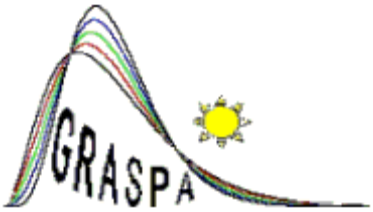
Web Working Papers
by

The Italian Group of Environmental Statistics



Gruppo di Ricerca per le Applicazione della Statistica
ai Problemi Ambientali

www.graspa.org

Hierarchical Space-time Modelling of PM10 Pollution in the Emilia-Romagna Region

Fedele Greco

GRASPA Working paper n.22, May 2005

Hierarchical Space-time Modelling of PM_{10} Pollution in the Emilia-Romagna Region

Fedele Greco
University of Bologna

Abstract: In this paper we propose a hierarchical spatio-temporal model for daily mean concentrations of PM_{10} measured in 11 monitoring sites located in the main cities of the Emilia-Romagna Region. Main aims of the proposed model are: the identification of the sources of variability characterising the PM10 process, the imputation of missing observations in order to obtain time series free from missingness that can be used in ecological regression studies, the estimation of pollution levels in unmonitored spatial locations. The modelling approach is fully Bayesian, the implementation has been performed via Monte Carlo Markov Chain algorithms. The model has been carefully checked using Bayesian *p*-values and graphical posterior predictive checks.

Keywords: Bayesian Hierarchical Models, Air Pollution, Dynamic Linear Models, Space-Time Modelling

1. Introduction

The analysis of the dynamics of airborne particulate matter (PM) concentration is a central issue in environmental monitoring. In fact, several epidemiological studies have shown association between daily levels of PM and adverse health effects (see Pope *et al.*, 1995, for a summary).

In recent years a number of papers has been devoted to spatio-temporal modelling of air pollutants data recorded from monitoring networks (Huerta *et al.*, 2004; Park *et al.*, 2004; Smith *et al.*, 2003; Shaddick and Wakefield, 2002; Tonellato, 2001). Models for air pollutants aim at studying several aspects of the generating process. The detection of space and time long term trend, the location of major air pollution sources in a study area, the design of the positioning of a new monitoring site are the most relevant. Moreover, such models have been devoted to determine whether environmental standards are being met for regulatory purposes. The need for an effective measurement of air pollutants and for regulatory policies arises from the estimated relationship between pollution events and adverse health effects. Although the long term effect of air pollutants are of main interest, the great majority of studies consider short-term or acute effects of air pollution on population health. However, air pollution adverse health effects have been found both in cohort studies (see Pope *et al.*, 1995; Dockery *et al.*, 1993) and in time series studies detecting short-term effects (Samet *et al.*, 2000; Biggeri *et al.*, 2004).

In several of these studies an effect of PM on human health was found, particularly with regard to circulatory and respiratory diseases. For these studies, besides the understanding of the observed processes dynamics, spatio-temporal modelling of PM concentrations can be useful to produce exposure variables useful in ecological risk models by: a) cleaning observed time series from confounding effects and measurement errors; b) adjusting observed time series for missing data; c) estimating the values of exposure variables for sites where data are not available.

In this paper we propose a hierarchical model for daily mean concentrations of PM_{10} measured in 11 monitoring sites located in the main cities of the Emilia-Romagna Region from January 1st 2000 to December 31st 2002. Data are characterized by a considerable presence of missing values. A number of meteorological variables are available in each monitoring site. The proposed model explicitly takes into account the spatial relationship among data collected in each monitoring site, the temporal structure of the observed time series, and the relationships between PM_{10} and meteorological variables. Among the aims of this model there is the identification of the different sources of variability of observed data (spatial variability, temporal variability, variability due to dependence on meteorological conditions, unexplained variability). The model is built in order to allow prediction in spatial locations where data are not available, while less attention is given to prediction in temporal points out of the study period.

As regards inference, we adopt a fully Bayesian approach. Posterior distributions are not obtainable in analytical form because of the complexity of the distributions involved in the hierarchical model. Posterior summaries of model parameters are computed by means of Gibbs sampling routines, as they are implemented in the WinBUGS software (Spiegelhalter *et al.*, 1998).

