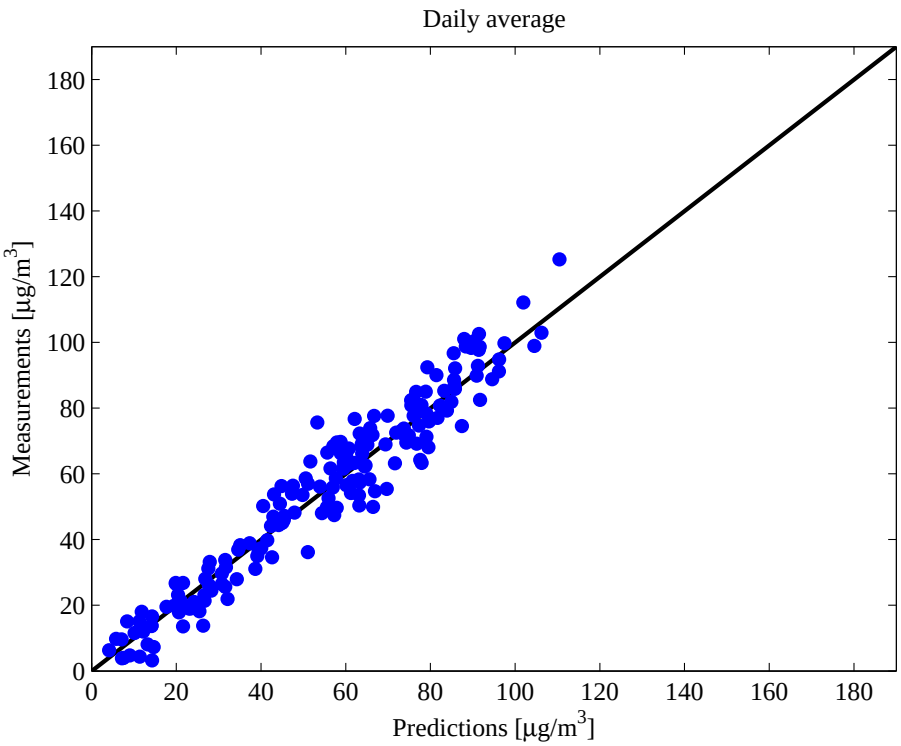

Figure 6: Measured O_3 values versus the predicted values obtained by the model with the most informative ten regressors obtained with a partial correlation.

Table 4: Regressors for the three final models.

	Maximum con.	8.00-20.00 average	Daily average
1	$O_3(k-1)$	$O_3(k-1)$	$O_3(k-1)$
2	$AirTemp(k)$	$O_3(k-2)$	$O_3(k-2)$
3	$AirTemp(k-1)$	$AirTemp(k)$	$AirTemp(k)$
4	$GlSolRad(k)$	$AirTemp(k-1)$	$AirTemp(k-1)$
5	$WindSpd(k)$	$AirTemp(k-3)$	$WindSpd(k)$
6	$DifSolRad(k)$	$RelHum(k)$	$GlSolRad(k)$
7	$RelHum(k-2)$	$RelHum(k-1)$	$RelHum(k)$
8	$WindDir(k)$	$NO_2(k)$	$NO_{\mathrm{x}}(k)$
9	$NO_{\mathrm{x}}(k)$	$NO_2(k-1)$	$NO_2(k)$
10	$NO_2(k)$	$GlSolRad(k)$	$PM10(k)$
11		$WindSpd(k)$	
12		$WindDir(k)$	
13		$PM10(k)$	
14		$Precip(k)$	
15		$NO_{\mathrm{x}}(k)$	

5. Conclusions

The paper provides a comparison of the different methods for regressor ranking for a prediction model for ozone concentrations in the city of Nova Gorica, Slovenia. The case study confirms that regressor selection as well as modelling depends on the location and it cannot be generalised. It was shown that for three different sorts of averaging interval of the ozone concentration three different models with three different sets of regressors obtained by three different regressor-ranking methods had to be developed.

