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.

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

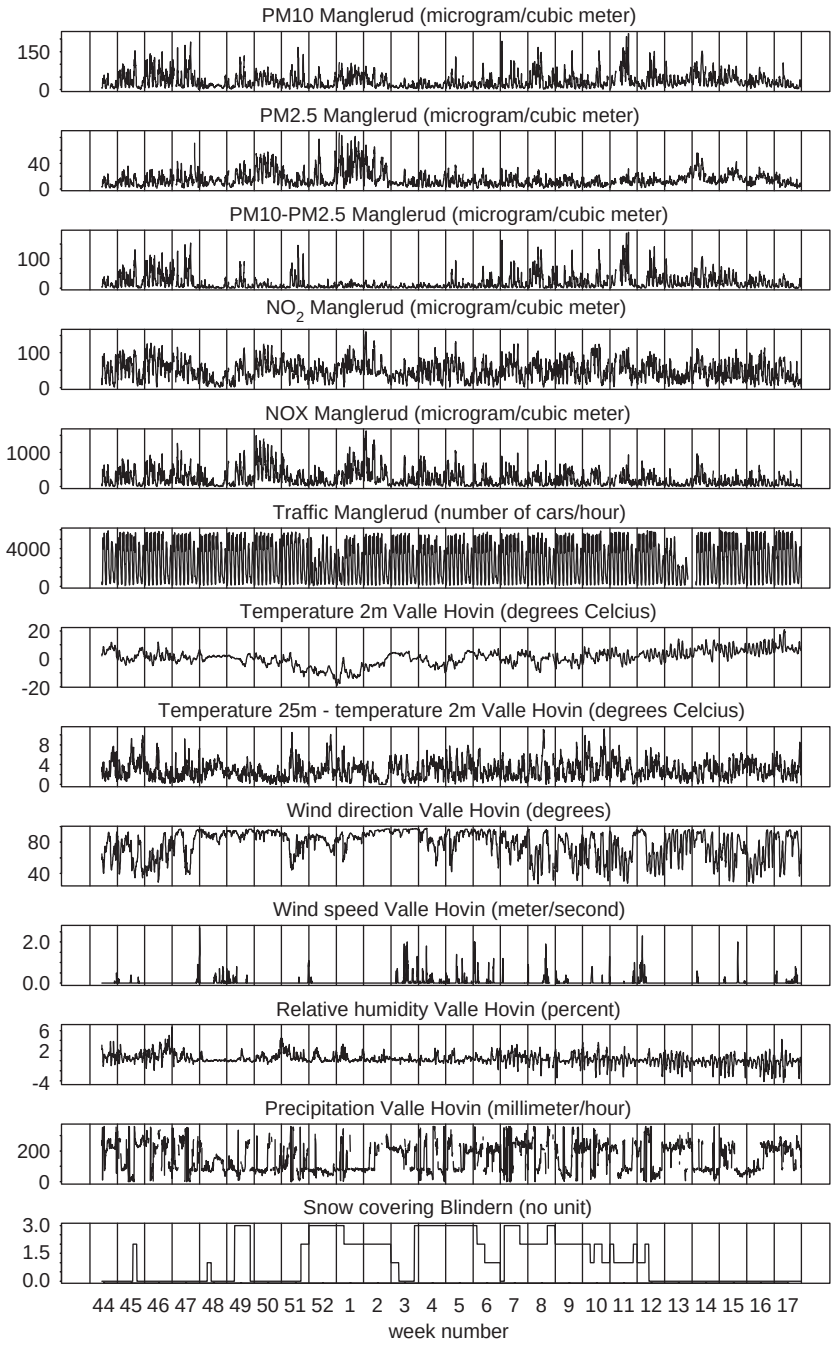

Fig. 1. Pollution data from Manglerud with corresponding traffic and meteorological data from week number 44, 2001 to week number 17, 2002.

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.

We have calculated the squared correlation coefficient R^2 for each model on the log-scale. In addition, we have calculated a measure of relative importance of the predictor variables as described in Appendix C.

4. Results

Based on the data presented in Section 2, the additive model (1) was estimated for the various pollution variables at the four different locations in Oslo.