A further major aim of the model is to impute missing observations in order to obtain time series free from missingness that can be used as covariates in studies of the short term effect of PM_{10} exposure on public health. In the Bayesian context, missing values can be treated as parameters and inference on such values is obtained by integrating out model parameters from the distribution of the missing data given the observed values. This approach to dealing with missing values may be easily implemented in WinBUGS.

The paper is organised as follows. In Section 2 we describe the analysed dataset and the meteorological variables selection criteria. In Section 3 the space-time hierarchical model is presented. In Section 4 we show the results of our application including parameter estimates, missing values imputation, spatial prediction and characterisation of the sources of variability. A discussion of the proposed methods and future developments is then presented in Section 5.

2. Particulate Matter and Meteorological data: preliminary analysis

The analysed data set contains time series of PM_{10} daily means ($\mu g/m^3$) collected at 11 monitoring sites within the Emilia-Romagna Region from January 1st 2000 to December 31st 2002; the spatial location of the monitoring sites is displayed in Figure 1. At least one monitoring site is available for each of the 9 provinces of the Region. This allows achieving a reasonable spatial representativeness: by the way we stress that such representativeness is limited to urban areas. Spatial information at non-urban locations is not available in the dataset.

Percentage of missing values varies from 7% to 40% in the monitoring sites. The monitoring sites have to be distinguished according to their location: 4 of them are located in background urban areas such as parks (Type A) while the remaining 7 are located in zones with high population density or high traffic density (Type B and C). PM_{10} levels are in average lower in Type A monitoring sites while Type B and C monitoring sites show comparable levels. The time series seasonality is very similar regardless of the Type.

Logarithmic transformation has been applied to PM_{10} data in order to obtain a symmetric distribution of the dependent variable in each monitoring site and to stabilize

the mean variance relationship. A strong correlation has been observed among site measurements, ranging from 0.86 for nearest sites to 0.6 for those further away. Even if a slight decrease of correlation with distance is observed, a great amount of the between-sites correlation is due to the common time process generating the data, in fact a strong correlation is observed in very distant monitoring sites time series measurements.

It is well known that measured PM_{10} levels are heavily influenced by the measurement instrument (Ayers *et al.*, 1999): in our data set, the 11 monitoring stations are equipped with automatic samplers, known to produce comparable results.

Meteorological variables for each site are obtained from the mass-consistent model CALMET, implemented by the Emilia-Romagna Regional Meteorological Service. Such model provides estimates on a regular grid of $10km \times 10km$ for daily mean temperature, daily mean mixing height (*MH*) and daily mean wind speed (*WS*). Temperature is highly correlated among monitoring sites (the correlation between time series is always greater than .98). Moreover temperature and *MH* show the same seasonal trend in each site and are highly correlated: inclusion of both variables in a regression model could give rise to collinearity problems. Despite temperature is the most used meteorological variable in space-time modelling of air pollutants, we choose to include only *MH* in the model because of its greater spatial variability that could be useful in explaining different time behaviours of the monitoring sites. Dependence of PM_{10} levels by *MH* has a physical justification since, when *MH* is low, the particulate matter does not spread in the atmosphere, and thus a negative relationship is expected between *MH* and PM_{10}. On the contrary, the relationship between particulate matter and temperature is likely to be spurious: an increase in pollution level when temperature decreases can be explained by anthropic factors such as traffic and emission due to heating. Finally we stress that a crucial meteorological variable in explaining particulate matter variations is precipitation, but unfortunately such variable is not available in our data set.

MH and *WS* values have been centred and divided by their range in order to speed up convergence of the Monte Carlo Markov Chain algorithm used for parameters estimation.

