An evaluation and comparison of the various statistical and deterministic techniques for forecasting the concentration of criteria air pollutants

Manju Mohan* and Anurag Kandya

Centre for Atmospheric Sciences,
Indian Institute of Technology,
Hauz Khas,
New Delhi 110016, India
Fax: 0091-11-26591386
E-mail: mmohan6@hotmail.com
E-mail: mmohan65@yahoo.com
E-mail: akandya@yahoo.com
*Corresponding author

Manish Yadav

531, Dr. Mukherjee Nagar,
New Delhi, India
E-mail: ymanish3@gmail.com

Abstract: A statistical approach is about working through the historical data in a general way and finding guides to future behaviour. In the present study, forecasting of the criteria pollutants has been done using simple statistical techniques and attempt has been made for an inter-comparison of these techniques with various advanced statistical and deterministic techniques. The inter-comparison analysis leads us to the conclusion that there is no single modelling approach which generates optimum results. 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 pollution forecasting.

Keywords: statistical modelling; deterministic models; concentration prediction; criteria pollutants; model evaluation.

Reference to this paper should be made as follows: Mohan, M., Kandya, A. and Yadav, M. (2011) 'An evaluation and comparison of the various statistical and deterministic techniques for forecasting the concentration of criteria air pollutants', *Int. J. Environment and Pollution*, Vol. 44, Nos. 1/2/3/4, pp.96–105.

Biographical notes: Manju Mohan is actively engaged in research and training in the field of Air Pollution Dispersion Modelling for more than 25 years now. She has been working in the areas related to Air Pollution Meteorology, Environmental Impact Assessment (EIA), Coastal and Urban Air Quality

Management and Quantitative Risk Assessment (QRA), etc. She has to her credit about 100 research papers published in international journals and the proceedings of various international and national conferences. She became Fellow, Institute of Environmental Engineers in 1998. Institution of Engineers (India) awarded her Nawab Zain Yar Jung Bahadur Memorial medal for the best paper in Environmental Sciences.

Anurag Kandya is currently pursuing PhD from Centre of Atmospheric Sciences, Indian Institute of Technology Delhi. He did his MTech in Environmental Engineering and Management at the Department of Civil Engineering, Indian Institute of Technology Delhi and BE (Civil) at L.D. College of Engineering, Ahmedabad, Gujarat, India. He has to his credit research papers published in international journals and the proceedings of various international and national conferences. He is also the author of the book titled “Elements of Civil Engineering”. He is a member of various professional bodies like ISTE, IGS and ISG.

Manish Yadav had done his MTech in Environmental Engineering and Management from the Department of Civil Engineering, Indian Institute of Technology Delhi after obtaining BTech in Civil Engineering.

1 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 major types of forecasting models, viz. time series (statistical models) and explanatory models (for example deterministic air quality models). According to the explanatory forecasting, any change in inputs will affect the output of the system in a predictable way, assuming the explanatory relationship will not change. However, time series forecasting treats the system as a black box and makes no attempt to discover the factors affecting its behaviour. The objective of such time series forecasting methods is to discover the pattern in the historical data series and extrapolate that pattern into the future. There are two main reasons to treat a system as a black box. First, the system may not be understood, and even if it was understood it may be difficult to measure the relationships assumed to govern the behaviour of the system. Second, the main concern in forecasting is to predict what will happen with reasonable accuracy and not to know why it happens. The modellers often look for such techniques that provide reasonable/better forecast than most other techniques with an advantage of not involving complex and often less-accurate input parameters such as emissions and meteorology. It is difficult to predict or extrapolate meteorological parameters spatially and temporally accurately and similarly the emission inventory in urban areas is not correctly reported. Thus, the errors involved with these inputs in a forecast model will certainly lead to errors in the predictions as well. Thus, it is proposed here to attempt the application of simple statistical techniques for air-pollution forecasting and compare the performance measures with the other techniques as per the available literature.

2 Methodology

There are two types of mathematical models used in Air Quality Modelling, i.e., Deterministic Models and Statistical Models. Deterministic models for forecasting the air quality are based on the physical and chemical behaviour of pollutants in the atmosphere. As mentioned earlier, these models require several inputs dealing with the emission and meteorology. Alternatively, statistical techniques do not consider individual physical and chemical processes and so the input parameters are related to emissions and meteorology and use mostly historical pollution data, which are easily accessible.