On the basis of the presented experiments, no particular regressor-selection method can be claimed to be superior for all of the models in our case. Regressor selection should be based on the consideration of different methods and the important inputs should be selected for every case study, when one notices consistent results among the methods. Case-specific models proved to be the best for our purpose.

In the next step, which is outside the scope of this paper, is the selection of the type of modelling method that is going to be used for studies at this and other geographical locations. Even though the GP regression modelling is used in this study, this might not be the final solution. The decisions about whether the modelling is on-line, recursive and which regression method is used are still to be made.

Future work on the modelling of different sorts of averaging intervals for ozone concentrations is envisaged not only for the present geographical location, but also for other locations with different geographical and meteorological conditions.

Appendix A. Results of the regressor selection

Table A.5: First 20 ranking results for the used methods of ranking and cumulative results for the maximum daily value of the ozone concentration. Regressors: ozone concentration (O_3), solid particles (PM10), nitrogen oxides concentration (NO_x), nitrogen dioxide concentration (NO_2), air temperature (AirTemp), relative humidity (RelHum), global solar radiation (GlSolRad), diffuse solar radiation (DifSolRad), wind speed (WindSpd), wind direction (WindDir) and precipitation (Precip). $k-i, i=0,\ldots,3$ denotes consecutive time instants.

	CCorr	PCorr	MI	PMI	LIP	nnSA	ANOVA	dCorr	Cumulative
1	O_3(k-1)	O_3(k-1)	O_3(k-1)	O_3(k-1)	O_3(k-1)	AirTemp(k-0)	O_3(k-1)	O_3(k-1)	O_3(k-1)
2	GlSolRad(k-0)	GlSolRad(k-0)	GlSolRad(k-0)	AirTemp(k-0)	GlSolRad(k-0)	O_3(k-1)	AirTemp(k-0)	GlSolRad(k-0)	GlSolRad(k-0)
3	AirTemp(k-0)	WindSpd(k-0)	AirTemp(k-0)	O_3(k-2)	WindSpd(k-0)	GlSolRad(k-0)	WindSpd(k-0)	AirTemp(k-0)	AirTemp(k-0)
