Environment Protection Engineering

Vol. 32 2006 No. 4

SZYMON HOFFMAN*

MISSING DATA COMPLETING IN THE AIR MONITORING SYSTEMS BY FORWARD AND BACKWARD PROGNOSIS METHODS

The possibility of modelling the concentration of air pollutants measured in the air monitoring systems by means of time-series neural models was examined. The data set analysed was built of hourly averages, gathered at the air monitoring station in Zabrze (the south of Poland) in the four-year period. The analysis was carried out in order to compare the prediction accuracies of hourly concentrations in the missing data blocks depending on prognosis. The results of O_3 and NO hourly concentration modelling for different anticipations were discussed. The results show that the prediction quality could be improved when both prognostic methods, the forward one and the backward one, are used simultaneously, each one for another part of the missing data.

1. INTRODUCTION

The data sets gathered continuously in the air monitoring systems create natural time-series, but they are never complete. The loss of data can make the assessment of the air quality required by low standards impossible [1]. Air quality standards allow us to reproduce the missing data by modelling when the monitoring set is not sufficient. Time-series analysis is suggested to be one of the common methods for data gaps completing.

In conventional statistics, mathematical functions of two main features of the time series, trend and seasonality, have to be initially suggested. The choice of those functions often influences the modelling error. In the case of short-time prognosis, the methods based on conventional analysis of trend and seasonality could be useless. Artificial neural networks allow us to overcome the problem of the choice of the mathematical relationship within time series.

In many cases, the accuracy of the neural models of time series is satisfactory. Their generation runs fast, especially if they are linear. Model creation does not have to be preceded by introductory analysis [2]–[4]. It has been proved that time-series

* Department of Chemistry, Water and Wastewater Technology, Technology University of Częstochowa, Dąbrowskiego 69, 42-200 Częstochowa, Poland.

analysis may be useful as the method of completing the missing data [5]. However, its accuracy quickly worsens when a prognosis horizon is broadened. If the values of modelling accuracy are arranged in a chronological order, the latest values in the gaps are decidedly worst than the first values predicted.

In this paper, a specific solution of the problem mentioned was tested. A main concept consisted in using the time-series analysis backward, in the reverse direction. The aim of the research was to determine and compare modelling accuracies in both prognosis directions. The examination was based on artificial neural networks, which were employed to create all time-series models. The data set analysed was built of O_3 and NO hourly averages, gathered in a long-term period at the air monitoring station.

2. DATA DESCRIPTION AND COMPUTATION METHODS

Investigations were carried out based on the sets of hourly data collected in the 4-year period (1994–1997) at the air monitoring station in Zabrze (the south of Poland), situated in the centre of the city and being greatly affected by pollutants emitted from local sources. Two pollutants were chosen for analysis: O_3 and NO. Ozone is a secondary pollutant whose diurnal cycle shows regular changes of concentration. NO is a primary pollutant. Contrary to O_3, NO diurnal changes greatly depend on local pollution and are the least regular among other air pollutants.

The analysis was carried out with the program Statistica Neural Networks. The linear type of neural networks was used in time-series analysis to predict the concentrations of NO and O_3. The pseudo-inverse algorithm was employed as the method of training the models [6]. The set of data was divided into three different subsets: the training subset (50% of cases), the verification subset (25% of cases) and the test subset (25% of cases). Each model had one output – the concentration value of specified pollutant at specified time (hour), dependent on the anticipation assigned. The input concentration introduced into the model was recorded earlier. In all of the models, the number of inputs (steps) were the same, i.e., 24. Twenty-four steps supply the model with the necessary knowledge of diurnal concentration changes. Based on the author's previous experience it has been found that an increase in the number of steps does not improve the models.

The time-series analysis was carried out separately for two data sets: the chronological one and the inverse one. Both sets were exploited in different groups of models, the first one in the forward models, and the second one in the backward models. In each group, the models differed in prognosis horizons (anticipations).

The following criteria were assumed for the model accuracy assessment:

1. The value of the correlation coefficient (r) for the best linear fitting of the real concentration values to the model output.
2. The value of the root mean square error ($RMSE$).
3. The value of the mean absolute error ($|e|$).