In this 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., SES, ARRSES, HLM, ARX and ARIMA. The details of the techniques are explained in Makridakis et al. (1998). Except for the last two techniques (ARX and ARIMA), first three techniques can be grouped as exponential smoothing methods. Here, an unequal set of weights is applied to the past data because the weights typically decay in an exponential manner from the most recent to the most distant data point. All methods in this group require certain parameters (β or α) whose values lie between 0 and 1. These parameters are first calculated for a given data set and then applied for the data set for which forecasting is performed. In this study, air pollution data for the years 1998–2003 are used for obtaining these parameters and forecasting is performed for the year 2004. In addition, to study the impact of length of time series on the predictions, shorter time series with data for 2002–2003 are used to forecast the concentrations for the year 2004. Further, performance measures are compared for both the cases.

The SES forecasting method requires the specification of α value. ARRSES may have an advantage over SES in that it allows the value of 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.

SES is extended to linear exponential smoothing to allow forecasting of data with trends. Thus, 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. In ARX model, the pollutant concentration at a certain instance is expressed as linear combinations of present and previous physical inputs, plus the noise term.

3 Data used

This study uses 7 years daily data, i.e., from 1998 to 2004 of all the four pollutants at Income Tax Office (ITO) Air-Quality Monitoring Station (AQMS) in Delhi for evaluating the accuracy of the above-mentioned statistical techniques. This AQMS is a highly polluted traffic intersection in Delhi where Central Pollution Control Board (CPCB) is regularly monitoring the ambient air quality on daily basis. The data availability for this monitoring station was more than 90% in the study period.

4 Results and discussions

In this study, air pollution data for the first six years (1998–2003) is used for obtaining the statistical coefficients for various techniques and subsequently forecasting is performed for the year 2004. In addition, to study the impact of length of test data series on the predictions, shorter time series with data for two years (2002–2003) are used to forecast the concentrations for the year 2004. Further, performance measures are compared for both the cases.

Table 1(a) and (b) shows the various performance measures of the five statistical techniques used in this study. It is observed that in almost all the cases, the contribution of Root Mean Square Error Unsystematic (RMSEU) to RMSE was much more than RMSES (RMSE Systematic), which is a desirable feature of a good model. Almost 100% of the time the value of Fraction within a factor of two (FAC2) was more than 0.8, all the values of Fractional Bias (FB) were less than 0.12, 50% of the time the coefficient of correlation (*r*) was more than 0.7 and 64% of the time the index of agreement (*d*) was more than 0.8. Similarly, Geometric Variance (VG) also is close to 1 most of the time, which is its ideal value. The details of these performance measures are described by Cheremisinoff et al. (1989). From this discussion, it is inferred that the performance of the statistical models is highly satisfactory and by and large comparable. As shown in these tables, one-day predictions are always superior to fourth- and seventh-day predictions.

The observations made in the study further reveal that for one-day prediction for the pollutants SO_2, NO_2 and RSPM, ARIMA technique scores well over the other four statistical techniques, i.e., SES, ARRSES, HLM and ARX whereas SES performs better for SPM. For fourth- and seventh-day prediction, ARIMA technique was found suitable for SO_2, NO_2 and RSPM whereas ARX technique was found suitable for SPM. On the basis of the pollutants, ARIMA scores well for SO_2, NO_2 and RSPM for 1, 4 and 7-day predictions whereas for SPM its SES for 1 day and ARX for 4th-and 7th-day prediction.

Almost 100% of the time, the value of the coefficient of correlation (*r*) was more than or equal to 0.5. When 2 years data set (2002–2003) was used for the estimation of the statistical parameters, the *r* value never crossed 0.7, however when this data set was increased to 6 years (1998–2003), around 50% of the time, the *r* value was more than or equal to 0.7. The index of agreement value (*d*) was most of the time more than or equal to 0.7. With 2 years data set, only 14% of the time the *d* value was more than 0.8 but by increasing this data set to 6 years, 64% of the time the *d* value was more than or equal to 0.8. In most of the cases, RMSE has reduced upon increasing the data set from 2 to 6 years. Thus, there is an overall improvement in the quality of forecast upon increasing the data from 2 to 6 years. It thus confirms the importance of long time series data for forecasting the concentration of air pollutants using statistical techniques.

