International Conference
AUTOMATICS AND INFORMATICS'09
29.09 – 4.10. 2009, Sofia, Bulgaria

APPLICATION OF GAUSSIAN PROCESSES TO THE PREDICTION OF OZONE CONCENTRATION IN THE AIR OF BOURGAS[1]

A. Grancharova[1], D. Nedialkov[1], J. Kocijan[2, 3], H. Hristova[1], A. Krastev[1]

[1] Institute of Control and System Research, Bulgarian Academy of Sciences, Acad G. Bonchev str., Bl.2, P.O.Box 79, Sofia 1113, Bulgaria, e-mail: alexandra.grancharova@abv.bg, dnnedialkov@abv.bg, tinardie@abv.bg, aikrastev@yahoo.com

[2] Department of Systems and Control, Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia, e-mail: jus.kocijan@ijs.si

[3] University of Nova Gorica, School of Engineering and Management, Vipavska 13, 5000 Nova Gorica, Slovenia

Abstract: Ozone is one of the main air polluters with harmful influence over human health. Therefore, predicting the ozone concentration and informing the population when the air quality standards have been exceeded is an important task. In this paper, different Gaussian process models for 1-hour ahead prediction of ozone concentration in the air of Bourgas, Bulgaria are identified and verified. For this purpose, the hourly measurements of the concentrations of ozone, SO_2, NO_2, phenol and benzene in the air, collected at the automatic measurement station in the center of Bourgas for year 2008, are used.

Key words: Ozone concentration prediction, Gaussian process models

INTRODUCTION

Ozone is one of the main air polluters with harmful influence over human health. Standards which guarantee the 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; *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 have been exceeded is an important task.

It has been shown in [2], that the ozone concentration has a strong daily cycle. Thus, the formation and collection of ozone in the air starts after 7 o'clock and it reaches its maximum between 13 and 16 o'clock. In [3], the relation between the ozone concentration and three meteorological parameters have been investigated using data about the region of Hessen in Germany. Based on these data, a linear regression model to predict the maximal daily ozone concentration in the air has been obtained.

However, until now no models to predict the ozone concentration in the air in any regions in Bulgaria have been developed. Furthermore, it may be expected that using Gaussian process models [4] would allow to obtain more accurate prediction models where the presence of different air polluters can be taken into account and more complex relations between their concentrations and the ozone concentration can be incorporated. The region of Bourgas city is among the regions with highest ozone pollution of the air and thus it is of primary interest to obtain a prediction model for this region. To achieve this, data provided by the Executive Environmental Agency of Bulgaria are used.

The following notation will be used in the paper. For a random variable y with Gaussian distribution, $\mathcal{N}(\mu(y),\sigma^2(y))$ denotes its probability distribution, and $\mu(y)$ and $\sigma^2(y)$ are respectively its mean and variance.

MODELLING OF DYNAMIC SYSTEMS WITH GAUSSIAN PROCESSES

The Gaussian process model is an example of a *non-parametric* probabilistic black-box model which, beside model predictions, inherently provides also the uncertainty of predictions. Its use and properties for modelling are reviewed in [4]. The use of Gaussian processes in the modelling of dynamic systems is a relatively recent development [5, 6, 7, 8, 9] and a retrospective review of dynamic systems modeling with Gaussian process models can be found in [10].

A Gaussian process is a collection of random variables which have a joint multivariate Gaussian distribution. Assuming a relationship of the form $y = f(z)$ between an input $z \in \mathbb{R}^D$ and output $y \in \mathbb{R}$, we have $y(1), y(2), \ldots, y(M) \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\Sigma_{pq} = \mathrm{Cov}(y(p), y(q)) = C(z(p), z(q))$ gives the covariance between the output points $y(p)$ and $y(q)$ corresponding to the input points $z(p)$ and $z(q)$. Thus, the mean $\mu(z)$ (usually assumed to be zero) and the covariance