3. The hierarchical model

Let Y_{ts}, MH_{ts}, WS_{ts} denote respectively the log of PM_{10} concentration, the mixing height and the wind speed at spatial location s (s=1,…,S) on day t (t=1,…,T) and let (X_{1s}, X_{2s}) be the spatial coordinates of site s. In the analysed dataset, S=11 and T=1096. We assume that:

$$\begin{aligned} &Y_{ts} \mid \mu_{ts}, \sigma_s^2 \sim N\left(\mu_{ts}, \sigma_s^2\right) \\ &\mu_{ts} = \alpha Z_s + \beta_1 X_{1s} + \beta_2 X_{2s} + \beta_3 MH_{ts} + \beta_4 WS_{ts} + \theta_t + \varepsilon_{ts} \end{aligned} \quad (1)$$

In this model σ_s^2 represents the residual variance in site s. Such parameter includes the measurement error variance as well as the unexplained variability: it is assumed that the error variance does not depend on time, but different monitoring sites are allowed to have different unexplained variances and thus different measurement errors. At the first level of the hierarchy, given μ_{ts} and σ_s^2, observations are modelled as they are

independent: spatial and temporal dependence is introduced in the second level of the hierarchy.

The variable Z is defined as follows: Z_s=1 if the site s is of Type A while Z_s=-1 otherwise; hence the parameter α, a Type-specific intercept, measures the effect of the monitoring site Type on the average log-PM_{10} concentration. We observe that, in order to obtain prediction at unmonitored locations, some information about the level of urbanisation and traffic in such location is needed to classify the spatial location as a Type A or a Type B-C site.

Parameters β_1 and β_2 capture the large-scale spatial trend while coefficients β_3 and β_4 capture the dependence of log-PM_{10} concentrations on the considered meteorological variables. Here we assume that the relationship between meteorological variables and PM_{10} is linear and the same effect is hypothesised in each monitoring site. From preliminary analysis the hypothesis of comparable effect in each monitoring site seems to hold. Moreover, using site-specific slope parameters to describe the relationship between meteorological conditions and PM_{10} prevents from obtaining predictions in spatial locations where data are not available.

As for θ_t, which can be regarded as a time-dependent intercept measuring the mean log-PM_{10} regional level at day t, we assume that:

$$\theta_t = \theta_{t-1} + \omega_t, \quad \omega_t \sim N\left(0, \sigma_\theta^2\right) \tag{2}$$

This represents a first-order smoothing non-stationary temporal model. In terms of Dynamic Linear Models, equation (1) is known as the observation equation, equation (2) is the system equation and θ_t is the state (see West and Harrison, 1997). The model is a limiting form of the autoregressive first order model and provides a non-stationary temporal model. This term has the effect of shrinking observed values at day t toward the regional mean: the deviation of the level in monitoring site s is captured from the information about meteorological conditions, Type of the monitoring site, spatial location and the random effects ε_{ts}. We observe that the atmospheric lifetime of particulate matter can be high, particularly for the smaller size particles, then a strong daily dependence is expected.

The terms ε_{ts} represent spatially correlated random effects. We assume that at each time t, the random effects $\boldsymbol{\varepsilon}_{t\bullet} = \left(\varepsilon_{t1}, \varepsilon_{t2}, ..., \varepsilon_{tS}\right)$ arise from a multivariate normal distribution with mean vector 0 and SxS correlation matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\varepsilon}_t \sim MVN\left(\boldsymbol{0}_s, \sigma_\varepsilon^2 \boldsymbol{\Sigma}\right)$$

A zero-mean constraint for the random effects at each time t has to be used for model identifiability. The parameter σ_ε^2 plays the role of the between site variance. The *ss'* entry of the correlation matrix represents the correlation between site s and s' and is specified as follows:

$$\Sigma_{ss'} = \exp\left(-\phi d_{ss'}\right)$$

This is an isotropic covariance model, where the correlation between two generic sites depends only on their distance. Since the Emilia-Romagna Region exhibits spatially stable meteorology, the isotropy assumption seems not too restrictive. The correlation is assumed to be a decreasing function of the distance $d_{ss'}$: more precisely, under this model the logarithm of the correlation decreases linearly with distance. The parameter $\phi > 0$ describes the decay's rate of correlation with distance. Moreover the spatial structure is considered constant over time: the underlying assumption is that the spatial correlation among random effects does not depend on time, that is spatial and temporal processes are separable. From this point of view, posterior estimates of parameters σ_ε^2 and ϕ can be viewed as estimates of the parameters of the same spatial process over repeated observations.