Table 1(a) Performance measures for assessment of the 4th day and 7th day predicted concentrations using ARX and ARIMA techniques (*a* = Years 2002–2003, *b* = 1998–2003)

		ARX		*ARIMA*		*ARX*		*ARIMA*		
		4th day prediction				*7th day prediction*				
Performance measures	*Pollutants*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*Ideal values*
r	SO_2	0.56	0.72	0.56	0.73	0.57	0.68	0.57	0.68	1
	NO_2	0.62	0.75	0.66	0.76	0.64	0.74	0.66	0.74	
	SPM	0.62	0.72	0.62	0.68	0.62	0.63	0.62	0.55	
	RSPM	0.60	0.60	0.60	0.60	0.61	0.53	0.61	0.53	
FB	SO_2	0.08	0.14	0.00	0.03	0.08	0.15	0.00	0.02	0
	NO_2	0.02	0.09	0.00	0.02	0.02	0.09	0.00	0.01	
	SPM	0.08	0.07	0.07	0.00	0.08	0.08	0.07	0.00	
	RSPM	0.09	0.11	0.03	0.06	0.09	0.12	0.03	0.06	
RMSEU	SO_2	1.76	4.46	1.15	4.88	1.73	4.70	1.16	5.24	0
	NO_2	12.92	12.92	14.11	13.52	12.41	13.07	14.28	13.98	
	SPM	136.44	131.36	138.68	171.71	133.25	143.53	135.58	195.50	
	RSPM	57.93	87.21	49.86	89.33	56.98	89.56	49.63	94.63	
RMSES	SO_2	1.70	4.20	2.16	3.10	1.71	4.41	2.14	3.32	0
	NO_2	10.65	10.61	11.73	7.90	10.71	11.04	11.22	8.18	
	SPM	123.36	108.41	124.02	81.55	124.63	127.06	124.93	109.94	
	RSPM	55.11	72.43	45.07	66.41	55.48	80.55	45.08	74.41	
RMSE	SO_2	2.45	6.13	2.45	5.78	2.43	6.44	2.43	6.20	0
	NO_2	16.74	16.71	18.35	15.66	16.39	17.11	18.16	16.19	
	SPM	183.94	170.32	186.05	190.09	182.45	191.69	184.37	224.29	
	RSPM	79.96	111.83	67.21	111.31	79.53	120.45	67.05	120.38	
d	SO_2	0.71	0.81	0.39	0.84	0.71	0.78	0.04	0.81	1
	NO_2	0.77	0.84	0.72	0.86	0.78	0.83	0.71	0.86	
	SPM	0.75	0.82	0.75	0.82	0.76	0.77	0.75	0.73	
	RSPM	0.74	0.76	0.72	0.76	0.75	0.70	0.73	0.71	
FAC2	SO_2	0.98	0.967	1.00	0.975	0.98	0.972	1.00	0.975	1
	NO_2	1.00	0.985	1.00	0.985	1.00	0.980	1.00	0.982	
	SPM	1.00	0.969	0.97	0.950	1.00	0.949	0.98	0.925	
	RSPM	0.97	0.872	0.96	0.882	0.97	0.845	0.96	0.857	
VG	SO_2	1.08	1.10	1.06	1.07	1.08	1.11	1.06	1.09	1
	NO_2	1.04	1.05	1.04	1.05	1.04	1.07	1.04	1.06	
	SPM	1.09	1.09	1.09	1.10	1.09	1.13	1.09	1.16	
	RSPM	1.11	1.23	1.10	1.22	1.11	1.26	1.10	1.25	

Table 1(b) Performance measures used to assess the quality of the one-day predicted concentrations of SO_2, NO_2, SPM and RSPM using SES, ARRSES, HLM, ARX and ARIMA techniques (*a* = Years 2002–2003, *b* = 1998–2003)