[1] This work was financed by the National Science Fund of the Ministry of Education and Science of Republic of Bulgaria, contract №DO02-94/14.12.2008 and the Slovenian Research Agency, contract №BI-BG/09-10-005 ("Application of Gaussian processes to the modeling and control of complex stochastic systems")

function $C(z(p),z(q))$ fully specify the Gaussian process. Note that the covariance function $C(z(p),z(q))$ can be any function with the property that it generates a positive definite covariance matrix. A common choice is:

$$C(z(p),z(q)) = v_1 \exp\left[-\frac{1}{2}\sum_{i=1}^{D} w_i (z_i(p) - z_i(q))^2\right] + v_0 \alpha_{pq} \quad (1)$$

where $\Theta = [w_1, \ldots, w_D, v_0, v_1]$ are the 'hyperparameters' of the covariance function, z_i denotes the i-th component of the D-dimensional input vector z, and α_{pq} is the Kronecker operator. The covariance function (1) is composed of two parts: the Gaussian covariance function for the modeling of system function and the covariance function for the modelling of noise. The noise, in our case, is presumed to be white. Other forms of covariance functions suitable for different applications can be found in [11]. For a given problem, the hyperparameters are learned (identified) using the data at hand. After the learning, one can use the w parameters as indicators of 'how important' the corresponding input components (dimensions) are: if w_i is zero or near zero it means that the inputs in dimension i contain little information and could possibly be removed.

Consider a set of M D-dimensional input vectors $\mathbf{Z} = [z(1), z(2), \ldots, z(M)]^T$ and a vector of output data $Y = [y(1), y(2), \ldots, y(M)]^T$. Based on the data $(\mathbf{Z}, Y)$, and given a new input vector z^*, we wish to estimate the probability 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 consists in *tuning* the parameters of the covariance function. This is done by maximizing the log-likelihood of the parameters, which is computationally relatively demanding since the inverse of the data covariance matrix ($M \times M$) has to be calculated at every iteration.

The described approach can be easily utilized for regression calculation. Based on a training set $\mathbf{Z}$, a covariance matrix $\mathbf{K}$ of size $M \times M$ is determined. As already mentioned before, the aim is to estimate the probability distribution of the corresponding output y^* at some new input vector z^*. For a new test input z^*, the predictive distribution of the corresponding output is $y^* \mid z^*, (\mathbf{Z}, Y)$ and is Gaussian, with mean and variance:

$$\begin{aligned} \mu(z^*) &= k(z^*)^T \mathbf{K}^{-1} Y \\ \sigma^2(z^*) &= k_0(z^*) - k(z^*)^T \mathbf{K}^{-1} k(z^*) \end{aligned} \quad (2)$$

where $k(z^*) = [C(z(1), z^*), \ldots, C(z(M), z^*)]^T$ is the $M \times 1$ vector of covariances between the test and training cases and $k_0(z^*) = C(z^*, z^*)$ is the covariance between the test input and itself.

Gaussian processes can be used to model static nonlinearities and can therefore be used for modelling of dynamic systems if delayed input and output signals are used as regressors [8]. In such cases an autoregressive model is considered, such that the current predicted output depends on previous estimated outputs, as well as on previous control inputs:

$$\begin{aligned} z(t) &= [\hat{y}(t-1), \hat{y}(t-2), \ldots, \hat{y}(t-L), u(t-1), \\ &\qquad u(t-2), \ldots, u(t-L)]^T \\ \hat{y}(t) &= f(z(t)) + \eta(t) \end{aligned} \quad (3)$$

where t denotes consecutive number of data sample, L is a given lag, and $\eta(t)$ is the prediction error. The quality of the mean values of predictions with a Gaussian process model can be assessed by computing the average squared error (ASE):

$$ASE = \frac{1}{M}\sum_{i=1}^{M} [\mu(\hat{y}(i)) - y(i)]^2 \quad (4)$$

and the log density error (LD) [4] is also a possible measure:

$$LD = \frac{1}{2M}\sum_{i=1}^{M} \log(2\pi) + \log[\sigma^2(\hat{y}(i))] + \frac{[\mu(\hat{y}(i)) - y(i)]^2}{\sigma^2(\hat{y}(i))} \quad (5)$$

In (4), (5), $\mu(\hat{y}(i))$ and $\sigma^2(\hat{y}(i))$ are the prediction mean and variance, $y(i)$ is the system's output and M is the number of the training points.

The Gaussian process model now not only describes the dynamic characteristics of the non-linear system, but at the same time provides information about the confidence in the predictions. The Gaussian process can highlight areas of the input space where prediction quality is poor, due to the lack of data, by indicating the higher variance around the predicted mean.