Table 4 shows the explained variation (the squared correlation coefficient) R^2 for the models on log-scale. The values are between 0.48 and 0.80, indicating that the models explain most of the variation in the pollution data, but that there still is considerable unexplained variation. The highest values of R^2 are found at Løren, which is the pollution measurement site that is closest to the meteorological measurement site Valle Hovin. Moreover, R^2 is larger for PM_{10} and PM_{10}–$PM_{2.5}$ than for $PM_{2.5}$, and larger for NO_x than for NO_2.

Fig. 2 shows the relative importance of each predictor variable for each model. It can be summarised as follows: The number of vehicles is very important for all pollutants. The temperature has some effect on all pollutants, whereas the temperature difference affects the NO variables, but not the PM variables. Wind direction and wind speed have large effect on all pollutants. Relative humidity is important for particulate matter, but not for nitrogen oxides. Most of the precipitation effects are due to the precipitation last week and is clear, though not very large, on the PM variables and more diffuse on the NO variables. Snow cover has some effect on PM_{10}–$PM_{2.5}$, but is of little importance for the other pollutants. Both day number and hour of day affect all pollutants to some extent.

Table 3
Predictor variables

Predictor variable	Definition
Number of vehicles	
Temperature	
Temperature difference	25–2 m
Wind direction	
Wind speed	
Relative humidity	
Precipitation last 4 h	$1/10(4P_t + 3P_{t-1} + 2P_{t-2} + 1P_{t-3})$
Precipitation last week	$1/(\sum_{j=1}^{j=168} w_j)\sum_{j=1}^{j=168} w_j P_{t-3-j}$, where $w_j = 169 - j$
Snow cover indicator	
Hour of day	values from 1 (00:00–01:00) to 24 (23:00–24:00)
Day number	values from 1 (1 November 2001) to 577 (31 May 2003)

The actual estimated smooth non-linear curves are shown in Figs. 3 and 4. The figures are organised as matrices with one column for each pollutant (named at the top of each column) and one row for each predictor variable (named to the right of each row). Results are presented as relative effects on the original scale. Relative effects are easy to compare between pollutants as well as between predictor variables. More specifically, the displayed curves are $100S(x)/S(x^{\text{ref}})$, where x is the predictor variable of interest, $S(\cdot)$ is the corresponding estimated smooth function in model (2) at original scale, and x^{ref} is a chosen reference value of x. Thus the curve is set to 100 at the reference value. The predictor variable of interest is varied from the lowest to the highest observed value, while the other explanatory variables are kept fixed. Confidence intervals could be calculated for the curves, but are not shown here, since they would make the plots too confusing. Instead, the spread between the four measurement stations gives an impression of the variability.

The upper left panel of Fig. 3 shows how the concentrations of PM_{10} varies as the number of vehicles changes and all other conditions are kept fixed. Each curve corresponds to one of the measurement stations. The curve for Løren is the one going farthest to the right, meaning that it is the location where the highest number of vehicles was observed. The interpretation of the curve is as follows: when the traffic volume increases from 1000 (reference value) to 7000 cars per hour and all other conditions are constant, the concentration of PM_{10} at Løren grows from 100 to 200, i.e. doubles. The other panels on the upper row show corresponding curves for $PM_{2.5}$, PM_{10}–$PM_{2.5}$, NO_2 and NO_x.

The remaining rows of the figure present the estimated effects of the other predictor variables, in order with reference values in parenthesis, temperature (0 °C),

Table 4
The squared correlation coefficient R^2 for each model on log-scale

Measurement station	PM_{10}	$PM_{2.5}$	PM_{10}–$PM_{2.5}$	NO_2	NO_x
Manglerud	0.58	0.55	0.61	0.59	0.64
Løren	0.72	0.62	0.76	0.77	0.80
Furuset	0.63	0.56	0.65	0.65	0.69
Alnabru	0.48	—	—	0.59	0.70