Performance measures	*Pollutants*	*SES*		*ARRSES*		*HLM*		*ARX*		*ARIMA*		*Ideal values*
		a	*b*	*a*	*b*	*a*	*B*	*a*	*b*	*a*	*b*	
		One day prediction										
r	SO_2	0.58	0.77	0.57	0.73	0.47	0.74	0.60	0.69	0.60	0.80	1
	NO_2	0.66	0.78	0.63	0.76	0.66	0.77	0.66	0.68	0.67	0.79	
	SPM	0.64	0.72	0.61	0.72	0.64	0.71	0.66	0.52	0.66	0.59	
	RSPM	0.62	0.52	0.59	0.48	0.59	0.51	0.64	0.64	0.62	0.68	
FB	SO_2	–0.01	0.00	0.00	0.00	0.01	–0.01	0.04	0.08	0.01	0.03	0
	NO_2	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.00	0.02	
	SPM	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.04	0.04	0.01	
	RSPM	0.00	0.01	–0.02	0.00	0.02	0.00	0.05	0.06	0.03	0.05	
RMSEU	SO_2	1.94	4.84	1.97	5.65	2.91	5.66	2.12	5.56	1.58	4.24	0
	NO_2	13.53	13.41	14.41	15.45	15.25	14.37	14.17	16.46	12.83	12.20	
	SPM	148.80	148.46	171.22	166.67	169.06	162.42	160.57	187.51	156.85	156.17	
	RSPM	64.37	102.26	80.29	115.55	73.59	109.77	70.45	95.63	51.85	83.41	
RMSES	SO_2	1.38	2.51	1.41	2.22	1.21	2.02	1.21	3.27	1.90	2.68	0
	NO_2	8.83	6.66	8.99	5.76	7.53	6.09	8.45	8.81	12.10	7.82	
	SPM	100.90	86.69	93.77	67.53	84.10	76.86	84.27	127.44	90.87	126.07	
	RSPM	45.18	69.10	37.20	68.11	42.18	67.49	38.91	53.90	42.11	56.49	

Table 1(b) Performance measures used to assess the quality of the one-day predicted concentrations of SO_2, NO_2, SPM and RSPM using SES, ARRSES, HLM, ARX and ARIMA techniques (*a* = Years 2002–2003, *b* = 1998–2003) (continued)

Performance measures	*Pollutants*	*SES*		*ARRSES*		*HLM*		*ARX*		*AR1MA*		*Ideal values*
		A	*b*	*a*	*b*	*a*	*B*	*a*	*b*	*a*	*b*	
		One day prediction										
RMSE	SO_2	2.38	5.45	2.42	6.07	3.15	6.01	2.44	6.45	2.47	5.01	0
	NO_2	16.16	14.97	16.98	16.49	17.01	15.60	16.50	18.67	17.64	14.49	
	SPM	179.78	172.17	195.22	179.83	188.82	179.68	181.34	225.07	181.27	200.70	
	RSPM	78.55	123.42	88.49	134.13	84.82	128.86	80.48	109.77	66.80	100.74	
d	SO_2	0.75	0.87	0.74	0.85	0.68	0.86	0.77	0.81	0.53	0.88	1
	NO_2	0.81	0.88	0.79	0.87	0.81	0.88	0.81	0.82	0.70	0.88	
	SPM	0.78	0.84	0.77	0.84	0.79	0.83	0.80	0.70	0.80	0.74	
	RSPM	0.77	0.71	0.76	0.68	0.76	0.70	0.79	0.79	0.74	0.81	
FAC2	SO_2	0.98	0.982	0.98	0.976	0.91	0.962	0.97	0.954	1.00	0.980	1
	NO_2	1.00	0.991	1.00	0.985	1.00	0.989	1.00	0.977	1.00	0.995	
	SPM	0.99	0.974	1.00	0.966	1.00	0.968	1.00	0.944	0.97	0.964	
	RSPM	0.96	0.859	0.96	0.835	0.95	0.848	0.97	0.908	0.96	0.927	
VG	SO_2	1.08	1.07	1.08	1.09	1.25	–1.30	1.09	1.12	1.06	1.07	1
	NO_2	1.04	1.05	1.04	1.06	1.04	1.05	1.04	1.07	1.04	1.04	
	SPM	1.08	1.09	1.10	1.10	1.09	1.10	1.08	1.14	1.04	1.11	
	RSPM	1.11	1.25	1.12	1.35	1.13	1.28	1.12	1.19	1.10	1.16	

Giuseppe et al. (2004) carried out forecasting of SO_2 using various statistical techniques like ANN, MNN, etc., which is shown in Table 2. The results indicate that the simple statistical techniques perform comparably with the advanced statistical techniques. 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 is required.

