Modelling the multi-year maximum daily PM_{10} concentration in Edinburgh: an application of the variability decomposition transfer function model

H. Al-Madfai, A. J. Geens & D. G. Snelson
Faculty of Advanced Technology, University of Glamorgan, Wales, UK

Abstract

Understanding the temporal variability in the concentration of airborne PM_{10} can be of benefit as it would lead to more reliable models that can inform the monitoring and control of air pollution. Established forecasting approaches are generally data driven and offer little in terms of furthering the understanding of the dynamics of data. A variability decomposition (VD) based transfer function model can be used to decompose variability in time series data into inherent and external, thus concentrating on modelling only the external variability as a function of the model inputs.

The VD approach was used to model the multi-year maximum daily PM_{10} concentration recorded in St Leonards, Edinburgh using historic values and PM_{10} concentrations recorded at the Grangemouth monitoring station situated 19 miles to the North West using two established approaches as benchmarks. The results indicate that the transfer function models using the Grangemouth data were superior to the univariate model in terms of the RMSE and MAPE. The performance of the VD transfer function model was comparable to the benchmark in terms of forecast accuracy, but superior in providing improved physical interpretation of the model components

Keywords: air quality forecasting, variability decomposition, transfer function modelling, PM_{10}, ARIMA, Box-Jenkins, transfer function models.

WIT Transactions on Ecology and the Environment, Vol 136, © 2010 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR100311

1 Introduction

Poor air quality levels in our towns and cities are at their highest on busy streets, near to factories, and in inner-city areas. Poor air quality impacts on the young, the sick and elderly people's health and the environment [1]. However air quality data usually contains extreme values that cannot be explained by these factors. Nevertheless, the pollutant threshold/limit values for air quality are set out in the European Directives [2]. The United Kingdom has National Air Quality Standards which defines levels that avoid significant risks to health [3].

PM_{10} is a standard measure of particulate air pollution. It has been linked to respiratory illnesses, including asthma, and real time monitoring facilities of PM_{10} levels are now available to the general public through the Eye on Earth website, which is maintained by the European Environment Agency.

A number of existing works attempt to produce temporal models for PM_{10} concentrations and other pollutants using a variety of approaches (see for example [4–6]. Many of these are forecasting focused and tend to favour non-parametric approaches such as Neural Networks over statistical methods, however some of these suffer from the improper use of statistical methods in the time domain rendering any comparisons between the approaches of little value (as, for example, in the use of linear regression in [7]). Temporal forecasts for PM_{10} are made available through the UK Air Quality Archive website (8) that is maintained by UK Department for Environment, Food & Rural Affairs (Defra). However the forecasts produced are spatio-temporal and are at macro level and the forecasting methodology is essentially judgmental and non transparent:

"*Ultimately it is the expert judgement of the duty forecaster which determines how the available data are combined to form the forecast issued to the public.*" [9]

A notable development is the Openair project [10] which aims to develop an open source set of tool to measure, analyse and model air quality data. This NERC and Defra funded project encourages transparency and collaboration between members of the air quality community and is showing a great deal of promise in this area.

This research primarily aims to model the maximum levels of PM_{10} recorded daily at St Leonards monitoring station in Edinburgh using established time series methods and compares these with the model produced by a novel modelling approach known as the Variability Decomposition (VD) approach. The VD approach decomposes variability in time series data into inherent and external, and in a transfer function setup it aims to provide further understanding of the dynamics of the data through concentrating on modelling only the external variability. This approach is introduced in detail next.

2 The variability decomposition approach (VD)

The VD approach decomposes variability in a stable input-output (I/O) time series system according to its source or main influence. In this approach, it is proposed that variability in time series data can be due to three main sources,

external, inherent and noise. This decomposition is used within the VD framework to construct transfer function models to represent and describe the dynamics of the underlying relationships between the time series datasets. These models can then be used in descriptive, control and forecasting applications as they provide added insights into the univariate dynamics and cross variable relationships governing the I/O system.

2.1 Validation of the VD

Consider the following autoregressive moving average (ARMA) model [11]:

$$\phi(B)y_t = \theta(B)e_t \tag{1}$$

where y_t is the observed datum at time t, e_t is independently and identically distributed noise (i.e. white noise), B is the backshift operator such that $By_t = y_{t-1}$. $\phi(B)$ and $\theta(B)$ are polynomials in B with orders p and q, that represent the autoregressive and moving average components of the model, respectively. Note that this model assumes y_t being stationary.

The model in Equation (1) represents the standard time series model that the Box-Jenkins [11] and other approaches are based on. It has been shown that ARMA models are capable of modelling any stationary time series [12].

