Simulation Modelling Practice and Theory 54 (2015) 101–115

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

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

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

ARTICLE INFO

Article history:
Received 19 November 2014
Received in revised form 17 February 2015
Accepted 14 March 2015
Available online 1 April 2015

Keywords:
Regressor selection
Regression modelling
Black-box modelling
Prediction of ozone concentration
Dynamical systems

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 h, 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.

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

☆ 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.

* Corresponding author at: Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia.

http://dx.doi.org/10.1016/j.simpat.2015.03.004
1569-190X/

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 $\mu g/m^3$ eight hours mean concentration; 'informing the public level', 180 $\mu g/m^3$ one hour mean concentration; and 'warning the public level', 240 $\mu g/m^3$ 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 h, 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 [19–21,18]: 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