Table 2 Inter-comparison of various statistical techniques for SO_2. ANN (Artificial Neural networks with back-propagation training algorithm), MNN (artificial neural networks with maximum likelihood cost function and conjugate gradient training algorithm), WAG (wavelet functions with genetic algorithms), NFU (neuro-fuzzy techniques), GAM (Generalised Additive Models), LPH (Local Prediction in phase-space), LIN (linear time-series model) and PER (persistence model)

		Performance measures				
Study	*Technique*	*RMSE*	*RMSEu*	*RMSEs*	*d*	*r*
Giuseppei Nunnari study (MF models)	ANN	26.54	22.01	14.83	0.77	0.66
	WAG	24.51	22.03	10.74	0.78	0.64
	NFU	29.09	23.62	16.97	0.75	0.65
	LPH	21.65	11.43	18.39	0.71	0.65
	GAM	21.51	11.57	18.15	0.71	0.65
	MNN	21.74	12.26	17.95	0.72	0.65
	LIN	22.76	11.66	19.54	0.67	0.60
	PER	27.32	23.86	13.31	0.71	0.53
Giuseppei Nunnari study (NMF models)	ANN	32.62	26.45	19.09	0.68	0.55
	WAG	32.67	30.38	12.01	0.69	0.52
	NFU	32.55	27.40	17.56	0.69	0.54
	LPH	22.90	11.38	19.87	0.66	0.59
	GAM	21.87	11.62	18.52	0.70	0.63
	MNN	–	–	–	–	–
	LIN	–	–	–	–	–
	PER	27.32	23.86	13.31	0.71	0.53
Present study	SES	5.45	4.84	2.51	0.87	0.77
	ARRSES	6.07	5.65	2.22	0.85	0.73
	HLM	6.01	5.66	2.02	0.86	0.74
	ARX (1 day)	6.45	5.56	3.27	0.81	0.69
	ARIMA (1 day)	5.01	4.24	2.68	0.88	0.80
	ARX (4 day)	6.13	4.46	4.20	0.81	0.72
	ARIMA (4 day)	5.78	4.88	3.10	0.84	0.73
	ARX (7 day)	6.44	4.70	4.41	0.78	0.68
	ARIMA (7 day)	6.20	5.24	3.32	0.81	0.68

Hanna et al. (1999) evaluated the performance of various deterministic models like ADMS, AERMOD and ISC3, the results of which are compiled in Table 3. As the statistical techniques used in this study when compared for parameters such as MG (Geometric Mean Bias), VG and FAC2 with the deterministic models shown in Tables 1 and 3, reveals that the performance of the statistical models is often superior to the deterministic models.

Table 3 Performance measures of the three deterministic models used for pollutant forecasting (source: www.cerc.co.uk/software/pubs). Advanced Dispersion Modelling System (ADMS), AMS/EPA Regulatory Model (AERMOD) and Industrial Source Complex Model (ISC3)

Parameter	*Technique*	*OPTEX tanks*	*OPTEX matrix*	*Duke forest*	*Kincaid*	*Indianapolis*	*Lovett*
MG	ISC3	0.86	0.55	0.32	1.33	0.85	–1.68
	ADMS	2.12	0.89	1.41	–0.03	1.14	0.14
	AERMOD	2.47	1.02	1.83	0.75	1.54	–0.37
VG	ISC3	2.8	2.9	11.6	8.5	6.8	46
	ADMS	3.0	1.59	1.7	0.7	5.6	3.6
	AERMOD	4.4	1.8	2	2.2	13	3.6
FAC2	ISC3	0.8	0.64	0.17	0.13	0.49	0.06
	ADMS	0.8	0.76	0.63	0.59	0.42	0.3
	AERMOD	0.7	0.76	0.53	0.29	0.39	0.25

5 Conclusions

The results show that there is no single modelling approach that 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. This study reveals that the simple statistical techniques perform comparably with those of the advanced statistical techniques. However, the distinct advantage of the simple statistical techniques is that they use only single type of data (concentration) and no effort in training the data with meteorological, emission and other such data is required in comparison with other advanced statistical techniques. The statistical techniques used in this 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 advanced statistical techniques, the methods adopted here have great potential for air pollution forecasting.

Acknowledgement

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

References

Cheremisinoff, P.N. (Ed.) (1989) 'Encyclopedia of environmental control technology', *Air Pollution Control*, Vol. 2, pp.123–128.

Giuseppe, N., Stephen, D., Uwe, S., Gavin, C., Rob, F. and Tim, C. (2004) 'Modelling SO_2 concentration at a point with statistical approaches', *International Journal of Environmental Modelling and Software*, Vol. 19, pp.887–905.

Hanna, S.R., Egan, B.A., Purdum, J. and Wagler, J. (1999) *Evaluation of ISC3, AERMOD, and ADMS Dispersion Models with Observations from Five Field Sites*, HC Report P020, API, 1220 L St. NW, Washington, DC 20005-4070.

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