Adding y_t to both sides of Equation (1) above, and rearranging yields

$$y_t = y_t - \phi(B)y_t - \theta(B)e_t \tag{2}$$

hence,

$$y_t = \phi^*(B)y_{t-1} - \theta(B)e_t \tag{3}$$

where

$$\phi^*(B) = y_t - \phi(B)y_t. \tag{4}$$

Let

$$H_{t-1} = \phi^*(B)y_{t-1} \text{ and } E_t = \theta(B)e_t, \tag{5}$$

where H_t now contains only historic values of the observed series. Therefore, we have

$$y_t = H_{t-1} + E_t. \tag{6}$$

It can be shown, without loss of generality, that H_{t-1} is free of y_t and is a function of historic values of the series (i.e. $y_{t-1}, y_{t-2,\ldots}$).

Given the independence of the noise component, it can be shown that

$$v(y_t) = v(H_{t-1}) + v(E_t), \tag{7}$$

where $v(H_{t-1})$ is the inherent variability, representing the observed historic values' contribution to the variability at present, and $v(E_t)$ is the external variability representing the variability in the present that cannot be explained using the observed historic values. Hence, Equation (7) decomposes the

variability of the observed series, $v(y_t)$ into inherent, modelled via the Autoregressive component of the model; external, modelled as the Movingaverage component of the model and noise (which is built into E_t).

The VD was validated using simulated data in [13, 14] where it has been shown that the influence of an exogenous variable mainly influences the Movingaverage component of the transfer function model.

3 Data and modelling

Experiments were carried out to model the maximum daily PM_{10} readings measured at St. Leonards monitoring station in Edinburgh between 19/1/2004 and 15/9/2009 as a univariate ARIMA model and in a transfer function models using VD and the established Box-Jenkins transfer function approach using the maximum daily PM_{10} recorded at the Grangemouth monitoring station as input. Figure 1 shows a time plot of the raw data where it can be seen that there are a number of days in which the maximum PM_{10} recorded seems to deviate markedly from the general norm of the data. A number of missing values covering approximately 10 weeks in 2007 were also present in the data. These were replaced using the means of the nearest neighbouring points.

The univariate ARIMA model was considered as the naive benchmark for this data and the results from the three models were compared using the conventional Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) penalty functions.

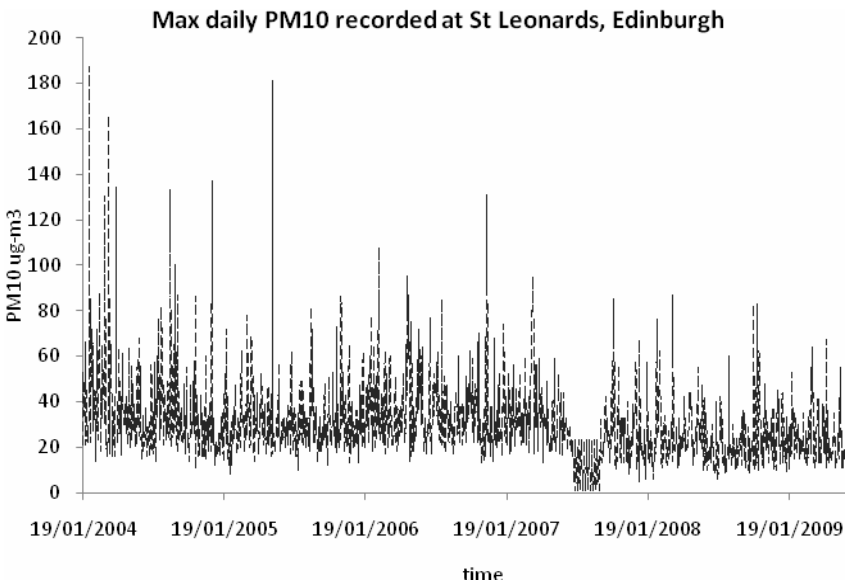

Figure 1: Daily maximum PM_{10} recorded in St Leonards monitoring station in Edinburgh.

4 Results

The univariate model obtained for the data was an ARIMA(1,1,1)(0,1,1):

$$(1-0.27B)(1-B^7)(1-B)y_t = (1+0.96B)(1-0.98B^7)e_t$$

The Box-Jenkins transfer function model was:

$$(1-B^7)(1-B)y_t = \frac{0.279}{1-0.28B}(1-B^7)(1-B)x_t + \frac{(1-0.966B)(1-0.98B^7)}{1-0.206B}e_t$$

while the VD model obtained for the data was:

$$(1-B)(1-B^7)y_t = \frac{(0.275+0.27B+0.006B^6+0.278B^7-0.279B^8)}{1-0.27B}x_t + \frac{(1-0.97B)(1-0.98B^8)}{1+0.64B}e_t$$

The forecast accuracy measures yielded by these models are shown in Table 1.

