Modelling the multi year air quality time series in Edinburgh: an application of the Hierarchical Profiling Approach

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

Abstract

Modelling and forecasting of time series of concentrations of air pollutants is essential in monitoring air quality and assessing whether set targets will be achieved. While many established time series modelling approaches transform the data to stationarity a priori, the explicit modelling and presentation of the non-stationary components of the series in this application is essential to allow for further understanding of variability and hence more informed policies. The Hierarchical Profiling Approach (HPA) was used to model the multi-year daily air quality data gathered at St. Leonards in Edinburgh, UK spanning from 1st of March 2004 to 15th July 2007. The HPA is an avant-garde approach that explicitly models the non stationary component of time series data at different levels depending on the span of the component, so that within-year disturbances are at Level 1 and year-long variability such as seasonality is at Level 2, and so on. HPA decomposes the variability into deterministic, stochastic and noise and uses continuous models to describe the non-stationary components using the deterministic part of the model. The stationary stochastic component is then modelled using established approaches. The dataset modelled was the total daily concentrations of carbon monoxide. After modelling within-year events at Level 1, a harmonic regression with trend model was used to describe the weekly aggregates of the data at Level 2. This model was then sampled back in the daily domain and no evidence of a larger cyclical component was found. Wind-speed and a dummy intervention-variable indicating the implementation of the smoking ban were considered in a transfer function model. The concluding model included the present and one lagged observations of wind speed. The smoking ban variable was not significant.

Keywords: monitoring air quality, Hierarchical Profiling Approach, smoking.

WIT Transactions on Ecology and the Environment, Vol 116, © 2008 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR080061

1 Introduction

Air quality in our towns and cities are at their highest levels 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]. The pollutant threshold/limit values for air quality are set out in the European Directives. The United Kingdom has National Air Quality Standards that defined levels that avoid significant risks to health.

Edinburgh is among the least-polluted of the European capitals but has a number of choke points where standing traffic causes pollutant levels to rise. Edinburgh has an Air Quality Management Area (certain areas of the city centre), which was declared on the 31st December 2000. The plan covered all the places where the annual average concentration of nitrogen dioxide is currently predicted not to attain the set target for 2005. Nitrogen dioxide is formed in a city from a build up of nitrogen oxides (NOx). Edinburgh studies have shown that 88 percent of nitrogen oxides come from road transport, with the remaining 12 percent coming from domestic heating and Edinburgh International Airport.

This investigation uses time series analysis to model and forecast air quality data, which was recorded at St. Leonards Air Quality monitoring station in the Southside of Edinburgh as a function of wind speed recorded at Edinburgh Gogarbank. It also attempts to evaluate the influence of the smoking ban introduced in Scotland on the 29th of March on the outdoor air quality recorded at that station. Specifically, this paper applies the Hierarchical Profiling Approach [2] to model the time series of CO levels recorded at the monitoring station. The HPA offers the advantage of explicitly quantifying and modelling the different components of a time series (e.g. trend and seasonality) and so provides an improved understanding of the underlying dynamics of the data.

The monitoring station is situated in a park adjacent to a medical centre car park. The nearest road is approximately 50 meters away which is a busy main road running into the city centre and out to the A7 South of Edinburgh. This Automatic point monitoring station produces high resolution measurements typically hourly or shorter period averages for particulates, oxides of nitrogen, sulphur dioxide carbon monoxide, benzene and 1,3-Butadiene. The air quality data used in this study is a United Kingdom wide monitoring network monitored by the Department for Environment, Food & Rural Affairs (defra). This data will be used for a larger study to measure background pollutant concentration levels for outdoor air quality where smokers are accommodated. Smoking causes pollutants to be released into the atmosphere. To measure the levels of the pollutants in the atmosphere from cigarettes background pollutant levels are required to be deducted to give a realistic concentration level.