GAUSSIAN PROCESS MODEL FOR PREDICTION OF OZONE CONCENTRATION IN THE AIR OF BOURGAS

Measurement data

Measurement data for the year 2008, collected at the automatic measurement station in the center of Bourgas, Bulgaria, are used. The data includes hourly measurements of the concentrations of ozone, SO_2, NO_2, phenol and benzene.

Gaussian process models of ozone concentration dynamics

Six Gaussian process models for prediction of ozone concentration are identified and verified based on the available measurement data. The average squared errors (ASE) corresponding to each model and computed both for the training data and for the validation data are given in Table 1.

The *training* data include the measurements from 9 till 16 o'clock at every 5-th day of the year 2008. The reason to consider this time interval is the following. In the previous research [2], it has been proved that the ozone concentration has a strong daily cycle. The formation and collection of ozone in the air starts after 7 o'clock and it reaches its maximum between 13 and 16 o'clock. After 16 o'clock, the ozone concentration decreases. Therefore, we are interested to obtain an accurate prediction of ozone concentration in the interval from 9 till 16 o'clock, where there is some risk to exceed the established air quality standards. Thus, the total number of the training data is 424 corresponding to 53 days of the year. Note that the days, for which there is not a full collection of are measurements in the interval 9 - 16 o'clock, are excluded from the data set.

The *validation* data include the measurements from 1 till 23 o'clock at all days of the year 2008. It should be noted that the days, for which there are measurements at some hours only, are excluded from the data set. Thus, the total number of the validation data is 5497 corresponding to 239 days of the year.

The identified Gaussian process models are the following:

Model A

The model has the form:

$$c_{O_3}(t+1) = f_A(c_{O_3}(t)) \quad (6)$$

where c_{O_3} is the concentration of ozone in the air and $t = 0,1,2,3, \ldots ,22$ are the hours of the day. For this model, the prediction of ozone concentration at a given hour of the day is based only on its value at the previous hour.

Model B

The model has the form:

$$c_{O_3}(t+1) = f_B(c_{O_3}(t), c_{NO_2}(t)) \quad (7)$$

where c_{NO_2} is the concentration of nitrogen dioxide in the air. It can be seen from Table 1, that the incorporation of c_{NO_2} as an input to the model increases its accuracy (the average squared error is decreased).

Model C

The model has the form:

$$c_{O_3}(t+1) = f_C(c_{O_3}(t), c_{NO_2}(t), c_{SO_2}(t)) \quad (8)$$

where c_{SO_2} is the concentration of sulphur dioxide in the air. The use of c_{SO_2} as an input parameter does not improve the model quality (see Table 1). Therefore, c_{SO_2} is excluded from the next models (models D and E).

Model D

The model has the form:

$$c_{O_3}(t+1) = f_D(c_{O_3}(t), c_{NO_2}(t), c_{C_6H_5OH}(t)) \quad (9)$$

where $c_{C_6H_5OH}$ is the concentration of phenol in the air. It can be seen from Table 1, that $c_{C_6H_5OH}$ does not contribute to the improvement of model accuracy and it is excluded from the next model (model E).

Model E

The model has the form:

$$c_{O_3}(t+1) = f_E(c_{O_3}(t), c_{NO_2}(t), c_{C_6H_6}(t)) \quad (10)$$

where $c_{C_6H_6}$ is the concentration of benzene in the air. This model has the largest average squared error for the validation data, i.e. it is least accurate among all models.

Model F

The model has the form:

$$c_{O_3}(t+1) = f_F(c_{O_3}(t), c_{NO_2}(t), c_{SO_2}(t), c_{C_6H_5OH}(t), c_{C_6H_6}(t)) \quad (11)$$

This model includes all measured parameters (ozone, SO_2, NO_2, phenol and benzene) and it is identified only for comparison purposes.

Table 1. Average squared errors for the different models.

MODEL	$ASE_{TRAINING}$	$ASE_{VALIDATION}$
Model A	122.59	177.56
Model B	115.33	**172.59**
Model C	113.26	176.31
Model D	115.33	172.59
Model E	109.47	184.75
Model F	117.93	179.66

From Table 1, it can be seen that model B provides the best quality of prediction, since it has the smallest average squared error for the validation data ($ASE_{VALIDATION}$ = 172.59). The accuracy of this model is further improved by enlarging the training data set. In Table 2, the average squared errors of four models of type B are given, which correspond to different number of the training data.

