*Research Article*

# Modeling the Short-Term Effect of Traffic and Meteorology on Air Pollution in Turin with Generalized Additive Models

**Pancrazio Bertaccini,[1] Vanja Dukic,[2] and Rosaria Ignaccolo[3]**

[1] *Department of Biella, Environmental Protection Agency of Piemonte, 13900 Biella, Italy*
[2] *Department of Applied Mathematics, University of Colorado-Boulder, Boulder, CO 80309, USA*
[3] *Department of Economics, University of Turin, 10123 Turin, Italy*

Correspondence should be addressed to Rosaria Ignaccolo, rosaria.ignaccolo@unito.it

Vehicular traffic plays an important role in atmospheric pollution and can be used as one of the key predictors in air-quality forecasting models. The models that can account for the role of traffic are especially valuable in urban areas, where high pollutant concentrations are often observed during particular times of day (rush hour) and year (winter). In this paper, we develop a generalized additive models approach to analyze the behavior of concentrations of nitrogen dioxide ($NO_2$), and particulate matter ($PM_{10}$), collected at the environmental monitoring stations distributed throughout the city of Turin, Italy, from December 2003 to April 2005. We describe nonlinear relationships between predictors and pollutants, that are adjusted for unobserved time-varying confounders. We examine several functional forms for the traffic variable and find that a simple form can often provide adequate modeling power. Our analysis shows that there is a saturation effect of traffic on $NO_2$, while such saturation is less evident in models linking traffic to $PM_{10}$ behavior, having adjusted for meteorological covariates. Moreover, we consider the proposed models separately by seasons and highlight similarities and differences in the predictors' partial effects. Finally, we show how forecasting can help in evaluating traffic regulation policies.

## 1. Introduction

The impact of air pollution on human health and environment has been one of the central issues in environmental public policy and decision making [1–5]. For example, European Union mission mandates yearly improvement of environmental quality, lower emission standards, and support of environmental technology and scientific research and development [6–9] while the recent air quality directive [10] 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 covary 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, for example, asthma, flu, or other 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 is 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 Turin metropolitan area, where air quality is a concern, previous analysis of pollution has examined carbon monoxide (CO) concentrations and traffic volume in the Turin metropolitan area, as in Bertaccini et al. [11] who used a seasonal linear regression model for each station monitoring CO. Subsequently, Fassó et al. [12] studied the same problem using a linear vectorial autoregressive model and carried out a sensitivity analysis to describe the relative roles of traffic and meteorology, by their respective principal components. Instead, Kim and Guldmann [13] evaluate the importance of wind direction in the air pollution concentration through land use regression models aided by geographical information systems. They analyze the pollution related to vehicular traffic (defined as vehicle-kilometers-traveled weighted by wind direction frequency) by fitting linear regression models for different sizes of the buffer zone.

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 are deterministic, local behavior can be very difficult to understand and to model properly. 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 [14], 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 covarying 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. Thus, while there are many drivers of air pollution in Turin (some observed and measured and some not measured), the flexibility of the GAM approach allows us to capture and quantify the role of a single driver (traffic) without the confounding effects of the other drivers (confounders).

While generalized additive models have been widely used as a standard method in studies of pollution and health (see, e.g., the pioneering work [15]), they have only recently been introduced into the air pollution modeling of impact of traffic and meteorological covariates, as in the work of Carslaw et al. [16]. 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 [17] use generalized additive models separately for several different pollutants in different locations over the Oslo urban area, using traffic- and meteorological-observed data. Then they apply GAM to evaluate the effect of salting the street with magnesium chloride (for ice condition in winter) on particulate matter concentrations [18], showing that it is a potentially useful measure to reduce PM in a road tunnel.

More recently, ultrafine particle concentration in Helsinki and Finland temporal trends were examined in the light of their relationship with rainfall and other meteorological variables in Clifford et al. [5] and Mlgaard et al. [4].

In this paper we focus specifically on quantifying the role of traffic on air pollution in Turin, Italy, in a way that could be useful to environmental policy makers. Air quality in Turin area is critical with respect to particulate matter, nitrogen dioxide, and ozone; in fact, air quality standards set by European directives are often exceeded for these three pollutants (in particular $PM_{10}$), whereas most other pollutant concentrations are below the limit values. 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. [19, 20], we propose the use of generalized additive models to analyze the space-averaged air pollutant concentration over Turin metropolitan area as a function of vehicular traffic, while adjusting for potential meteorological and other possibly unobserved confounders. Our goal is to quantify the effect of traffic and evaluate potential interventions (traffic reductions) specifically for Turin. While we cannot generalize the estimated magnitude of the effects in Turin to other geographic areas, we hope that our analysis contributes to the urban air quality research in three ways: (1) the results for Turin present additional information for completing the pollution picture for European cities; (2) provides a good reference for environmental policy for cities with similar geographic surroundings; (3) methodologically, provides evidence that a simple form of the traffic variable can often describe the behavior of pollutants sufficiently well.

The paper is organized as follows. Section 2 is devoted to data description related to traffic, pollution, and meteorology. In Section 3 we describe the basic theory and some advantages of using the generalized additive models and then discuss the selection of the best model and the predictor subset for pollutant concentration, aiming to balance complexity and goodness of fit. Specific models are proposed and results analyzed for two critical pollutants, $NO_2$ and $PM_{10}$ (Sections 3.1 and 3.2), both for the whole period December 2003–April 2005 and separately by season in a year. Moreover, we carry out a forecasting application of the proposed model for $NO_2$ in order to show how it can be used for traffic regulation policy assessment (Section 3.3). Conclusions are discussed in Section 4.

## 2. Data

*2.1. Traffic.* The traffic data are provided by 5T s.r.l., a company working in the Turin 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
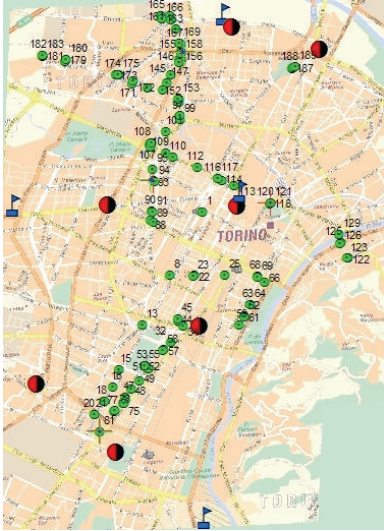
FIGURE 1: Turin vehicle traffic, air, and meteorology monitoring network: green disk are the 107 traffic counters, red and black disks are the pollution stations, and blue flags are the meteorological stations.

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 [21] and Wood [22]. 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 (Figure 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 nontrivial 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 passs 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 equal 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 furthers our study period from 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 Figure 2. The boxplots show typical features of the traffic trend at three different time scales: daily, weekly, and yearly. In the daily scale we can see the

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

| | Con | Riv | Reb | Cri | Gai | Lin | Gra |
|---|---|---|---|---|---|---|---|
| $NO_2$ | × | × | × | × | × | × | |
| $PM_{10}$ | × | × | | | × | | × |

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 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 holidays.

*2.2. Pollution.* Pollution data have been provided by *ARPA Piemonte* and *Regione Piemonte*. In this paper we focus on $NO_2$ and $PM_{10}$, which are measured on an hourly and daily scales, respectively. The measurements were recorded at a subset of the total of seven environmental stations across Turin (*Grassi, Rebaudengo, Rivoli, Consolata, Cristina, Gaidano,* and *Lingotto* stations are located as shown in Figure 1), while $NO_2$ and $PM_{10}$ measurement sensors are distributed as in Table 1. All the measurement stations are traffic ones, except *Lingotto* that is a background site; in order to have an average representative of all Torino area, we consider *Lingotto* data too.

In order to provide an example of $NO_2$ behavior over time, we summarize the $NO_2$ concentration measured at the "Consolata" station (Figure 3). As can be seen in Figure 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 Figure 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 (Figure 2(a)), motivating the importance of using the hourly time scale. As can be seen from the boxplots by day of the week (Figure 3(c)) the concentration seems to increase in the first few weekdays and decrease during the weekend. The box plots by month (Figure 3(c)) confirm that the lowest values happen in August, while the highest happen in the winter.

