# Modeling pollutant threshold exceedance probabilities in the presence of exogenous variables[1]

Rosaria Ignaccolo
Università di Torino, Italy, ignaccolo@econ.unito.it

Dana Sylvan
Hunter College of the City University of New York, USA

Michela Cameletti
Università di Bergamo, Italy

**Abstract:** Many studies link exposure to various air pollutants to respiratory illness, making it important to identify regions where such exposure risks are high. One way of addressing this problem is by modeling probabilities of exceeding specific pollution thresholds. In this paper, we consider particulate matter with diameter less than 10 microns ($PM_{10}$) in the North-Italian region Piemonte. The problem of interest is to predict the daily exceedance of 50 micrograms per cubic meter of $PM_{10}$ based on air pollution data, geographic information, as well as exogenous variables. We use a two-step procedure involving nonparametric modeling in the time domain, followed by spatial interpolation. Resampling schemes are employed to evaluate the uncertainty in these predictions.

**Keywords:** exceedance probability map, air pollution, space-time modeling

## 1 Motivation, background, data

It is well known that high levels of pollutants in the ambient air have adverse effects on human and environmental health. Environmental directives have been issued in order to account for such potential dangers, setting limit values for various air pollutants. By estimating the probability to exceed a fixed value of a given pollutant, we can identify areas where the risk to exceed such limit values is high. Past environmental studies focused on mean behavior revealed that inclusion of exogenous variables may lead to better estimators and predictors of pollutant concentrations. It seems therefore natural to expect that including additional information, such as meteorological and orographical variables might improve daily predictions of exceedance probabilities. In this study we extend the methodology introduced

[1]Ignaccolo's work was partially supported by Regione Piemonte, while Sylvan's research was funded in part by the PSC-CUNY Research Award 63147-00-41.

in Draghicescu and Ignaccolo (2009) by including exogenous variables. Our case study considers daily $PM_{10}$ concentrations (in $\mu g/m^3$) measured from October 2005 to March 2006 by the monitoring network of Piemonte region (Italy) containing 24 sites. As covariates we use daily maximum mixing height (HMIX, in $m$), daily mean wind speed (WS, in $m/s$), daily emission rates of primary aerosols (EMI, in $g/s$), altitude (A, in $m$) and coordinates (UTMX and UTMY, in $km$). Note that the time-varying variables are obtained from a nested system of deterministic computer-based models implemented by the environmental agency ARPA Piemonte. For a complete description and preliminary analysis of the data we refer to Cameletti et al. (2010).

# 2 Theoretical Framework

Let $D \subset \mathbf{R}^2$, and assume that at each location $s \in D$ we observe a temporal process $X_s(t) = G_s(t, Z_s(t))$, where $G_s$ is an unknown transformation, $Z_s$ is a standardized stationary Gaussian process with $\gamma_s(l) := cov(Z_s(t), Z_s(t+l))$, such that $\sum_{l=-\infty}^{\infty} |\gamma_s(l)| < \infty$. For fixed $x_0 \in \mathbf{R}$, define the exceedance probability

$$\mathbf{P}_{x_0}(t, s) = P(X_s(t) \geq x_0). \tag{1}$$

Clearly $\mathbf{P}_{x_0}(t, s)$ takes values in $[0, 1]$ and is non-increasing in $x_0$. The problem of interest is to predict $\mathbf{P}_{x_0}(t, s^*)$ at location $s^* \in D$ where there are no observations and at any time $t$, based on observations of the process $X_s(t)$ at $n$ time points and $m$ spatial locations.