Model hierarchy is completed by prior specification for the hyperparameters. A normal prior $N(0,1000)$ is assumed for the coefficients β_i, i=1,…,4. Small parameters inverse Gamma distributions ($IG(0.01,0.01)$) have been specified for the variance parameters σ_s^2, σ_θ^2 and σ_ε^2. A uniform distribution $U(0,2)$ is assumed for ϕ: this turns out in a prior belief for the spatial correlation ranging from .13 to 1 at a distance of 1 *km* and from 0 to 1 at the maximum distance of 250 *km*s. Finally a prior distribution is needed for parameter θ_0: we choose a normal prior with mean equal to the observed mean at 31th December 1999 and variance 10 in order to specify a fairly vague prior.

3.1 Model assessment and posterior predictive checks

In order to assess the plausibility of the posited model and the adequacy of model fitting, we make use of the posterior predictive Bayesian *p*-values (Gelman *et al.*, 1996; Rubin, 1984). The rationale on which posterior predictive checks are based is that observed data should look plausible under the posterior predictive distribution or, in other words, the replicated data generated under the model should look similar to the observed data. Thus we draw samples from the posterior predictive distribution of replicated data and compare these samples to the observed data by defining some test statistics $T(\cdot)$ in order to assess different features of model fitting. Let Y^{rep} denote the replicated data that could have been observed coherently with the model, and let ψ denote the set of model parameters. The posterior predictive distribution is obtained as follows:

$$[Y^{rep} \mid Y] = \int [Y^{rep} \mid \psi][\psi \mid Y] d\psi$$

The Bayesian *p*-value is defined as the probability that replicated data are more extreme than the observed data, as measured by the chosen test statistic:

$$p = \Pr\left(T\left(Y^{rep}\right) \geq T(Y) \mid Y\right)$$

In practice, the posterior predictive distribution is obtained via MCMC simulation: given G draws from the posterior density of model parameters $\psi^{(g)}$, g=1,…,G, we draw one Y^{rep} from the posterior predictive distribution for each $\psi^{(g)}$, computing the value assumed from

the test statistic $T(Y^{rep})$ for each draw. The estimated *p*-value is the proportion of these G simulations for which the test statistic $T(Y^{rep})$ equals or exceeds the statistic $T(Y)$. Another effective model checking tool is graphical predictive check: this checking procedure consists in displaying data versus simulated data from the fitted model and looking for systematic discrepancies between real and simulated data set.

4. Results

4.1 Discussion of estimation results

In Table 1 posterior distributions of model parameters are summarized. The posterior means of parameters β_1 and β_2 indicate a decreasing spatial trend in North-South and West-East directions. The effect of the monitoring site Type, as measured by α, has the expected sign, that is the mean level is significantly lower in Type A monitoring sites. A negative relationship has been estimated between the considered meteorological variables and the level of PM_{10} concentrations. In the original scale of the meteorological variables, when *MH* increases 100 *m*, a decrease of 0.02 in PM_{10} concentrations (in the log scale) is estimated and when and *WS* increases 1 *m/s*, a decrease of 0.04 in PM_{10} concentrations (in the log scale) is estimated.

Unexplained variability is quite low for all the monitoring sites but monitoring sites 1, 4 and 7. These monitoring sites show a greater variability with respect to the other monitoring sites, probably because of a greater measurement error and for the massive presence of outliers. Moreover monitoring site 7 is characterised by a period of three months of low measurements likely due to a failure of the measuring instrument.

The posterior mean of parameter ϕ (0.027, see Table 1) shows that the spatial correlation decreases to zero at a distance of approximately 90 *km*s, while spatial correlation is still considerable (greater than 0.5) at a distance of 25 *km*s, showing that a monitoring site can be considered as representative of a relatively wide area.

4.2 Model checking

The effectiveness of the model in predicting missing observations has been checked excluding a period of three weeks from the available data in monitoring site 5 and the predicted values obtained estimating the model with such data have been compared with the observed values excluded from the analysis. As displayed in Figure 3, the model show adequacy in predicting missing observations, since all the observations in the three weeks period are included in the Bayesian posterior credibility interval. We have repeated this exercise in several monitoring sites for different periods of the year, an the model confirmed its ability in predicting missing observations, except for data that can be classified as outliers.