Also for $PM_{10}$ we show, as an example, the concentration measured at the "Consolata" station in Figure 4. The weekly representation in Figure 4(b) shows, as usual with other pollutants, an increase in concentration during the first days of the week, followed by a decrease till the Sunday's lowest values. The whole period time series (Figure 4(a)) and the box plots by month (Figure 4(c)) point out the large difference of concentration observed over the seasons of the year (cold and warm ones), despite the relative constancy of traffic. For further explorative analysis on pollution features see [20, Chapter 1].
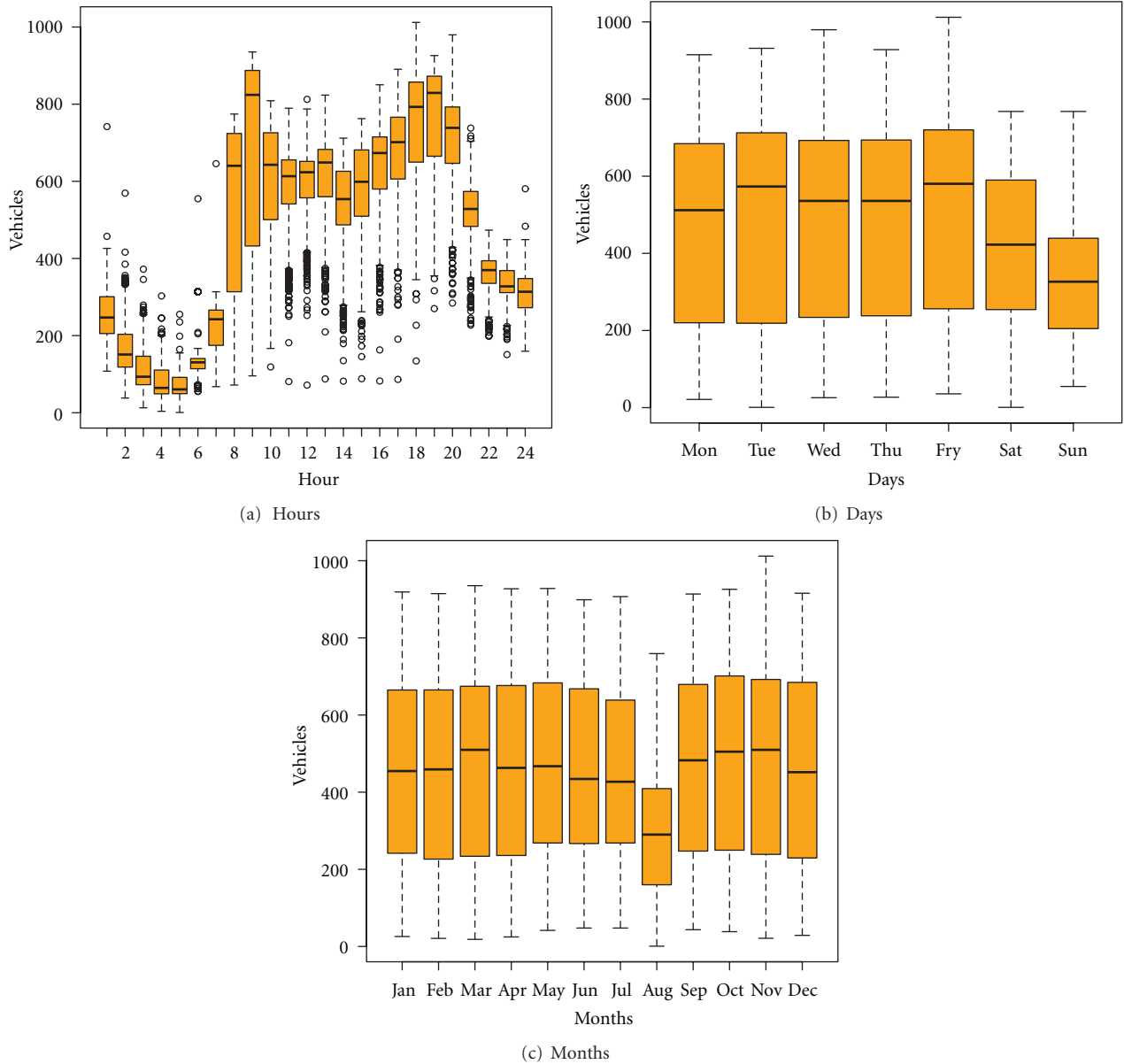
(a) Hours



(b) Days



(c) Months

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

*2.3. Meteorology.* Meteorological data are collected by four different stations by *ARPA Piemonte* and *Regione Piemonte*, shown in Table 2. The locations of the meteorological stations are shown in Figure 1, marked with the 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 Turin metropolitan area, so we can basically use the value measured by a single (*ReissRomoli* (*CSELT*)) station as representative of the city-wide pressure level.

Meteorology is reduced to the city-wide vector (ME) containing wind speed (wsp, in m/s), solar radiation (sun, in W/m$^2$), relative humidity (rh, in percentage), temperature (tmp, in degree Celsius), and pressure (press, in hPa).

TABLE 2: Available meteorological variables.

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

Precipitation has not been included due to being composed of relatively rare and localized events and to having a rather

(a) Time Series



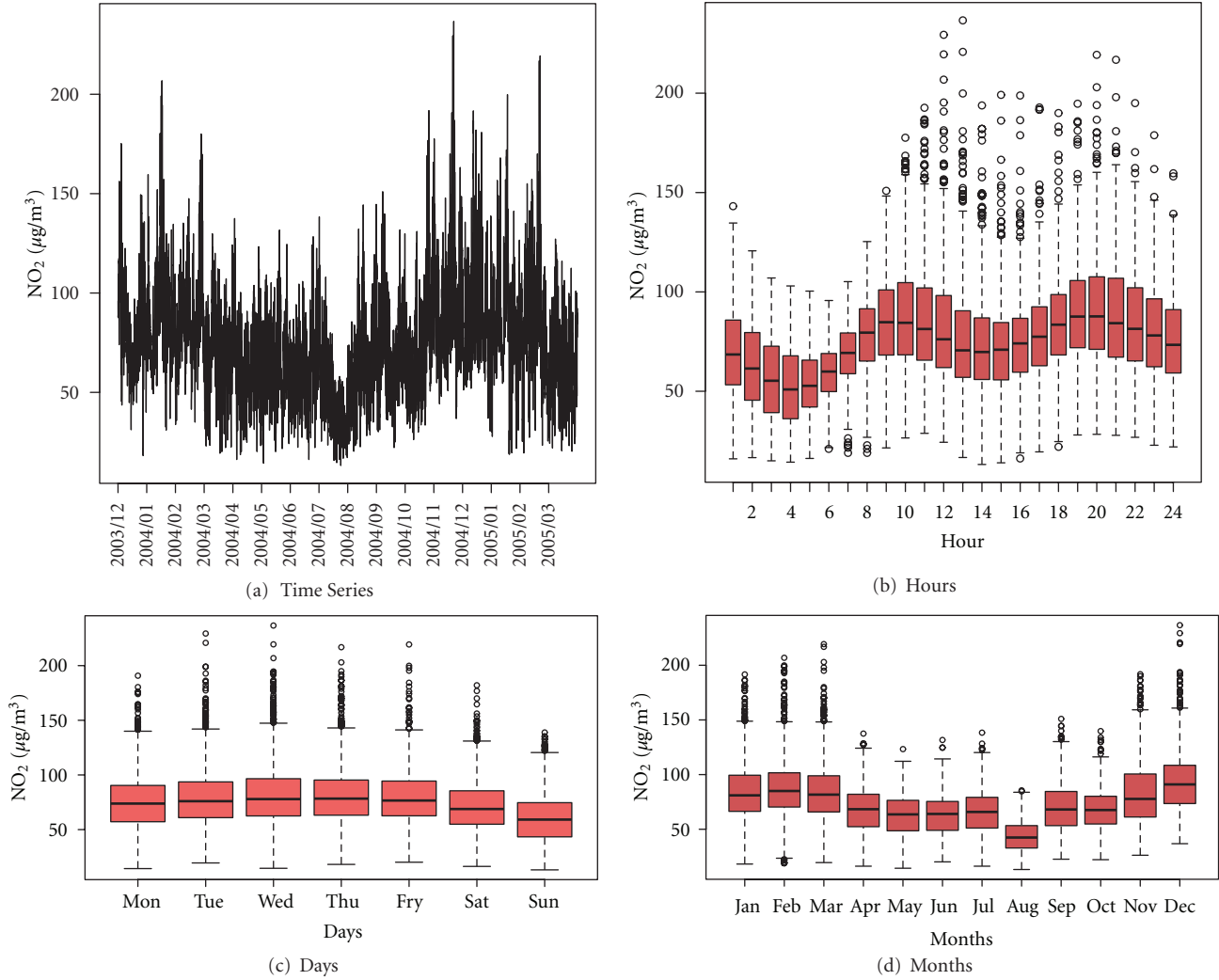(b) Hours



(c) Days



(d) Months

FIGURE 3: Time series and box plots for $NO_2$ concentration measured at Consolata Station.

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 our models 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.

In Figure 5 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.