4	O_3(k-2)	AirTemp(k-0)	AirTemp(k-1)	GlSolRad(k-0)	AirTemp(k-0)	AirTemp(k-1)	NO_x(k-0)	O_3(k-2)	AirTemp(k-1)
5	GlSolRad(k-1)	AirTemp(k-1)	O_3(k-2)	RelHum(k-0)	AirTemp(k-1)	AirTemp(k-2)	GlSolRad(k-0)	GlSolRad(k-1)	GlSolRad(k-2)
6	O_3(k-3)	DifSolRad(k-0)	GlSolRad(k-1)	AirTemp(k-1)	NO_2(k-0)	AirTemp(k-3)	NO_2(k-0)	GlSolRad(k-2)	GlSolRad(k-3)
7	GlSolRad(k-2)	NO_x(k-0)	GlSolRad(k-2)	WindSpd(k-0)	AirTemp(k-3)	NO_x(k-0)	AirTemp(k-1)	O_3(k-3)	O_3(k-2)
8	AirTemp(k-1)	NO_2(k-0)	O_3(k-3)	WindDir(k-1)	RelHum(k-2)	GlSolRad(k-2)	WindDir(k-0)	AirTemp(k-1)	O_3(k-3)
9	GlSolRad(k-3)	RelHum(k-2)	GlSolRad(k-3)	WindDir(k-0)	NO_x(k-0)	NO_2(k-0)	DifSolRad(k-0)	GlSolRad(k-3)	GlSolRad(k-1)
10	AirTemp(k-2)	WindDir(k-0)	WindSpd(k-0)	WindSpd(k-1)	WindDir(k-0)	NO_x (k-3)	DifSolRad(k-1)	AirTemp(k-2)	WindSpd(k-0)
11	AirTemp(k-3)	DifSolRad(k-1)	AirTemp(k-2)	GlSolRad(k-1)	Precip(k-0)	O_3(k-2)	RelHum(k-0)	AirTemp(k-3)	AirTemp(k-3)
12	NO_x(k-2)	GlSolRad(k-3)	AirTemp(k-3)	PM10(k-0)	DifSolRad(k-0)	NO_x(k-1)	NO_2(k-2)	NO_x(k-1)	AirTemp(k-2)
13	NO_x(k-1)	Precip(k-1)	RelHum(k-0)	NO_2(k-0)	PM10(k-0)	O_3(k-3)	NO_2(k-1)	WindSpd(k-0)	NO_x(k-0)
14	NO_x(k-3)	Precip(k-0)	RelHum(k-1)	RelHum(k-1)	DifSolRad(k-1)	DifSolRad(k-0)	Precip(k-0)	NO_x(k-2)	DifSolRad(k-0)
15	NO_x(k-0)	WindDir(k-3)	WindSpd(k-3)	NO_2(k-1)	RelHum(k-1)	WindDir(k-0)	NO_x(k-1)	NO_x(k-0)	NO_x(k-1)
16	DifSolRad(k-3)	RelHum(k-0)	WindSpd(k-2)	GlSolRad(k-2)	RelHum(k-0)	WindDir(k-1)	RelHum(k-2)	NO_x(k-3)	DifSolRad(k-1)
17	DifSolRad(k-0)	WindDir(k-1)	NO_x(k-3)	AirTemp(k-2)	NO_2(k-3)	GlSolRad(k-3)	GlSolRad(k-3)	DifSolRad(k-0)	RelHum(k-0)
18	DifSolRad(k-2)	NO_2(k-2)	DifSolRad(k-0)	DifSolRad(k-0)	O_3(k-2)	WindSpd(k-0)	DifSolRad(k-2)	DifSolRad(k-3)	NO_x(k-3)
19	DifSolRad(k-1)	NO_x(k-2)	Precip(k-1)	DifSolRad(k-2)	WindSpd(k-3)	GlSolRad(k-1)	Precip(k-1)	DifSolRad(k-2)	DifSolRad(k-2)
20	WindSpd(k-0)	NO_2(k-1)	Precip(k-0)	RelHum(k-2)	NO_x(k-3)	WindSpd(k-3)	WindDir(k-3)	DifSolRad(k-1)	WindDir(k-0)