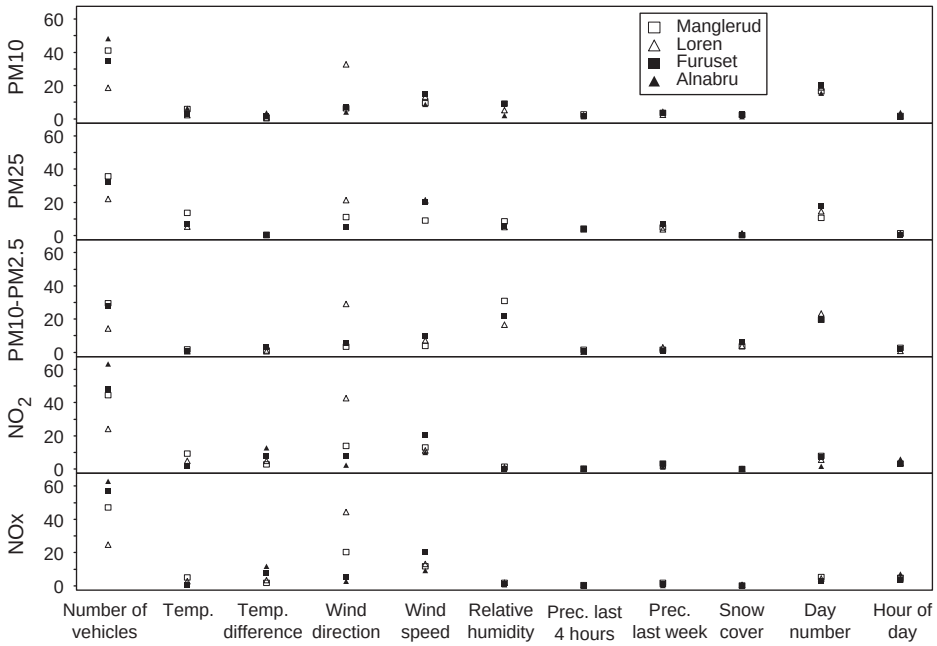

Fig. 2. Relative importance (in %) of each predictor variable within each model.

temperature difference (0 °C), wind direction (0°), wind speed (0 m s^{-1}) and relative humidity (97%). The rows of Fig. 4 display the effects of precipitation the last 4 h (0 mm h^{-1}), precipitation the foregoing week (0 mm h^{-1}), snow cover (0), day number (396 days, corresponding to the 31 September 2002) and hour of day (1). It should be noted that the range of the y-axis is the same in all the plots of a given predictor variable.

In the following we comment in Fig. 3 in more detail.

The upper row of the figure shows that increasing traffic volume corresponds to increasing air pollution. This result is obviously as expected, and is consistent to the data of Keary et al. (1998), showing positive correlation between PM_{10} and traffic volume. Apart from this main trend, there is a rather large variation in the estimated effects, both between the different locations and the pollution variables. NO_x is the pollutant that is most affected by the traffic, with an approximately linear relationship. Hence the concentration of NO_x is almost proportional to the number of vehicles. This is reasonable, since NO_x includes both NO, coming directly from the exhaust and NO_2 that is created when NO reacts with oxygen. The coarse fraction component PM_{10}–$PM_{2.5}$ mainly comes from particles whirled up from the road, whereas a large share of the fine particles in $PM_{2.5}$ originates from wood-burning in fire places. This explains why PM_{10}–$PM_{2.5}$ has a stronger relationship to the traffic volume than $PM_{2.5}$.

The effect of temperature seems to be rather similar at the different locations, but varies among the pollutants. Low temperatures are strongly associated with high concentrations of $PM_{2.5}$, but temperature seems to be rather unimportant for the difference PM_{10}–$PM_{2.5}$, which is consistent with the fact that wood-burning results in higher concentrations of $PM_{2.5}$. Keary et al. (1998) found a negative correlation between PM_{10} concentrations and temperature, which is also seen in our results. For NO_2 and NO_x, the tendency is that the concentrations are lowest around 0°.

The temperature difference seems to be of little importance for the PM variables. However, for NO_2, and even more so for NO_x, a positive temperature difference corresponds to high concentrations. A positive temperature difference means that the air is hotter 25 m above ground than 2 m above, such that the air stays near the ground. The results are therefore sensible.

The wind direction appears to have a large influence on the air pollution, which is reasonable, and also found in Levy et al. (2003). The estimated curves are however quite different for the various locations. This is as expected, since the effect of wind direction is very local, depending on the exact location of the measurement equipment relative to the road. For all four measurement sites, the highest concentrations are found when the wind blows from the road towards the measurement equipment (cfr. Table 1). $PM_{2.5}$ is the pollutant that is least affected by the wind direction. A plausible explanation is again that $PM_{2.5}$ is caused by wood-burning rather than traffic.