With regard to posterior predictive checks, 1000 post-convergence replications of the data set under the posited model have been drawn. Model fitting has been evaluated separately in each monitoring site with respect to the 5^{th} percentile, the median and the 95^{th} percentile.

While the posterior predictive check on the median describes the adequacy of the model in reproducing the mean level in each monitoring site, the 5th and 95th percentiles give information about model fitting in summer and winter respectively, since PM_{10} levels are sensibly lower in summer than in winter.

When the Bayesian *p*-value is contained between 0.05 and 0.95, the observed data can be considered coherent with the model, with respect to the proposed test statistic. A Bayesian *p*-value lower than 0.05 denotes a systematic underestimation of the test statistic computed on the observed data. On the contrary, a Bayesian *p*-value greater than 0.95 denote a systematic overestimation of the test statistic computed on the observed data.

The Bayesian *p*-values resulting from the estimated model are shown in Table 2. With regard to the median test statistic, we observe a satisfactory model performance for all monitoring sites, since all *p*-values are included in the interval 0.05-0.95. Analysis of the results for the 5^{th} and 95^{th} percentile suggests a better performance of the model in winter: this could be in part explained by a different relationship between PM_{10} levels and meteorological conditions in this season, suggesting a non-linear relationship. Moreover during summer the PM_{10} levels are quite low and explaining the variations in such period is more difficult since a considerable amount of the total summer variability could be due to random fluctuations.

With regard to the model ability in reproducing the between sites correlation, in Figure 4 we display the observed correlations (dots) and the median of the correlations computed on the replicated data sets (triangles): the posterior marginal correlation among the replicated series reproduces adequately the marginal correlation observed in the analysed data set.

4.3 Characterisation of the sources of variability

One of the main aims of the proposed model is the identification of the different sources of variability of the PM_{10} generating process, evaluating the contribution to the total variability due to spatial and temporal trend, as well as dependence from meteorological variables. The strategy we adopt for attributing posterior variability to the different sources is quite similar to the method proposed in Gelman and Pardoe (2005) for evaluating explained variance in multilevel (hierarchical) model. We stress that our model can be considered as a multilevel model in which levels are determined from spatial and temporal dimensions.

In order to characterise the variability of the process generating particulate matter in each monitoring site, we observe that the posterior variance at the second level of the hierarchy can be decomposed as follows:

$$V(\mu_{\bullet s} \mid Y) = V(\theta \mid Y) + V(M_{\bullet s} \mid Y) + V(\varepsilon_{\bullet s} \mid Y) + \\ + Cov(\theta, \varepsilon_{\bullet s} \mid Y) + Cov(\theta, M_{\bullet s} \mid Y) + Cov(M_{\bullet s}, \varepsilon_{\bullet s} \mid Y)$$

where $M_{\bullet s} = \beta_3 MH_{\bullet s} + \beta_4 WS_{\bullet s}$ consists e model components included to take account of the meteorological variables effect. These quantities can be easily evaluated via MCMC sampling.

In Table 3, the components of the linear predictor $\mu_{\bullet s}$ posterior variance are shown. The posterior variance $V(\mu_{\bullet s} \mid Y)$ of the monitoring sites is largely due to the time process θ, that contributes for about 68% to the total posterior variance. Spatial random effects account for a smaller portion of the posterior variability (about 18%), while contribution of the meteorological variables is about 5%. Covariances between spatial, temporal and meteorological component are ignorable. We conclude that a large amount of the variability, given the estimated model, is explained by a common temporal behaviour. Spatial information plays a remarkable role in explaining deviations from such temporal behaviour, while little information is added by the inclusion of meteorological covariates.

Finally, we remark that observed time series can be broadly thought as replications of the same temporal process, with a feeble large scale spatial trend and a spatial correlation that vanishes at a distance of 90 *km*s.