Table 2. Average squared errors for models of type B.

NUMBER OF TRAINING DATA	$ASE_{TRAINING}$	$ASE_{VALIDATION}$
424	115.33	172.59
456	95.13	185.16
664	99.60	**169.82**
960	93.59	176.67

From Table 2, it can be seen that the Gaussian process model B, whose identification is based on 664 training data has smallest average squared error for the validation data ($ASE_{VALIDATION}$ = 169.82). Its hyperparameters are the following:

$$\Theta = [w_1, w_2, v_0, v_1] = [165.96,\ 710.20,\ 10.02,\ 115.94] \quad (12)$$

In Figures 1 to 5, the mean value and 95% confidence interval of the ozone concentration predicted with this model are shown for some days of year 2008.

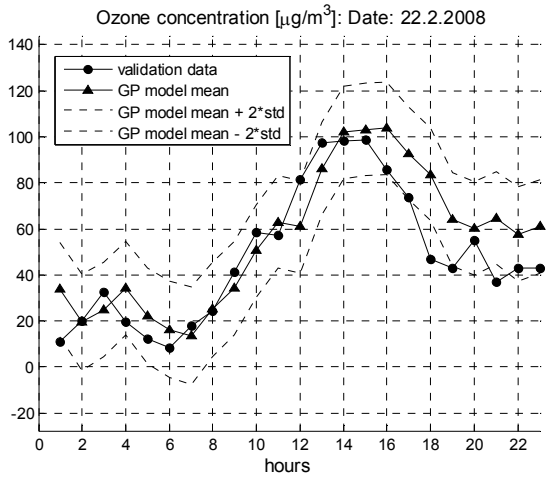

Fig. 1. The predicted mean value and 95% confidence interval of the ozone concentration for 22-nd February, 2008.

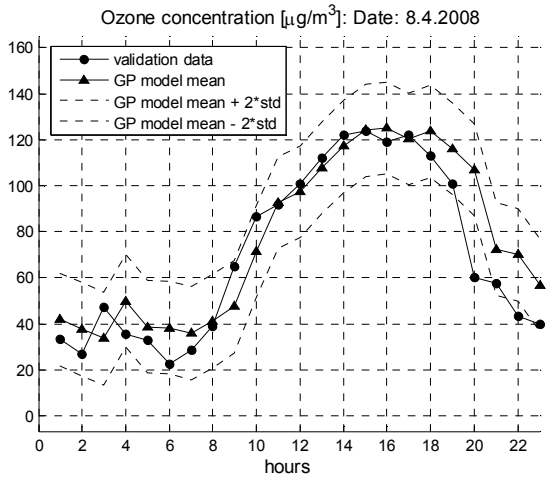

Fig. 2. The predicted mean value and 95% confidence interval of the ozone concentration for 8-th April, 2008.

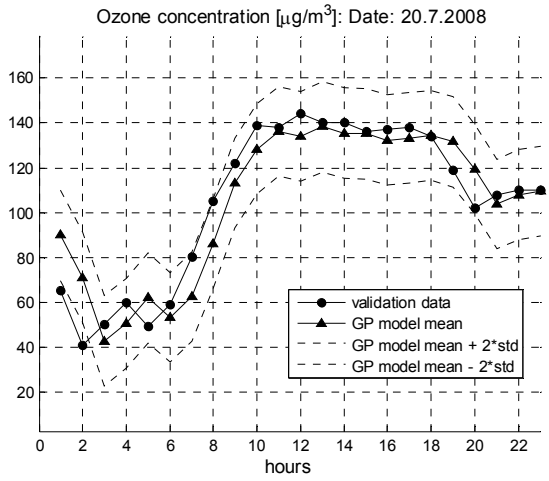

Fig. 3. The predicted mean value and 95% confidence interval of the ozone concentration for 20-th July, 2008.

