Доклади на Българската академия на науките
Comptes rendus de l'Académie bulgare des Sciences

Tome 64, No 1, 2011

SCIENCES ET INGENIERIE

Automatique et informatique

ON-LINE GAUSSIAN PROCESS MODEL FOR THE PREDICTION OF THE OZONE CONCENTRATION IN THE AIR

Dejan Petelin*, Juš Kocijan*,, Alexandra Grancharova*****

(*Submitted by Corresponding Member Ch. Roumenin on July 7, 2010*)

Abstract

Ozone is one of the main air pollutants with harmful influence on 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 a method for prediction of the ozone concentration based on an on-line updated dynamic model obtained from measurement data is proposed and evaluated as a first- and third-order model. For this purpose hourly measurements of the concentrations of ozone and nitrogen dioxide in the air of the town of Burgas, Bulgaria are used.

Key words: ozone concentration prediction, dynamic systems modelling, on-line gaussian process model

1. Introduction Ozone is one of the main air pollutants with a harmful influence on human health. The European standards that guarantee human-health protection are as follows [1]: *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. This paper describes a method for the prediction of the ozone concentration based on an on-line updated dynamic model obtained from measurement data.

This work was financed by the Slovenian Research Agency, Grants Nos P2-0001 and J2-2099 and contract No BI-BG/09-10-005, and by the National Science Fund of the Ministry of Education, Youth and Science of Republic of Bulgaria, Contract No DO02-94/14.12.2008.

Ozone concentration has a pronounced daily cycle [10], which can be modelled and forecasted using a variety of methods, and methods that describe the nonlinear dynamics from available data are particularly useful. In [6] neural networks models and Gaussian process models for ozone-concentration forecasting in some regions of Slovenia were developed and evaluated.

The town of Burgas is among the regions with the highest levels of ozone pollution in the air in Bulgaria, which makes it important to obtain a prediction model for this region. In [5] Gaussian process 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 Burgas. These models are learned *off-line* using only a subset of the available data due to the high computational burden of modelling Gaussian process models. However, this limitation and, consequently, the quality of Gaussian process models can be improved with on-line updating using the most recent measurements.

The Gaussian process model is a probabilistic, non-parametric, black-box model. 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 Gaussian process 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 Gaussian process models from other methods.

The purpose of this paper is to propose a solution to the problem of ozone prediction that is operational throughout the year and is based on on-line learning of the model used for the prediction. The proposed solution is based on an *on-line* Gaussian process model that can be utilised for *few-hour-ahead* predictions of the ozone concentration. For this purpose the sparse, on-line, Gaussian processes learning method [3] is taken and modified in such a way that it also makes on-line predictions.

2. Modelling dynamic systems with Gaussian processes. ***2.1. Theoretical basis.*** A Gaussian process model is a flexible, probabilistic, non-parametric model with uncertainty predictions. Its uses and properties for modelling are reviewed in [14]. The use of Gaussian processes for modelling dynamic systems is a relatively recent development [2, 4, 8]. A retrospective review can be found in [7].

A Gaussian process is a collection of random variables that have a joint multivariate Gaussian distribution. The mean $\mu(\mathbf{x})$ and the covariance function $C(\mathbf{x}_p, \mathbf{x}_q)$ fully specify the Gaussian process. Note that the covariance function $C(\cdot, \cdot)$ can be any function that has the property of generating a positive semi-definite covariance matrix.

The covariance function $C(\mathbf{x}_p, \mathbf{x}_q)$ can be interpreted as a measure of the distance between the input points $\mathbf{x}_p$ and $\mathbf{x}_q$. For systems modelling it is usually composed of two main parts, representing the functional part and the noise part.

A common choice is

$$C_f(\mathbf{x}_p, \mathbf{x}_q) = v_1 \exp\left[-\frac{1}{2}\sum_{d=1}^{D} w_d (x_{dp} - x_{dq})^2\right] + \delta_{pq} v_0, \tag{1}$$

where $\mathbf{\Theta} = [w_1 \dots w_D \ v_0 \ v_1]^T$ are the 'hyperparameters' of the covariance functions, D is the input dimension and $\delta_{pq} = 1$ if $p = q$ and 0 otherwise. Other possible covariance functions are given in [9,14].

Consider a set of N D-dimensional input vectors $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$ and a vector of output data $\mathbf{y} = [y_1, y_2, \dots, y_N]^T$. Based on the data $(\mathbf{X}, \mathbf{y})$, and given a new input vector $\mathbf{x}^*$, we wish to find the predictive distribution of the corresponding output y^*. Unlike other models, there is no model-parameter determination as such, within a fixed model structure. With this model, most of the effort involves *tuning* the parameters of the covariance function. This is done by maximising the log marginal likelihood ($\log(p(\mathbf{y}|\mathbf{X})) = -\frac{1}{2}\log(|\ \mathbf{K}\ |) - \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} - \frac{N}{2}\log(2\pi)$), where $\mathbf{K}$ is the $N \times N$ training covariance matrix. The number of parameters to be optimized is small ($D+2$, see equation (1)), which means that the optimization convergence might be faster and that the 'curse of dimensionality' so common to black-box identification methods is circumvented or at least decreased.