4.4 Spatial prediction

Spatial prediction in unmonitored points can be easily obtained given the proposed model: to obtain such prediction we need to classify the new point as a Type A or Type B/C location and to know meteorological variables in such point during the study period. Let *s'* denote the new monitoring site, prediction at this site at time *t* is obtained by sampling from the posterior predictive distribution $[\mu_{ts'} \mid y]$ whose components are:

$$\begin{aligned} \mu_{ts'} \mid y = {} & (\alpha \mid Y) Z_{s'} + (\beta_1 \mid Y) X_{1s'} + (\beta_2 \mid Y) X_{2s'} + \\ & + (\beta_3 \mid Y) MH_{ts'} + (\beta_4 \mid Y) WS_{ts'} + (\theta_t \mid Y) + (\varepsilon_{ts'} \mid Y) \end{aligned} \tag{3}$$

In model specification we assumed that, at each time *t*, the random effects $\boldsymbol{\varepsilon}_{t\bullet} = (\varepsilon_{t1}, \varepsilon_{t2}, ..., \varepsilon_{tS})$ arise from a multivariate normal distribution with zero mean vector and *S*x*S* correlation matrix $\boldsymbol{\Sigma}$. For predicting spatial random effect in point *s'* we observe that the vector of spatial random effects $(\varepsilon_{t1}, \varepsilon_{t2}, ..., \varepsilon_{tS}, \varepsilon_{ts'})$ follows a multivariate normal distribution with zero mean and $(S+1) \times (S+1)$ variance covariance matrix, with the final row containing the correlations $\sigma_\varepsilon^2 \exp(-\phi d_{ss'})$. Let Ω denote the first 11 elements of this row. The conditional distribution of the random effect in the new monitoring site given the random effects $\boldsymbol{\varepsilon}_{t\bullet}$ is normal with mean and variance:

$$E(\varepsilon_{ts'} \mid \varepsilon_{t\bullet}) = \sigma_\varepsilon^{-2} \Omega' \Sigma^{-1} \varepsilon_{t\bullet} \qquad V(\varepsilon_{ts'} \mid \varepsilon_{t\bullet}) = \sigma_\varepsilon^2 \left(1 - \Omega' \Sigma^{-1} \Omega\right)^{-1}$$

Samples from the predictive distribution $[\mu_{ts'} \mid Y]$ are obtained via MCMC sampling from distribution (3).

For testing the model adequacy in predicting PM_{10} levels in point where data are not available, we have considered a spatial location not considered in model estimation in which data have been recorded for two years during the study period, using is as a test data. This monitoring site, classified as a Type B, is located in the neighbourhood of monitoring sites 7 and 8 included in the analysed data set. The distance between the new monitoring site and the nearest site included in the model is about 4 *km*s. As shown in Figure 5,

predicted values are very close to the true observed values: the correlation between observed values and predicted values is about 0.92 and the most extreme peaks of the test data are captured by the model. The good model performance in predicting test data is not surprising because of three main reasons: first of all, the additional time series is characterised by regular time behaviour and is not affected from high measurement error and massive presence of outliers. Secondly, no matter of the spatial information, the common time process is sufficient in predicting a very large amount of the variability of the time series in a point of the study region, independently from its spatial location. Finally, spatial information in this test is high because of its proximity to monitoring sites 7 and 8 included in the data set, then a considerable improve in prediction in such monitoring site is expected since, as stated in the previous Section, spatial correlation has a non-negligible effect in explaining the PM_{10} variability.

5. Discussion and future developments

We constructed a hierarchical Bayesian model for predicting missing observations, predicting PM_{10} levels where data are not available and characterising the PM_{10} generating process sources of variability. The model has been carefully checked via posterior predictive checks, including Bayesian *p*-values and graphical inspections displaying the data alongside with simulated data from the fitted model. Model adequacy in predicting missing observations has been checked by eliminating some periods of observation from each time series: predictive adequacy has been then evaluated comparing predicted values with observed values excluded from the analysis.

Results show that a large amount of the variability is due to a common temporal process; another important source of variability is small-scale spatial dependence. Variability due to dependence from meteorological variables is ignorable with respect to the other sources of variability. The observed time series in the Emilia-Romagna Region can be broadly considered as replications of the same temporal process: spatial random effects and the terms related with meteorological values take account of the deviations from the common temporal trend. This statement is confirmed by the ability of the model in predicting PM_{10} levels where data are not available (see Section 4.4): we are able to predict data in points out of the monitoring network just because the generating process is quite homogeneous in the study region and few monitoring sites allow obtaining an adequate representation of the process. This is consistent with the fact that Emilia-Romagna is characterised by a considerable homogeneity with respect to meteorological conditions, social behaviours and economic system.