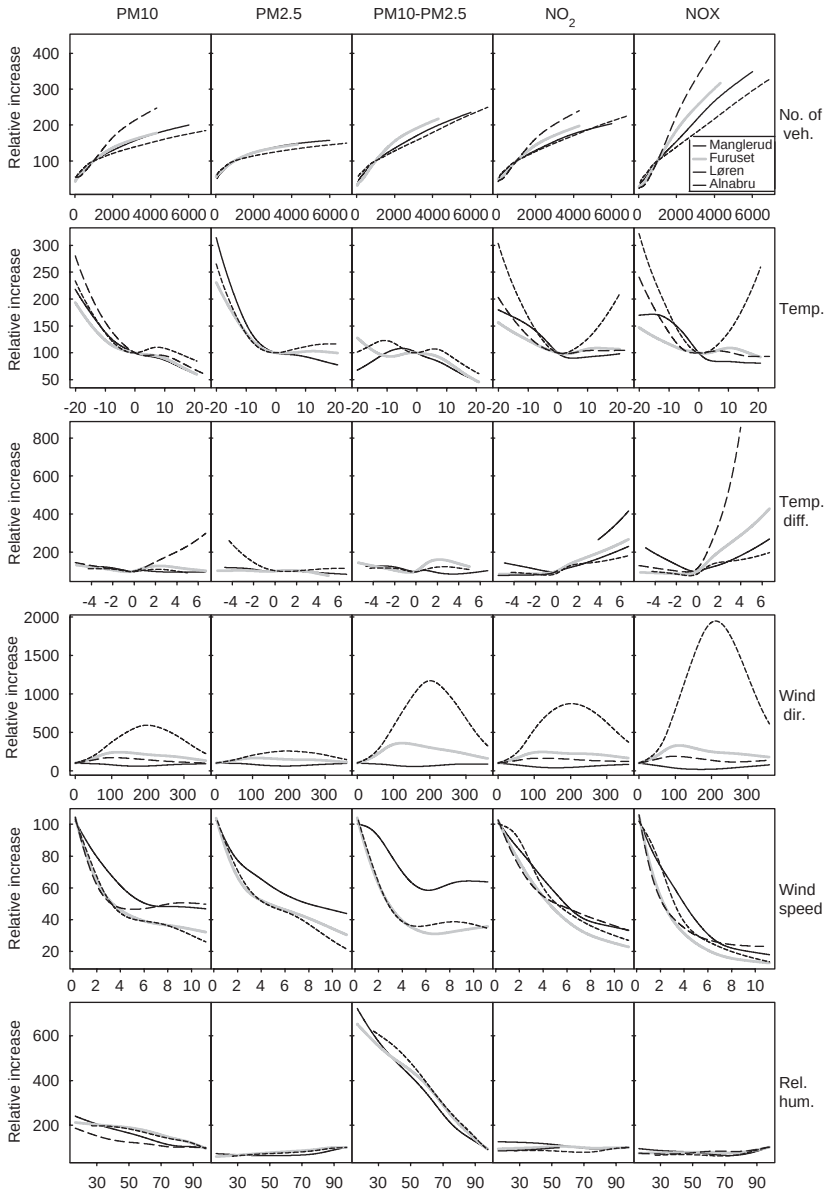

Fig. 3. Estimated effects (on the original scale) of traffic, temperature, temperature difference, wind direction, wind speed, and relative humidity.

For rather low values of the variable, wind speed has a pronounced effect; increasing speed corresponds to decreasing air pollution. The tendency is particularly marked for NO_x. This is as anticipated, and agrees with previous findings of Levy et al. (2003), Chaloulakou et al. (2003) and Keary et al. (1998). For the particulate matter, the curves tend to flatten out for higher wind speeds, especially for PM_{10}–$PM_{2.5}$. A possible explanation is that particles are whirled up because of a high wind speed, instead of being blown away.

The results for relative humidity are very clear. For PM_{10}, high humidity corresponds to low concentrations. In other words, the curves go downward for increasing humidity. For $PM_{2.5}$, the effect is opposite, and the curves are directed slightly upwards. This sums up to strongly downwards inclined curves for the difference PM_{10}–$PM_{2.5}$. Hence, the effect of humidity is considerable on the largest particles. However, it seems to have no particular influence on the NO_2 and NO_x concentrations.