## 3. GAM Models for Turin-Wide Pollution

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 (GAMs). 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 nonparametric terms but do not suffer from the curse of dimensionality like some other nonparametric methods such as kernel smoothing or polynomial modeling. For the outcome (e.g., logarithm of pollutant, $Y = \log(\text{Pollutant})$), we assume that it is additive in its predictors and normally distributed with mean $\mu_t$ and variance $\sigma^2$. The systematic part $\mu_t$ could include linear and nonlinear components, as well as potential
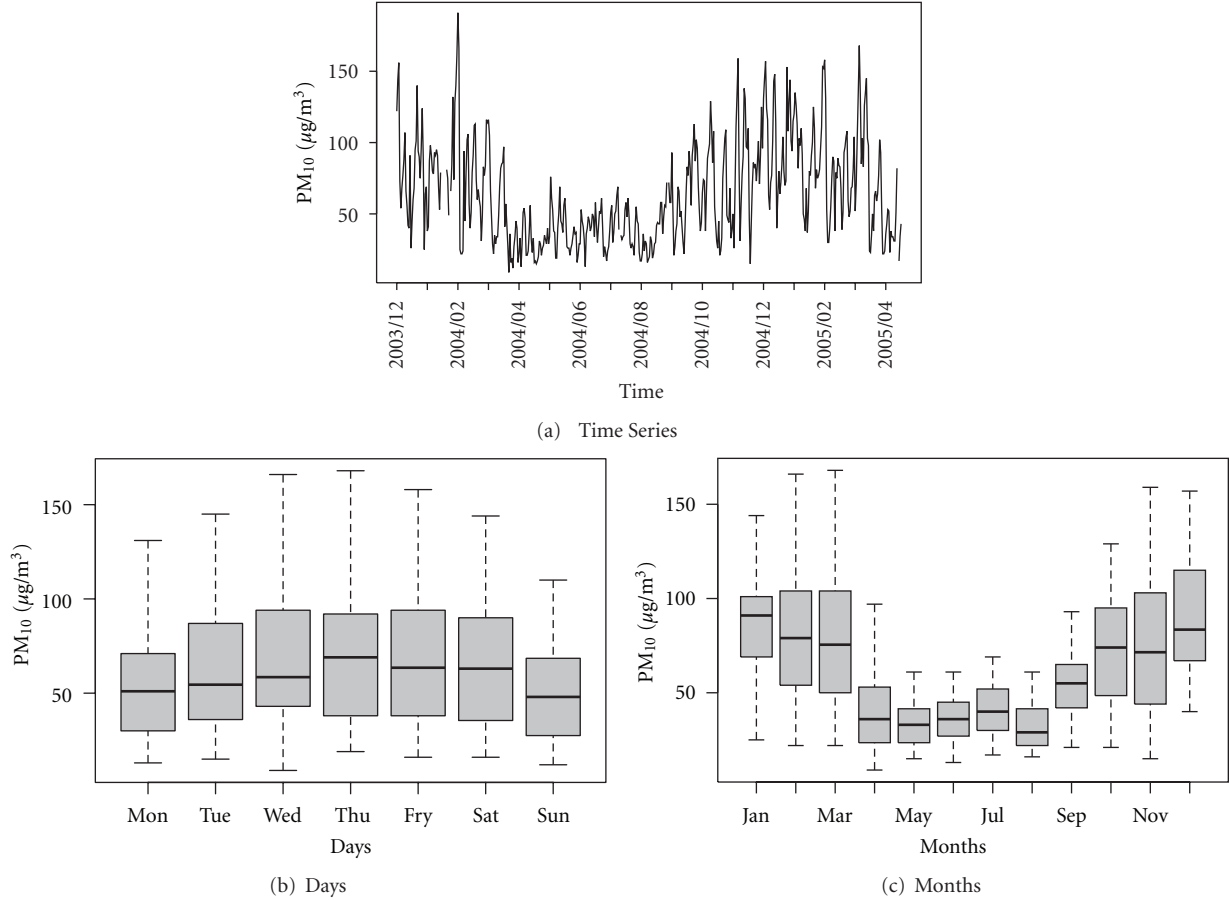
(a) Time Series



(b) Days



(c) Months

Figure 4: Time series and box plots for $PM_{10}$ concentration measured at Consolata Station.

confounders. A general model with additive components would then be

$$Y_t \sim \text{Gaussian}(\mu_t, \sigma^2),$$

$$\mu_t = \alpha + \sum_{f=1}^{l} \beta_f x_{f,t} + \sum_{g=1}^{m} \sum_{h \in H_g} \eta_{g,h} z_{g,t-h}$$

$$+ \sum_{i=1}^{p} s(k_{i,t}, \lambda_i) + \sum_{j=1}^{q} \sum_{h \in H_g} s(k_{j,t-h}, \lambda_j), \tag{1}$$

where $\alpha$ is the intercept, $\vec{x}_t$ are the current-time predictors, $\vec{\beta}$ are their (linear) 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 the linear effects $\eta_{g,h}$. Nonlinear effects of covariates $k_i$ (or their lagged version with lag times in $H_g$) are modeled nonparametrically trough smooth functions $s(\cdot, \lambda_i)$, where the smoothness is controlled by the scalar parameter $\lambda_i$.

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, Table 1). For each pollutant we consider the time series of the logarithm of the average pollutant concentration over Turin. 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

(a) Pressure

(b) Wind speed

(c) Temperature

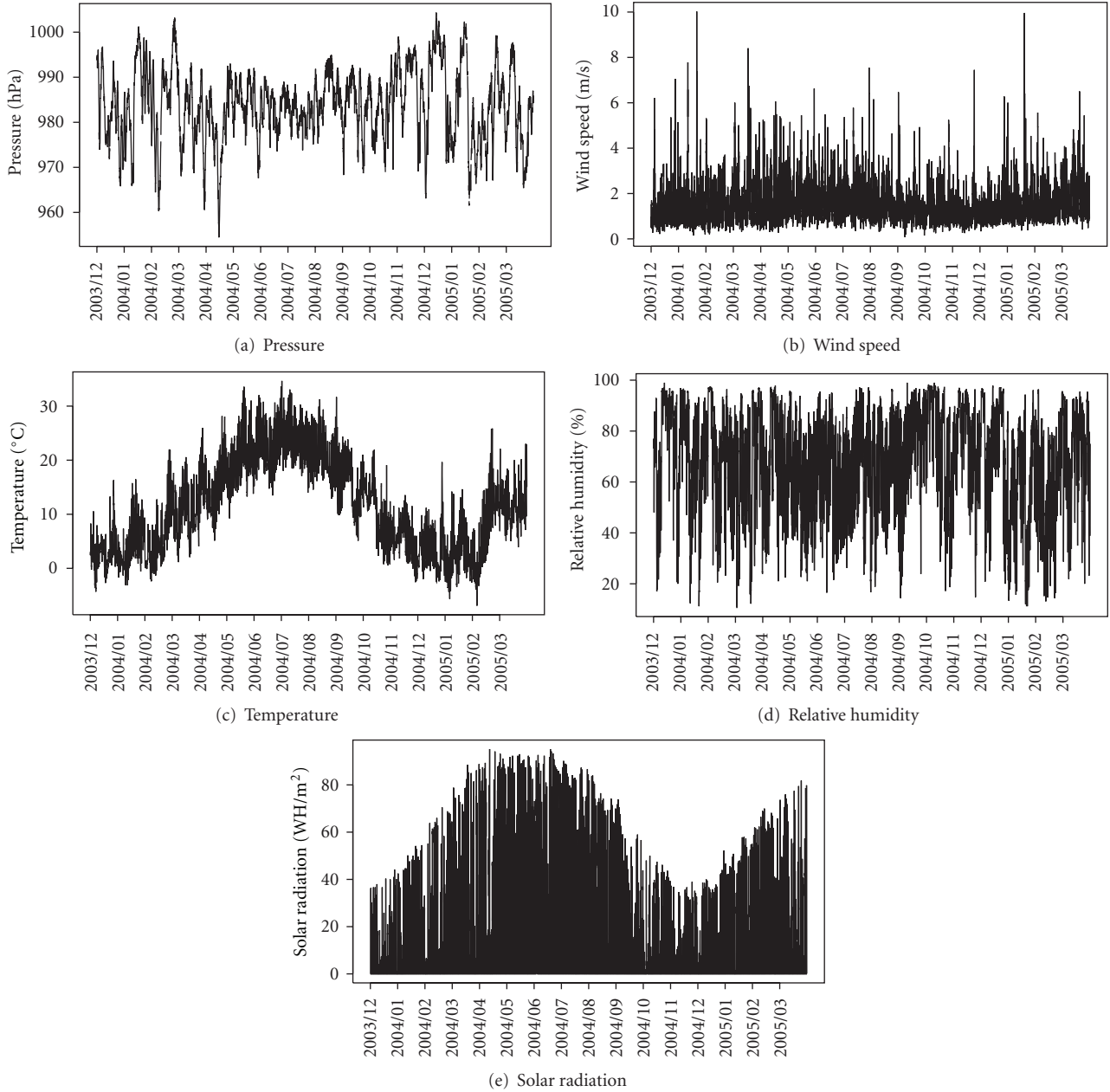(d) Relative humidity

(e) Solar radiation

FIGURE 5: Time series of the averaged meteorological variables.

crucial in choosing the optimal GAM model. To that extent, we use the Bayesian information criterion (BIC) [23]. The BIC is like the AIC [24] 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 $\theta$: BIC $= -2\ell_n(\hat{\theta}) + \log(n)\text{length}(\theta)$.