Table 1: Forecast accuracy measures yielded by the three modelling approaches. Both transfer function approaches seem to be superior to the univariate model.

model penalty function	RMSE	MAPE
Univariate	14.749	39.584
Box-Jenkins Transfer Function	13.982	36.738
Variability Decomposition TF	13.87	37.4

The residuals from the three models were homoscedastic, had no visually obvious patterns when plotted against time and showed further white-noise behaviour with no significant autocorrelations or partial autocorrelations.

While all the polynomials in the models above are stationary, the Movingaverage components of the models contain parameters that are close in value to the boundary of invertibility. This might be seen as a data preparation problem or as the data having a long memory whereby the PM_{10} particles seem to remain in the atmosphere for a long time - hence the model requiring a relatively large number of past observations to explain the variability in the data.

5 Discussion

Figure 2 shows the observed and within sample one step-ahead forecast obtained from the ARIMA, Box-Jenkins transfer function and the VD models, respectively.

It can be seen from Figure 2 that, visually, both transfer function models provide a better fit for the data compared to the univariate ARIMA. This improved fit is evident in the transfer function models predicting the exotic

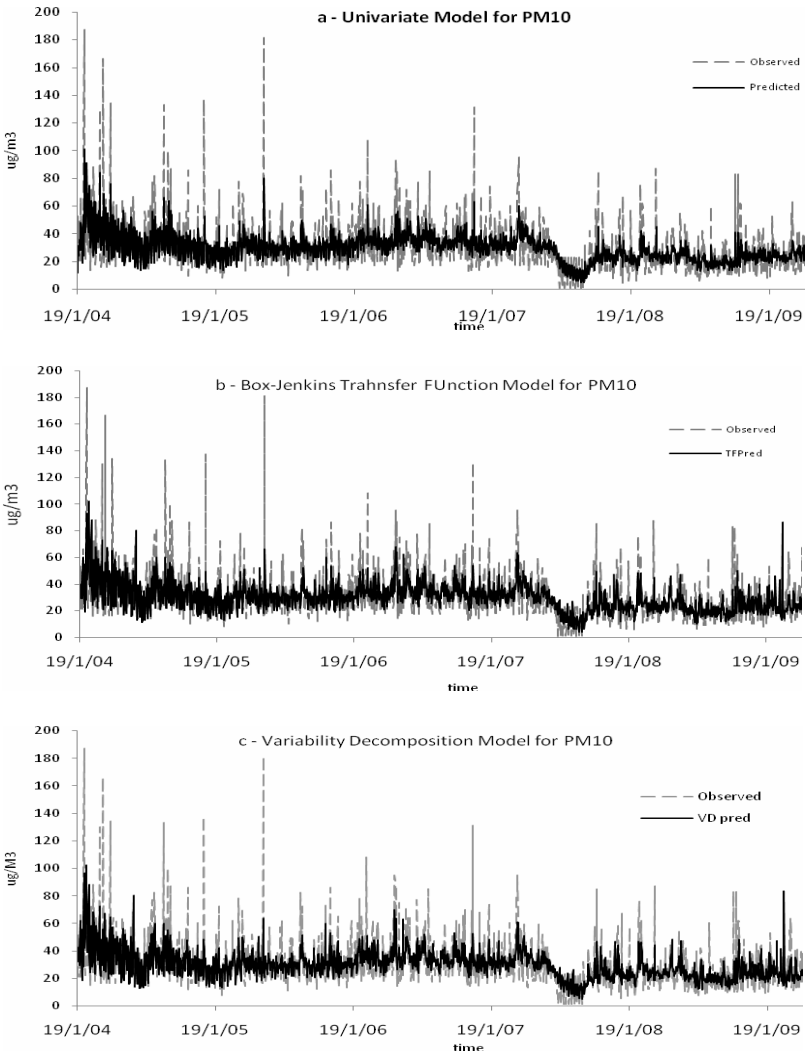

Figure 2: Within sample forecasts and observed data obtained from the three modelling approaches. Visually the transfer function approaches seem to provide a superior fit compared to the univariate model.

observations with relatively more accuracy compared to ARIMA. This finding is also supported by the RMSE and MAPE values which are smaller for the transfer function models compared to the univariate ARIMA. Physically, this can be interpreted to indicate that some of the PM_{10} measured at St Leonard can be linked to the PM_{10} measured at Grangemouth.