2 Methodology

The Hierarchical Profiling Approach (HPA) is an event-driven time series modelling approach and can be seen as a generalisation of the Box-Jenkins intervention analysis. It has been developed in [3] and has been used in analysing

difficult datasets in a number of disciplines including energy forecasting and crime modelling [4–7]. Assuming additivity, the HPA is based on decomposing the variability in time series into deterministic and stochastic and noise components. Analytically it model a time series, y_t, as

$$y_t = f(t) + Z_t, \ldots. \quad (1)$$

where $f(t)$, the deterministic component, is a collection of continuous functions built and fitted by the modeller and Z_t is the variable that holds the stochastic and noise components and can be modelled using a stochastic approach.

The deterministic function in Equation (1), $f(t)$ additively models the changes of the behaviour time series corresponding to identified and known events as well as the typical annual pattern of variability and the trend of the data. Hence, the first level of building $f(t)$ starts at the highest resolution of the data looking for changes in the time series corresponding to known events. The second level models the pattern at the next resolution level, and so on. For a daily dataset, for example, $f(t)$ at Level One models the within-year disturbances that are associated with salient exogenous events, at Level Two it models the weekly seasonality, at Level Three it models the annual seasonality of the data and so on. However, when estimating higher level profiles it may be necessary to aggregate the data to a lower resolution to smooth the volatility that is often observed at higher resolutions.

In theory, any deterministic function can be used in modelling the profiles and hence in building $f(t)$, so long as it is continuous and representative of the profile it aims to model. The parameters for the profile function can be estimated based on a number of criteria including least squares. Hence, the continuity of the profile functions allows the profiles to be estimated at a resolution different to the data's original and then be resampled at higher resolutions.

Assuming successful profiling of all levels, the profile-adjusted stochastic component, $Z_t = y_t - f(t)$, now likely to be weakly stationary, can then be modelled using established approaches such as ARIMA, State Space or Transfer Function models if explanatory variables are to be included in the analysis.

The HPA offers a number of advantages over other approaches in that it builds a catalogue of salient events and quantifies the corresponding changes in the time series corresponding to these events. This allows for an improved understanding of the underlying dynamics of the time series as well as the bulk forecasting of future values should the event occur again. The HPA models trend and seasonality explicitly thus allowing for further investigations of these components to be carried out. The HPA is also capable of dealing with time series with multiple seasonal components by modelling each seasonally component as an independent profile. And, the HPA can act as a powerful prewhitening technique to transform difficult datasets to stationarity.

3 Data

The raw dataset used in this research are the hourly measurements of carbon monoxide (CO) levels readings at St. Leonards Air Quality Monitoring Station in Edinburgh. These readings are made available in the public domain shortly after measurement on an Air Quality website [http://www.airquality.co.uk/archive/data_selector.php?u=7092d2de63c5e5fa319474c479f490d3]. The observations made available on the website are initially labelled as 'provisional' until they undergo a process of inspection and correction, if needed, to then be labelled as 'ratified'. This ratification process is in three stages and involves human intervention at its final stage. Consequently, it takes provisional observations up to 15 weeks to be 'ratified'. It is the assumption of this work that ratified data is representative and reliable.