The main goal of this paper is to assess the effective role of vehicular traffic on two different pollution species. In order to do that thoroughly, we propose two approaches to represent traffic: the first approach is to model the nonlinear effect of traffic using splines, while the second one models a linear effect of simply transformed traffic variables. We select

the most appropriate functional form for each pollutant and 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 issue is related to cross-correlation between pollutants and some meteorological variables. This cross-correlation, when strong, suggests possible use of lagged variables in the model. In fact, this often allows a substantial improvement of fit. 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 splines only for those individual lagged variables that have

been selected based on the highest correlation with the pollutant. Since the latter procedure always yielded a better BIC score, we will only present results based on it for modeling pollution in our study.

All computation was done in the *R* package `mgcv` [25, 26] that allows to estimate penalized generalized additive models, based on penalized regression splines with automatic smoothness estimation [27].

*3.1. Modeling Hourly NO$_2$.* We now describe the global model for the behavior of hourly NO$_2$ (averaged over the city of Turin), during the period of December 2003 through April 2005. We show how to select 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, given the hourly scale, 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. [16], 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. Moreover, a dummy for rush hour was not found significant when all other variables, including traffic and lagged traffic, were in the model.

Then the proposed model for the average hourly log concentration of NO$_2$ is given as follows:

$$M_1: \mu_t = \alpha + s(t, \lambda_t) + \vec{\beta}\text{DoW} + s(\text{tr}, \lambda_C) + s(\text{lag}(\text{tr}, 1), \lambda_C)$$
$$+ s(\text{wsp}, \lambda_C) + s(\text{lag}(\text{wsp}, 1), \lambda_C)$$
$$+ s(\text{lag}(\text{wsp}, 2), \lambda_C) + s(\text{sun}, \lambda_C)$$
$$+ s(\text{lag}(\text{sun}, 1), \lambda_C) + s(\text{lag}(\text{sun}, 12), \lambda_C)$$
$$+ s(\text{rh}, \lambda_C) + s(\text{tmp}, \lambda_C) + s(\text{press}, \lambda_C).$$
$$(2)$$

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 turn out to contribute greatly to quality of fit. The other covariates are vehicular traffic (tr) and its lagged version lag(tr, 1) that is traffic at the previous hour; 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).

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 penalties $\lambda_t$ and $\lambda_C$) are estimated using the generalized cross validation (GCV). Since time has a quite

TABLE 3: Coefficients of the parametric part of the additive model for NO$_2$ (bold indicates 5% significance).

|          | Est.         | Std. Er. |
|----------|--------------|----------|
| (Interc.) | **4.157781** | 0.008705 |
| Mon.     | **0.131553** | 0.008831 |
| Tue.     | **0.122523** | 0.013177 |
| Wed.     | **0.108691** | 0.015761 |
| Thu.     | **0.112780** | 0.015928 |
| Fri.     | **0.112615** | 0.013356 |
| Sat.     | **0.072378** | 0.008766 |

different trend with respect to the other covariates, we fit several models, each with a different number of knots and select the functional form for the time predictor and for the other covariates separately. The resulting smallest BIC is equal to −7085.848 and is obtained in correspondence of 248 and 6 spline basis for time and for meteorological covariates, respectively. 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 our model is 0.825, in agreement with those reported in Aldrin and Hobæk Haff [17] and Carslaw et al. [16].

Table 3 and Figure 6 summarize the main effects of the predictors under consideration, where linear effects are described with the estimated coefficient values, and the main nonlinear effects are presented 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 confounders on the pollutant. The first plot shows the estimated spline of time with around 6 knots per week. 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 of time is more pronounced during wintertime (winter 2003-04), see Figure 6(a). Concentrations are generally lower and more stable otherwise, reflecting the usual seasonal behaviour normally associated with the atmospheric boundary layer. Days of week (DOW) always have positive effects with respect to the baseline (Sunday), see Table 3, with Saturdays having the lowest contribution among the six days.

We can observe that *traffic* is, as expected, an important factor (see Figure 6(b) for partial traffic effect with relative standard error), being one of the most important atmospheric nitric oxides generator. Nitric oxides seem to be
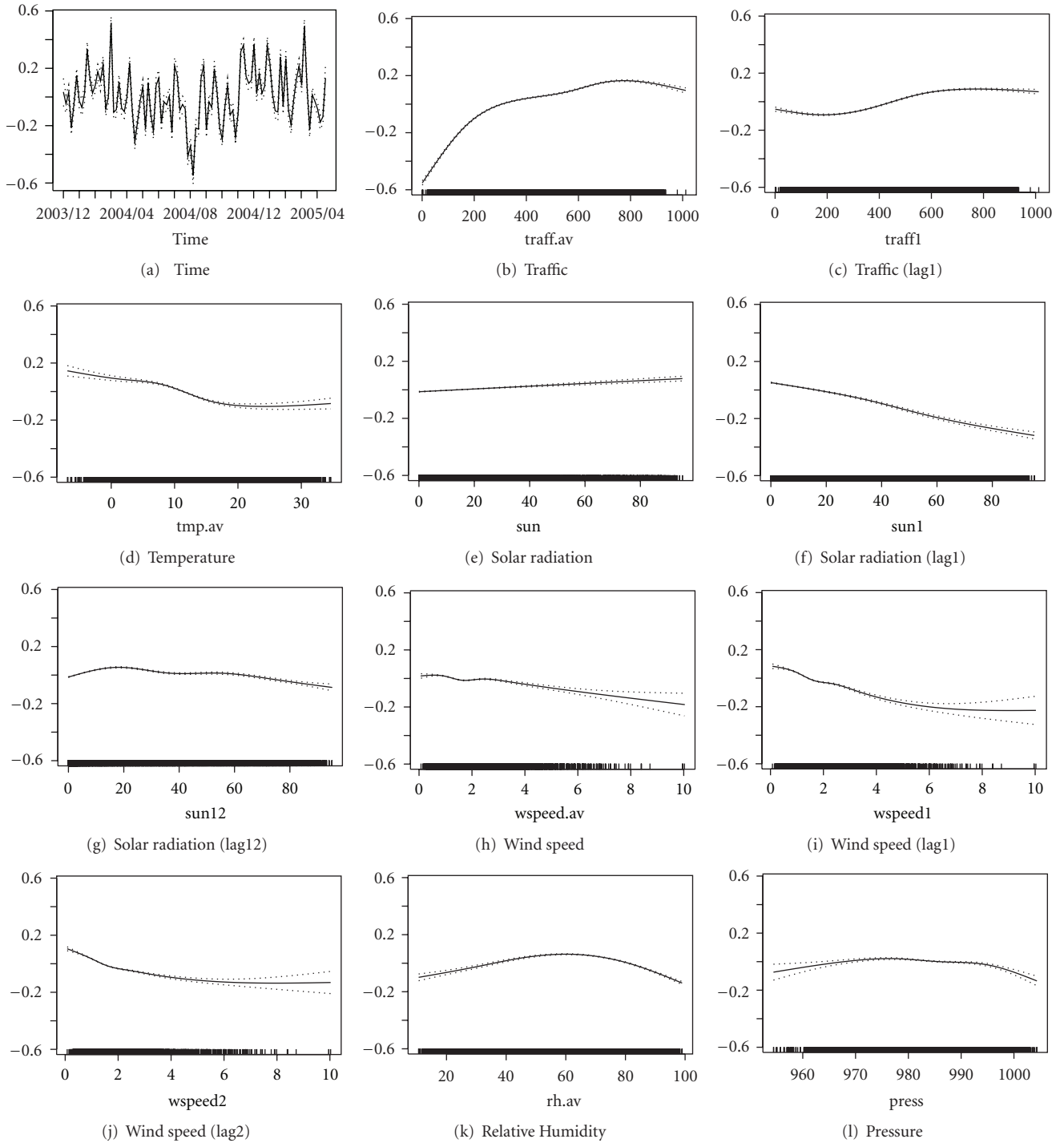
(a) Time

(b) Traffic

(c) Traffic (lag1)

(d) Temperature

(e) Solar radiation

(f) Solar radiation (lag1)

(g) Solar radiation (lag12)

(h) Wind speed

(i) Wind speed (lag1)

(j) Wind speed (lag2)

(k) Relative Humidity

(l) Pressure

FIGURE 6: Estimated effects of traffic and meteorological variables for $NO_2$.

especially related to traffic as the average log concentrations keep increasing rapidly with the number of vehicles at lower counts (below the median), ultimately almost leveling off to a saturation level after about 700 vehicles per hour. We can highlight a threshold between 200 and 300 vehicles, corresponding to the night-versus-day time traffic (see Figure 2(a)). Below this threshold the relationship between

the average log concentration and traffic is generally steeper than above it.

