INTERNATIONAL CENTRE FOR ECONOMIC RESEARCH

ICER

WORKING PAPER SERIES

P. Bertaccini, V. Dukic, R. Ignaccolo

MODELING THE SHORT-TERM EFFECT OF TRAFFIC ON AIR POLLUTION IN TORINO WITH GENERALIZED ADDITIVE MODELS

Working Paper No.10/2009

Modeling the short-term effect of traffic on air pollution in Torino with generalized additive models*

P. Bertaccini[†], V. Dukic[‡], R. Ignaccolo[§]

May 2009

Abstract

Vehicular traffic typically plays an important role in atmospheric pollution. This is especially true in urban areas, where high pollutant concentrations are often observed. In this paper, we consider hourly measures of concentrations of nitrogen oxides (NO, NO_2 and NO_x), carbon oxide (CO) and particulate matter (PM), collected at the stations distributed throughout the city of Turin. To help explain the short-term behavior of the concentrations of these pollutants, we propose using generalized additive models (GAM), focusing in particular on traffic along with the meteorological predictors. All the data are collected during the period from December 2003 to April 2005.

Keywords: urban area, air quality, vehicular traffic, CO, NO_2, NO_x, NO, PM, generalized additive models.

---

*The work of Dr. Bertaccini and Prof. Ignaccolo was partially supported by the Regione Piemonte CIPE project 2004 and the PRIN project n.2006131039. Prof. Dukic is very grateful to ICER for having hosted her in 2009 while this work was written. The authors would also like to thank 5T s.r.l., that provided support and free access to the vehicle traffic database, Regione Piemonte and Arpa Piemonte for everything that concerns Meteorological and Chemical variables.

[†]Dipartimento di Statistica e Matematica Applicata, Università degli Studi di Torino (Italy), C.so Unione Sovietica 218/bis, 10134, Torino, Italy. Tel: +39 011 670 5761; Fax: +39 011 670 5783 (bertaccini@econ.unito.it)

[‡]Department of Health Studies, University of Chicago (IL), 5841 S. Maryland Ave., MC 2007, Rm. W260, Chicago, Illinois 60637, Phone: 773-834-2172; Fax: 773-702-1979 (vdukic@health.bsd.uchicago.edu)

[§]Dipartimento di Statistica e Matematica Applicata, Università degli Studi di Torino (Italy), C.so Unione Sovietica 218/bis, 10134, Torino, Italy. Tel: +39 011 670 5758; Fax: +39 011 670 5783 (ignaccolo@econ.unito.it)

1 Introduction

The impact of air pollution on human health (Barnett et al. (2005); Bell et al. (2004); Samet et al. (2000)) and environment has been one of the central issues in environmental public policy and decision making. For example, European Union mission mandates yearly improvement of environmental quality, lower emission standards, and support of environmental technology and scientific research and development (Dir96/62/EC (1996); Dir99/30/EC (1999), Dir00/69/EC (2000) and Dir02/3/EC (2002)) while the recent air quality directive Dir50/08/EC (2008) requires that information on air quality for current day, with trend and forecast for the next days be publicly available. Similarly, the United States policy makers and industry leaders have recently begun instituting renewable energy and environmental protection research programs at universities and state agencies across the country.

Understanding the behavior of pollutants, and understanding the components of variation in pollutant concentrations are arguably the most important goals of air quality research for public policy purposes. For example, understanding how pollutant concentrations vary with respect to intensity and patterns of traffic would allow policy makers to assess the consequences of implementing certain traffic regulation measures. However, if an intervention such as traffic measure is being considered or evaluated, it is crucial to also account for those processes which co-vary with the outcome (pollutant) as well as with the regulatory (traffic) variable. In the studies of traffic and air pollution such confounding processes could include meteorological, health, social and other societal-level processes that affect both pollution and traffic volume. Those confounders are unfortunately often unobserved – such as flu or other infectious disease activity that makes people stay at home more and drive less, and also happens to occur in winter when smog and air pollution are high – and thus the level of their covariation with traffic patterns and also with the pollution are difficult to ascertain. However not accounting for those confounders at all would hide the true effects of interest and yield biased estimates of the regulatory effects.

In the Torino region, previous analysis of pollution have examined carbon monoxide (CO) concentrations and traffic volume in the Torino metropolitan area, as in Bertaccini et al. (2007) who used a seasonal linear regression model for each station monitoring CO. Subsequently, Fassó et al. (2007) studied the same problem using a linear vectorial auto-regressive model and carried out a sensitivity analysis to describe the relative roles of traffic and meteorology, by their respective principal components.

However, sometimes in modeling city-level processes, (generalized) linear models are not the most adequate ones to use. Although chemical and

physical dynamics of processes is deterministic, local chaotic behavior can be very difficult to understand and to model properly. However, local processes, such as wind and topography of the city structures, can impact the pollutant distributions around the city significantly. Therefore, it would be advantageous to consider a statistical alternative to the deterministic differential equation based modeling of pollution. To that extent, generalized additive models, or GAM (Hastie and Tibshirani (1990)) offer an alternative which is capable of not only flexibly modeling relationship between pollution concentration and predictors, but also relationships between predictors. This approach could flexibly approximate complex physical and chemical relationships between processes co-varying with traffic and pollution. In addition, GAM can account for the smooth time-varying processes reflecting the confounders which vary slowly relative to the predictor of interest, by including "time" as a flexibly (but smoothly) modeled predictor.

While generalized additive models have been widely used as a standard method in studies of pollution and health (see for example the pioneering work by Schwartz (1994)), they have only recently been introduced into the air pollution modeling with traffic and meteorological covariates (Carslaw et al. (2007)). The authors find that one of the most important factor is the flexible interaction between wind speed and wind direction, due to the canyon effect of the nearby buildings. Their analysis has confirmed the important role of wind in pollutant dispersion and in describing the variation in pollutant concentration due to changes in meteorological conditions. Similarly, Aldrin and Hobæk Haff (2005) use generalized additive models for several different pollutants in different locations over the Oslo urban area, using traffic and meteorological observed data.

In this paper we present a set of models that are able to realistically explain much of the variation in the pollutant concentration while still yielding precise estimates of the effects of meteorology and traffic on pollution concentration. More specifically, building on the work in Bertaccini et al. (2008) and Bertaccini (2009), we propose the use of generalized additive models to analyze the space-averaged air pollutant concentration over Torino metropolitan area as a function of vehicular traffic, while adjusting for potential meteorological and other possibly unobserved confounders.

The paper is organized as follows: after a brief data description (Sec. 2), we describe the basic theory and some advantages of using the generalized additive models (Sec. 3). In Section 4 we introduce the model for the whole city area, and discuss the selection of the best model and the predictor subset for pollutant concentration. In addition, as traffic is measured in numerous sites throughout the city, we also discuss an optimal way to summarize traffic data (Sec. 4.1). Finally, specific models are proposed and results analyzed

for common pollutants, CO, NO, NO_2, NO_x, and PM (Sec. 4.2 and 4.3).