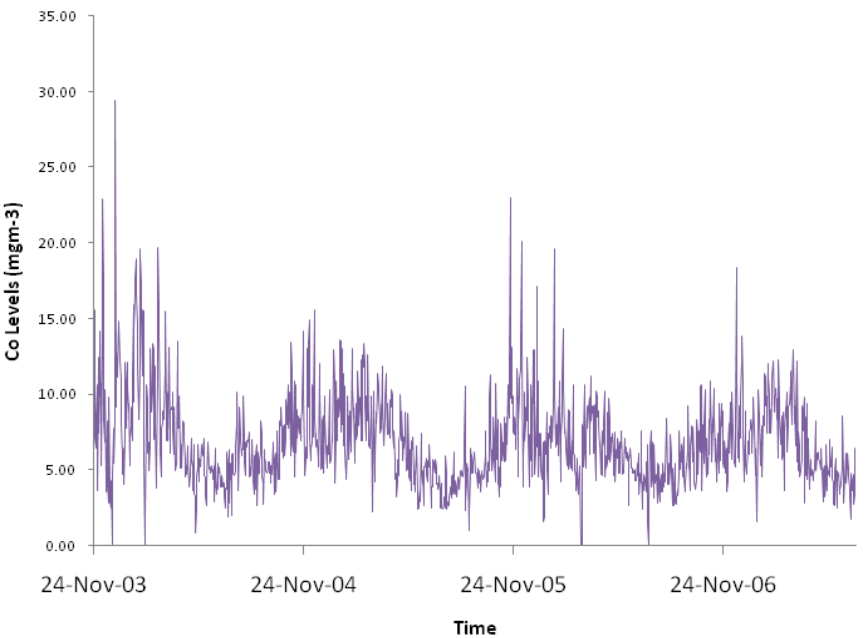

Figure 1: Daily CO levels reading at St. Leonard's Station. Annual seasonality and a number of exotic observations can be seen in the data.

The raw data is made available as hourly readings of CO Levels spanning from 24th November 2003 to 15th July 2007, a total of 31920 hourly observations, aggregated into the daily domain as the total CO readings per day. Table 1 shows a time plot of the daily CO time series in which a seasonal pattern and a number of exotic observations can be seen. The graph also shows that the first few weeks of data was erratic and seemingly ill behaved. This was confirmed by the data

collection agency since these first few weeks were used to calibrate the equipment used in the measurements. Consequently, the data used in this work is the ratified daily aggregates of CO readings from 1st March 2004 to 15th July 2007 – a total of 1232 observations.

A number of exotic observations and time windows (identified subjectively) were noted in the data and great efforts have been put into trying to link these observations to events on the ground to allow for these events to be profiled. Unfortunately, to-date very little information was made available to this research about such events and information applied for under the Freedom of Information Act is still awaited leaving this application and work in progress.

In order to further the understanding of the underlying dynamics of the time series under investigation, two exogenous variables were included in the study. The maximum wind speed recorded on the day (in knots) and a binary intervention variable indicating the date of the smoking ban in Scotland (contained 0s up to the 26th of March 2006, the date of the smoking ban, and 1s afterwards) were included as explanatory variables in the study.

4 Analysis

The HPA was applied to create a profile for the normal behaviour of the data. To this end, exotic observations were subjectively identified at the dates 18th November 2005, 8th December 2005, 4th January 2006, 3rd February 2006, 17th & 18th December and 27th & 28th December 2006. In order to establish a reliable 'norm' for the data, these exotic observations were excluded from this stage of the analysis and replaced as missing values using the average of the available clean observations made at the same dates but different years.

The cleansed dataset was then aggregated to the weekly domain to reduce the volatility in the data and the harmonic regression with polynomial trend model given in Equation (2) was estimated:

$$f_2(t) = a + bt + ct^2 + \sum_{i=1}^{26}\left(a_i \sin(\alpha i t) + b_i \cos(\alpha i t)\right), \ldots \quad (2)$$

where $f_2(t)$ is the profile for the annual seasonality, t the time index variable, a, b and c are the trend parameters (to be estimated), $\left(a_i \sin(\alpha i t) + b_i \cos(\alpha i t)\right)$ is harmonic i out of a total of 26, $\alpha = 2\pi / 52$ is the harmonic angle, a_i and $b_i; i = 1,\ldots,26$ are the harmonic regression parameter (to be estimated).

Using the Levenberg-Marquardt procedure to fit the model in Equation (2) to the data, the following model was obtained:

$$f_2(t) = 47.975 - 0.023t + 12.206\sin(\alpha t) - 5.078\cos(\alpha t) + 2.68\sin(2\alpha t) + 3.11\sin(3\alpha t) + 1.759\sin(6\alpha t) \quad (3)$$

This profile was resampled in the daily domain (using *t/7* as the time index) as shown in Figure 2. The deviations from this profile (i.e. $y_t - f_2(t)$) were inspected and no evidence for a further pattern in the data was observed. Hence, this concluded the application of the HPA with *f(t)* set to be equal to the model in Equation (3).

