FORECASTING THE URBAN AIR QUALITY USING VARIOUS STATISTICAL TECHNIQUES

Anurag Kandya* and Manju Mohan
Centre for Atmospheric Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India

Abstract

A statistical approach is about working through the historical data and finding guides to future behaviour. In the present study, five statistical techniques i.e. Single Exponential Smoothing (SES), Adaptive Response Rate Single Exponential Smoothing (ARRSES), Holt's Linear Method (HLM) ARX (Auto Regressive eXogenous) Model and Auto Regressive Integrated Moving Averages (ARIMA) are adopted for predicting the urban air quality over Delhi. Considering the uncertainty and unavailability of most of the inputs of deterministic and advance statistical techniques, the methods adopted here are proposed to have great potential for air quality forecasting.

Keywords: statistical techniques, forecasting air quality

1.0 INTRODUCTION

In air pollution problems, the air quality models are used to predict concentrations of one or more species in space and time as related to the dependent variables. They form one of the most important components of an urban air quality management plan. There are two types of mathematical models used in Air Quality Modelling i.e. Deterministic Models and Statistical Models. Deterministic models for air quality assessment are based on the physical and chemical behaviour of pollutants in the atmosphere. These models require several inputs dealing with the emission and meteorology. Alternatively, statistical techniques do not consider individual physical and chemical processes and use such historical pollution data which are easily accessible. Also, statistical technique is a more pragmatic approach as the performance of these models is comparable and many times could be superior to the deterministic models without involving elaborate input requirements. The present study thus undertakes air pollution forecasting for criteria pollutants from simple statistical techniques and attempts to compare the performance measures with other statistical and deterministic models. The processes are ignored and instead tests are done on previous data to look for patterns that can be used for prediction. Here an attempt has been made to study the performance of various simple statistical techniques which require no effort in training the data in comparison to other advanced statistical techniques.

2.0 METHODOLOGY

In the present study the forecasting of the four pollutants i.e. NO_2, SO_2, SPM and RSPM is performed using the five statistical techniques i.e. Single Exponential Smoothing (SES), Adaptive Response Rate Single Exponential Smoothing (ARRSES), Holt's Linear Method (HLM) Auto Regressive eXogenous (ARX) and Auto Regressive Integrated Moving Averages (ARIMA) as outlined by Mohan et. al (2007). Each technique has it's own advantages and limitations (Makridakis, 1998). SES method is quite simple as in this method the new forecast value is the old forecast plus an adjustment for the error that occurred in the last forecast. The past forecast error is used to correct the next forecast in a direction opposite to that of the error. There will be an adjustment until the error is corrected. Adaptive Response Rate Single Exponential Smoothing (ARRSES) may have an advantage over SES in that it allows the value of smoothing parameter (α) to be modified, in a controlled manner, as changes in the pattern of the data occur. Thus we can say that ARRSES method is an SES method where α value is systematically, and automatically, changed from period to period to allow for changes in the pattern of the data even when the data are non-seasonal and show no trend. This characteristic seems attractive when hundreds or even thousands of items require forecasting. Single exponential smoothing is extended to linear exponential smoothing to allow forecasting of data with trends. Thus Holt Linear Method (HLM) is similar to the basic form of the single smoothing given by the equation of SES but applies to updating of the trend. In an ARIMA model, the concentrations at a certain instant are expressed as linear combinations of previous concentrations values and random terms (noise), which are specified in a statistical sense. Thus, in ARIMA models the physical causes of phenomena are not distinguished in the input. Such models represent a 'black box' approach. The ARIMAX

* Corresponding Author: Anurag Kandya, Centre for Atmospheric Sciences, Indian Institute of Technology Delhi, New Delhi 110016, Email: akandya@yahoo.com

(ARIMA with exogenous input) model represents a 'gray-box' approach. In ARIMAX modelling, the pollutant concentration at a certain instance is expressed as linear combinations of present and previous physical inputs, plus the noise term.

3.0 DATA USED

For the present study, seven year daily data (1998-2004) of all the four pollutants at ITO (Income Tax Office) air quality monitoring station in Delhi was used for evaluating the accuracy of the above mentioned statistical techniques. The first six year data was used for optimizing the various statistical coefficients. The optimized statistical coefficients were then used for predicting the pollutant concentration of the next year.

4.0 RESULTS AND DISCUSSIONS

Various performance measures like the Correlation Coefficient (r), Fractional Bias (FB), Root Mean Square Error (RMSE) both systematic and unsystematic, Index of agreement (d), Factor of 2 (FAC 2) and Geometric Variance (VG) were to evaluate the various statistical techniques used for predicting the urban air quality in terms of the pollutant concentration. For one day prediction, the correlation coefficient for the various pollutants varied from 0.58 to 0.66 for SES, 0.57 to 0.63 for ARRSES, 0.47 to 0.65 for HLM, 0.60 to 0.66 for ARX and 0.59 to 0.67 for ARIMA. For 4 day prediction, the correlation coefficient for ARX ranged from 0.56 to 0.62 while for ARIMA, it was 0.56 to 0.66 and that for the 7 day prediction it was 0.56 to 0.63 for ARX and 0.56 to 0.66 for ARIMA. The Index of Agreement was as high as 0.81 for one day prediction, 0.77 for four day prediction and 0.78 for seven day prediction which was obtained while using ARX technique. Based on the performance measures, ranking of the statistical models used in the present study have been done in Table 1. The observations made in the study reveals that for 1 day prediction ARIMA technique scores well over the other four statistical techniques i.e. SES, ARRSES, HLM and ARX. For 4 and 7 day prediction both ARIMA and ARX techniques were found suitable. For 1 day prediction of NO_2, SES method is the most suitable statistical technique while for that of SO_2, RSPM and SPM, ARIMA is the best statistical technique. For 4 day prediction of NO_2, ARX technique is comparatively better than the rest of the four statistical techniques while for the remaining 3 pollutant i.e. SO_2, SPM and RSPM, ARIMA is the best statistical technique. For 7 day prediction of NO_2 and SPM, ARX technique is comparatively better than the rest of the four statistical techniques while for the remaining 2 pollutant i.e. SO_2, and RSPM, ARIMA is the best statistical technique (Mohan et al. 2009).