4. The ratio *RMSE*/*s*, where *s* is the standard deviation.

The values of *RMSE* and |*e*| are useful criteria for assessing the models constructed for the same pollutant. To compare the models for different pollutants some dimensionless criteria should be used, e.g., correlation coefficient and the *RMSE*/*s* ratio.

3. RESULTS AND DISCUSSION

It is known that the quality of prediction depends on the prognosis horizon. In this analysis, some kinds of prediction error were assumed as the criteria of that quality. The values of *r*, *RMSE*, |*e*| and *RMSE*/*s* for the fixed number of steps (24) and different anticipation (prognosis horizon) were presented in tables 1 and 2 for O_3 and NO concentrations, respectively. The error values for both forward and backward prognoses were shown in the separate parts of the tables. Prediction accuracies in forward and backward directions worsen gradually in the same way, provided that anticipation improves for O_3 (table 1) and for NO (table 2) concentrations. The results obtained for the same anticipations are almost identical in both prognosis directions. This prognosis symmetry means that autocorrelation is equally strong in both directions of the time series. The results show how to optimize prediction when the set of missing data contains more cases. The first half of the missing cases should be predicted by the forward prognosis, whereas the cases in the second part of the set should be predicted by backward prognosis. The poorest prediction accuracy is expected in the middle of of the cases subjected to modelling. When the predicted gaps are wide, other alternative methods of modelling should be considered. A multiple regression seems to be the most promising alternative method [4].

Table 1

The estimation of modelling accuracy of hourly ozone concentration for various prognosis horizons, Zabrze, 1994–1997, 24 steps

Prognosis horizon (hours)	Forward prognosis				Backward prognosis			
	r	*RMSE* (μg/m^3)	|*e*| (μg/m^3)	*RMSE*/*s*	*r*	*RMSE* (μg/m^3)	|*e*| (μg/m^3)	*RMSE*/*s*
1	0.960	7.30	4.83	0.281	0.960	7.30	4.84	0.281
2	0.911	10.74	7.54	0.413	0.910	10.77	7.55	0.415
3	0.867	12.96	9.38	0.499	0.866	13.01	9.35	0.501
4	0.829	14.52	10.70	0.559	0.828	14.56	10.68	0.560
5	0.801	15.57	11.60	0.599	0.799	15.62	11.62	0.601
6	0.779	16.28	12.25	0.626	0.778	16.31	12.21	0.628
8	0.755	17.05	12.94	0.656	0.753	17.09	12.91	0.658
12	0.735	17.61	13.51	0.678	0.735	17.63	13.39	0.678
24	0.748	17.25	13.31	0.664	0.747	17.27	13.23	0.664
72	0.672	19.24	14.96	0.740	0.668	19.33	14.91	0.744
240	0.617	20.45	15.98	0.787	0.612	20.56	15.92	0.791

The models for NO concentration (table 2) are apparently less precise than their analogues for ozone concentration (table 1). The prediction accuracy of NO concentration is poor. The possibility of practical implementation of forward or backward time-series analysis is rather doubtful, especially when wider gaps should be completed. However, the prediction of NO concentration seems to be most difficult, also by other methods.

Table 2

The estimation of the modelling accuracy of hourly NO concentration for various prognosis horizons, Zabrze, 1994–1997, 24 steps

Prognosis horizon (hours)	Forward prognosis				Backward prognosis			
	r	*RMSE* (μg/m^3)	|*e*|, (μg/m^3)	*RMSE*/*s*	*r*	*RMSE* (μg/m^3)	|*e*| (μg/m^3)	*RMSE*/*s*
1	0.860	10.08	4.36	0.511	0.860	10.10	4.33	0.511
2	0.714	13.82	6.29	0.700	0.714	13.82	6.41	0.700
3	0.615	15.56	7.51	0.788	0.615	15.57	7.53	0.789
4	0.554	16.44	8.16	0.833	0.554	16.44	8.21	0.833
5	0.514	16.94	8.63	0.858	0.514	16.94	8.64	0.858
6	0.487	17.24	8.91	0.873	0.488	17.22	8.91	0.873
8	0.456	17.57	9.17	0.890	0.458	17.54	9.23	0.889
12	0.418	17.94	9.57	0.909	0.419	17.92	9.54	0.908
24	0.373	18.32	9.89	0.928	0.374	18.29	9.77	0.927
72	0.288	18.91	10.44	0.958	0.288	18.86	10.32	0.958
240	0.245	19.16	10.65	0.970	0.253	18.78	10.49	0.968

4. SUMMARY AND CONCLUSIONS

Time-series analysis by means of artificial neural networks appears to be a useful tool for modelling the surface concentration of air pollutants. This method can be employed in missing data completing in the air monitoring systems.