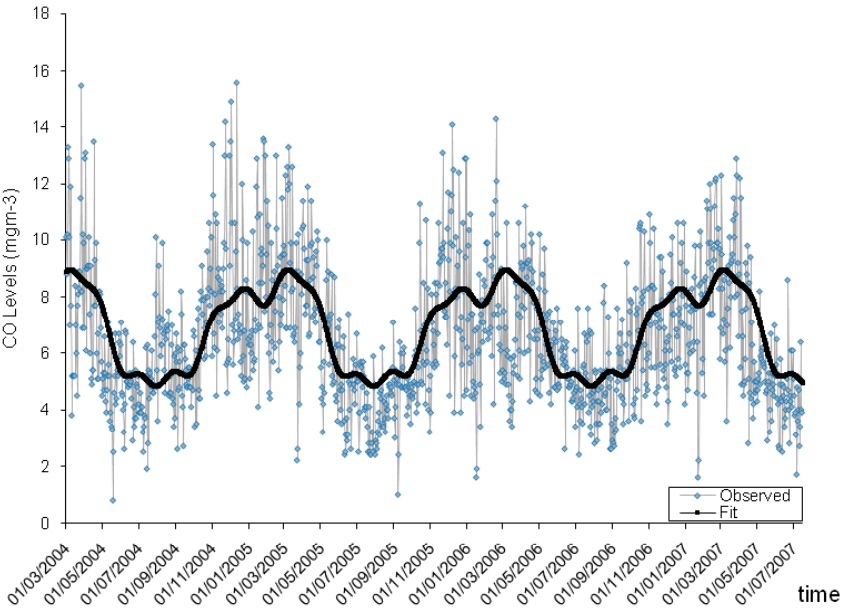

Figure 2: Daily total CO levels and HPA profile.

In the next step, the profile-adjusted stochastic component *Zt* was calculated and modelled as a transfer function ARMA model using the Box-Jenkins approach with the maximum daily wind speed the smoking ban variables as inputs and the following model was obtained:

$$(1+0.65B^7)(1-B)Z_t = -.131(1-B)w_t + (1+0.71B+0.14B^2+0.089B^4)(1-0.74B^7)e_t \quad (4)$$

where B is the backshift operator so that $B^k Z_t = Z_{t-k}$, w_t is the maximum wind speed recorded at time t and e_t is white noise. The within sample forecasts for the original data $y_t(1)$ were then obtained by reintroducing the profile to the forecasts obtained from Equation (4) $Z_t(1)$ as:

$$y_t(1) = Z_t(1) + f(t), \ldots \quad (5)$$

yielding a Residual Mean Square Error of 1.751 and Mean Absolute Error = 1.35, showing the observed and one step-ahead within sample forecasts for the total daily CO readings.

The time plot of the residuals e_t looked random with no evidence of a pattern remaining in the data and both the ACF and PACF of e_t show just one significant spike at lag 20. Figure 3 shows the observed CO readings and one step-ahead forecasts obtained from the model in Equation (4).