On the other hand, the effect of lagged traffic seems close to the zero line, see Figure 6(c). For that reason, we assess the utility of a simpler model for $NO_2$ with log of traffic with a linear effect and no lagged traffic in the model. All other predictors are kept in the same form. This simpler model

with logarithmic transformation of traffic can be used for policy evaluation and fast prediction. The estimated linear effect of log(traffic) was 0.26 (se = 0.003). The BIC of the simpler model was −6596.97, while the BIC of the spline model was −7085.85. This yields support for the model with splines over the model with simpler forms of traffic, but we nonetheless emphasize the potential utility of the simpler model.

Having a model with a linear effect of log-transformed traffic is greatly appealing from the policy evaluation standpoint. The GAM framework allows us to estimate the net effect of traffic, without other confounders, and therefore, having a linear effect facilitates direct estimation of the overall pollutant reduction as a result of a reduction in traffic. For example, our estimated effect of the log of traffic on log $NO_2$ was 0.26. From the policy point of view, this means that a 10% decrease in traffic would result in approximately 3% reduction in $NO_2$ concentration, on average.

At low *temperatures* the average log concentration tends to be higher and almost constant below the 10 degree Celsius (Figure 6(d)). After that it slightly decreases at higher temperatures, levelling off above a temperature around 20 degree Celsius. In fact, the pollutant does not seem to be really conditioned by the temperature and shows an almost linear trend at two different levels. The higher values at low temperatures are apparently 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 Figures 6(e)–6(g), 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 a little increasing in the concentration at the same hour, but the lagged variables show negative effects, particularly for the first lag. 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, persistent at different lags, and—as expected—it generally reduces the concentrations considerably as it increases. Lagged variables show that a strong wind may influence $NO_2$ pollution for many hours (Figures 6(h)–6(j)). The pollutant reduces its concentration for wind speed above 2 m/s suddenly, but lower wind speed could have some effect after one hour or more. The stronger effect of the wind is recognized as a delayed effect, and the lagged covariates have an increase of the effect between 2 and 6 m/s levelling off above that intensity.

Peculiar decrease observed in the partial effect of *relative humidity* (Figure 6(k)) at high values could be associated to rainfall events that usually accompany it. In fact, during rainfall events the humidity that goes to saturation and precipitation is generally effective in pollution reduction. The behaviour at low values could be associated with the increase of wind intensity, when pollution and humidity are normally blown away.

The variation observed in *pressure* (Figure 6(l)) is very small, this is unusual since high pressure is normally related to atmospheric stability, except in the event of atmospheric inversion, and it could be due to the use of hourly scale for a variable that usually changes more slowly in time.

### 3.1.1. Modeling $NO_2$ Separately by Season.

We further estimate model $M_1$ separately by season, to examine whether there are any seasonal differences. The four seasons are defined as follows:

  (i) Winter: 19th December 2003 to 18th March 2004,

 (ii) Spring: 19th March 2004 to 18th June 2004,

(iii) Summer: 19th June 2004 to 18th September 2004,

(iv) Fall: 19th September 2004 to 18th December 2004.

As with the global model, the estimated function of time and the days of week (DoW) capture the adjusted effect of unobserved confounders on the pollutant. The day-of-the-week effects appear significant mainly during winter and less so for the other seasons.

Also in separate seasons *traffic* shows important effects on pollution (see Figure 7(a)) but in two very different ways. During springtime and summertime, the traffic partial effect sharply increases for small traffic volumes, till 200–300 vehicles per hours, while it does not vary a lot for higher volumes. For winter and autumn, we observe a traffic partial effect that is almost constant when traffic volume changes and this suggests that the reduction of pollution concentration during wintertime is hard, even through deep traffic regulations. This result is also supported by the traffic role in the conjunction with the industry or domestic heater emissions. In fact, during the cold seasons the effect of typical atmospheric stability on pollutant emissions makes traffic be just one of the agents determining accumulation of pollution (so that high concentrations occur even with low traffic). To the contrary, during warm seasons traffic becomes the most important source of pollution and $NO_2$ concentration, anyway below limit values, and steeply increases with traffic values until it reaches the saturation condition. The partial effect of the previous hour traffic differs, when comparing cold and warm seasons, only for small traffic volumes (see Figure 7(b)). For volumes higher than 200–300 vehicles per hour, for all seasons the estimated spline is close to the zero line although it is significant.

The seasonal analysis of the meteorological covariates allows us to highlight the sensible differences that characterize the role of a variable in more homogeneous environmental conditions. In fact fitting the model $M_1$ to any single season allows us to separate the effect of the meteorological variation during the cold seasons or the warm seasons, when pollutants could behave very differently in relation to the meteorological variables. These differences are particularly evident for temperature, the solar radiation, the wind speed (with special attention for the delayed effects), and pressure.

Starting with *temperature* (Figure 7(c)), we can see that during the warm seasons its contribution to $NO_2$ concentration is low (anyway $NO_2$ concentration has low values in these seasons). Instead during winter and fall seasons, we can observe that at low temperatures the average log
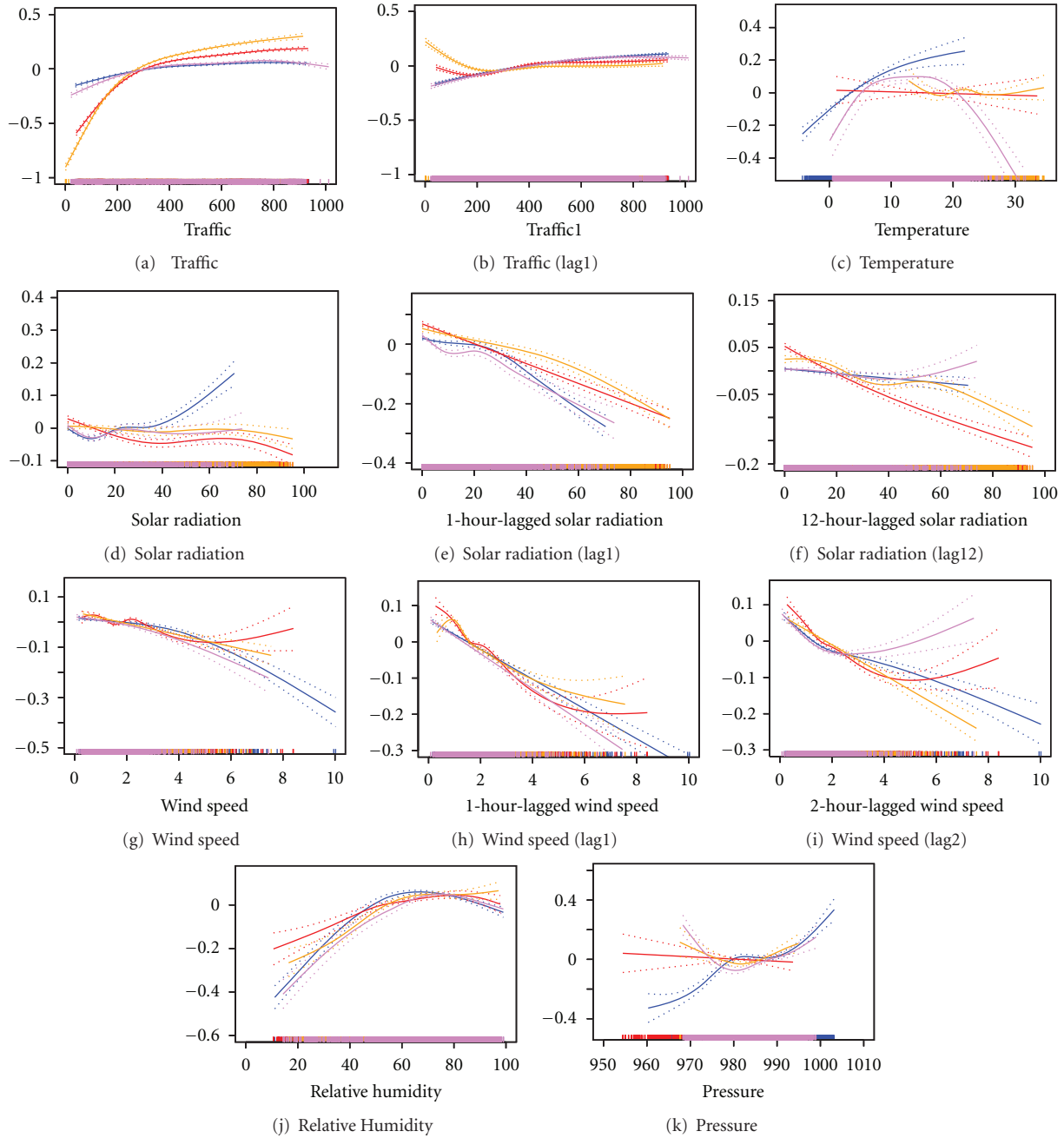
(a) Traffic

(b) Traffic (lag1)

(c) Temperature

(d) Solar radiation

(e) Solar radiation (lag1)

(f) Solar radiation (lag12)

(g) Wind speed

(h) Wind speed (lag1)

(i) Wind speed (lag2)

(j) Relative Humidity

(k) Pressure

FIGURE 7: Estimated effects of traffic and meteorological variables for $NO_2$ during different seasons. Blue is Winter, red is Spring, orange is Summer, and violet is Fall.

concentration tends to increase with temperature until 10 degree Celsius. After that value, it slightly decreases at higher temperatures during the autumn, while it keeps to increase during the winter. In this last case, the model shows the same result observed in other studies (e.g., [16, 17]) and it could be explained with the activation of photochemical reactions due to higher radiation (coherently the same situation is presented by the partial effect due to direct solar radiation in Figure 7(d)). Temperature effect during the autumn shows two apparently contrasting behaviours: the concentration

increases for low values of temperature and it decreases for high values. This result can be interpreted by considering that the autumn is a transition season moving from summer to winter, so that for a period the partial effect is similar to the winter one and for another period to the summer one.

