Available online at www.sciencedirect.com
SCIENCE DIRECT®
Atmospheric Environment 39 (2005) 2145–2155
ATMOSPHERIC ENVIRONMENT
www.elsevier.com/locate/atmosenv

Generalised additive modelling of air pollution, traffic volume and meteorology

Magne Aldrin*, Ingrid Hobæk Haff

Norwegian Computing Center, P.O. Box 114 Blindern, N-0314, Oslo, Norway

Received 22 September 2004; received in revised form 2 December 2004; accepted 17 December 2004

Abstract

We present a general model where the logarithm of hourly concentration of an air pollutant is modelled as a sum of non-linear functions of traffic volume and several meteorological variables. The model can be estimated within the framework of generalised additive models.

Although the model is non-linear, it is simple and easy to interpret. It quantifies how meteorological conditions and traffic volume influence the level of air pollution. A measure of relative importance of each predictor variable is presented.

Separate models are estimated for the concentration of PM_{10}, $PM_{2.5}$, the difference PM_{10}–$PM_{2.5}$, NO_2 and NO_x at four different locations in Oslo, based on hourly data in the period 2001–2003. We obtain a reasonably good fit, in particular for the largest particles, PM_{10} and PM_{10}–$PM_{2.5}$, and for NO_x. The most important predictor variables are related to traffic volume and wind. Further, relative humidity has a clear effect on the PM variables, but not on the NO variables. Other predictor variables, such as temperature, precipitation and snow cover on the ground are of some importance for one or more of the pollutants, but their effects are less pronounced.
© 2005 Elsevier Ltd. All rights reserved.

Keywords: Air quality modelling; Urban air quality; Particulate matter; Nitrogen oxides; Forward validation

1. Introduction

The growing health problems caused by traffic-related air pollution has resulted in an increased interest in analysis and prediction of the air quality. Several methodologies, both deterministic and statistical, have been proposed. These are often based on linear or non-linear regression models where the concentration of an air pollutant at a specific site is related to traffic volume and meteorological variables.

Levy et al. (2003) relates the concentrations of $PM_{2.5}$, ultra-fine particles and polycyclic aromatic hydrocarbons to traffic volume, wind direction and distance from the road, using linear mixed effects regression models. Chaloulakou et al. (2003) use linear regression to relate PM_{10} and $PM_{2.5}$ concentrations to predictor variables as temperature, wind speed, wind direction, time of year and day of week. They recognise that the meteorological variables are non-linearly related to the concentrations of PM_{10} and $PM_{2.5}$. To handle this, they convert the meteorological predictor variables into binary variables which are used as predictor variables in a modified linear model. Several authors use non-linear methods. Gardner and Dorling (1999), Kukkonen et al. (2003) and Schlink et al. (2003) all conclude that neural networks (see for instance Ripley (1996) for a general reference) are superior to linear techniques in predicting PM_{10}, NO_2,

*Corresponding author. Tel.: +47 22 85 25 00.
E-mail address: magne.aldrin@nr.no (M. Aldrin).

1352-2310/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2004.12.020

NO_x or ozone concentrations from several meteorological variables. In addition, Schlink et al. (2003) compares neural networks to several other methods, including generalised additive models (GAM, Hastie and Tibshirani, 1990). They conclude that the precision of neural networks and generalised additive models is comparable, and that both methods outperform linear ones due to their ability to model static non-linearities.

Our aim is to present a general statistical model to approach two important issues so far left unresolved: quantifying the effects of various predictor variables on the concentration of air pollution variables, and showing how the results from several sites and for several pollutants can be presented simultaneously in a comprehensive way. Our basic model is a generalised additive model with Gaussian response. Because we want to assess the specific contributions to the pollutant of various variables, we prefer a generalised additive model having a simple and explicit formulation of the response–predictor relationships to for instance a neural network model. The model is

$$\log(y_t) = s_1(x_{1t}) + \cdots + s_p(x_{pt}) + \varepsilon_t, \quad (1)$$

where y_t is a univariate pollution variable, $s_i(\cdot)$ are unknown, but smooth functions that must be estimated, and x_{it} are the predictor variables, i.e. traffic volumes, meteorological conditions and time-related variables. Finally, ε_t is the residual, i.e. the part of $\log(y_t)$ that is unexplained by the model. The logarithmic transformation applied to the air pollutants is also used by Chaloulakou et al. (2003) and Schlink et al. (2003). It makes the data more homoscedastic and ensures that all predicted values are positive on the original scale.

Separate models are estimated for hourly measurements of concentrations of PM_{10}, $PM_{2.5}$, the difference PM_{10}–$PM_{2.5}$, NO_2 and NO_x for four different locations in Oslo in the period from 2001 to 2003. The degree of smoothness of the s-functions is controlled by tuning appropriate smoothing parameters. Less smoothness gives better fit to data, but may result in over-fitting. To ensure a reasonable degree of smoothness, our choice of smoothing parameters is guided by forward validation, a modification of cross-validation.