The described approach can be easily utilized for regression calculations. Based on the training set $\mathbf{X}$ a covariance matrix $\mathbf{K}$ of size $N \times N$ is determined. As already mentioned, the aim is to find the distribution of the corresponding output y^* at some new input vector $\mathbf{x}^* = [x_1(N+1), x_2(N+1), \dots, x_D(N+1)]^T$.

For a new test input $\mathbf{x}^*$, the predictive distribution of the corresponding output is $y^*|(\mathbf{X}, \mathbf{y}), \mathbf{x}^*$ and this is Gaussian, with a mean and a variance

$$\mu(\mathbf{x}^*) = \mathbf{k}(\mathbf{x}^*)^T\ \mathbf{K}^{-1}\ \mathbf{y}, \tag{2}$$

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*) - \mathbf{k}(\mathbf{x}^*)^T\ \mathbf{K}^{-1}\ \mathbf{k}(\mathbf{x}^*), \tag{3}$$

where $\mathbf{k}(\mathbf{x}^*) = [C(\mathbf{x}^1, \mathbf{x}^*), \dots, C(\mathbf{x}^N, \mathbf{x}^*)]^T$ is the $N \times 1$ vector of covariances between the test and training cases, and $k(\mathbf{x}^*) = C(\mathbf{x}^*, \mathbf{x}^*)$ is the covariance between the test input and itself.

Gaussian processes can, like other machine learning methods, e.g. neural networks, be used to model static nonlinearities and can therefore be used for the modelling of dynamic systems [8] if the delayed input and output signals are fed back and used as regressors. In such cases an autoregressive model is considered, such that the current output depends on the previous outputs, as well as on the

previous control inputs.

$$\begin{aligned} \mathbf{x}(k) &= [y(k-1), y(k-2), \ldots, y(k-L), \\ &\quad u(k-1), u(k-2), \ldots, u(k-L)]^T, \\ y(k) &= f(\mathbf{x}(k)) + \epsilon, \end{aligned} \tag{4}$$

where k denotes the consecutive number of the data sample. Let $\mathbf{x}$ denote the state vector composed of the previous outputs y and inputs u up to a given lag L, and ϵ is white noise.

As it can be seen from the presented relations, the obtained model not only describes the dynamic characteristics of the nonlinear system, but also provides information about the confidence in these predictions by means of the prediction variance. The Gaussian process can highlight areas of the input space where the prediction quality is poor, due to lack of data, by indicating a higher variance around the predicted mean.

2.2. On-line modelling. A noticeable drawback of system identification with Gaussian process models is the computation time necessary for the modelling. Gaussian process regression involves several matrix computations in which the load 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 also to make use of the method for large-scale dataset applications, numerous authors have suggested various sparse approximations [12, 13] as well as on-line modelling [3], which is a special kind of sparse approximate method. A common property to all sparse approximate methods is that they try to retain the bulk of the information contained in the full training dataset, but reduce the size of the resultant covariance matrix so as to facilitate a less computationally demanding implementation of the GP model.

The selected on-line learning method [3], suited to our problem, is based on a combination of a Bayesian on-line approach [11] and a sequential construction of a relevant subsample of the data on which an approximation of the GP model is based. This approximation is obtained by using parametrisation and projection techniques. To keep the subset of the most relevant data a fixed size there are two types of update to the GP model: a *basic* update that is performed when the error of a new approximation is smaller than a defined threshold, and a *full* update, which is performed otherwise. While a basic update only updates parameters that present the approximation, without increasing their number, a full update, besides updating parameters, also adds current data to the subset of most relevant data. If this operation results in the maximum size of the subset being exceeded, the least relevant data is removed.

Originally, this method was only capable of sequential training, so it did not make predictions in an on-line fashion. Therefore we modified it in such a

way that it also makes k-step-ahead predictions based on the newly updated GP model with every new data entry. We believe that this modification improves this method to the level where it can be applied for ozone forecasting.

3. Case study. The used data are the same as those in [5]. The data includes hourly measurements of the concentrations of ozone, sulfur dioxide, nitrogen dioxide, phenol and benzene for the year 2008 collected at the automatic measurement station in the centre of Burgas, Bulgaria. Since the situation changes depending on the period of the year, it is necessary to use the data acquired throughout the year for the development of the model for prediction.

It should be noted that for the training of the Gaussian process models the mean hourly concentrations of ozone are used.