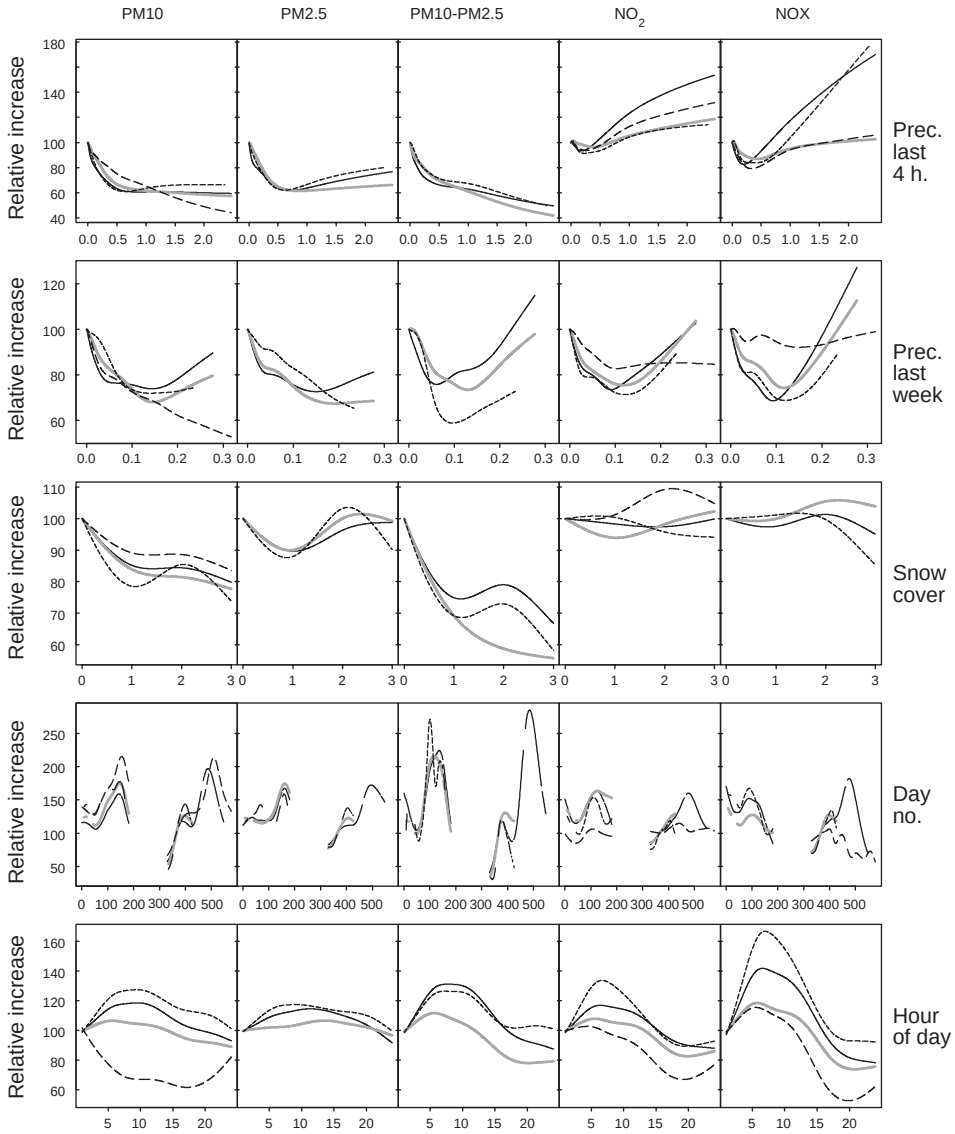

Fig. 4. Estimated effects (on the original scale) of precipitation last 4 h, precipitation last week, snow cover, day number, and hour of day.

The overall impression is that increasing precipitation induces lower concentrations of particulate matter. This is again consistent with the data of Keary et al. (1998), who found a negative correlation between the concentration of PM_{10} and rainfall. However, the effect seems to be the opposite for high values of precipitation last week, which may be difficult to explain. The curves for NO_2 and NO_x are rather widespread and difficult to interpret. Large amounts of precipitation appear to correspond with higher concentrations, and we have found no obvious reason for this.

When the ground is covered with snow or ice, the concentrations of PM_{10} and particularly PM_{10}–$PM_{2.5}$, tend do be smaller, which indicates that large particles are captured into the snow. The other concentrations are apparently not much influenced by the snow cover, which seems to be reasonable at least for the nitrogen oxides.

