Simulation Modelling Practice and Theory 33 (2013) 68–80

Contents lists available at SciVerse ScienceDirect

Simulation Modelling Practice and Theory

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

Evolving Gaussian process models for prediction of ozone concentration in the air

Dejan Petelin [a], Alexandra Grancharova [c,*], Juš Kocijan [a,b]

[a] Department of Systems and Control, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia
[b] University of Nova Gorica, Centre for Systems and Information Technologies, Vipavska 13, SI-5000 Nova Gorica, Slovenia
[c] Institute of System Engineering and Robotics, Bulgarian Academy of Sciences, Acad G. Bonchev Str., Bl.2, P.O. Box 79, Sofia 1113, Bulgaria

ARTICLE INFO

Article history:
Available online 25 May 2012

Keywords:
Ozone concentration prediction
Dynamic systems modelling
Evolving Gaussian process model

ABSTRACT

Ozone is one of the main air pollutants with harmful influence to human health. Therefore, predicting the ozone concentration and informing the population when the air-quality standards are not being met is an important task. In this paper, various first- and high-order Gaussian process models for prediction of the ozone concentration in the air of Bourgas, Bulgaria are identified off-line based on the hourly measurements of the concentrations of ozone, sulphur dioxide, nitrogen dioxide, phenol and benzene in the air and the meteorological parameters, collected at the automatic measurement stations in Bourgas. Further, as an alternative approach an on-line updating (evolving) Gaussian process model is proposed and evaluated. Such an approach is needed when the training data is not available through the whole period of interest and consequently not all characteristics of the period can be trained or when the environment, that is to be modelled, is constantly changing.

1. Introduction

Ozone (O_3), a form of oxygen, is a highly unstable and poisonous gas that can form and react under the action of light and that is present in two layers of the atmosphere. The ozone is a very specific air substance, which is present in the whole Earth's atmosphere – from the ground level to the top of the atmosphere. The stratospheric ozone prevents the harmful solar ultraviolet radiation to reach the Earth's surface. However, in the tropospheric layer, which is at ground level, the ozone is an air pollutant, which damages human health and the ecosystem equilibrium. Exposure to ozone can cause serious health problems in plants and people, thus ozone pollution is a major problem in some regions of the world. It tends 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 are the reasons why the atmospheric ozone dynamics is an object of intensive research.

The most direct way to obtain accurate air quality information is from measurements made at surface monitoring stations across countries. Fixed measurements of hourly ozone concentrations in compliance with the European Directive on ambient air quality and cleaner air for Europe [1] give continuous information about the evolution of surface ozone pollution at a large number of sites across Europe. In several Member States, they are more and more supplemented by numerical model outputs delivered at a regional or local scale, in keeping with the European Directive. The European standards that guarantee human-health protection are as follows: *health protection level*, 120 μg/m^3 8 h mean concentration; *informing the public level*,

* Corresponding author.
E-mail address: alexandra.grancharova@abv.bg (A. Grancharova).

1569-190X/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.simpat.2012.04.005

180 μg/m^3 1 h mean concentration; and *warning the public level*, 240 μg/m^3 1 h mean concentration. Therefore, predicting the ozone concentration and informing the population when the air-quality standards are not being met are important tasks.

Ozone concentration has a pronounced daily cycle [22], which can be modelled and forecasted using a variety of methods, and methods that describe the non-linear dynamics from available data are particularly useful. Thus, there exists a number of methods for ozone concentration prediction based on various modelling techniques, e.g. based on neural network NARX models [2,15,31], polynomial NARX models [26], fuzzy systems [20,21], support vector machines [7], ARIMA stochastic models [9], Gaussian processes (GP) [15,14,13]. There are also methods which are based on a combination of some of the mentioned techniques, e.g. the approach in [10] combines the use of neural networks, support vector machines and genetic algorithms.