Considering the analysis of the regressors for this data from [5], the following model structure is used

$$c_{\mathrm{O}_3}(t+1) = f(c_{\mathrm{O}_3}(t), c_{\mathrm{NO}_2}(t)), \tag{5}$$

where c_{O_3} is the concentration of ozone in the air, c_{NO_2} is the concentration of nitrogen dioxide in the air and t are the hours of the day. A prediction of the ozone concentration for a given hour based on this model depends on the values of the ozone concentration and the nitrogen dioxide concentration only for the previous hour.

A third-order model is also used for comparison

$$\begin{aligned} c_{\mathrm{O}_3}(t+1) \;=\; & f(c_{\mathrm{O}_3}(t), c_{\mathrm{NO}_2}(t), \\ & c_{\mathrm{O}_3}(t-1), c_{\mathrm{NO}_2}(t-1), \\ & c_{\mathrm{O}_3}(t-2), c_{\mathrm{NO}_2}(t-2)). \end{aligned} \tag{6}$$

In this case the prediction depends on the values of the ozone concentration and the nitrogen dioxide concentration for all three previous hours.

It was shown in [5] that a satisfactory level of prediction can be achieved with an off-line trained Gaussian process model. It should be noted, however, that the computational demand associated with computing the mean and the variance of a new prediction (equations (2), (3)) with a standard Gaussian process model would be high if the model contains data throughout the year. The solution to this problem is to utilise an on-line learning method that uses only a subset of the data with satisfactory information content and that can cope with changes in the conditions of the ozone-formation process. As it was already mentioned, the method [3] was modified so that predictions can be made concurrently, while the model is updating itself with incoming data.

The entire set of available data (the number of measurements is 6,105) is used for testing the on-line modelling and prediction. The predictions are validated with the standard mean relative square error (MRSE $= \sqrt{\frac{\sum_{i=1}^{N} e(i)^2}{\sum_{i=1}^{N} y(i)^2}}$,

where $y(i)$ and $e(i) = \hat{y}(i) - y(i)$ are the ozone concentration measurements and the prediction error in the i-th prediction) and the logarithm of the predictive density error (LPD $= \frac{1}{2}\log(2\pi) + \frac{1}{2N}\sum_{i=1}^{N}\left(\log(\sigma^2(i)) + \frac{e(i)^2}{\sigma^2(i)}\right)$, where $\sigma^2(i)$ is the prediction variance in the i-th step), which is the measure more suitable for validating the Bayesian model predictions. The obtained validation measures are given on Table 1.

T a b l e 1

Values of the mean relative square error (MRSE) and the log predictive density error (LPD) for the first-index 1 – and the third-index 3-order model for one-, two-, three- and four-steps-ahead predictions

Steps ahead	$MRSE_1$	$MRSE_3$	LPD_1	LPD_3
1	**0.0480**	0.0652	**5.2140**	5.7838
2	0.0913	0.1299	17.2909	18.3163
3	0.1322	0.1848	35.2957	38.6553
4	0.1712	0.2541	56.8864	61.0204

Predictions for up to four-steps ahead are made at every time sample. The first-order model is compared with the third-order model to evaluate the quality of the model predictions against the higher-order model. Note that all the predictions were taken into account including the predictions from the beginning of the sequence when the on-line learned model contains a small amount of input data. The number of data contained in the on-line updated Gaussian process model never exceeds the 50 data points that are the carriers of information about the dynamics of ozone concentration.

It is clear from Table 1 that the first-order model predictions, depicted in bold style, are more accurate than those of the third-order model. The values of the validation measures are increasing with the increasing steps of the prediction, which is to be expected, but the first-order model provides consistently better predictions than the third-order model. The predictions of the first-order model for one-hour ahead for four days from different seasons of the year 2008 are depicted in Fig. 1.

It is clear that the accuracy of the predictions for the time interval of interest – between 9 a.m. and 4 p.m. – when the ozone concentration is the highest, is good enough for practical forecasting of when the ozone safety limits are being exceeded.

4. Conclusions. A modification of the on-line learning and prediction of GP models is proposed in this paper for on-line, few-hours-ahead predictions of