Giuseppei et al (2003) carried out forecasting of SO_2 using various statistical techniques like ANN, MNN, etc. The comparison of the performance measure viz; index of agreement obtained by the various advanced statistical techniques as used by Giuseppei et al for forecasting SO_2 concentration and that used in the present study indicate that the advanced statistical techniques perform slightly better. However, the distinct advantage of the simple statistical techniques is that they require only single type of data (concentration) and no effort in training the data with meteorological, emission and other such data in comparison to other advanced statistical techniques.

Hanna et. al evaluated the performance of various deterministic models like ADMS, AERMOD and ISC3. The statistical techniques used in the present study when compared with the deterministic models for parameters such as VG and FAC2 reveal that the performance of the statistical models is often superior to that of the deterministic models.

5.0 CONCLUSIONS

The results show that there is no single modelling approach, which generates optimum results in terms of full range of performance indices considered. Amongst the five statistical techniques considered in this study, ARIMA technique scores well over the other techniques The present study reveals that the advanced statistical techniques perform slightly better than those of the simple statistical techniques. However, the distinct advantage of the simple statistical techniques is that they only single type of data (concentration) and no effort in training the data with meteorological, emission and other such data in comparison to other advanced statistical techniques. The statistical techniques used in the present study when compared with the deterministic techniques show that the performance of the statistical models is often superior to the deterministic models without involving elaborate input requirements. It is also suggested that the performance of the simple statistical models can be the benchmark for evaluating the performance of deterministic models and advanced statistical methods. Given the uncertainty and unavailability of most of the inputs of deterministic and advance statistical techniques, the methods adopted here have great potential for air pollution forecasting (Mohan et al. 2009, Mohan et al. 2007).

Table 1 Ranking of the five statistical models based on the performance measures (Mohan et al. 2009)

Performance Measures	Best Statistical Technique for 1 day prediction				Best Statistical Technique for 4 day prediction				Best Statistical Technique for 7 day prediction			
	SO_2	NO_2	SPM	RSPM	SO_2	NO_2	SPM	RSPM	SO_2	NO_2	SPM	RSPM
r	4	5	4	4	5	5	5	5	4	5	4	4
FB	2	1,2,3,5	1	1	5	5	5	5	5	5	5	5
MG	5	3	5	5	5	4	5	5	5	4	5	5
VG	5	all	5	5	5	4, 5	5	5	5	4, 5	4	5
NMSE	5	1, 4	5	5	5	4, 5	5	5	5	4	4,5	5
SD	1	1	5	5	4	4	4	5	4	4	4	5
RMSE	1	1	5	5	4	4	4	5	4	4	4	5
d	1	1, 3, 4	4	4	4	4	4, 5	4	4	4	4	4
FAC 2	5	5	4	4	5	5	4	4	5	5	4	4
Overall Best Technique	5	1	5	5	5	4	5	5	5	4	4	5

Where 1=SES, 2=ARRSES, 3=HLM, 4=ARX and 5=ARIMA

REFERENCES:

Manju Mohan, Anurag Kandya and Manish Yadav, (2009) 'An Evaluation and Comparison of the Various Statistical and Deterministic Techniques for Forecasting the Concentration of Criteria Air Pollutants' (accepted for publication in the International Journal of Environment and Pollution)

Manju Mohan, Anurag Kandya and Manish Yadav, (2007) 'Air Quality Index Estimation from measurements and statistical modelling techniques over Delhi', Proceedings of the 11th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes held at Queens College, Cambridge University, UK from July 2-5, 2007, pp 93-96

Mage D., G. Ozolins, P. Peterson, A. Webster, R. Ortherfer, V. Vandeveerd and M. Gwynne, (1996) 'Urban air pollution in megacities of the world', *Atmospheric environment* 30, pp. (681–686).

Makridakis, S., S.C. Wheelwright and Hyndman, R.J. (1998) *Forecasting,* 3rd edition, New York, John willey & Sons, Inc

Giuseppe Nunnari, Stephen Dorling, Uwe Schlink, Gavin Cawley, Rob Foxall, Tim Chatterton, (2004) 'Modelling SO2 concentration at a point with statistical approaches', International Journal of Environmental Modelling and Software, Vol 19, pp. (887 – 905).

Steven R. Hanna et al, Evaluation of the ADMS, AERMOD and ISC3 Dispersion Models with Optex, Duke, Forest, Kincaid, Indianpolis and Lovett Field Data sets (www.cerc.co.uk/software/pubs)

ACKNOWLEDGEMENT

We thank Central Pollution Control Board, New Delhi for providing the needful data for the study.