The day number is supposed to describe varying pollution level over time, including seasonal variation, that is not explained by the other variables. Most of the curves seem to have a top near day number 150 (March/April 2002) and a corresponding top around day number 520, that is 1 year after. This suggests that there is a phenomenon in spring at all four locations that the other parts of the model do not account for. One

possible explanation is that pollutants have been bound to the snow, and are released when the snow is melting. However, 2 years of data is too little to draw any conclusions about seasonal effects of this type. The gap in the curves between roughly 200 and 350 are caused by lack of meteorological data between week 19 and 39 year 2001.

The variable "hour of day" is not in itself interpretable, but should describe the diurnal variations not explained by the rest of the model. It appears to have some effect, especially on NO_x, for which the concentrations in the morning tend to be higher than in the evening, correcting for all other variables. Hence, although other variables (traffic in particular) model most of the diurnal variation, there is some left. One of the reasons may be interactions between some of the variables, and, as earlier mentioned, interactions are not accounted for in this model.

5. Summary and conclusion

This work presents a way to estimate the relationship between each of five pollution variables, namely concentrations of PM_{10}, $PM_{2.5}$, PM_{10}–$PM_{2.5}$, NO_2 and NO_x, and traffic, as well as a set of meteorological variables. The model used is additive on the log-scale, resulting in a multiplicative model on the original scale. The estimates were made based on hourly data collected during a period of one and a half year at four different locations in Oslo.

The estimated models gave a reasonably good fit in terms of the squared correlation coefficient. For all stations, the models for PM_{10} and PM_{10}–$PM_{2.5}$ and for NO_x had more explanatory power (higher values of R^2) than the ones for $PM_{2.5}$ and NO_2, respectively.

Even though some of the estimates from different locations are rather spread, there are some general trends. Traffic volume has, as expected, a highly significant effect on air pollution, especially on NO_x. For temperatures below 0 °C, concentrations, of PM_{10} and $PM_{2.5}$ in particular, increase as the temperature drops. A positive temperature difference between 25 and 2 m above ground raises the concentrations of the two gases NO_2 and NO_x. Wind direction has a very significant, but local effect on the pollution. Higher wind speed reduces the air pollution. Relative humidity decreases the concentration of PM_{10} and PM_{10}–$PM_{2.5}$, and increases the concentration of $PM_{2.5}$ as it increases, but has no effect on NO_2 and NO_x. Overall, heavier precipitation tends to reduce the PM concentrations, and have a more diffuse effect on the NO concentrations. The concentration of PM_{10}–$PM_{2.5}$, and to a lesser extent PM_{10}, is smaller when the ground is covered with snow or ice. Diurnal and seasonal variations of the concentrations were mostly described by variables such as traffic and temperatures. However, some systematic temporal variations remained.

Everything taken into consideration, the relations between air pollution, traffic and meteorological variables are quite well estimated using generalised additive models. These models combine non-linearity, which is to be found in these relations, and interpretability. The estimates for different locations have many traits in common, hence similar models may be estimated for other locations.

However, potential interactions have been ignored. For instance, in reality, there will obviously be an interaction between wind direction and wind speed, since wind direction will have no effect when the wind speed is very low. Fortunately, pairwise interactions can be handled within the same framework and the same software, by replacing the two terms $s_i(x_{it}) + s_j(x_{jt})$ with an interaction term $s_{ij}(x_{it}, x_{jt})$ (Cleveland and Devlin, 1988; Currie et al., 2004). Interpretation of the effect of the predictor variables included in the interactions becomes slightly more difficult, however.

Our focus has been on interpretation of empirical relationships, but the models might also be useful for forecasting, for instance one day ahead. The residuals ε_t from our models are clearly autocorrelated. This correlation should be modelled, for instance by some autoregressive model, in order to reduce the forecast uncertainty. Further, forecasting the air pollution concentrations requires reliable forecasts of the predictor variables. Forecasts of the meteorological variables are often available on a routinely basis, but their uncertainty will probably contribute considerably to the total uncertainty. On the other hand, the traffic patterns are rather stable over time, making the number of vehicles easy to predict with reasonable precision. Finally, the future values of the variables day number and hour of day are known exactly, so forecasts for these are not required.

Acknowledgements

This work was sponsored by the Norwegian Public Roads Administration and the Norwegian Research Council, Project 154079/420. We thank Arnoldo Frigessi for helpful comments.