Table A.6: First 20 ranking results for the used methods of ranking and cumulative results for the average value of the ozone concentration between hours 8.00 and 20.00. Regressors: ozone concentration (O_3), solid particles (PM10), nitrogen oxides concentration (NO_x), nitrogen dioxide concentration (NO_2), air temperature (AirTemp), relative humidity (RelHum), global solar radiation (GlSolRad), diffuse solar radiation (DifSolRad), wind speed (WindSpd), wind direction (WindDir) and precipitation (Precip). $k-i, i=0,\ldots,3$ denotes consecutive time instants.

	CCorr	PCorr	MI	PMI	LIP	nnSA	ANOVA	dCorr	Cumulative
1	O_3(k-1)	O_3(k-1)	O_3(k-1)	O_3(k-1)	O_3(k-1)	NO_x(k-0)	O_3(k-1)	GlSolRad(k-0)	O_3(k-1)
2	GlSolRad(k-0)	GlSolRad(k-0)	GlSolRad(k-0)	GlSolRad(k-0)	GlSolRad(k-0)	O_3(k-1)	NO_2(k-0)	O_3(k-1)	GlSolRad(k-0)
3	O_3(k-2)	NO_2(k-0)	WindSpd(k-0)	RelHum(k-0)	WindSpd(k-0)	AirTemp(k-0)	NO_2(k-1)	NO_2(k-0)	NO_2(k-0)
4	GlSolRad(k-1)	NO_2(k-1)	AirTemp(k-0)	WindSpd(k-0)	NO_2(k-1)	AirTemp(k-3)	AirTemp(k-0)	NO_x(k-0)	AirTemp(k-0)
5	NO_2(k-0)	RelHum(k-0)	NO_x(k-0)	NO_2(k-0)	NO_2(k-0)	NO_2 (k-0)	RelHum(k-0)	O_3(k-2)	NO_x(k-0)
6	AirTemp(k-0)	NO_x(k-1)	GlSolRad(k-1)	AirTemp(k-0)	PM10(k-0)	GlSolRad(k-1)	GlSolRad(k-0)	GlSolRad(k-1)	O_3(k-2)
7	O_3(k-3)	PM10(k-0)	O_3(k-2)	NO_x(k-0)	AirTemp(k-0)	NO_2 (k-1)	WindSpd(k-0)	AirTemp(k-0)	AirTemp(k-1)
8	GlSolRad(k-2)	Precip(k-0)	NO_2(k-0)	RelHum(k-1)	AirTemp(k-1)	GlSolRad(k-2)	Precip(k-0)	O_3(k-3)	RelHum(k-0)
9	NO_x(k-0)	WindSpd(k-0)	AirTemp(k-1)	WindDir(k-0)	RelHum(k-0)	GlSolRad(k-0)	PM10(k-0)	GlSolRad(k-2)	NO_2(k-1)
10	AirTemp(k-1)	WindDir(k-0)	AirTemp(k-2)	AirTemp(k-1)	NO_x(k-0)	AirTemp(k-1)	RelHum(k-1)	AirTemp(k-1)	WindSpd(k-0)
11	GlSolRad(k-3)	AirTemp(k-0)	GlSolRad(k-3)	DifSonSev (k-0)	Precip(k-0)	AirTemp(k-2)	O_3(k-2)	WindSpd(k-0)	GlSolRad(k-2)
12	AirTemp(k-2)	AirTemp(k-1)	RelHum(k-0)	PM10(k-0)	AirTemp(k-3)	RelHum(k-2)	Precip(k-1)	NO_x(k-1)	GlSolRad(k-1)
13	AirTemp(k-3)	AirTemp(k-3)	DifSolRad(k-0)	GlSolRad(k-1)	O_3(k-2)	O_3(k-2)	WindDir(k-0)	GlSolRad(k-3)	O_3(k-3)
14	NO_2(k-1)	DifSolRad(k-0)	O_3(k-3)	WindDir(k-1)	RelHum(k-1)	NO_x(k-3)	DifSolRad(k-0)	AirTemp(k-2)	NO_x(k-1)
15	NO_x(k-1)	DifSolRad(k-1)	NO_x(k-1)	NO_2(k-1)	WindDir(k-0)	O_3(k-3)	AirTemp(k-1)	RelHum(k-0)	DifSolRad(k-0)
16	DifSolRad(k-0)	Precip(k-2)	GlSolRad(k-2)	RelHum(k-2)	WindDir(k-3)	NO_2 (k-3)	NO_x(k-1)	NO_2(k-1)	AirTemp(k-2)
17	RelHum(k-0)	O_3(k-2)	AirTemp(k-3)	GlSolRad(k-2)	PM10(k-2)	RelHum(k-3)	RelHum(k-2)	AirTemp(k-3)	AirTemp(k-3)
18	NO_2(k-2)	NO_x(k-2)	NO_2(k-1)	AirTemp(k-2)	WindDir (k-1)	DifSolRad(k-3)	NO_x(k-3)	DifSolRad(k-0)	NO_x(k-3)
19	NO_x(k-2)	GlSolRad(k-2)	DifSolRad(k-1)	GlSolRad(k-3)	Precip (k-1)	RelHum(k-0)	GlSolRad(k-2)	NO_x(k-2)	DifSolRad (k-1)
20	DifSolRad(k-1)	Precip(k-1)	WindSpd(k-1)	DifSolRad(k-1)	GlSolRad(k-3)	GlSolRad(k-3)	NO_2(k-3)	DifSolRad(k-1)	RelHum(k-1)

Table A.7: First 20 ranking results for the used methods of ranking and cumulative results for the average daily value of the ozone concentration. Regressors: ozone concentration (O_3), solid particles (PM10), nitrogen oxides concentration (NO_x), nitrogen dioxide concentration (NO_2), air temperature (AirTemp), relative humidity (RelHum), global solar radiation (GlSolRad), diffuse solar radiation (DifSolRad), wind speed (WindSpd), wind direction (WindDir) and precipitation (Precip). $k-i, i=0,\ldots,3$ denotes consecutive time instants.