The estimated *solar radiation* splines shown in Figures 7(d)–7(f) suggest that the partial effect of this variable has generally a reducing effect on the average log concentration of nitrogen dioxide. The effect is stronger one hour after the exposition and also 12 hours after during the warm

seasons; this persistency is likely explained by the fact that a strong radiation tends to delay a new rise in pollution concentration. The steep increase of the direct radiation effect during wintertime can be connected with the activation of photochemical reactions due to higher radiation that acts on a large amount of pollutant generally present in the winter metropolitan atmosphere such as nitric monoxide.

Observing the estimated *wind speed* splines for different seasons and lags in Figures 7(g)–7(i), we can see that wind speed generally keeps its important effect in reducing $NO_2$ concentrations, considerably when it increases: a strong wind may influence the pollution for many hours. In particular, during winter and summer the decreasing effect is clear and it starts even at small wind speed intensity; moreover, during the winter we observe the higher wind intensities. In the other two intermediate seasons the wind effect is smaller and for the 2 hours lagged variable seems to have a positive effect that probably needs further analysis.

The partial effect of *relative humidity* (Figure 7(j)) maintains a very similar behaviour during all the seasons that looks like the observed one for the global model. In this case we can highlight that higher values of humidity cause an increase in the average $NO_2$ log concentration.

As for the *pressure* (Figure 7(k)) we can see that the main effects are visible during the cold season, coherently with the fact that the concentration increases when the pressure increases too. Instead, the positive effect of low pressure values during the fall season needs further study to be explained.

*3.2. Modeling Daily $PM_{10}$.* In order to understand the extent to which the behavior of daily $PM_{10}$ (after the logarithm transformation) depends on traffic intensity, we began with a flexible model that incorporates splines to capture the effects of average daily traffic, as well as the average daily traffic during the previous day, in addition to meteorological predictors. The initial model was thus as follows:

$$
\begin{aligned}
M_2: \mu_t = {} & \alpha + s(t, \lambda_t) + \vec{\beta}(\text{DoW}) \\
& + s(\text{tr}, \lambda_C) + s(\text{wsp}, \lambda_C) \\
& + s(\text{lag}(\text{wsp}, 1), \lambda_C) + s(\text{lag}(\text{wsp}, 2), \lambda_C) \\
& + s(\text{rh}, \lambda_C) + s(\text{press}, \lambda_C),
\end{aligned}
\tag{3}
$$

where $\mu_t$ is the average daily log $PM_{10}$ at day $t$, and tr is the total traffic during day $t$ in the city of Turin. Furthermore, wsp denotes average daily wind speed, lag(wsp, 1) and lag(wsp, 2) are the lagged versions of the average daily wind speed from one and two days prior to day $t$, respectively. These lagged variables have been chosen based on their high pairwise correlation with the pollutant. Similarly, rh and press denote the average daily relative humidity and pressure, respectively. Note that given that $PM_{10}$ data are daily, we will use daily averages for all the covariates in the model aside from traffic.

However, upon examination of the results from the above model, we see that the estimated effects of traffic appear nearly linear (see Figure 8). In fact, the role of traffic appears
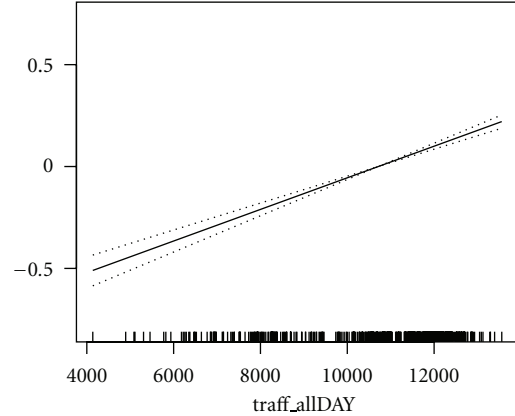


FIGURE 8: Estimated effects of traffic for $PM_{10}$ (model $M_2$).

TABLE 4: Estimated coefficients and standard errors of the parametric part of the additive predictors for $PM_{10}$ (bold indicates 5% significance).

| | PM | |
|---|---|---|
| | Est. | Std. Er. |
| (Intercept) | **3.05** | 0.100 |
| Daily traffic | **0.00008** | 0.00001 |
| Monday | −0.01 | 0.059 |
| Tuesday | 0.09 | 0.062 |
| Wednesday | 0.11 | 0.060 |
| Thursday | 0.08 | 0.060 |
| Friday | 0.05 | 0.066 |
| Saturday | 0.06 | 0.053 |

to be purely linear, without any saturation effects like those observed in the case of $NO_2$. This is expected to some degree, since particulate matter can be produced in large quantities through tire ablation and black carbon smoke, implying that increased traffic leads to increased $PM_{10}$ production. Motivated by this observation, we opted to also fit a simpler model with a simple linear effects of traffic and lagged traffic. Thus, the proposed simpler model is as follows:

$$
\begin{aligned}
M_3: \mu_t = {} & \alpha + s(t, \lambda_t) + \vec{\beta}(\text{DoW}) \\
& + \gamma \, \text{tr} + s(\text{wsp}, \lambda_C) \\
& + s(\text{lag}(\text{wsp}, 1), \lambda_C) + s(\text{lag}(\text{wsp}, 2), \lambda_C) \\
& + s(\text{rh}, \lambda_C) + s(\text{press}, \lambda_C).
\end{aligned}
\tag{4}
$$

The simpler model's BIC was almost identical to the BIC of the model with the splines, motivating us to present the estimated linear effects of the simpler model only in Table 4. The visual results are shown in Figure 9. We can observe that the coefficient of the traffic variable is positive, indicating a positive linear relationship between traffic and daily $PM_{10}$ log concentration.

Figure 9(a) shows a strong relative increase of $PM_{10}$ during wintertime, reflecting confounders like social (e.g., heating) or meteorological (e.g., boundary layer thickness

(a) Time

(b) Temperature

(c) Wind speed

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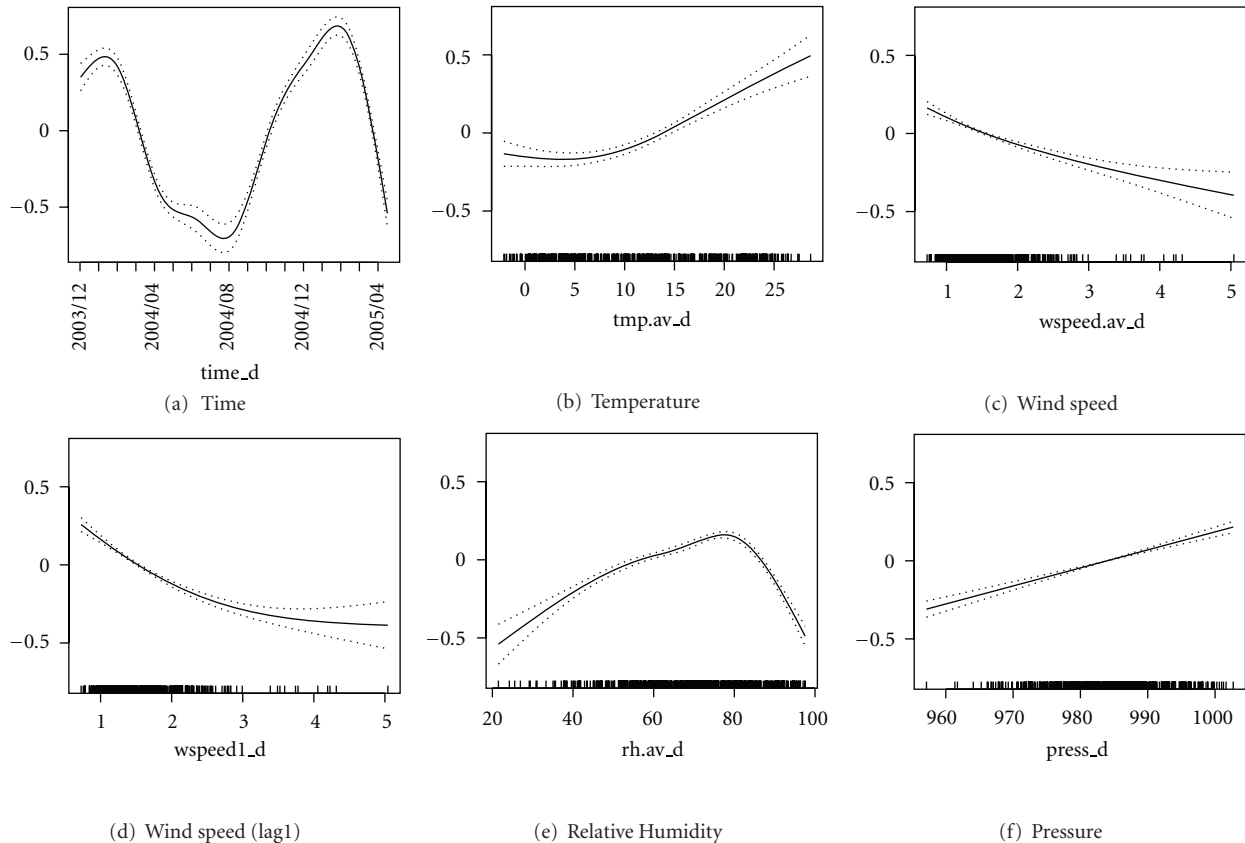
(e) Relative Humidity