2 Data

Traffic data are provided by 5T s.r.l., working in the Torino city area with a widely distributed set of 500 "inductive loop " sensors (i.e. flow counting points), embedded in the surface of the roads. Inductive loops work by a simple principle of sensing the change in inductance – when a car (or another large metal object) passes over a loop, the car's presence changes the total inductance, and the loop sensor count goes up by one. Loop network is a part of the monitoring system UTOPIA/SPOT (Urban Traffic Optimization by Integrated Automation/System for Priority and Optimization of Traffic), designed to serve as an urban traffic control system as described in Kronborg and Davidsson (2000) and Wood (1993). Such a system operates as a framework implemented to improve both private and public transportation efficiency in the Turin metropolitan area. The network of available sensors is set up to monitor the vehicular traffic at the main intersections of the city road graph (Fig.1).

This extensive network allows us to observe the behavior of traffic over time at multiple points throughout the city. However, having so many measuring devices also means that many of the individual time series will have a non-trivial fraction of missing data, sometimes over large continuous periods of time. These "gaps"in the measurement series are most often due to road maintenance or to the repair of the sensors themselves. In such cases, the missingness can be treated as missing at random (independent of the pollutant levels).

Our traffic data, the number of vehicles that pass over a certain monitor within 5-minute intervals, have been aggregated into hourly counts. Specific subsets of all traffic time series have been chosen so that they all correspond to the outflow of traffic at any given crossroads (which also equals to the influx of traffic to the same crossroads), in order to avoid double counting of the vehicles. The availability of meteorological and chemical data constrains further our study period to December 19th, 2003, to April 27th, 2005, and the final dataset is thus composed of 107 hourly measurement time series.

In the analyses in this paper we use hourly city-wide averaged variables, focusing on the average traffic behavior of the city, as shown in Fig.2. The boxplots show typical features of the traffic trend at three different timescales: daily, weekly and yearly. In the daily scale we can see the strong difference in traffic magnitude between day-time and night-time; as well as high traffic intensity due to the morning and evening rush hour. The weekly

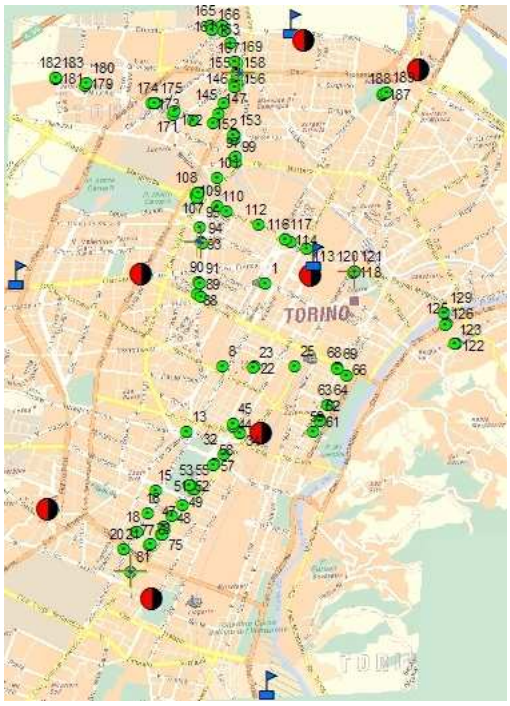
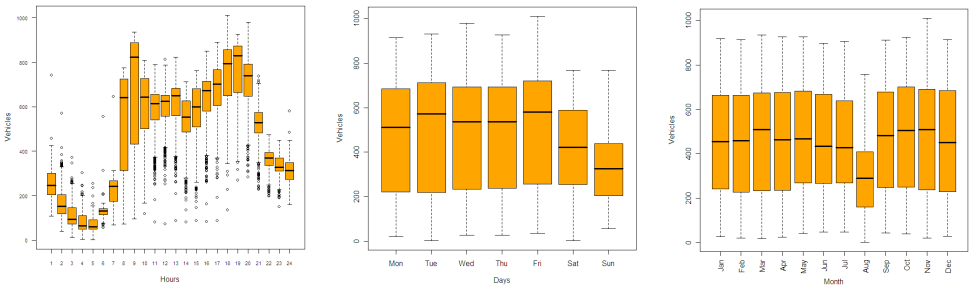


Figure 1: Turin vehicle traffic, air and meteorology monitoring network: green disk are the 107 traffic counters, red and black disks are the pollution stations and blue flags are the meteorological stations



(a) Hours (b) Days (c) Months

Figure 2: Box-plots of the city-wide average traffic volume for different time-scale.

scale shows the differences between weekdays and weekends: Saturday and Sunday traffic differs both in the total number of vehicles and the timings of the peak volume hours. Observing the yearly representation we can see that traffic is almost constant during the year except for the month of August where a sharp reduction is due to the summer holydays.

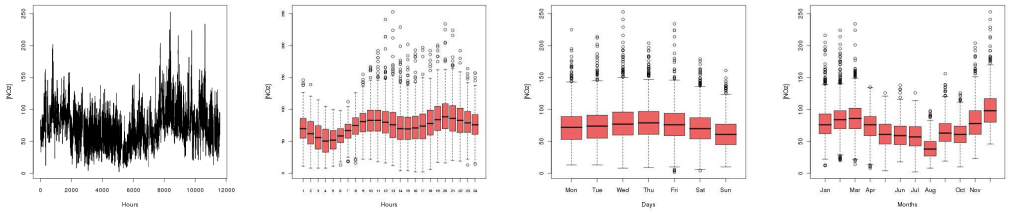
Arpa Piemonte and Regione Piemonte have provided pollution data. Hourly records at six different stations for NO_x, NO_2, NO and CO have been provided, while PM was measured by four sensors on a daily basis. The seven stations are *Grassi*, *Rebaudengo*, *Rivoli*, *Consolata*, *Cristina*, *Gaidano* and *Lingotto* located as shown in Fig.1, while sensors are distributed as in Tab.1.

Table 1: Available chemical sensors at the sites: Consolata (Con.), Rivoli (Riv.), Rebaudengo (Reb.), Cristina (Cri.), Gaidano (Gai.), Lingotto (Ling.), Grassi (Gra.).

	Con	Riv	Reb	Cri	Gai	Lin	Gra
CO	x	x	x	x	x	x	
NO	x	x	x	x	x	x	
NO_2	x	x	x	x	x	x	
NO_x	x	x	x	x	x	x	
PM_{10}	x	x			x		x

Pollution concentrations for the five different pollutants exhibit relatively similar behavior. In order to show a typical behavior of the pollutants, we summarize as an example the NO_2 concentration measured at the "Consolata" station (Fig.3). As can be seen in Fig.3.a, the lowest values happen during the middle of the month of August while the highest are during the two winters (recall that the study period is December 2003 through April 2005). The hourly box-plots of the concentration shown in Fig.3.b allow us to see that the concentration decreases during the night and has two peaks: one in the morning and one in the evening, related to commuter behavior. Note that this shape is pretty similar to the one observed for vehicular traffic (Fig.2.a), motivating the importance of using the hourly time-scale. As can be seen from the boxplots by day of the week (Fig.3.c) the concentration seems to increase in the first few weekdays and decrease during the weekend. The box-plots by month (Fig.3.c) confirm that the lowest values happen in August while the highest happen in the winter. Further explorative analyses on pollution features are available in Bertaccini (2009), Chapter 1.