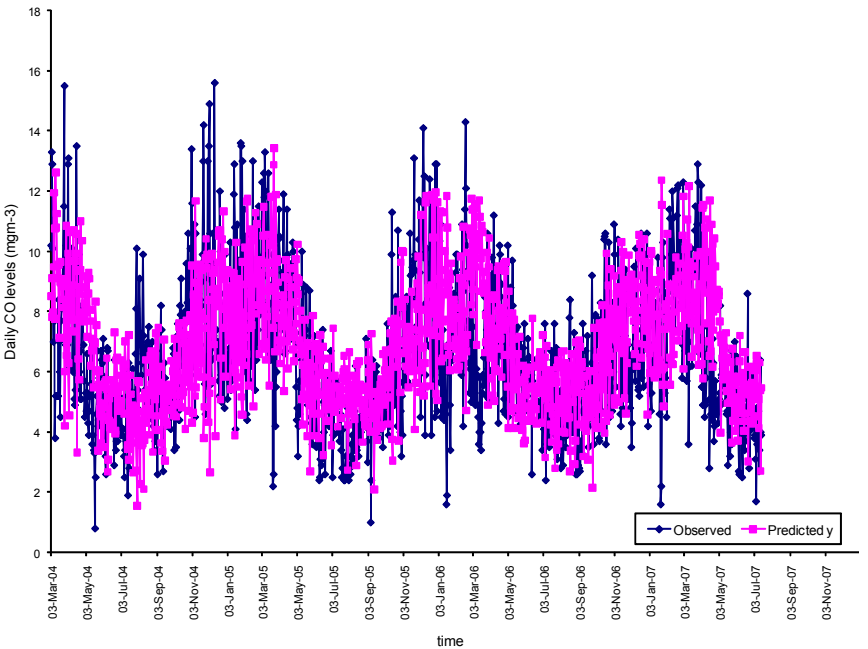

Figure 3: Observed and within sample forecasts for CO levels.

5 Discussion

Using the HPA the different components that make up the variability in CO levels were quantified and explicitly modelled. Alongside the obvious annual seasonality in the data the HPA has identified a weak downwards trend in the data that is statistically significant. It was possible to prewhiten the data using the HPA and hence a successful transfer function model for the data was obtained. Therefore the HPA was a useful in modelling the data.

The Transfer Function model obtained for the data seems reliable since the model diagnostics all seem satisfactory and the significant spike observed at lag 20 in the ACF and PACF of the model's errors does not correspond to a logical or calendar cycle. This provided sufficient evidence to conclude that the CO levels recorded at this site are inversely related to the maximum wind speed

recorded on the day. It is reasonable to speculate that this relationship relates to the role of the wind in the dispersion of airborne particles. However, there was insufficient evidence to conclude that the smoking ban had a significant effect on the levels of CO recorded at this station.

As in any statistical investigation, the reliability of the results of this study depends mainly on the reliability of the data used in the analysis. There was very little information available to this study for any Level one profiles to be constructed despite the efforts put in to identify any salient events that could be associated with the exotic observations identified in the data. In addition, there are only three seasonal cycles in the span of the data with only one complete cycle post the smoking ban. Therefore, while it is the authors' belief that the results presented in this work are as reliable as can be achieved given the available data, further analysis needs to be carried when more data becomes available. It is expected that the downwards trend that was just significant in this study (with 95% CI -0.046 to 0.000) will become stronger. In addition, having one or two more seasonality cycles in the data would most certainly yield a more reliable estimate of the annual seasonal component of the data.

Acknowledgements

This study was commissioned by the Scottish Licensed Trade Association with funding support from the UK Tobacco Manufacturers' Association.

References

[1] K-J. Chuang, C-C. Chan, T-C. Su, C-T. Lee and C-S. Tang, The effect of urban air pollution on inflammation, oxidative stress, coagulation, and autonomic dysfunction in young adults. *Am. J. Respir. Crit. Care Med.,* **176**, pp. 370–376, 2007.

[2] Al-Madfai, H., Ameen J., Ryley, A., Daily electricity demand forecasting: a Hierarchical Profiling Approach. *ETK/NTTS*, Crete, 2001.

[3] Al-Madfai, H., Weather corrected electricity demand forecasting. School Of Technology. 2002, University of Glamorgan: Pontypridd.

[4] Al-Madfai, H., Ivaha, C. & Ware A., The Dynamic Spatial Disaggregation Approach to Geo-Temporal Crime Forecasting. *The Ninth Crime Mapping Research Conference*, Pittsburgh, USA, US Department of Justice, 2007.

[5] Ivaha, C., Al-Madfai, H., Higgs, G., Ware A. & Corcoran J., The simple satial disaggregation approach to satio-temporal crime forecasting. *International Journal of Innovative Computing, Information and Control (IJICIC)*, **3**(3), pp. 509–523, 2007.

[6] AL-Madfai, H., Ivaha, C. & Ware, A., Hierarchical Profiling of daily crime time series data as a precursor to modelling. *International Symposium on Forecasting*, San Antonio, Texas, 2005.

[7] Al-Madfai, H., Ameen, J. & Ryley, A., The Hierarchical profiling approach to STLF of multi-year daily electricity demand in South Wales. *International symposium on forecasting*, Sydney, 2004.