In the *first step* we use the methodology proposed in Draghicescu and Ignaccolo (2009). For each site $s$, we model the temporal risks non-parametrically, by using the Nadaraya-Watson kernel estimator

$$\hat{\mathbf{P}}_{x_0}(t, s) = \frac{\sum_{i=1}^{n} K\left(\frac{t_i - t}{b_t}\right) 1_{\{X_s(t_i) \geq x_0\}}}{\sum_{i=1}^{n} K\left(\frac{t_i - t}{b_t}\right)}, \tag{2}$$

where $K$ is a kernel function. The temporal bandwidth $b_t$ should not depend on the threshold $x_0$, in order for the resulting estimator to be non-increasing. In what follows, the threshold $x_0$ is considered fixed and, to keep notation simple, we write $b$ instead of $b_t$. In the *second step*, we use universal kriging with exogenous variables to predict the exceedance probability field at a location $s^* \in D$ where there are no observations. Since linear interpolation does not guarantee that the resulting exceedance probability estimator takes values in the interval $[0, 1]$, we first apply a $1:1$ transformation and consider $\hat{Q}_{x_0}(t, s) = \Phi^{-1}(\hat{\mathbf{P}}_{x_0}(t, s))$ which is defined on $\mathbf{R}$, where $\Phi(\cdot)$ is the standard Normal cumulative distribution function. After performing kriging on the transformed field $\hat{Q}_{x_0}(t, s)$, we obtain the desired exceedance probability maps by inversion: $\hat{\mathbf{P}}_{x_0}(t, s) = \Phi(\hat{Q}_{x_0}(t, s))$. For fixed time point $t$ and location $s_i$, we consider the model

$$\hat{Q}_{x_0}(t, s_i) = \beta E(t, s_i) + w(t, s_i), \tag{3}$$

where $E(t, s_i)$ is a vector of exogenous variables at time $t$ and location $s_i$, $\beta$ is the vector of "slopes", and $w(t, s)$ is a zero-mean second-order stationary spatial process for any $s \in D \subset \mathbf{R}^2$. Time point $t$ is fixed, and the spatial covariance is denoted by $C(t, ||s_i - s_j||) := Cov\big(w(t, s_i), w(t, s_j)\big)$. We then use the Matèrn class to model this covariance function: $C(t, ||s_i - s_j||) = \frac{\sigma_t}{2^{\nu_t - 1}\Gamma(\nu_t)} \Big(\frac{2\sqrt{\nu_t}||s_i - s_j||}{\rho_t}\Big)^{\nu_t} \mathcal{K}_{\nu_t}\Big(\frac{2\sqrt{\nu_t}||s_i - s_j||}{\rho_t}\Big)$. The parameter $\nu_t > 0$ characterizes the smoothness of the process, $\sigma_t$ denotes the variance, and $\rho_t$ measures how quickly the correlation decays with distance. For each $t$, the parameters of the Matèrn covariance are estimated by weighted least squares. The best linear unbiased predictor of the transformed field at location $s_0 \in D$ is obtained via universal kriging (Gaetan and Guyon 2010, p. 44) as

$$\hat{Q}^*_{x_0}(t, s_0) = \hat{\beta} E(t, s_0) + w^*(t, s_0). \tag{4}$$

Here $\hat{\beta}$ is the generalized least squares estimate of the trend coefficients and $w^*(t, s_0) = \sum_{i=1}^{m} \lambda_i \hat{w}(t, s_i)$ is the simple kriging predictor, with $\hat{w}(t, s_i) = \hat{Q}_{x_0}(t, s_i) - \hat{\beta} E(t, s_i)$. The weights $\lambda_i$, $1 \leq i \leq m$ are completely determined by the covariance function parameters $\nu_t, \rho_t$, and $\sigma_t$. The standard error of $\hat{Q}^*_{x_0}(t, s_0)$ can be also expressed in terms of the interpolation parameters $\lambda_i$. However, this standard error may not be completely accurate since the Matèrn parameters are estimated from the same data thus adding uncertainty, and the error induced by the first step of our procedure is not considered. For these reasons, we use block bootstrap (Buhlmann, 2002) to take into account all the uncertainty sources.