Meteorological data are collected by four different stations as shown in Tab.2, the data are provided by Arpa Piemonte and Regione Piemonte. The locations of the meteorological stations are shown in Fig.1, marked with the



(a) Time Series (b) Hours (c) Days (d) Months

Figure 3: Time series and box-plots for NO_2 concentration measured at Consolata Station.

blue flags. For each variable we generally have at least three locations providing data at any given time. Hence, we have a rather reliable description of the meteorological conditions around the city. In addition, pressure generally differs very little across the entire Torino metropolitan area, so we can basically use the value measured by a single (*ReissRomoli* (*CSELT*)) station as representative of the city-wide pressure level.

Table 2: Available meteorological variables

	RRom (North)	Cons (Center)	Alenia (West)	Vall (South)
Press.	x			
Temp.	x	x	x	x
Rel. Hum.	x	x		x
Wind Sp.	x	x	x	
Wind Dir.	x	x	x	
Solar Rad.	x	x	x	
Rain	x	x	x	

Averages of the available variables are presented as time series in the following figures (Fig.4). As can be seen, meteorology is reduced to the city-wide vector (ME) containing wind speed (`wsp`), solar radiation (`sun`), relative humidity (`rh`), temperature (`tmp`) and pressure (`press`). Precipitation has not been included due to being composed of relatively rare and localized events and to having a rather limited impact on our results of interest (the sensitivity analysis was examined separately and is not shown in this paper). Moreover, wind direction has been omitted from the model due to the lack of a meaningful single "average" direction for the whole city, and the negligible effect observed on the model results (again examined separately and not shown). Finally in all our model we also consider the lagged (delayed) effects of some of the crucial meteorological variables, to account for the amount

time it takes for certain chemical and physical processes to realize and have an impact.

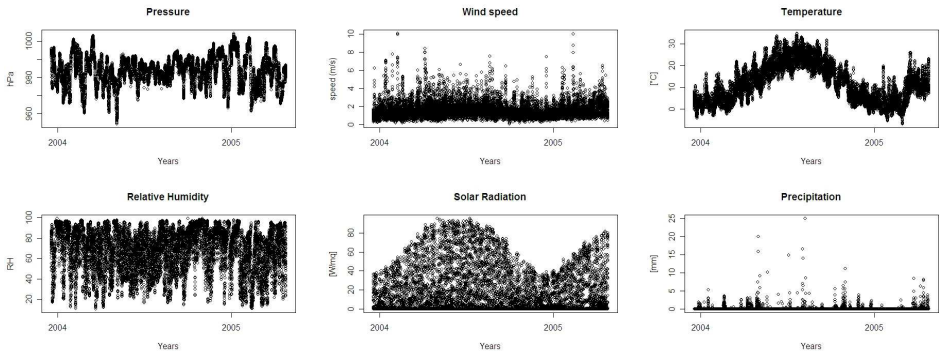


Figure 4: Time series of the averaged meteorological variables

In Fig.4 we present the time series of the averaged collected meteorological variables. Pressure generally shows variability over time which seems to have a shorter range during the summer. Wind speed is generally low, with some strong events that will turn out to be important in influencing the quality of air. Temperature as well as solar radiation shows the typical seasonal behavior with high values during the summer and low values during the winter. Relative humidity is generally conditioned by rainfall or wind events. Precipitation is relatively rare, with many days without rainfall and some occasional events.

3 Generalized additive models (GAM)

In modeling of air pollution, we will assume that transformed average outcome is additive in predictors, and can be appropriately modeled using Generalized Additive Models (GAM). GAMs have the advantage that they are able to describe nonlinear effects over time, and still be easily interpretable due to their additive structure. Moreover GAMs provide some flexibility via nonlinear or non-parametric terms but do not suffer from the curse of dimensionality like some other non-parametric methods such as kernel smoothing or polynomial modeling. For the outcome (eg., logarithm of pollutant, $Y = log(Pollutant)$), we assume that it is additive in its predictors and normally distributed with mean μ_t and variance σ^2. The systematic part μ_t could include linear and nonlinear components, as well as potential confounders.

A general model with additive component would then be

$$\begin{aligned} Y_t &\sim Gaussian(\mu_t, \sigma^2) \\ \mu_t &= \alpha + \sum_{f=1}^{l} \beta_f x_{f,t} + \\ &+ \sum_{g=1}^{m} \sum_{h \in H_g} \eta_{g,h} z_{g,t-h} + \sum_{i=1}^{p} s(k_{i,t}, \lambda_i) \end{aligned}$$

where α is the intercept, $\vec{x_t}$ are the current-time predictors, $\vec{\beta}$ are their effects, $z_{g,t-h}$ is the value of variable z_g h hours prior to the current time (with lag times taking values in set H_g), with an effect size $\eta_{g,h}$. Nonlinear effects are modeled non-parametrically trough smooth functions $s(\cdot, \lambda)$, where the smoothness is controlled by the scalar parameter λ.

In this study we model the aforementioned pollutants as time series representing the average level of pollution measured hourly or daily, where averaging is done over the available stations (the number of stations at each time changes depending on the pollutant under observation, Tab. 1). For each pollutant we consider the time series of the logarithm of the average pollutant concentration over Torino. Given that we wish to estimate the effect on pollution solely due to traffic, we pay special attention to potential confounders, which are related to both the concentration of the pollutant in the atmosphere and to the traffic volume itself. Meteorological variables are the typical confounders, and are routinely adjusted for in the pollution analyses. In GAM, we have the added flexibility of considering smooth functions of the meteorological variables, $s(ME, \lambda_{me})$. However, there are also potential unmeasured confounders which we have not observed, such as for example health and behavior patterns related to weather (and therefore pollution) and to traffic volume. Though these confounders are unobserved, we can assume that they are varying rather smoothly over time, or at least more smoothly than the predictor of interest (in this case traffic). In cases where such assumption is appropriate, we can proxy these unobserved confounders via a smooth function of time.

On the one hand, not adjusting for these unmeasured confounders will result in bias in the estimates of the effect of traffic. On the other, if we adjust too much (using a highly varying function of time), the effect of traffic may be conditioned away. Thus, a sensible model selection criterion which is capable of balancing goodness of fit with penalty due to complexity and high variability of confounder functions is crucial in choosing the optimal GAM model. Following the common trend, we use the Bayesian information criterion (BIC) Schwartz (1978). The BIC is like the AIC (Akaike (1978))

but with more severe penalization related to the complexity of the model. It takes the form of the penalized log-likelihood where the penalty is equal to the logarithm of the sample size n times the number of estimated parameters θ: $BIC = 2\ell_n(\hat{\theta}) - \log(n)\ length(\theta)$.