2. Data

The data set consists of pollution data from four different locations in Oslo, namely Manglerud, Furuset, Løren and Alnabru, for the period from 1 November 2001 to 31 May 2003, with corresponding measurements of traffic volume and meteorological conditions. These four locations are situated near roads with rather heavy traffic, see Table 1 for an overview. All data were collected hourly. All data series contain periods of missing observations. Withdrawing these periods, there are left between 4000 and 9000 h of observations for the different pollution variables at the various locations. The data have been collected by The Norwegian Public Roads Administration.

The pollution variables were measured as concentrations with unit $\mu g\,m^{-3}$, and are presented in Table 1. The traffic volumes are the total number of vehicles passing the measurement site in both directions every hour. These were counted directly at two of the pollution measurement sites, whereas the two other sites were

Table 1
Summary of pollution and traffic data

Pollution measurement site	Manglerud	Løren	Furuset	Alnabru
PM_{10}	Yes	Yes	Yes	Yes
$PM_{2.5}$	Yes	Yes	Yes	No
PM_{10}–$PM_{2.5}$	Yes	Yes	Yes	No
NO_2	Yes	Yes	Yes	Yes
NO_x	Yes	Yes	Yes	Yes
Description	4 m south-east of highway E6	4 m north-east of highway Ring 3	10 m north-west of highway E6	4 m west of a municipal main road in an area with heavy traffic
Distance from Valle Hovin (met. station)	3 km	1 km	5 km	3 km
Corresponding traffic count site	Manglerud	Løren	Karihaugen	Karihaugen

Table 2
Summary of meteorological data

Variable	Unit	Comment
Temperature 2 m above ground	°C	Average
Temperature 25–2 m above ground	°C	Average
Wind direction 10 m above ground	deg	Average, 0 = wind from north
Wind speed 10 m above ground	$m\,s^{-1}$	Average
Relative humidity	%	Average
Precipitation	$mm\,h^{-1}$	Sum
Snow cover indicator		Take values from 0 to 3: 0: no snow, 1: 1–50% snow coverage, 2: 50–99% snow coverage, 3: 100% snow coverage.

related to the traffic volume at a nearby count site (Karihaugen), which is situated at the same highway as the Furuset site. This is summarised in Table 1.

The meteorological variables are listed in Table 2. All of these except one were observed at Valle Hovin, situated between 1 and 5 km from the four pollution measurement sites. The snow cover indicator was observed at Blindern, 5–10 km west of the pollution measurement sites.

Fig. 1 shows the pollution and traffic data from Manglerud along with the meteorological data from the first 26 weeks of the data period.

Some minor preprocessing was made to the data, to clear them for negative values and other incoherences. The details are given in Appendix A.

3. Methods

We have modelled each of the five pollution variables, PM_{10}, $PM_{2.5}$, the difference PM_{10}–$PM_{2.5}$, NO_2 and NO_x, separately using model (1) with the predictor variables given in Table 3, with one exception: snow cover was not included as a predictor variable in the NO_x model for Alnabru, since data were available for one winter period only, giving an unstable estimate of the snow cover effect.

The two precipitation variables in Table 3 need some comments. The precipitation last 4 h is intended to take care of the actual effect of precipitation in the air as well as the effect of a dry or wet road. The precipitation last week is meant to describe the effect of abundant precipitation, assuming that it may wash the polluting particles away from the road. The weights w_j (see right column of Table 3) are linearly decreasing to ensure that the hours closest in time have most influence.

The predictor variable "hour of day" should take into account diurnal variation that is not explained by the other variables, such as traffic and temperature, but has no interpretation of its own. The predictor variable "day number" is meant to take care of long-term variation, including seasonal variation, that is not explained by the other predictor variables.

The predictor variables are moderately correlated. The correlation between the snow cover indicator and the temperature is −0.56, between the relative humidity and temperature, it is −0.43 and between the number of vehicles and "hour of day", it is 0.39. All other correlation coefficients are 0.31 or less in absolute values. Based on these moderate correlations, we do not expect any serious problems with confounding between predictor variables.

For given values of the smoothing parameters (see Appendix B), the s-functions in model (1) are estimated by the method of least squares within the framework of generalised additive models (Hastie and Tibshirani, 1990), using the software package Splus (version 6.1.2, Insightful corporation, Seattle, WA).

The residuals ε_t will in practice be autocorrelated. This could be handled by some time series model. It is important for prediction, but has little effect on the estimation of the s-functions. According to the theory of generalised estimating equations (see for instance Liang and Zeger, 1986), the estimates are consistent even though the autocorrelation is ignored. We have therefore chosen not to include a model for the residuals. If forecasting were the purpose, appropriate modelling of ε_t would be necessary.