# 3 Results

In this case study on the North Italian region Piemonte we used data at $m = 24$ sites and $n = 182$ days. The $PM_{10}$ threshold was set to $x_0 = 50\ \mu g/m^3$. The computations were done in `R`, using the `gstat` package (Pebesma, 2004). Regarding the bootstrap, we sampled with replacement $k = 13$ blocks of length $l = 14$ from the $(n - l + 1)$ possible overlapping blocks. We chose $l = 14$ empirically. A temporal window of two weeks captures the meteorological and air pollution patterns well. Also, by trying other values we did not get significantly different results. In future research we plan to generalize the methodology of Li et al. (2007) to more complex dependencies. The block sampling was then repeated $B$ times, yielding the $B$ bootstrap samples. Bootstrap replicated exceedance probability maps were obtained by performing the first and second steps on each bootstrap sample. In the spatial interpolation step we used a $56 \times 72$ regular grid covering Piemonte. Based on the distribution of the $B$ bootstrap replications, we obtained the quantile maps together with the standard errors of the exceedance probability predictions. In our computations we used $B = 500$ bootstrap replications. Maps of the 10th, 50th and 90th percentiles of the exceedance probability bootstrap distribution for March 5, 2006 are showed in Figure 1, identifying increased risks around the metropolitan area of Torino.

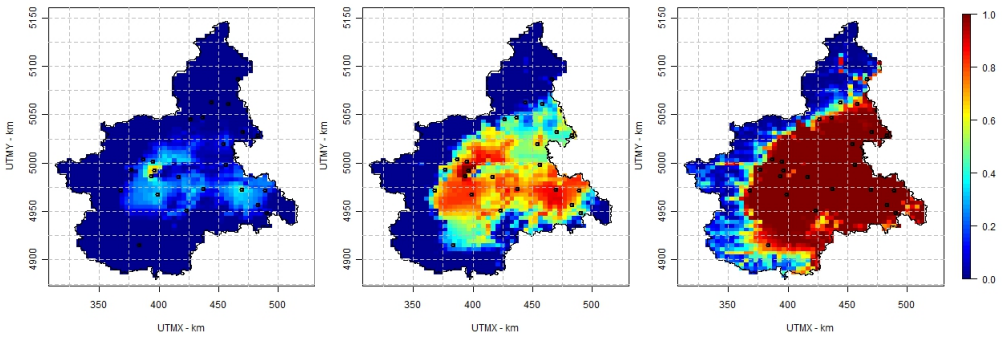


Figure 1: Maps of the bootstrap predicted 50 $\mu g/m^3$ $PM_{10}$ exceedance probabilities on March 5th, 2006: 10th (left), 50th (center) and 90th percentile (right).

# 4 Discussion

This work is a continuation of Draghicescu and Ignaccolo (2009), where preliminary exceedance probability maps were obtained based on a two-step procedure. Seasonal (winter and summer) maps were quite good, however, the daily exceedance probability maps did not seem to reflect the true air pollution spatial patterns well. By introducing exogenous variables we were now able to obtain more reasonable spatial patterns for air pollution risks in Piemonte. In addition, we obtained confidence regions by estimating uncertainty in our predictions through bootstrap. It seems though that the standard errors might be too large, possibly because the shuffling in the block bootstrap did not respect the temporal evolution of the process. Our ongoing research is focused on improving these confidence bands by considering seasonal time series bootstrap.

# References

Buhlmann P. (2002). Bootstraps for Time Series, *Statistical Science*, 17, 1, 52-72.

Cameletti M., Ignaccolo R., Bande S. (2010). Comparing air quality statistical models, *GRASPA Working Papers*, 40 (downloadable at `www.graspa.org`).

Draghicescu D., Ignaccolo R. (2009). Modeling threshold exceedance probabilities of spatially correlated time series, *Electronic Journal of Statistics*, 3, 149-164.

Gaetan C., Guyon X. (2010). *Spatial Statistics and Modelling*. Springer.

Li, B., Genton, M.G., Sherman, M. (2007). Nonparametric assessment of properties of space-time covariance functions, *JASA*, 102, 736–744.

Pebesma E.J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30, 683–691.