With regard to the modellisation of the relationship between meteorological variables and particulate matter, the most restrictive hypothesis is linearity. Such assumption has been checked by the inclusion of a squared term for both the mixing height and the wind speed. This addition does not improve model fitting, while credibility intervals for the squared term coefficients include zero. The inclusion of time dependent regression parameters, which could be modelled as random walks, gives very instable estimates and poor performances of the MCMC algorithm, that shows convergence problems even after many iterations. The main reason behind this is that for each time point, the regression coefficient has to be estimated on a maximum of 11 data (when no missing data are recorded in day t for every monitoring site). However a non-linear relationship is suggested, mainly because such relationship seems to hold in winter, when PM_{10} levels are

higher, while a weaker relationship is observed in summer when random fluctuations weight more significantly in the total variability. Further modelling of PM_{10}-meteorological conditions relationship would be useful.

It is well known that the measurement instrument heavily influences measured PM_{10} levels: in our analysis we consider only monitoring sites equipped with automatic samplers that produce comparable results. Monitoring sites equipped with measurement instruments (TEOM samplers) that are known to underestimate PM_{10} levels have been excluded from the analysis. Our work will be extended in order to obtain calibration for such measuring instruments.

References

Ayers, G.P., Keywood, M. D. and Gras, J. L., (1999) TEOM vs. manual gravimetric methods for determination of $PM_{2.5}$ aerosol mass concentrations. *Atmospheric Environment*, **33**, 3717-3721.

Biggeri A., Bellini P., Terracini B. *eds*, (2004) Metanalisi Italiana degli studi sugli effetti a breve termine dell'inquinamento atmosferico. *Epidemiologia e Prevenzione*, Special issue (**28**).

Dockery D.W., Pope C.A., Spengler J.D., Ware J.H., Fay M.E., Ferris B.G., Speizer F.E., (1993) An association between air pollution and mortality in six U.S. cities. *The New England Journal of Medicine,* **24**, 1753-1759.

Gelman A., Meng X., Stern H., (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, **6**, 733-807.

Gelman A., Carlin J., Stern H., Rubin D., (2004) *Bayesian data analysis*.2nd Edition Chapman and Hall.

Gelman A., Pardoe I., (2005) Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. To appear in: *Technometrics*.

Huerta G., Sansò B., Stroud J.R., (2004) A spatiotemporal model for Mexico City ozone levels. *Applied Statistics*, **53**, 231-248.

Park E. S., Guttorp P., Kim H. (2004) Location major PM_{10} source areas in Seul using multivariate receptor modelling. *Environmental and Ecological Statistics*, **11**, 9-19.

Pope C. A. et al (1995) Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *American Journal of Respiratory and Critical Care Medicine,* **151**, 669-674.

Rubind D.B., (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, **4**, 1151-1172.

Samet M. *ed*, (2000) *The National Morbidity, Mortality, and Air Pollution Study"* Healt Effect Institute, Cambridge, MA.

Shaddick G., Wakefield J.C. (2002) Modelling multiple pollutant data at multiple sites, *Applied Statistics*, **51**, 351-372.

Smith R.L., Kolenikov S., Cox L.H. (2003) Spatio-temporal modelling of PM2.5 data whit missing values. *Journal of Geophysical Research*, **108** D24.

Spiegelhalter D., Thomas A., Best N., (1998) *WinBugs: Bayesian inference using Gibbs sampler, Manula Version 1.2*. Imperial College, London and Medical Research Council Biostatistics Unit, Cambridge.

Tonellato S.F., (2001) A multivariate time series model for the analysis and prediction of carbon monoxide atmospheric concentrations. *Applied Statistics*, **50**, 187-200.
West M., Harrison J., (1997) *Bayesian forecasting and dynamic linear models*. Springer-Verlag, New York.