In this paper, we focus on the use of GP modelling techniques for development and comparison of various models for prediction of ozone concentration in the air. The GP model is a probabilistic, non-parametric model based on the principles of Bayesian probability. It differs from most of the other black-box identification approaches in that it does not try to approximate the modelled system by fitting the parameters of the selected basis functions, but rather by searching for relationships among the measured data. The output of the GP model is a normal distribution, expressed in terms of the mean and the variance. The mean value represents the most likely output and the variance can be interpreted as a measure of its confidence. The obtained variance, which depends on the amount and the quality of the available identification data, is important information when it comes to distinguishing the GP models from other computational intelligence methods. Because of their properties GP models are especially suitable for modelling of uncertain processes or when modelling data are unreliable, noisy or missing. GP models fit well for modelling of environmental systems as well as for ozone pollution modelling. Thus, GP models have been developed for prediction of ozone concentration in the air of Bourgas, which is among the regions in Bulgaria with the highest levels of ozone pollution in the air. In [14] *first-order* GP models based on measurements of the air-pollutant concentrations are identified and verified for *one-step-ahead* predictions of the ozone concentration in the air of Bourgas. Furthermore, in [13] *high-order* GP models by using measurements of both the air pollutants and the meteorological parameters are identified and verified. In both cases GP models are trained *off-line* using only a subset of the available data due to the high computational burden of modelling GP models. However, this limitation and, consequently, the quality of GP models can be improved with *on-line* updating using the most recent measurements.

A noticeable drawback of system identification with GP models is the computation time necessary for the modelling. Regression based on GP models involves several matrix computations in which the computational complexity increases with the third power of the number of input data, such as matrix inversion and the calculation of the log-determinant of the used covariance matrix. This computational greed restricts the amount of training data, to at most a few thousand cases. To overcome the computational-limitation issues and to also make use of the method for large-scale dataset applications, numerous authors have suggested various sparse approximations [27,28]. All sparse approximate methods try to retain the bulk of the information contained in the full training dataset, but reduce the size of the covariance matrix to facilitate a less computationally demanding implementation of the GP model. The special kind of sparse approximate method is *on-line* modelling method Sparse On-line Gaussian Processes (OGP) [8] which tries to incorporate all information of the data by projecting to the reduced covariance matrix.

The OGP method was already implemented for modelling the ozone concentration in the air [25]. As the weather and its characteristics are constantly changing, the model should be updated and adjusted as well. That means it should not only update the model with information contained in streaming data, but should concurrently optimize hyperparameter values as well. As we experienced the OGP method has problems with numerical instability, therefore we propose, by our opinion, more robust method for *on-line* updating (*evolving*) of a GP model and compare its performance with an *off-line* trained GP models. The proposed method is based on the concept described in [24], but it is implemented differently as the method used in the experimental part of the paper.

The paper is structured as follows. In Section 2, the use and properties of Gaussian processes for modelling are reviewed. In Section 3, first- and high-order GP models for prediction of ozone concentration in the air of Bourgas, Bulgaria are identified *off-line*. A method for prediction of ozone concentration based on an *on-line* updated (*evolving*) GP model is proposed and evaluated in Section 4. The concluding remarks end the paper.

2. Modelling of dynamic systems with Gaussian processes

A GP model is a flexible, probabilistic, non-parametric model with uncertainty predictions. Its uses and properties for modelling are reviewed in [29]. The use of Gaussian processes for modelling dynamic systems is a relatively recent development [6,12,18]. A retrospective review can be found in [17].

A Gaussian process is a collection of random variables which have a joint multivariate Gaussian distribution (Fig. 1). Assuming a relationship of the form $y = f(\mathbf{x})$ between input $\mathbf{x}$ and output y, we have $y_1, \ldots, y_N \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\Sigma_{pq} = \mathrm{Cov}(y_p, y_q) = C(\mathbf{x}_p, \mathbf{x}_q)$ gives the covariance between output points corresponding to input points $\mathbf{x}_p$ and $\mathbf{x}_q$. Thus, the mean $\mu(\mathbf{x})$ and the covariance function $C(\mathbf{x}_p, \mathbf{x}_q)$ fully specify the Gaussian process.

The value of covariance function $C(\mathbf{x}_p, \mathbf{x}_q)$ expresses the correlation between the individual outputs $f(\mathbf{x}_p)$ and $f(\mathbf{x}_q)$ with respect to inputs $\mathbf{x}_p$ and $\mathbf{x}_q$. Note that the covariance function $C(\cdot,\cdot)$ can be any function that generates a positive semi-definite covariance matrix. It is usually composed of two parts,

$$C(\mathbf{x}_p, \mathbf{x}_q) = C_{\mathrm{f}}(\mathbf{x}_p, \mathbf{x}_q) + C_{\mathrm{n}}(\mathbf{x}_p, \mathbf{x}_q), \tag{1}$$