4 GAM models for Torino-wide pollution

Following the general principle mentioned in the section 3 we now develop semiparametric GAM models of pollutant concentration over the whole urban Torino area. Estimation of the penalty parameter in the model is obtained using the generalized cross validation method (Wood (2006), Wood and Augustin (2002)), and the best number of basis is found by comparing different models, using the BIC values.

Our main goal is to assess what the effective role of vehicular traffic on five different pollution species. In order to do that thoroughly, we propose different approaches to summarize traffic and select the most appropriate functional form for each pollutant. The selection of the suitable models is based on the information criterion BIC. We use this criterion to select the most important variables as well as the optimal number of spline basis for each covariate in the model.

Another important issues is related to cross-correlation between pollutants and some meteorological variables. This cross-correlation, when strong, suggests that one should use lagged variables into the model. In fact this often allows a substantial improvement of fit. In this paper, lagged variables have been dealt with in two ways: a) using a spline of the average of up to twelve previous values (lags 1-12), and b) using the splines for those individual lagged variables selected based on the highest correlation with the pollutant. Since the latter procedure almost always yields a better fit, we prefer it for modeling pollution in our study.

4.1 Assessing the effect of traffic via model selection

As we stated before, in this work we consider the log-pollution concentration as the outcome. Covariates are traffic and meteorology, moreover time is used as a proxy for unmeasured confounders. To assess the effect of traffic, we propose different models and compare the appropriateness of each of the model via BIC. The models we consider are those with or without meteorological or traffic variables, and adjusting for the unmeasured confounders with a varying-degree spline function of time (i.e. $s(t, \lambda)$). The three considered

models have an additive form as follows:

$$M_1 : \mu_t = \alpha + s(t, \lambda) + s(ME, \lambda_{ME}) \quad (1)$$
$$M_2 : \mu_t = \alpha + s(t, \lambda) + s(tr, \lambda_{tr}) \quad (2)$$
$$M_3 : \mu_t = \alpha + s(t, \lambda) + s(tr, \lambda_{tr}) + s(ME, \lambda_{ME}) \quad (3)$$

where ME indicate the involved meteorological variables: wind speed, solar radiation, relative humidity, temperature and pressure as $s(ME, \lambda_{ME}) = s(wsp, \lambda_{wsp}) + s(sun, \lambda_{sun}) + s(RH, \lambda_{RH}) + s(temp, \lambda_{temp}) + s(press, \lambda_{press})$; tr is the number of vehicles per hour. Time t is defined as $(julian\ day) + (h/24)$, where h is the hour of the day.

The results shown in Tab.3 indicate that traffic seems to be generally more important than meteorology for the model fit. This is most visible in models for NO_2 and NO_x. However, in models for PM the meteorology seems to be more effective then traffic, which is not unexpected given the more physical and granular nature of PM. Finally, as we can see by the BIC values of model M_3, the use of meteorological and traffic variables together gives, obviously, better performances for every pollutant.

Table 3: BIC values for models of traffic and logarithm of pollutant concentration

	NO	NO_2	NO_x	CO	PM
M1	26492	1908	14817	4386	439
M2	25636	-1576	12605	2390	598
M3	19572	-5010	5764	-4543	337

Given that the network of 107 traffic induction loops is rather dense, we propose summarizing traffic data so that its key features (with respect to pollution) are preserved. We have analyzed and compared four different ways of "summarizing" traffic data: 1) a simple average of the whole network, 2) time decomposition of traffic using moving average, 3) spline of traffic average and 4) principal component analysis. In the last case we analyze the effect of using 1, 3 or 10 principal components (one component is enough to explain over 90% of the variability in traffic, while 10 components are enough to explain over 96% of traffic variability and are denoted as $PCA(tr, 10)$); basically we can observe that the first component of traffic PCA behaves like the average of the traffic both when used as is or as a spline (even though the simple average is slightly favored by BIC). Similarly, using the first three components slightly improves on the previous case, while using the first 10 components yields a rather good fit, as shown below.

Following that general principle, we define the models as follows

$$\begin{aligned} M_4 : \mu_t &= \alpha + s(t,\lambda) + \beta_M tr_{MtC} + \beta_W tr_{WtC} + \beta_D tr_{DtC} + \beta_R tr_R \\ &+ s(ME, \lambda_{ME}) \quad (4) \\ M_5 : \mu_t &= \alpha + s(t,\lambda) + \beta_{tr} tr + s(ME, \lambda_{ME}) \quad (5) \\ M_6 : \mu_t &= \alpha + s(t,\lambda) + \vec{\beta}_{PCA} PCA(tr, 10) + s(ME, \lambda_{ME}) \quad (6) \end{aligned}$$

where tr_{MtC}, tr_{WtC}, tr_{DtC} and tr_R are respectively the monthly, weekly, daily and residual components in a time scale decomposition, tr is the average traffic variable and $PCA(tr, 10)$ indicates the first ten principal components from the analysis of the whole network (107 loops).

Tab.4 shows the BIC values of the models. The performance of the model M3 remains the best when considering pollutants NO_2, NO_x and CO, while in two cases (i.e. NO and PM) model M4 seems to fit slightly better. This may imply that in these cases the decomposition of traffic in seasonal components is better (more succinctly) able to explain the traffic pattern. These results will be taken into account for the definition of the further models described below.

Table 4: BIC values in models with different parametrizations of traffic

	NO	NO_2	NO_x	CO	PM
M4	19167	-3710	6810	-2830	288
M5	20806	-3688	7590	-3033	337
M6	19838	-4843	6416	-4153	385

4.2 Modeling of hourly NO, NO_2, NO_x and CO

We now describe how to select more carefully the predictors to use in models, which are related to the chemical and physical dynamics of the measured pollutant. This theory-based approach to selecting variables may not necessarily result in a better fit, but it will help incorporate scientific reasoning, physics and chemistry, behind the behavior of the pollutants.

First, lagged values of wind speed and solar radiation are expected to play an important role in the chemistry and physical transport of the pollution throughout the city. Following Carslaw et al. (2007), the wind direction was considered in the preliminary phases of this analysis but no important effects on pollutant concentration have been observed. This result is likely related to the fact that we are working with the average of the variables over the whole city, which may cancel out any directional effects. These models use

hourly average concentrations of NO, NO_2, NO_x and CO, and are given as follows:

$$\begin{aligned} M_7 : \mu_t &= \alpha + s(t, \lambda_t) + \vec{\beta} DoW + s(tr, \lambda_C) \\ &\quad + s(wsp, \lambda_C) + s(lag(wsp, 1), \lambda_C) + s(lag(wsp, 2), \lambda_C) \\ &\quad + s(sun, \lambda_C) + s(lag(sun, 1), \lambda_C) + s(lag(sun, 12), \lambda_C) \\ &\quad + s(rh, \lambda_C) + s(tmp, \lambda_C) + s(press, \lambda_C) \end{aligned} \tag{7}$$

Here, social and generally unmeasured confounders are recognized with the smooth function of time $s(t, \lambda_t)$, and to some extent also with the vector of variables indicating the days of the week DoW which turns out to contribute greatly to quality of fit. The other covariate are vehicular traffic (tr); wind speed (wsp) and the lagged values at one hour (*lag(wsp,1)*) and two hours (*lag(wsp,2)*); solar radiation (sun) and the lagged values at one hour (*lag(sun,1)*) and twelve hours (*lag(sun,12)*); relative humidity (rh); temperature (tmp) and pressure ($press$). Despite the fact that the results in the previous section indicate that the use of traffic seasonal decomposition is the most suitable for Nitric oxide, we will advocate using the spline functions of traffic because it will turn out that it performs better with the additional predictors in the model.