(f) Pressure

FIGURE 9: Estimated effects of meteorological variables for $PM_{10}$.

variation) processes. Increase in temperature seems to be associated with an almost linear increase in average $PM_{10}$ log concentration (Figure 9(b)). Increase in wind speed is related to reduced $PM_{10}$ concentration, both for current time (Figure 9(c)) and its one-day-lagged values (Figure 9(d)). Increase in relative humidity is associated with a reduction in average $PM_{10}$ log concentration at high values and with an increase in average $PM_{10}$ log concentration at low values (Figure 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 $PM_{10}$ log concentration (Figure 9(f)).

The linear effects of traffic are good news from the policy point of view, implying that simpler models with linear effects of traffic could be used to replace the more complex ones. In fact, the estimated linear effect of 0.00008 per day implies that a reduction in traffic of 1,300 cars each day would lead to an approximate reduction in average concentration of Turin's $PM_{10}$ of about 10%. Note that 1,300 cars are approximately 10% of the average daily traffic in Turin, so effectively a reduction of 10% in traffic intensity would result in the reduction of 10% in Turin's $PM_{10}$. This is a remarkable result, which could allow for simple and fast implementation and evaluations of policy decisions.

### 3.2.1. Modeling $PM_{10}$ Separately by Seasons.
The analysis stratified by season for log concentration of $PM_{10}$ shows

similar predictor effects and reveals few difference between seasons, as shown in Figure 10. Most predictors show similar behavior across the four seasons. The only exception is summer, with several notable differences. Traffic seems to have a roughly linear effect in all seasons, except in the summer where a slight saturation effect is observed at very high values. Analogously, relative humidity effects are similar in all seasons except for the summer: in all seasons, $PM_{10}$ log-concentration relationship to relative humidity seems quadratic, rising at first and then declining after a certain threshold is passed. This is expected, as relative humidity would be related to precipitation that tends to happen at high relative humidity values and has suppressing effect on particulate matter in the air. However, in the summer, relative humidity seems to have a purely linear effect on $PM_{10}$ log concentration. This too is expected, as relative humidity in the summer tends not to be related to rain but to "hot and humid" days with little wind.

### 3.3. Forecasting for Traffic Regulation Assessment.
Traffic regulation is one of the most important action to reduce the pollution concentrations. The city of Turin lays in one of the most polluted area of Europe. This condition is basically due to the orographic shape of the plain surrounded by mountains and the high density of industry and population. A very common traffic regulation relies on imposing a general reduction of the number of vehicles selecting them by the European pollution category (Euro stages) or by the numbers of the plate (even or odd for "alternate plates").

(a) Traffic

(b) Temperature

(c) Wind speed

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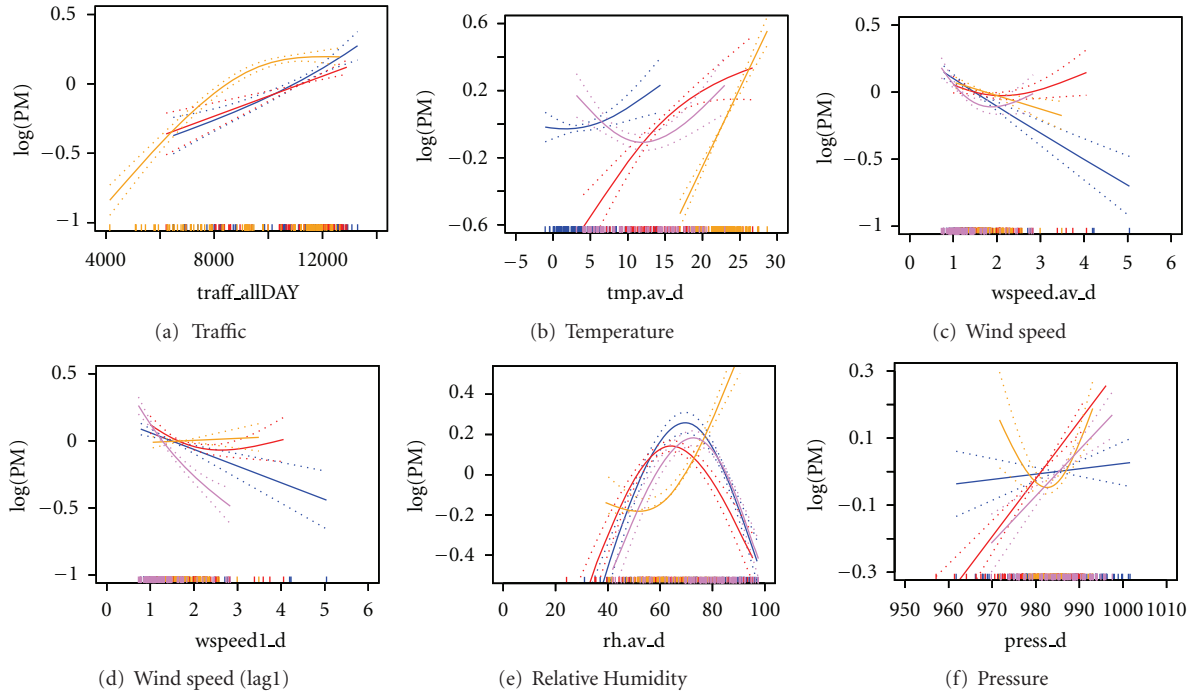
(e) Relative Humidity

(f) Pressure

FIGURE 10: Estimated effects of meteorological variables for $PM_{10}$, by season. Blue is Winter, red is Spring, orange is Summer, and violet is Fall.

In this section we assess the effect of a traffic regulation scenario using GAM prediction. We consider the model $M_1$ for $NO_2$ concentration that was fitted on the whole available data (starting on the 19th December 2003 and ending the 27th April 2005 that is about 11904 hours) and was found to generally fit well, with $R^2 = 0.83$ (Section 3.1). To make prediction on a new dataset, we consider a week during the winter time starting on Tuesday 18th January 2005 and ending on Monday 24th January 2005; we choose this part of the week in order to check possible delayed effect on traffic regulation on the next days. To have a new scenario, we choose to evaluate the type of traffic regulation common in Turin, which controls the movement of cars based on whether the last number of their license plate is odd or even. This policy is generally applied during the most polluted days of the week (i.e., Wednesday and Thursday) and is meant to reduce the circulation of around 50% of the vehicles. Figure 11(a) illustrates what a 50% reduction on two days would look like in the week January 18–24, 2005. When we predict $NO_2$ concentration with the original dataset, we observe in Figure 11(b) that our GAM model is able to describe the variation of the concentration of $NO_2$ (blue line) with respect of the original data (black line), generally following the hourly variation of the measured concentration. Then we use the "new" traffic values, under the reduction scenario, and it is clearly possible to see the short-term effect of this reduction during the two days in the red line in Figure 11. The numerical impact in reduction of pollution concentration, on a weekly basis, is around the 6% of $NO_2$, according to the prediction of the model. During the two regulated days alone, the reduction is around 12%.