Model (1) is additive on log-scale, and can be transformed back to a model with multiplicative effects on the original scale as

$$y_t = S_1(x_{1t}) \cdot S_2(x_{1t}) \cdots S_p(x_{pt}) \cdot E_t, \quad (2)$$

where $S(\cdot)$ is $\exp(s(\cdot))$ and $E_t = \exp(\varepsilon_t)$.

Generalised additive models are well suited for this type of application, due to their ability to describe the

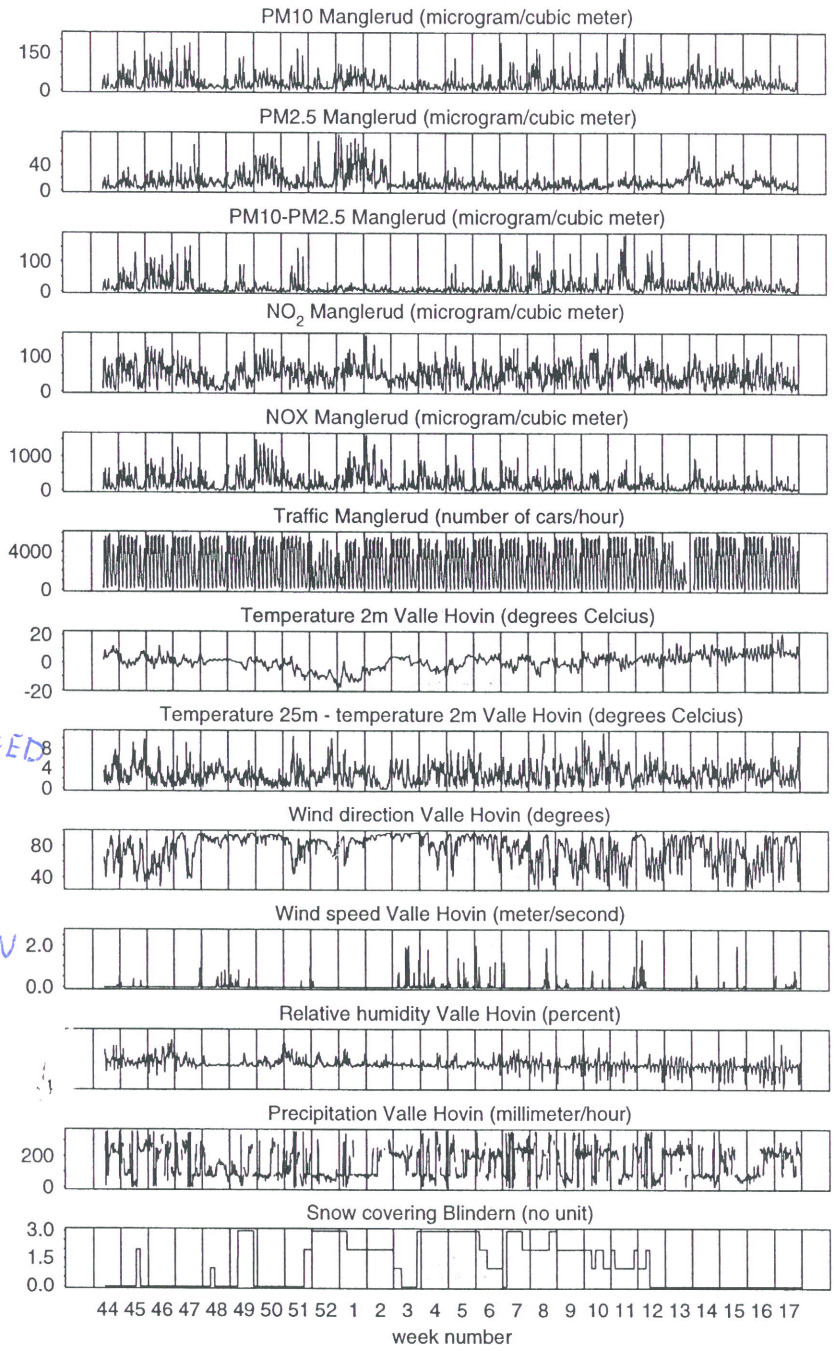

Fig. 1. Pollution data from Manglerud with corresponding traffic and meteorological data from week number 44, 2001 to week number 17, 2002.

so-called static non-linearities, i.e. non-linear effects which are stable over time. They are easy to interpret, since each predictor variable enters the model separately in an additive structure. However, interactions are more difficult to handle. Our model (1) contains only main effects. Potentially important interactions, for instance between wind direction and wind speed, are ignored. In comparison, neural-network models may describe both non-linearities and interactions, but it is difficult to sort out and quantify the separate effect of each predictor variable in such models.