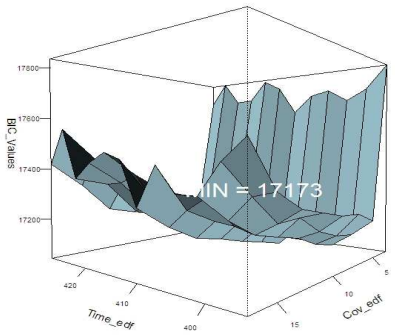
To select the best model supported by the available data, we first choose the suitable number of basis for the covariate smooth functions according to the BIC. The actual degrees of freedom (the penalty λ) are estimated using the generalized cross validation (GCV). The resulting performances of the models are given in Tab. 5, where we can observe that the models are able to explain a large fraction of variation in each of the processes. Although we do not advocate using the coefficients of determination statistic for assessing goodness of fit, we report for consistency with previous published work that the coefficient of determination in all our models is above 0.8, in agreement with those reported in Aldrin and Hobæk Haff (2005) and Carslaw et al. (2007). Since time has a quite different trend respect to the other covariates we propose several models with different number of knots, and select the number of the basis for time and for the other covariates separately. Looking at variations of these models with other numbers of degrees of freedom and number of time basis functions, yields the surface plot shown in Fig.5 where the "potential well" indicates the minimum BIC obtained.

Tab. 6 and Fig.6 summarize the main effects of the predictors under consideration, where linear effects are described with the coefficients and the main non linear effects are represented graphically as smooth functions.

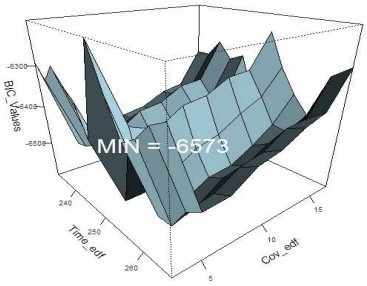
The estimated function of time and the days of week (DoW) are, as mentioned above, supposed to capture the adjusted effect of unobserved con-

Table 5: Performance indexes and number of basis for model M_7 (BIC values, Time-bs = basis dimension of time, Cov-bs = basis dimension of other covariates)

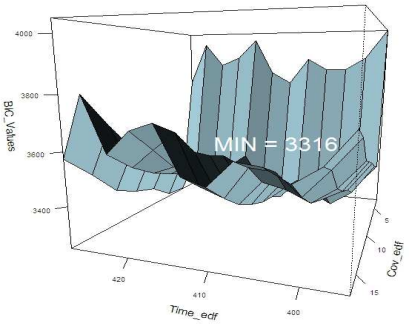
	NO	NO2	NOx	CO
BIC	17173	-6573	3316	-6681
Time-bs	410	248	410	410
Cov-bs	6	6	6	6



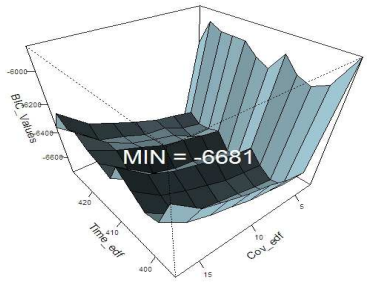
(a) NO



(b) NO_2



(c) NO_x



(d) CO

Figure 5: BIC distribution with respect to degrees of freedom of Time and other covariates

founders on the pollutant. The first plot shows the estimated spline of time with around 4 or 6 knots per week (depending on the pollutant). This relative large number of knots could explain the daily and weekly cyclical social behaviour (i.e. heat during the day, or heavy traffic in specific hours of the day or the week) that is related to traffic and pollution. It is reasonable to expect that the number of knots should have some influence on BIC and on the importance we attribute to the unmeasured variables, and that it should have an effect on the other estimates. However, comparing this model with others with smaller number of knots, we observe that this model is still better with respect to the BIC criterion, while the other predictors' estimated spline coefficients change only negligibly.

The smooth effect curves of time for all pollutants show stronger effects during wintertime (winter 2003-04 corresponds to hours 12400-12500 and winter 2004-05 to 12740-12840), see Fig.6.a. Concentrations are generally lower and more stable otherwise, reflecting the usual seasonal behaviour normally associated with the atmospheric boundary layer. Only NO seem to have a slightly different behaviour during summer. Beside the quite large sensitivity due to the typical instability of the pollutant, it seems to increase during summers when other pollutants decrease and it is likely related to certain photochemical processes. Days of week (DOW) seems to always have positive effects relative to baseline (Sunday), see Tab.6, with Saturdays having the lowest average concentrations among the other six days.

We can observe that traffic tr is, as expected, an important factor in pollution (see Fig.7 for partial traffic effect with relative standard error, detailed for every pollutant). In fact traffic is one of the most important atmospheric nitric oxide generator. Nitric oxides seem to be especially related to traffic as the average log-concentrations keep increasing rapidly with the number of vehicles at lower counts (below the median), ultimately leveling off after about 700 vehicles per hour per loop. The average CO log-concentration has an almost linear behaviour with a slow increase associated with the increase in number of vehicles. Like NO, CO pollutant is associated with vehicular traffic but the range of variation is far less significant (the carbon monoxide is no longer a critical pollutant in Torino). For all the pollutants we can highlight a threshold between 300 and 400 vehicles, corresponding to the night-vs-day time traffic. Below this threshold the relationship between the average log-concentration and traffic is generally steeper than above it.

At low temperatures (tmp) the average log-concentration tends to be high for majority of pollutants, and then it decreases at higher temperatures, given other factors in the model. This effect is more clear for NO and NO_x than for NO_2 and CO, see Fig.6.c. In fact, the last two pollutants seem to be scarcely conditioned by the temperature and show an almost linear trend. The higher

values at low temperatures are related to the seasonal atmospheric situation: generally low temperatures are during the winter, when the solar radiation and boundary layer are reduced too.