## 4. Conclusions

In this paper we have presented a study of air pollution in the city of Turin through the framework of generalized additive models. We have used the generalized additive models (GAMs) to model the behavior of two species of pollutants ($NO_2$ and $PM_{10}$) averaged over the city of Turin as a function of traffic, while controlling for the main meteorological variables as well as an unobserved confounding process. GAMs allow flexible modeling of pollution processes which has traditionally been done in a classical style of differential-equation-based models. In our study, the GAMs have been able to capture the relationship between pollutants and predictors flexibly, using semiparametric 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 helps 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. The results show that for $NO_2$, traffic, in its log-transformed form, is adequate for explaining the log-pollution concentration, while for $PM_{10}$ traffic, in its linear form, turns out to be adequate. We also estimate the relationships between other covariates and the pollutants. An increase in traffic volume is clearly associated
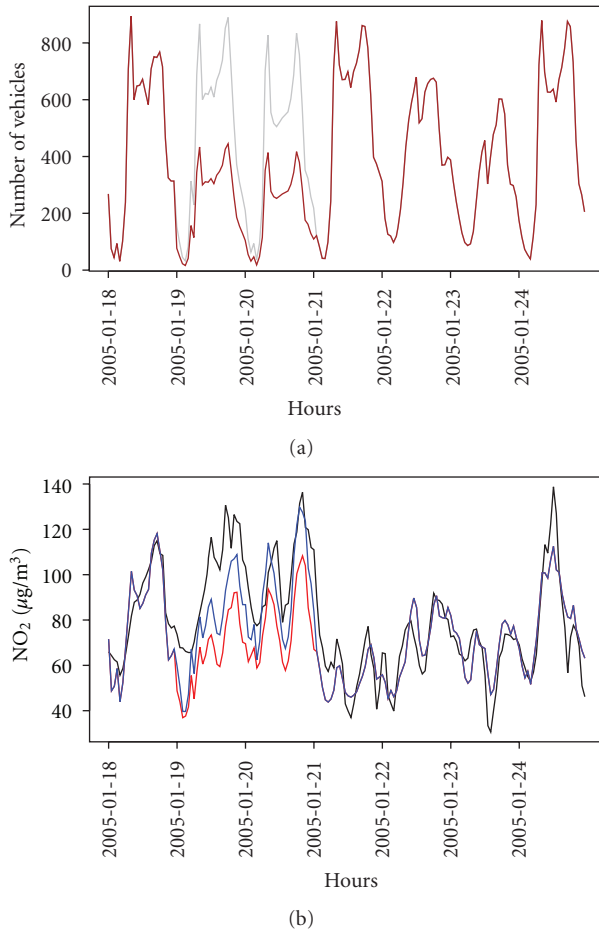
(a)



(b)

Figure 11: (a) Traffic volume in January, 18–24, 2005; original data in grey, policy scenario traffic volume in red. (b) NO$_2$ concentration in January, 18–24, 2005; measured data in black, predicted values without (blue) and with (red) traffic policy scenario.

with increase in the pollutants adjusted for other factors, while temperature, solar radiation, and wind speed have positive partial effects in the pollution reduction, especially in the winter. The nonlinearities found in other estimated effects confirm that the generalized additive models are a useful framework to estimate and interpret the relations between pollution, traffic, and meteorology.

A seasonal analysis provides a detailed description of the predictors' partial effects, where traffic, temperature, pressure, and solar radiation show the more interesting variations. In particular with respect to traffic that is the covariate that can be controlled, our result shows that the variation of the number of vehicles during the cold seasons (and especially in winter) is less effective than within the warm ones. This behavior can be explained by the general higher presence of pollution during cold seasons and by the presence of other sources of emissions, particularly building heaters that during warm seasons disappear. Hence during the winter, an hypothetical traffic regulation certainly helps to reduce the pollution concentration, but an effective reduction of pollution can be reached only working on all the other sources.

Although forecasting in near future time is possible, we suggest to use our proposed models to evaluate traffic reduction policies by predicting pollutant concentrations with policy-modified traffic data, taking into account the meteorological information.

Moreover, during the last year in Turin a progressive increase of the district heating has been undertaken that should reduce the heating-related pollution problem aforementioned. When new data will be available, the models we propose can be useful to obtain new insights and evaluate the effect of this intervention in the city.

## Acknowledgments

## References

[1] A. G. Barnett, G. M. Williams, J. Schwartz et al., "Air pollution and child respiratory health: a case-crossover study in Australia and New Zealand," *American Journal of Respiratory and Critical Care Medicine*, vol. 171, no. 11, pp. 1272–1278, 2005.

[2] M. L. Bell, A. McDermott, S. L. Zeger, J. M. Samet, and F. Dominici, "Ozone and short-term mortality in 95 US urban communities, 1987–2000," *Journal of the American Medical Association*, vol. 292, no. 19, pp. 2372–2378, 2004.

[3] J. M. Samet, F. Dominici, F. C. Curriero, I. Coursac, and S. L. Zeger, "Fine particulate air pollution and mortality in 20 U.S. cities, 1987–1994," *New England Journal of Medicine*, vol. 343, no. 24, pp. 1742–1749, 2000.

[4] B. Mlgaard, T. Hussein, J. Corander, and K. Hämeri, "Forecasting size-fractioned particle number concentrations in the urban atmosphere," *Atmospheric Environment*, vol. 46, pp. 155–163, 2012.

[5] S. Clifford, S. Low Choy, T. Hussein, K. Mengersen, and L. Morawska, "Using the Generalised Additive Model to model the particle number count of ultrafine particles," *Atmospheric Environment*, vol. 45, no. 32, pp. 5934–5945, 2011.

[6] Dir96/62/EC, "Council directive n. 62 of 27 September 1996 on ambient air quality assessment and management," 1996.

[7] Dir99/30/EC, "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," 1999.

[8] Dir00/69/EC, "Council directive n. 69 of 16 November 2000 relatingto limit values for benzene and carbon monoxide in ambient air," 2000.

[9] Dir02/3/EC, "Council directive n. 3 of 12 February 2002 relating to ozone in ambient air," 2002.

[10] Dir50/08/EC, "Council directive n. 50 of 21 May 2008 on ambient air quality and cleaner air for Europe," 2008.

[11] P. Bertaccini, M. Cameletti, and A. Fassó, "Urban mobility and atmosphericpollution within the Torino metropolitan area,"

in *Proceedings of the Atti della Conferenza (SIS '07)*, pp. 607–608, 2007.

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

[13] Y. Kim and J. M. Guldmann, "Impact of traffic flows and wind directions on air pollution concentrations in Seoul, Korea," *Atmospheric Environment*, vol. 45, no. 16, pp. 2803–2810, 2011.

[14] T. J. Hastie and R. Tibshirani, *Generalized Additive Models*, Chapman & Hall, London, UK, 1990.

[15] J. Schwartz, "Non parametric smoothing in the analysis of air pollutionand respiratory illness," *Canadian Journal of Statistics*, vol. 22, no. 4, pp. 471–488, 1994.

[16] D. C. Carslaw, S. D. Beevers, and J. E. Tate, "Modelling and assessing trends in traffic-related emissions using a generalised additive modelling approach," *Atmospheric Environment*, vol. 41, no. 26, pp. 5289–5299, 2007.

[17] M. Aldrin and I. Hobæk Haff, "Generalised additive modelling of air pollution, traffic volume and meteorology," *Atmospheric Environment*, vol. 39, no. 11, pp. 2145–2155, 2005.

[18] M. Aldrin, I. Hobæk Haff, and P. Rosland, "The effect of salting with magnesium chloride on the concentration of particular matter in a road tunnel," *Atmospheric Environment*, vol. 42, no. 8, pp. 1762–1776, 2008.

[19] P. Bertaccini, V. Dukic, and R. Ignaccolo, "Modelling traffic related airpollution in Torino with generalized additive models," in *Proceedings of the 4th International Workshop on Spatio-Temporal Modelling (METMA '08)*, pp. 123–128, Alghero, Italy, 2008.

[20] P. Bertaccini, *Methods and models for vehicular traffic-related atmosphericpollution in the turin urban area [Ph.D. thesis]*, Università degli Studi di Torino, 2009.

[21] P. Kronborg and F. Davidsson, *ImprovementS for a Scandinavian Spoturban Traffic Signal Control SyStem*, TFK—Transport Research Institute, Stockholm, Sweden, 2000.

[22] K. Wood, *Urban Traffic Control, Systems Review*, PR41, Transport Research Laboratory, Crowthorne, UK, 1993.

[23] G. Schwartz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[24] H. Akaike, "A new look at the Bayes procedure," *Biometrika*, vol. 65, no. 1, pp. 53–59, 1978.

[25] R Development Core Team, *R: A Language And Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012, http://www.r-project.org./.

[26] S. N. Wood, "mgcv: GAMs with GCV/AIC/REML Smoothness Estimation and GAMMs by PQL," R package version 1. 7-13, 2012.

[27] S. N. Wood, *Generalized Additive Models: An Introduction With R*, Chapman & Hall, London, UK; CRC Press, Boca Raton, Fla, USA, 2006.