Appendix A. Modifications of data

Due to measurement errors, some negative values of pollution concentrations appear in the raw data. Values less than $-5\,\mu g\,m^{-3}$ are regarded as errors and are replaced by missing values, whereas negative values above that limit are accepted. However, since the pollution data are modelled at log-scale, both negative

values and values at or near 0 cause problems. Therefore, measured concentrations between a small positive value δ and $-5\,\mu g\,m^{-3}$ are set to δ, where $\delta = 2\,\mu g\,m^{-3}$ for PM_{10}, NO_2 and NO_x, and $\delta = 1\,\mu g\,m^{-3}$ for $PM_{2.5}$ and PM_{10}–$PM_{2.5}$.

Further, measurements of particulate matter around midnight on New Year's Eve, more specifically 4 h before and 20 h after midnight, have been removed. The reason is that the values measured during those hours were extremely high, due to the fireworks.

The snow cover indicator is measured daily. Hourly values are constructed by repeating the daily value 24 times.

Regarding the traffic volume, counts with 0 vehicles per hour are treated as missing values, since 0-values on these large roads probably are recorded due to errors on the measurement equipment.

The traffic volume and the two precipitation variables were actually transformed before their non-parametric s-functions were estimated (note that the results are shown on the original scale). The aim was to help the non-parametric smoother by a preliminary transformation. The traffic were log-transformed, whereas for precipitation we used log(precipitation+0.1) to avoid problems with 0 precipitation.

Appendix B. Choice of smoothing parameters

The smoothness of each function s_i in model (1) is controlled by a smoothness parameter, here expressed by the number of degrees of freedom or effective parameters for each function. This must be chosen before the function is estimated. As the number of degrees of freedom increases, the function becomes less smooth, but more flexible and it gives better fit to data. Hence, choosing the smoothness parameter is a trade-off between good fit and smoothness, or in other words, between bias and variance.

Even though prediction is not our primary aim, measuring prediction performance is still useful for model selection. Therefore, we have chosen the number of degrees of freedom by means of forward validation, which is a modification of cross-validation useful for time series. The forward validation is based on one-day-ahead hourly predictions of the concentration of PM_{10} at Alnabru for the period from 1 January 2003 to 31 July 2003. If the focus were on precise forecasting, it would probably be worthwhile repeating the forward validation experiment for other measurement stations and other pollutants. However, we found this unnecessary for our application, since we only use the results as a rough guide to get reasonably smooth estimates and prevent over-fitting.

For each day, and for a given set of degrees of freedom, the model is re-estimated using the data up to the day before. Then the hourly log PM_{10} concentrations for the next day is predicted, assuming that the predictor variables that day are known. The prediction is compared to the actual value, and the hourly prediction errors are calculated. This is repeated for each day from 1 January 2003 to 31 July 2003, and the root mean squared error of prediction (RMSE) is calculated. The whole procedure is repeated for various choices of degrees of freedom.

Since it would have been computationally too extensive to vary the degrees of freedom for all the functions independently, we restricted all the functions but the one corresponding to the variable day number to have the same number of degrees of freedom. Moreover, instead of varying the number of degrees of freedom for both day number and the other variables simultaneously, we first fixed it (to 10) for the day number, while we varied it for the other variables. Then, we fixed the number of degrees of freedom for the other variables at the value giving the minimum RMSE (10), and varied it for day number.

The results from the forward validation are shown in Fig. 5. The upper panel shows that the minimum RMSE occurs at 10 degrees of freedom for all variables but the day number, whereas the minimum RMSE is found at 15 degrees of freedom for the day number. However, using 10 degrees of freedom for the remaining variables resulted in strange curvatures in some of the estimated functions. Since we want smooth, easily interpretable functions, we reduced the number of degrees of freedom for these variables to 4 in the final models presented in Section 4, whereas the number of degrees of freedom for day number were kept at 15. This gives smoother functions, without increasing the RMSE considerably, according to the upper panel of Fig. 5.

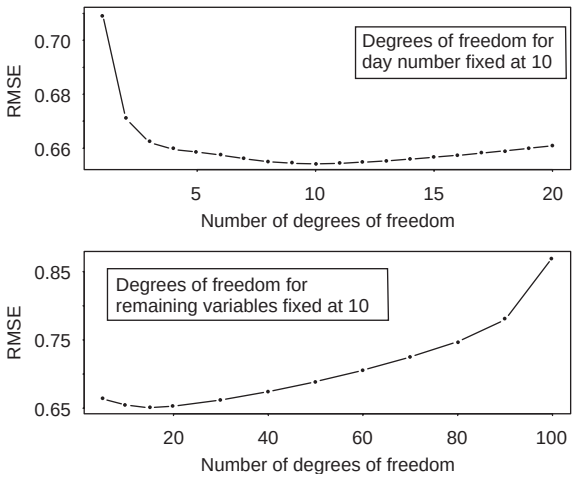

Fig. 5. Results from the forward validation on the number of degrees of freedom for all variables but the "Day number" (upper panel), and for the "Day number" (lower panel).