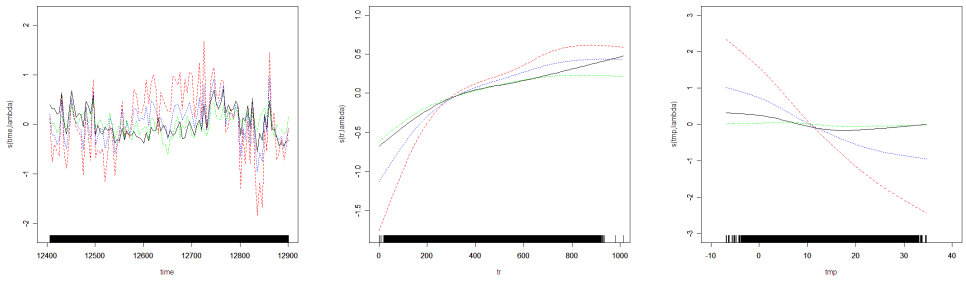
The estimated solar radiation splines, shown in Fig.6.d-f (sun, $sun1$, $sun12$), suggest that the partial effect of this variable has a generally different behavior in influencing the average concentration depending on the lag of the effect observed: in fact, high values of solar radiation cause an increase in the concentration in the next hour, but the lagged variables show negative effects, particularly for NO and NO_x. The persistent effect after many hours is likely explained by the fact that a strong radiation tends to delay a new rise in pollution concentration.

Wind speed has an important effect, given other variables in the model, especially when observed at different lags, and it generally reduces the concentrations significantly as it increases. Lagged variables show that a strong wind may influence the pollution for many hours (Fig.6.g-i). The four pollutants reduce the concentration in a similar way for wind speed below $2m/s$, with the functions being very different after that point but also more uncertain because of sparse data. The same effect is observed on the four pollutants when the wind measure from the preceding hour is used. The lagged values show that delayed effects normally have a greater effect especially on NO_2 and NO_x.

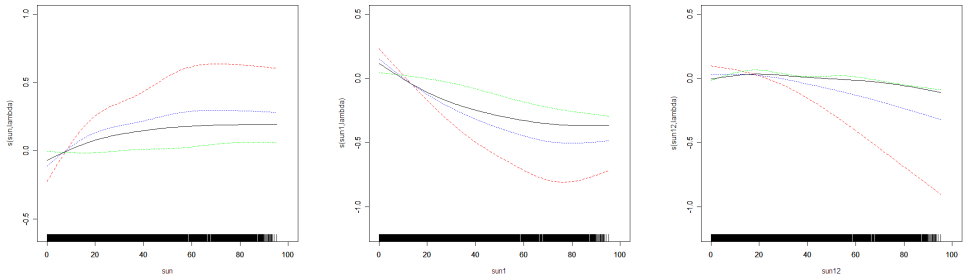
Peculiar decrease observed in the partial effect of relative humidity at high values (Fig.6.j) could be associated to rainfall events that usually accompany it. In fact, during rainfall events the humidity goes to saturation and precipitation is generally effective in pollution reduction. Such a behavior is common for all the pollutants even if CO decreases more slowly. The behaviour at low values could be associated with the increase of wind intensity, when pollution and humidity are normally blown away. Differently from other pollutants, NO has a quite constant trend, even if this situation could be due to an increase in solar radiation and the consequent chemical dynamics (low values of relative humidity are normally related to clear sky situations). The variation observed in pressure (Fig.6.k) is very small and it could be due to the use of hourly scale for a variable that usually changes more slowly in time. This result is unusual since high pressure is normally related to atmospheric stability, except in the event of atmospheric inversion; hence further analysis may be necessary.

4.3 Estimates for PM_{10} model

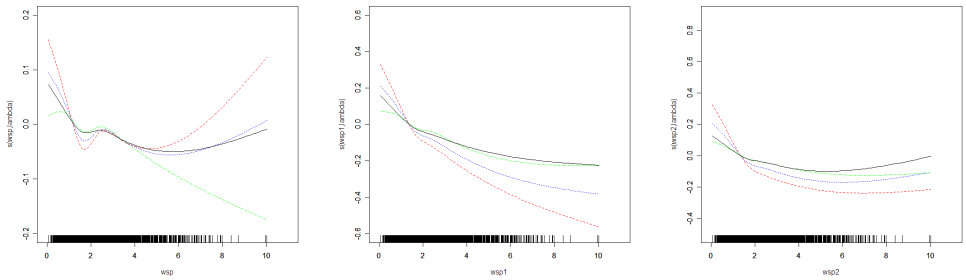
Given that PM data are daily, we will use the daily average for all the covariates in the model. As discussed in Sec. 4.1, time decomposition of



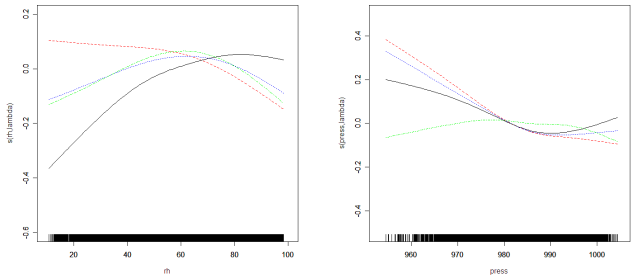
(a) Time (b) Traffic (c) Temperature



(d) Solar radiation (e) Solar radiation (lag 1) (f) Solar radiation (lag 12)



(g) Wind speed (h) Wind speed (lag1) (i) Wind speed (lag2)

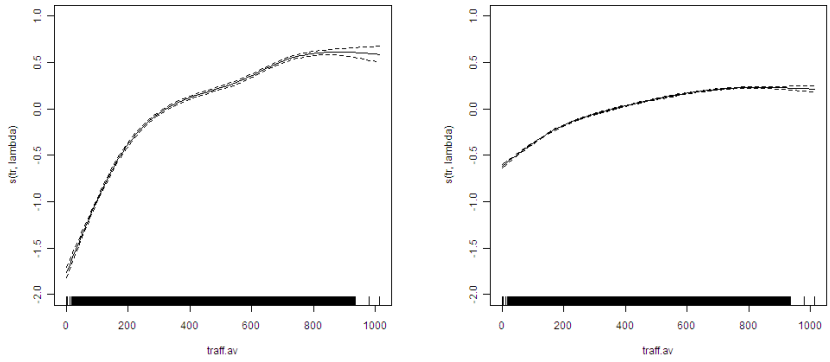


(j) Relative Humidity (k) Pressure

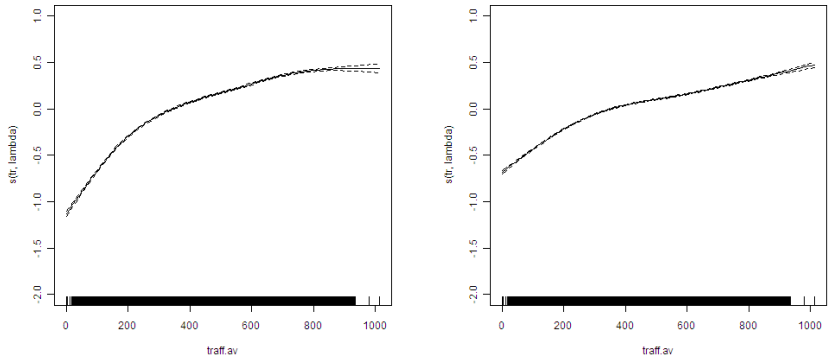
Figure 6: Estimated effects of traffic and meteorological variables for Nitric Monoxide (NO, red dashed), Nitric Dioxide (NO_2, green dash-dot), Nitric Oxides (NO_x, blue small dashed) and Carbon Monoxide (CO, black continuous).