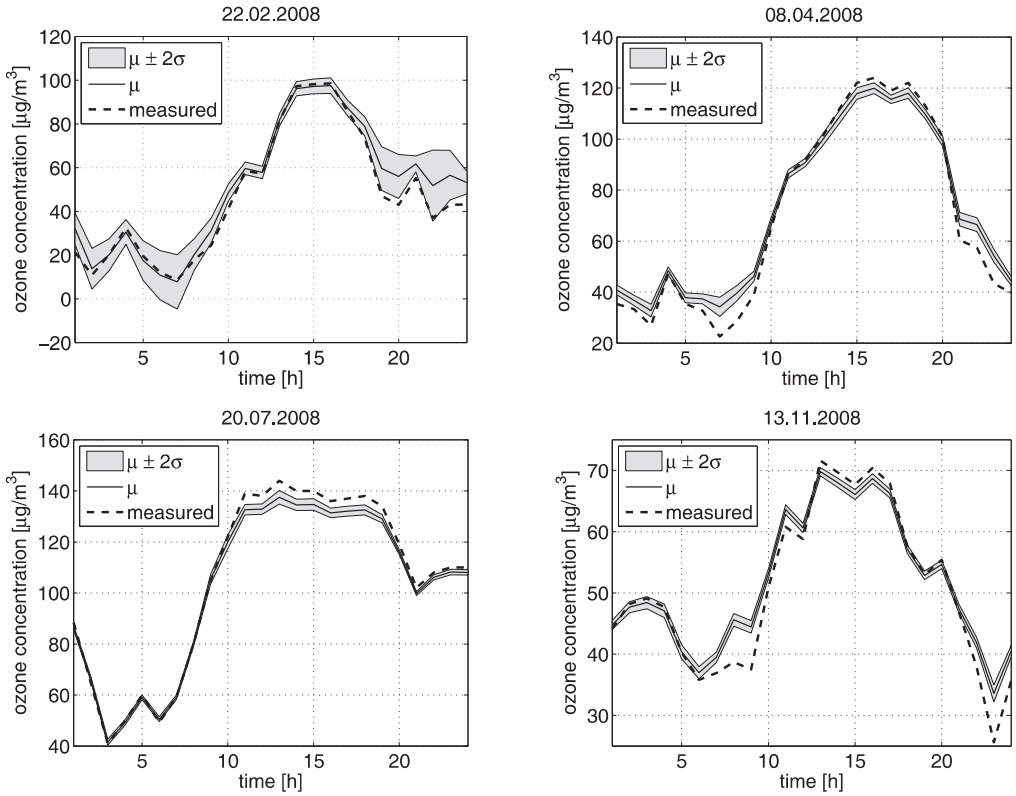

Fig. 1. One-step-ahead predictions for four days in various seasons of the year

the ozone concentration in the air of the town of Burgas, Bulgaria. The obtained results in the case study show that the predictions based on the first-order model are better than those based on the third-order model. Also, the predictions based on the first-order model for the critical time interval are sufficiently accurate.

Future work will be directed towards an improvement of the on-line modelling method for ozone-forecasting applications in practice and the utilisation of the method for the forecasting of other compounds in the air.

REFERENCES

[1] Directive 2002/3/EC of the European Parliament and of the Council of 12 February 2002 relating to ozone in ambient air.

[2] Ažman K., J. Kocijan. ISA Transactions, **46**, 2007, No 4, 443–457.

[3] Csató L., M. Opper. Neural Comput., **14**, 2002, No 3, 641–668.

[4] Grancharova A., J. Kocijan, T. A. Johansen. Automatica, **44**, 2008, No 6, 1621–1631.

[5] Grancharova A., D. Nedialkov, J. Kocijan, H. Hristova, A. Krastev. Proceedings of the International Conference on Automatics and Informatics, 2009, IV-17–IV-20.
[6] Grašič B., P. Mlakar, M. Božnar. Nuovo cimento Soc. ital. fis., C Geophys. space phys., **29**, 2006, No 6, 651–661.
[7] Kocijan J. Proceedings of the 9th International PhD Workshop on Systems and Control: young generation viewpoint, 2008.
[8] Kocijan J., A. Girard, B. Banko, R. Murray-Smith. Mathematical and Computer Modelling of Dynamic Systems, **11**, 2005, No 4, 411–424.
[9] Kocijan J., A. Grancharova. Compt. rend. Acad. bulg. Sci., **63**, 2010, No 4, 601–608.
[10] Nedialkov D., M. Angelova, G. Baldjiev, A. Krastev, H. Hristova. Proceedings of 19th International Symposium on Bioprocess Systems, Sofia, 2006.
[11] Opper M. On-line learning in neural networks, Cambridge University Press, 1998, 363–378.
[12] Quinonero-Candela J. J. Mach. Learn. Res., **6**, 2005, 1939–1959.
[13] Quinonero-Candela J., C. E. Rasmussen, C. K. I. Williams. Approximation Methods for Gaussian Process Regression, Microsoft Research, 2007.
[14] Rasmussen C. E., C. K. I. Williams. Gaussian Processes for Machine Learning, Cambridge, MA, MIT Press, 2006.

*Jozef Stefan Institute
39, Jamova
SI-1000 Ljubljana, Slovenia
e-mail: dejan.petelin@ijs.si

**University of Nova Gorica
13, Vipavska
SI-5000 Nova Gorica, Slovenia

***Institute of System Engineering and Robotics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 2, P.O.Box 79
1113 Sofia, Bulgaria