	CCorr	PCorr	MI	PMI	LIP	nnSA	ANOVA	dCorr	Cumulative
1	O_3(k-1)	O_3(k-1)	O_3(k-1)	O_3(k-1)	O_3(k-1)	O_3(k-1)	O_3(k-1)	O_3(k-1)	O_3(k-1)
2	O_3(k-2)	NO_2(k-0)	NO_x(k-0)	WindSpd(k-0)	NO_x(k-0)	NO_2(k-2)	GlSolRad(k-0)	NO_x(k-0)	GlSolRad(k-0)
3	GlSolRad(k-0)	GlSolRad(k-0)	GlSolRad(k-0)	GlSolRad(k-0)	GlSolRad(k-0)	AirTemp(k-2)	NO_2(k-1)	O_3(k-2)	NO_x(k-0)
4	NO_x(k-0)	NO_2(k-1)	O_3(k-2)	NO_x(k-0)	WindSpd(k-0)	AirTemp(k-0)	NO_2 (k-0)	GlSolRad(k-0)	NO_2(k-0)
5	O_3(k-3)	WindSpd(k-0)	AirTemp(k-0)	AirTemp(k-0)	NO_2(k-1)	GlSolRad(k-1)	RelHum(k-0)	NO_2(k-0)	O_3(k-2)
6	NO_2(k-0)	NO_x(k-0)	GlSolRad(k-1)	NO_2(k-0)	NO_2(k-0)	NO_x(k-0)	Precip(k-0)	O_3(k-3)	AirTemp(k-0)
7	GlSolRad(k-1)	RelHum(k-0)	WindSpd(k-0)	RelHum(k-0)	PM10(k-0)	AirTemp(k-1)	AirTemp(k-0)	NO_x(k-1)	WindSpd(k-0)
8	GlSolRad(k-2)	Precip(k-0)	O_3(k-3)	PM10(k-0)	AirTemp(k-1)	GlSolRad(k-3)	WindSpd(k-0)	GlSolRad(k-1)	O_3(k-3)
9	NO_x(k-1)	PM10(k-0)	AirTemp(k-1)	AirTemp(k-1)	Precip(k-0)	RelHum(k-1)	O_3(k-2)	GlSolRad(k-2)	NO_x(k-1)
10	GlSolRad(k-3)	RelHum(k-1)	DifSolRad(k-0)	O_3(k-2)	NO_x(k-1)	AirTemp(k-3)	RelHum(k-1)	DifSolRad(k-0)	AirTemp(k-1)
11	AirTemp(k-0)	Precip(k-1)	AirTemp(k-2)	WindDir(k-2)	AirTemp(k-0)	NO_x(k-1)	Precip(k-1)	AirTemp(k-0)	DifSolRad(k-0)
12	DifSolRad(k-0)	O_3(k-2)	NO_x(k-1)	RelHum(k-1)	RelHum(k-0)	NO_2(k-1)	RelHum(k-2)	GlSolRad(k-3)	NO_2(k-1)
13	AirTemp(k-1)	RelHum(k-2)	NO_2(k-0)	DifSolRad(k-0)	RelHum(k-1)	O_3(k-2)	DifSolRad(k-0)	AirTemp(k-1)	GlSolRad(k-1)
14	AirTemp(k-2)	NO_x(k-2)	GlSolRad(k-2)	WindDir(k-0)	PM10 (k-2)	WindSpd(k-0)	AirTemp(k-1)	DifSolRad(k-1)	RelHum(k-0)
15	AirTemp(k-3)	DifSolRad(k-0)	AirTemp(k-3)	NO_2(k-1)	Precip(k-1)	RelHum (k-2)	O_3(k-3)	AirTemp(k-2)	NO_2(k-2)
16	NO_2(k-1)	NO_x(k-1)	DifSolRad(k-1)	NO_x(k-1)	DifSolRad(k-0)	NO_x(k-2)	NO_x(k-1)	NO_2(k-1)	GlSolRad(k-3)
17	NO_x(k-2)	PM10(k-1)	RelHum(k-0)	GlSolRad(k-1)	Precip(k-2)	NO_2(k-3)	NO_2 (k-2)	NO_x(k-2)	DifSolRad(k-2)
18	DifSolRad(k-1)	AirTemp(k-0)	GlSolRad(k-3)	GlSolRad(k-3)	O_3(k-2)	DifSolRad(k-0)	NO_x(k-0)	WindSpd(k-0)	AirTemp(k-3)
19	DifSolRad(k-2)	AirTemp(k-1)	WindSpd(k-1)	WindDir(k-1)	DifSolRad(k-2)	PM10(k-0)	PM10(k-0)	DifSolRad(k-2)	DifSolRad(k-3)
20	NO_x(k-3)	DifSolRad(k-3)	NO_2(k-1)	DifSolRad(k-2)	NO_x(k-2)	DifSolRad(k-3)	NO_2(k-3)	AirTemp(k-3)	RelHum(k-1)