Table 6: Coefficients of the parametric part of the additive predictors for NO_x, NO, NO_2 and CO

	NO_x		NO		NO_2		CO	
	Est.	Std.Er.	Est.	Std.Er.	Est.	Std.Er.	Est.	Std.Er.
(Interc.)	4.81	0.014	3.41	0.025	4.14	0.008	0.19	0.009
Mon.	0.17	0.015	0.27	0.027	0.15	0.008	0.04	0.010
Tue.	0.17	0.022	0.27	0.039	0.14	0.012	0.05	0.014
Wed.	0.21	0.024	0.32	0.044	0.12	0.014	0.08	0.016
Thu.	0.23	0.025	0.38	0.044	0.12	0.014	0.09	0.016
Fri.	0.18	0.022	0.32	0.040	0.11	0.012	0.06	0.015
Sat.	0.11	0.015	0.17	0.027	0.07	0.008	0.04	0.010



(a) NO (b) NO_2



(c) NO_x (d) CO

Figure 7: Estimated effect of traffic and relative standard error for Nitric Monoxide, Nitric Dioxide, Nitric Oxides and Carbon Monoxide.

traffic data (calculated using moving average) seems to be the best way to summarize traffic (see Tab. 4). We start with the model M_4, and add "day of the week" (DoW), and different lagged covariates:

$$\begin{aligned} M_8 : \mu_t &= \alpha + s(t, \lambda_t) + \beta_1(DoW) \\ &\quad + \beta_2 tr_{MtC} + \beta_3 tr_{WtC} + \beta_4 tr_{DtC} + \beta_5 tr_{Res} \\ &\quad + s(wsp, \lambda_C) + s(lag(wsp, 1), \lambda_C) + s(lag(wsp, 2), \lambda_C) \\ &\quad + s(rh, \lambda_C) + s(press, \lambda_C) \end{aligned} \tag{8}$$

where t is time; tr_{MtC}, tr_{WtC} and tr_{DtC} are, respectively, the monthly, weekly and daily components in a time scale decomposition and tr_R is the residual. The *wsp* covariate denotes wind speed; *lag(wsp,1)* and *lag(wsp,2)* are the lagged wind speed at one and two hours, respectively. These lagged variables have been choosen based on the correlation between wind speed and the pollutant. Finally, *rh* and *press* denote relative humidity and pressure, respectively. In Figure 8 we show the BIC values for different models as we change the number of basis for covariates and for time. As can be seen from that figure the BIC preferes the model M_8, whose specific performance index values and number of bases are shown in Table 7.

The coefficients of the traffic seasonal components and days of week are presented in Table 8. We can observe that even though the coefficients of traffic are positive, indicating a positive effect of traffic on the pollution log-concentration, their values are generally small, and show little variation on the montly and weekly scale. This result is consistent with Tab. 3, where we have observed that traffic seems less effective then meteorology in explaining the behavior of PM concentration. The effect of daily factor (DoW) with respect to Sunday is again clear, and as expected weekdays have higher average log-concentration than Saturday. Figure 9.a shows a strong relative increase of PM during wintertime, reflecting confounders like social (e.g. heating) or meteorological (e.g. boundary layer thickness variation) processes. Increase in temperature seems to be associated with an increase in average PM log-concentration (Fig.9.b). Increase in wind speed is related to reduced PM concentration, both for current time (Fig.9.c) and its one day lagged values (Fig.9.d). Increase in relative humidity is associated with a reduction in average PM log-concentration at high and low values (Fig.9.e). This could be due to rain (high values) or strong wind (low values), although at low values the data are more sparse. Finally an increase in pressure is related in an almost linear way to the increase in the average PM log-concentration (Fig.9.f).

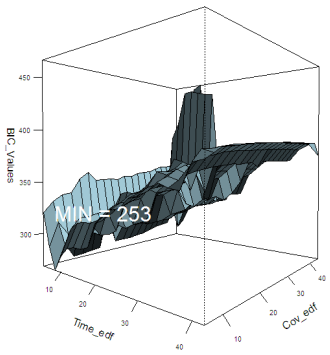


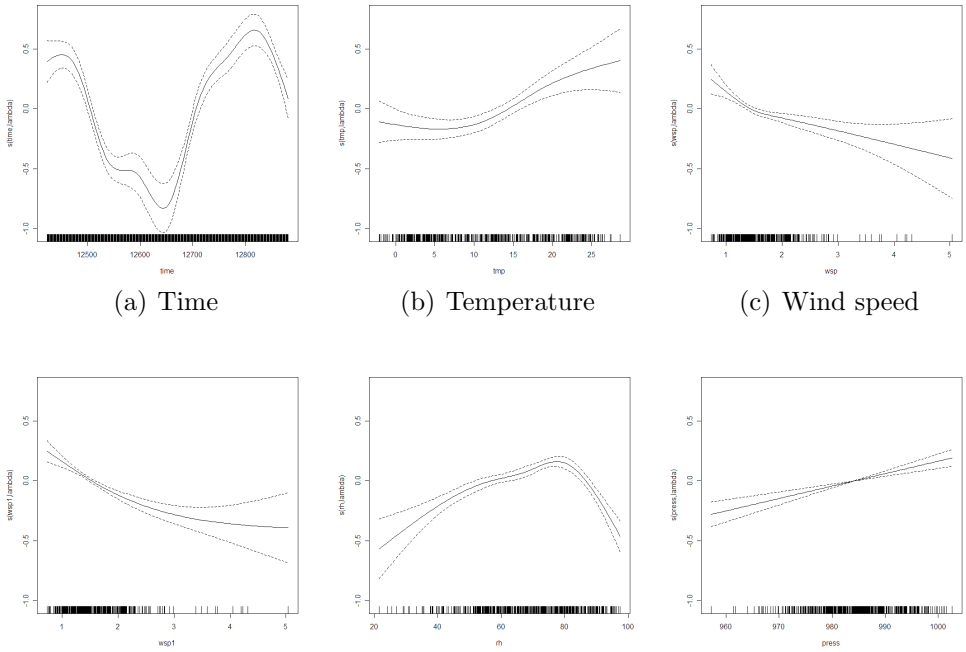
Figure 8: BIC values for PM models, respect to the number of knots of Time and other covariates

Table 7: Performance indexes and number of basis for model M_8 (BIC values, Time-bs = basis dimension of time, Cov-bs = basis of other covariates)