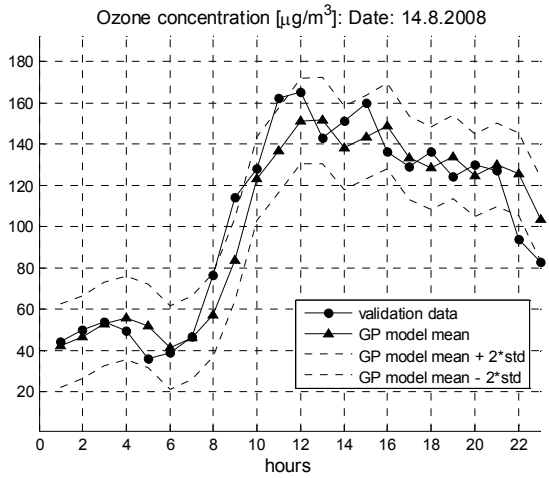

Fig. 4. The predicted mean value and 95% confidence interval of the ozone concentration for 14-th August, 2008.

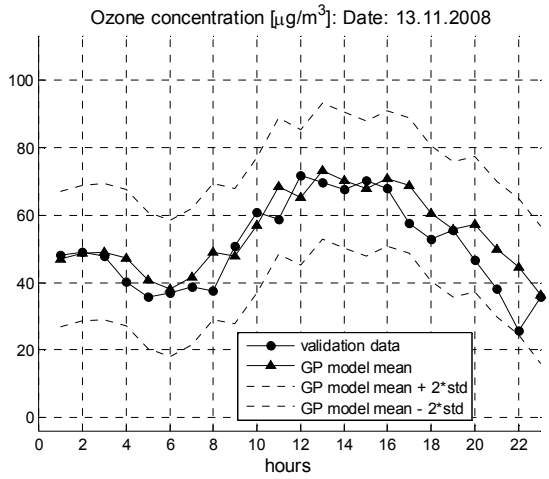

Fig. 5. The predicted mean value and 95% confidence interval of the ozone concentration for 13-th November, 2008.

It can be seen that the accuracy of prediction is highest for the time interval of the training data (from 9 till 16 o'clock), where the ozone concentration is maximal. The less accuracy of prediction outside this interval is acceptable.

CONCLUSIONS

In this paper, different Gaussian process models for 1-hour ahead prediction of ozone concentration in the air of Bourgas are identified. The verification analysis shows that the model with the best quality of prediction is the one which uses as input parameters only the concentrations of ozone and nitrogen dioxide in the air. The obtained results prove that the prediction is accurate enough and can be used for public warning in cases of high health risk.

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. Nedialkov D., M. Angelova, G. Baldjiev, A. Krastev, H. Hristova. Ozone concentracions in the air – standards and daily cycle. Proceedings of 19-th International Symposium on Bioprocess Systems, Sofia, 2006.
3. Nedialkov, D., M. Angelova, A. Krastev, H. Hristova. Prognostication of ozone concentration in the air. Proceedings of 20-th International Symposium on Bioprocess Systems, Sofia, November 6-7, 2007, pp. II.1-II.8.
4. Rasmussen, C. E., C. K. I. Williams. Gaussian processes for machine learning, MIT Press, Cambridge, MA, London, 2006.
5. Ažman K., J. Kocijan. Application of Gaussian processes for black-box modelling of biosystems. ISA Transactions, vol. 46, No. 4, pp. 443-457, 2007.
6. Girard, A., C. E. Rasmussen, J. Quinonero Candela, R. Murray-Smith. Gaussian process priors with uncertain inputs & application to multiple-step ahead time series forecasting. Proceedings of NIPS 15, Vancouver, Canada, MIT Press, 2003.
7. Grancharova, A., J. Kocijan, T. A. Johansen. Explicit stochastic predictive control of combustion plants based on Gaussian process models. Automatica, vol. 44, No. 6, pp. 1621-1631, 2008.
8. Kocijan, J., A. Girard, B. Banko, R. Murray-Smith. Dynamic systems identification with Gaussian processes. Mathematical and Computer Modelling of Dynamic Systems, vol. 11, No. 4, pp. 411-424, 2005.
9. Likar, B., J. Kocijan. Predictive control of a gas-liquid separation plant based on a Gaussian process model. Computers & Chemical Engineering, vol. 31, pp. 142-152, 2007.
10. Kocijan, J. Gaussian process models for systems identification. Proceedings of the 9-th International PhD Workshop on Systems and Control: young generation viewpoint, Izola, Simonov zaliv, 2008, 8 pages.
11. Rasmussen, C. E. Evaluation of Gaussian Processes and other Methods for Non-Linear Regression, Ph.D. Dissertation, Graduate Department of Computer Science, University of Toronto, Toronto, 1996.