[1] Guideline for developing an ozone forecasting program, Tech. Rep. EPA-454/R-99-009, United States Environmental Protection Agency (1999).

[2] S. M. Al-Alawi, S. A. Abdul-Wahab, C. S. Bakheit, Combining principal component regression and artificial neural-networks for more accurate predictions of ground-level ozone, Environmental Modelling & Software 23 (2008) 396–403.

[3] A. B. Chelani, Prediction of daily maximum ground ozone concentration using support vector machine, Environ Monit Assess 162 (2010) 169–176.

[4] C.-H. Cheng, S.-F. Huang, H.-J. Teoh, Predicting daily ozone concentration maxima using fuzzy time series based on a two-stage linguistic partition method, Computers and Mathematics with Applications (2011) 2016–2028.

[5] C. Duenas, M. Fernandez, S. Canete, J. Carretero, E. Liger, Stochastic model to forecast ground-level ozone concentration at urban and rural areas, Chemosphere 61 (2005) 1379–1389.

[6] Y. Feng, W. Zhang, D. Sun, L. Zhang, Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and SVM data classification, Atmospheric Environment 45 (2011) 1979–1985.

[7] T. Fontes, L. Silva, M. Silva, N. Barros, A. Carvalho, Can artificial neural networks be used to predict the origin of ozone episodes?, Science of the Total Environment 488–489 (2014) 197–207.

[8] G. G. Garner, A. M. Thompson, Ensemble statistical post-processing of the national air quality forecast capability: Enhancing ozone forecasts in Baltimore, Maryland, Atmospheric Environment 81 (2013) 517–522.

[9] Y. Lin, W. G. Cobourn, Fuzzy system models combined with nonlinear regression for daily ground-level ozone predictions, Atms. Environment 41 (2007) 3502–3513.

[10] K. P. Moustris, P. T. Nastos, I. K. Larissi, A. G. Paliatsos, Application of multiple linear regression models and artificial neural networks on the surface ozone forecast in the greater Athens area, Greece, Advances in Meteorology 2012 (2012) 1–8.

[11] A. Nebot, V. Mugica, A. Escobet, Ozone prediction based on meteorological variables: a fuzzy inductive reasoning approach, Atmospheric Chemistry and Physics Discussions 8 (2008) 1234312370.