	PM
BIC	253
Time-bs	9
Cov-bs	5

Table 8: Estimated coefficients and standard errors of the parametric part of the additive predictors for PM

	PM	
	Est.	Std.Er.
(Intercept)	3.74	0.031
tr_{MtC}	0.00009	0.0007
$trWtC$	0.00005	0.0008
$trDtC$	0.002	0.0006
$trRes$	0.005	0.0023
$Monday$	0.19	0.071
$Tuesday$	0.24	0.083
$Wednesday$	0.24	0.082
$Thursday$	0.19	0.085
$Friday$	0.22	0.089
$Saturday$	0.11	0.066



(a) Time (b) Temperature (c) Wind speed

(d) Wind speed (lag1) (e) Relative Humidity (f) Pressure

Figure 9: Estimated effects of traffic and meteorological variables on particulate matter (PM)

5 Conclusions

In this paper we present a study of the air pollution in the city of Turin through the framework of generalized additive models. We have used the generalized additive models (GAM) to model the behavior of five species of pollutants (CO, NO, NO_2, NO_x and PM) averaged over the city of Torino as a function of traffic, while controling for the main meteorological variables as well as an unobserved confounding process. GAM allow flexible modeling of pollution processes which has traditionally been done in a classical style of differential equation-based models. In our study, the GAM models have been able to capture the relationship between pollutants and predictors flexibly, using semi-parametric components modeled with penalized cubic regression splines, where the penalty (the smoothing parameter) is estimated using generalized cross validation (GCV). One of the main advantages of GAM is perhaps their ability to extend this flexibility to unobserved confounders, by allowing "time" to act as a proxy for them. Including a smoothly varying function of time to capture the behavior of relatively slowly-varying unobserved confounders help address the bias in estimates of the effects of interest, such as traffic.

We have used the Bayesian Information Criterion (BIC) to select the optimal number of knots for the splines, and choose among several different models. First, we compare simple models containing only meteorological variables and traffic to check the relative importance of the effect of traffic on pollution. The results show that for CO, NO, NO_2 and NO_x traffic is more important than meteorology in explaining the log-pollution concentration, while for log-PM traffic turns out to be less important than the observed meteorological covariates. Second, between the four different ways to summarize traffic data, in three cases (CO, NO_2 and NO_x) the splines of traffic fit better, while in the remaining two cases (NO and PM) the time decomposition of traffic gives better fit. Third, we estimate the relationship between the different covariates for all pollutants. In fact, increase in traffic volume is associted with increase in the pollutants adjusted for other factors, while temperature, solar radiation and wind speed have positive partial effects in the pollution reduction. The nonlinearities found in the estimated effects confirm that the generalized additive models are a useful framework to estimate and interpret the relations between pollution, traffic and meteorology.

There are several possible extensions to this work. We have observed that better model selection techniques and effective degree of freedom computation are necessary for high-degree spline models, perhaps along the lines of work by Simonoff and Tsai (1999) and Shi and Tsai (1999). Another natural

extension is to consider a joint spatial or spatio-temporal model for pollutants measured by the individual monitoring stations throughout Torino, using the vectorized version of the generalized additive mixed models (GAMM) (Wood, 2006), a work which is currently ongoing.

References

Akaike, H., 1978. A new look at the bayes procedure. Biometrika 65, 53–59.

Aldrin, A., Hobæk Haff, I., 2005. Generalised additive modeling of air pollution, traffic volume and meteorology. Atmospheric Environment 39, 2145–2155.

Barnett, A. G., Williams, G. M., Schwartz, J., Neller, A. H., Best, T. L., Petroeschevsky, A. L., W., S. R., 2005. Air pollution and child respiratory health: a case-crossover study in Australia and new Zealand. American Journal of Respiratory and Critical Care Medicine 171, 1272–1278.

Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M., Dominici, F., 2004. Ozone and mortality in 95 U.S. cities from 1987 to 2000. Journal of American Medical Association 292, 2372–2378.

Bertaccini, P., 2009. Methods and models for vehicular traffic-related atmospheric pollution in the turin urban area. Ph.D. thesis, Universitá degli Studi di Torino.

Bertaccini, P., Cameletti, M., Fassó, A., 2007. Urban mobility and atmospheric pollution within the Torino metropolitan area. Atti della Conferenza SIS '07, 607–608.

Bertaccini, P., Dukic, V., Ignaccolo, R., 2008. Modelling traffic related air pollution in Torino with generalized additive models. Proceedings of the 4^{th} International Workshop on Spatio-Temporal Modelling (METMA'08), Alghero, Italy, 123–128.

Carslaw, D. C., Beevers, S. D., Tate, J. E., 2007. Modelling and assessing trends in traffic-related emission using a generalised additive modelling approach. Atmospheric Environment 41, 5289–5299.

Dir00/69/EC, 2000. Council directive n. 69 of 16 november 2000 relating to limit values for benzene and carbon monoxide in ambient air.

Dir02/3/EC, 2002. Council directive n. 3 of 12 february 2002 relating to ozone in ambient air.

Dir50/08/EC, 2008. Council directive n. 50 of 21 may 2008 on ambient air quality and cleaner air for europe.

Dir96/62/EC, 1996. Council directive n. 62 of 27 september 1996 on ambient air quality assessment and management.

Dir99/30/EC, 1999. Council directive n. 30 of 22 april 1999 relating to limit values for sulphur dioxide, nitrogen dioxide and oxides of nitrogen, particulate matter and lead in ambient air.

Fassó, A., Cameletti, M., Bertaccini, P., 2007. Uncertainty decomposition in environmental modelling and mapping. Proceedings of Summer Computer Simulation Conference (SCSC'07), 867–874.

Hastie, T. J., Tibshirani, R., 1990. Generalized Additive Models. Champman and Hall, London.

Kronborg, P., Davidsson, F., 2000. Improvements for s scandinavian spot urban traffic signal control system. TFK - Transport research institute. Stockholm.

Samet, J. M., Dominici, F., Curriero, F., Coursac, I., Zeger, S. L., 2000. Fine particulate air pollution and mortality in 20 u.s. cities: 1987-1994. New England Journal of Medicine 343, 1742–1757.

Schwartz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Schwartz, J., 1994. Non parametric smoothing in the analysis of air pollution and respiratory illness. Canadian Journal of Statistics 22, 471–488.

Shi, P., Tsai, C.-L., 1999. Semiparametric regression model selections. Journal of Statistical Planning and Inference 77, 119–139.

Simonoff, J. S., Tsai, C.-L., 1999. Semiparametric and additive model selection using an improved akaike information criterion. Journal of Computational and Graphical Statistics 8, 22–40.

Wood, K., 1993. Urban traffic control, systems review. PR41. Crowthorne. TRL.

Wood, S., Augustin, N., 2002. Gam with integrated model selection using penalized regression splines and applications to environmental modeling. Ecological Modeling 157 (2-3), 157–177.

Wood, S. N., 2006. Generalized Additive Models: An Introduction with R. Champman and Hall/CRC, London/Boca Raton, FL.