The structure of the VD model is more complex compared to the Box-Jenkins transfer function model. While model parsimony is generally desired, the components of the VD model are more pertinent in relating to the problem at

hand. The numerator polynomial of the transfer function term in the model indicates that the variability in maximum PM_{10} data at St Leonards is related to (and can be partially explained by) the maximum PM_{10} measured on the same day and on one, six, seven and eight days before.

This finding is facilitated through the flexible structure of the VD model which provides a regression-like numerator polynomial in the numerator of the transfer function term. By contrast, while the established Box-Jenkins transfer function model yielded comparable forecast accuracy measures, it cannot provide similar insights into the underlying dynamics of the data due to structural limitations. It is worth noting that as the distance between the two sites is approximately 19 miles (31 km, as the crow flies) some of the relationships inferred through the VD model may not necessarily be causal since ambient conditions may be similar at both sites due to geographical proximity.

All models considered in this study failed to produce accurate predictions of the peaks in maximum PM_{10} in St Leonard. Future models can look at using other regional PM_{10} data as inputs to the VD model to enhance predictability and forecast accuracy.

As noted earlier, a few of the models' parameter estimates lie close to the boundaries of invertibility. One possible explanation for this is that these long memory model components represent the slow dispersion nature of airborne PM_{10}. Nevertheless, future work can look into refining these models.

In conclusion, the VD approach produced a superior transfer function model for the maximum daily PM_{10} measured at St Leonard monitoring station. The VD model indicated that some of the variability in St Leonard's maximum PM_{10} can be explained using the max PM_{10} measured at Grangemouth however other datasets may be considered in order to better model the extreme observation in the PM_{10} series.

References

[1] *The Effect of Urban Air Pollution on Inflammation, Oxidative Stress, Coagulation, and Autonomic Dysfunction in Young Adults.* K-J. Chuang, C-C. Chan, T-C. Su, C-T. Lee and C-S. Tang. s.l. : Am. J. Respir. Crit. Care Med, 2007, Vol. Vol. 176, pp. 370-376.

[2] *Council Directive 1999/30/EC of 22 April 1999 relating to limit values for sulphur dioxide, nitrogen dioxide and oxides of nitrogen, particulate matter and lead in ambient air . 1999.*

[3] *Air Quality Standards (Scotland) Regulations 2007 . 2007.*

[4] *Modelling the multi year air quality time series in Edinburgh:.* Al-Madfai, Hasan, Snelson, David and Geens, Andrew. Skiathos : WIT Press, 2008.

[5] *A neural network forecast for daily average PM10 concentrations in Belgium.* Hooyberghs, Jef, et al. 18, 2005, Atmospheric Environment, Vol. 39, pp. 3279-3289 .

[6] *Quality and performance of a PM10 daily forecasting model.* Stadlobera, Ernst, Hörmannb, Siegfried and Pfeiler, Brigitte. 6, 2008, Atmospheric Environment, Vol. 42, pp. 1098-1109 .

[7] *PM10 forecasting for Thessaloniki, Greece.* Slini, T, et al. s.l. : Elsevier, 2005, Environmental Modelling and Software, pp. 559-565.
[8] Air Quality Archive. *Air Quality Data and Statistics.* [Online] [Cited: 2 3 2010.] http://www.airquality.co.uk/data_and_statistics_home.php.
[9] UK Air Quality Archive. *How the Air Pollution Forecasts are produced.* [Online] Defra. [Cited: 2 3 2010.] http://www.airquality.co.uk/uk_forecasting/forecast_is_made.php.
[10] Openair Project. [Online] University of Leeds. [Cited: 2 3 2010.]
[11] Box, George E. P., Jenkins, Gwilym M and Reinsel, Gregory C. *Time Series Analysis: Forecasting and Control.* [ed.] 4. s.l. : Wiley, 2008.
[12] Sir Kendall, Maurice and Ord, Keith. *Time series.* 3. s.l. : Hodder Arnold, 1990.
[13] Al-Madfai, H. *Weather Corrected Electricity Demand Forecasting.* Pontypridd. : School Of Technology, University of Glamorgan, 2002.
[14] *The Variability Decomposition (VD) approach to transfer function modelling: ap-.* Al-Madfai, Hasan, Ameen, Jamal and Ryley, Alan. Sydney : s.n., 2004. 24th International Symposium on Forecasting.
[15] *Daily Electricity Demand Forecasting: A hierarchical Profiling Approach.* Al-Madfai, H., Ameen J., Ryley, A. Crete : ETK/NTTS, 2001.