[12] D. Petelin, A. Grancharova, J. Kocijan, Evolving Gaussian process models for the prediction of ozone concentration in the air, Simulation Modelling Practice and Theory 33 (1) (2013) 68–80.

[13] T. A. Solaiman, P. Coulibaly, P. Kanaroglou, Ground-level ozone forecasting using data-driven methods, Air Quality, Atmosphere & Health 1 (2008) 179–193.

[14] D. Sundaramoorthi, A data-integrated simulation model to forecast ground-level ozone concentration, Annals of Operations Research 216 (2014) 53–69.

[15] R. A. Hites, Elements of environmental chemistry, Wiley-Intersc., Hoboken, 2007.

[16] A. Bytnerowicz, K. Omasa, E. Paoletti, Integrated effects of air pollution and climate change on forests: a northern hemisphere perspective, Environmental Pollution 147 (3) (2006) 438–445.

[17] Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe.
URL http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:152:0001:0044:EN:PDF

[18] R. May, G. Dandy, H. Maier, Artificial Neural Networks - Methodological Advances and Biomedical Applications, InTech, Rijeka, 2011, Ch. Review of Input Variable Selection Methods for Artificial Neural Networks, pp. 19–44.

[19] A. L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artificial Intelligence 97 (1997) 245–271.

[20] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[21] R. Kohavi, G. H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1–2) (1997) 273–324.

[22] I. Lind, L. Ljung, Regressor selection with the analysis of variance method, Automatica 41 (2005) 693–700.

[23] M. Glavan, D. Gradišar, M. Atanasijević-Kunc, S. Strmčnik, G. Mušič, Input variable selection for model-based production control and optimisation, The Int. Journal of Advanced Manufacturing Technology 68 (9-12) (2013) 2743–2759. doi:10.1007/s00170-013-4840-1.

[24] G. J. Szekely, M. L. Rizzo, N. K. Bakirov, Measuring and testing dependence by correlation of distances, The Annals of Statistics 35 (6).

[25] K. Torkkola, Feature extraction by non-parametric mutual information maximization, Journal of Machine Learning Research 3 (2003) 1415–1438.

[26] S. Frenzel, B. Pompe, Partial mutual information for coupling analysis of multivariate time series, Physical Review Letters 99 (2007) 204101–1–204101–4.

[27] K. Li, P. J.-X, Neural input selection-a fast model-based approach, Neurocomputation 70 (1) (2007) 762–769.

[28] H. Niska, M. Heikkinen, M. Kolehmainen, Inteligent Data Engineering and Automated Learning, Vol. 4224 of Lecture Notes in Computer Science, Springer, 2006, Ch. Genetic algorithms and sensitivity analysis applied to select inputs of a multi-layer perceptron for the prediction of air pollutant time-series, pp. 224–231.

[29] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, MIT Press, Cambridge, MA, 2006.

[30] R. M. van Aalst, F. A. A. M. de Leeuw, National ozone forecasting system and international data exchange in northwest Europe, European topic centre on air quality, Tech. Rep. 9, European Environment Agency (1998).

[31] P. Mlakar, Determination of features for air pollution forecasting models, in: H. Adeli (Ed.), Proceedings of Intelligent Information Systems IIS97, Grand Bahama Island, Bahamas, IEEE Computer Society, Los Alamitos, 1997, pp. 350–354.

[32] K. Ažman, J. Kocijan, Application of Gaussian processes for black-box modelling of biosystems, ISA Transactions 46 (2007) 443–457.

[33] J. Q. Shi, T. Choi, Gaussian process regression analysis for functional data, Chapman and Hall/CRC, Taylor & Francis group, Boca Raton, FL, 2011.

[34] D. J. C. MacKay, Introduction to Gaussian processes, in: C. M. Bishop (Ed.), Neural Networks and Machine Learning, NATO ASI Series, Kluwer, 1998, pp. 133–166.

[35] J. Kocijan, B. Likar, Gas-liquid separator modelling and simulation with Gaussian-process models., Simulation Modelling Practice and Theory 16 (8) (2008) 910–922.