In the case of ozone concentration modelling, satisfactory results of prediction were obtained for different prognosis horizons. Ozone is a secondary pollutant whose concentration changes quite regularly in a diurnal cycle. That is why its concentration is more predictable than the concentrations of other pollutants and the time-series analysis can be recommended as a modelling method for missing data of ozone.

The results of modelling NO concentration by means of the time-series analysis are apparently less satisfactory. NO is a primary pollutant and its diurnal changes are the least regular among other air pollutants.

In this paper, the effect of the changes of anticipation, one of the time-series parameters, was analysed. The accuracies of different models were compared separately for O_3 and NO concentrations. The following conclusions were drawn:

1. Artificial neural networks adapted for the time-series analysis can be used for missing data completing in the air monitoring systems.

2. The accuracy of prognosis worsens when its horizon is broadened.

3. Short-time forecasting of ozone concentration gives satisfactory results, even for the long-term anticipation.

4. The modelling of the levels of primary pollutants, such as NO, gives worse results because their diurnal cycles are less regular. The possibilities of practical implementing the time-series analysis in NO prediction are limited, especially for the long-term anticipations.

5. The quality of prediction in the second half of any modelled set of data can be improved by running of time series analysis in reverse order, i.e., by backward prognosis application.

ACKNOWLEDGEMENTS

This work was carried out in the frame of the research project number 1 T09D 037 30 supplied by research budget of Polish Government for the years 2006–2008.

REFERENCES

[1] HAUCK H., KROMP-KOLB H., PETZ E., *Requirements for the completeness of ambient air quality data sets with respect to derived parameters*, Atmos. Environ., 1999, Vol. 33, 13, 2059–2066.

[2] GARDNER M.W., DORLING S.R., *Artificial neural networks (the multilayer perceptron) – a review of applications in the atmospheric sciences*, Atmos. Environ., 1998, Vol. 32, 14/15, 2627–2636.

[3] BALLESTER E.B. et al., *Effective 1-day ahead prediction of hourly surface ozone concentrations in eastern Spain using linear models and neural networks*, Ecological Modelling, 2002, 156, 27–41.

[4] HOFFMAN S., JASIŃSKI R., *Studies on NO_x concentration modeling in the air monitoring systems*, [in:] *Pathways of pollutants and mitigation strategies of their impact on the ecosystems*, edited by M.R. Dudzińska, M. Pawłowska, Monografie Komitetu Inżynierii Środowiska PAN, Lublin, 2004, 185–192.

[5] HOFFMAN S., *Zastosowanie sieci neuronowych w krótkoterminowym prognozowaniu stężeń zanieczyszczeń powietrza*, [in:] *Emisje, zagrożenie, ochrona powietrza*, edited by A. Musialik-Piotrowska and J.D. Rutkowski, PZITS, Wrocław, 2004, 115–122.

[6] Statistica Neural Networks, StatSoft 1998.

PROGNOZOWANIE WSTECZNE W UZUPEŁNIANIU BRAKUJĄCYCH DANYCH W SYSTEMACH MONITORINGU POWIETRZA

Przeanalizowano możliwość modelowania stężeń zanieczyszczeń, mierzonych na stacjach monitoringu powietrza, z wykorzystaniem neuronowych modeli szeregów czasowych. Przedstawiono rezultaty modelowania chwilowych (jednogodzinnych) stężeń zanieczyszczeń powietrza na podstawie wieloletnich danych pomiarowych zarejestrowanych na stacji monitoringu powietrza w Zabrzu. Celem analiz było określenie i porównanie dokładności predykcji chwilowych stężeń wybranych zanieczyszczeń powietrza w blokach brakujących danych, w zależności od kierunku prognozy. Stwierdzono, że jakość prognozy można poprawić, stosując w drugiej części modelowanego bloku danych wsteczną analizę szeregów czasowych.