Appendix C. Measure of relative importance

Let $z_t = \log(y_t)$ be the observed response on the log-scale, and let $\hat{z}_t$ be the predicted response from the model. Then

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(z_t - \hat{z}_t)^2, \tag{3}$$

is the unexplained variation, i.e. the variation not explained by the model, where n is the number of observations. Further, let $\hat{z}_{t,(-i)}$ be the predicted response from a modified model where $s_i(x_{it})$ is replaced by the constant $s_i(\bar{x}_i)$, where $\bar{x}_i$ is the average of x_{it} over time.

Then

$$\hat{\sigma}^2_{(-i)} = \frac{1}{n}\sum_{t=1}^{n}(z_t - \hat{z}_{t,(-i)})^2 \tag{4}$$

is the unexplained variation when the effect of the ith predictor variable is ignored. The variation $\hat{\sigma}^2_{(-i)}$ will be larger than $\hat{\sigma}^2$, and the difference $\hat{\sigma}^2_{(-i)} - \hat{\sigma}^2$ can be interpreted as the variation explained by ith predictor variable. Summing the individual differences over all predictor variables gives

$$\sum_{i=1}^{p}(\hat{\sigma}^2_{(-i)} - \hat{\sigma}^2) = \sum_{i=1}^{p}\hat{\sigma}^2_{(-i)} - p\cdot\hat{\sigma}^2. \tag{5}$$

Then we define the proportion (in percent) of variation explained by the ith predictor variable is given by

$$100\,\frac{\hat{\sigma}^2_{(-i)} - \hat{\sigma}^2}{\sum_{i=1}^{p}\hat{\sigma}^2_{(-i)} - p\cdot\hat{\sigma}^2}. \tag{6}$$

There are alternative measures of relative importance, see for instance Doksum and Samarov (1995). In general, the various measures may give slightly different answers when the functions $s_j(x_j)$ are correlated.

If model (1) were restricted to be linear and the predictor variables were uncorrelated, the measure of relative importance given by Eq. (6) would be equivalent to the ith squared standardised regression coefficient divided by the sum of the squared standardised regression coefficients.

References

Chaloulakou, A., Kassomenos, P., Spyrelli, N., Demokritou, P., Koutrakis, P., 2003. Measurements of PM_{10} and $PM_{2.5}$ particle concentrations in Athens, Greece. Atmospheric Environment 37, 649–660.

Cleveland, W.S., Devlin, S.J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American Statistical Association 83, 597–610.

Currie, I.D., Durban, M., Eilers, P.H.C., 2004. Smoothing and forecasting mortality rates. Statistical Modelling 4, 279–298.

Doksum, K., Samarov, A., 1995. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. Annals of Statistics 23, 1443–1473.

Gardner, M.W., Dorling, S.R., 1999. Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. Atmospheric Environment 33, 709–719.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London.

Keary, J., Jennings, S.G., O'Connor, T.C., McManus, B., Lee, M., 1998. PM_{10} concentration measurement in Dublin city. Environmental Monitoring and Assessment 52, 3–18.

Kukkonen, J., Pratanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Cawley, G., 2003. Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmospheric Environment 37, 4539–4550.

Levy, J.I., Bennett, D.H., Melly, S.J., Spengler, J.D., 2003. Influence of traffic patterns on particulate matter and polycyclic aromatic hydrocarbon concentrations in Roxbury, Massachusetts. Journal of Exposure Analysis and Environmental Epidemiology 13, 364–371.

Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Schlink, U., Dorling, S., Pelikan, E., Nunnari, G., Cawley, G., Junninen, H., Greig, A., Foxall, R., Eben, K., Chatterton, T., Vondracek, J., Richter, M., Dostal, M., Bertucco, L., Kolehmainen, M., Doyle, M., 2003. A rigorous intercomparison of ground-level ozone predictions. Atmospheric Environment 37, 3237–3253.