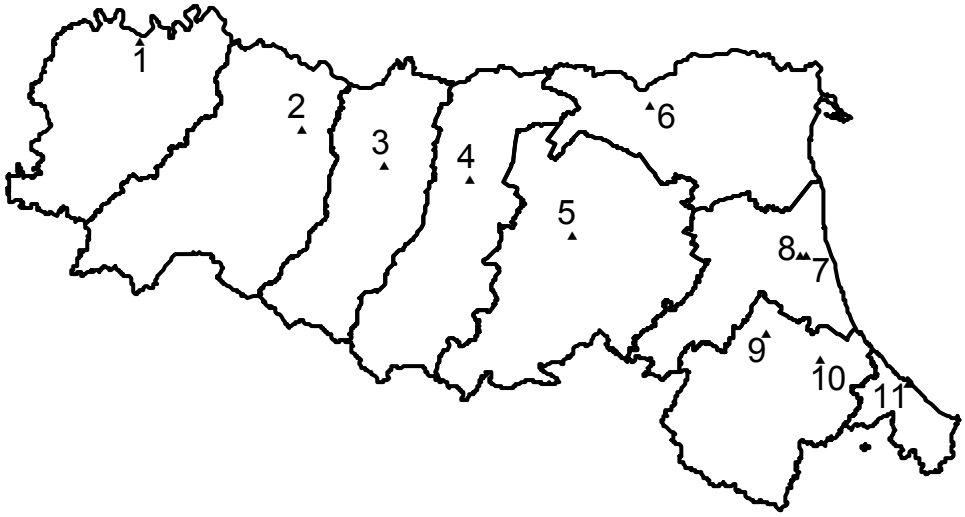


Figure 1: *Monitoring sites spatial locations.*

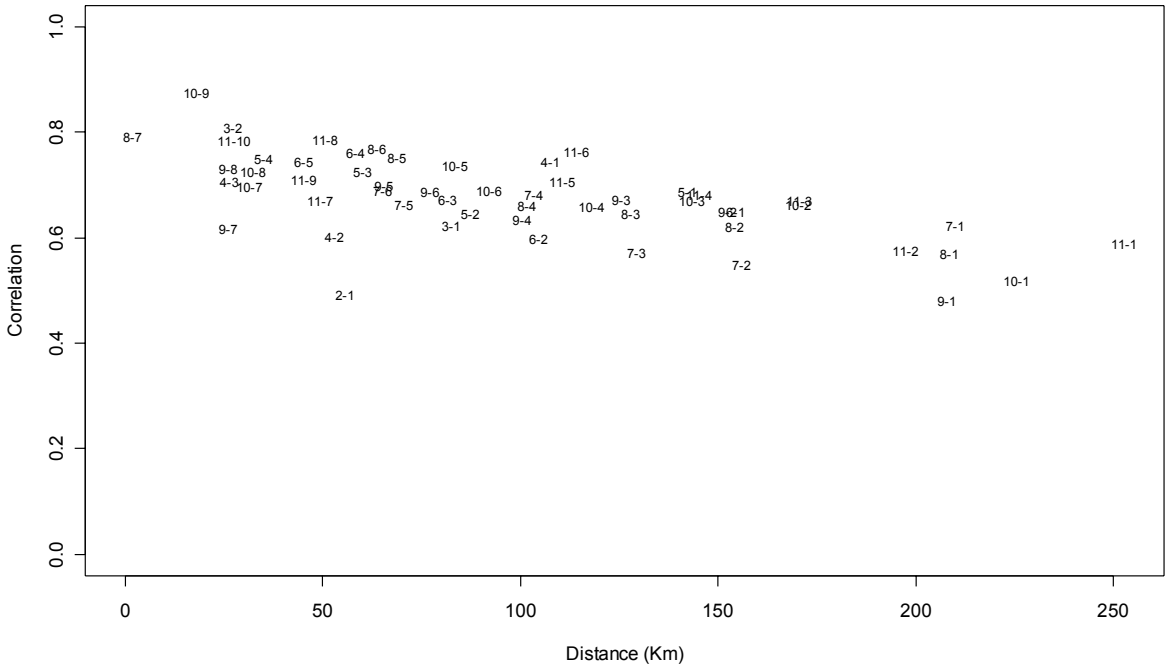


Figure 2: *Between-sites correlations vs distances.*

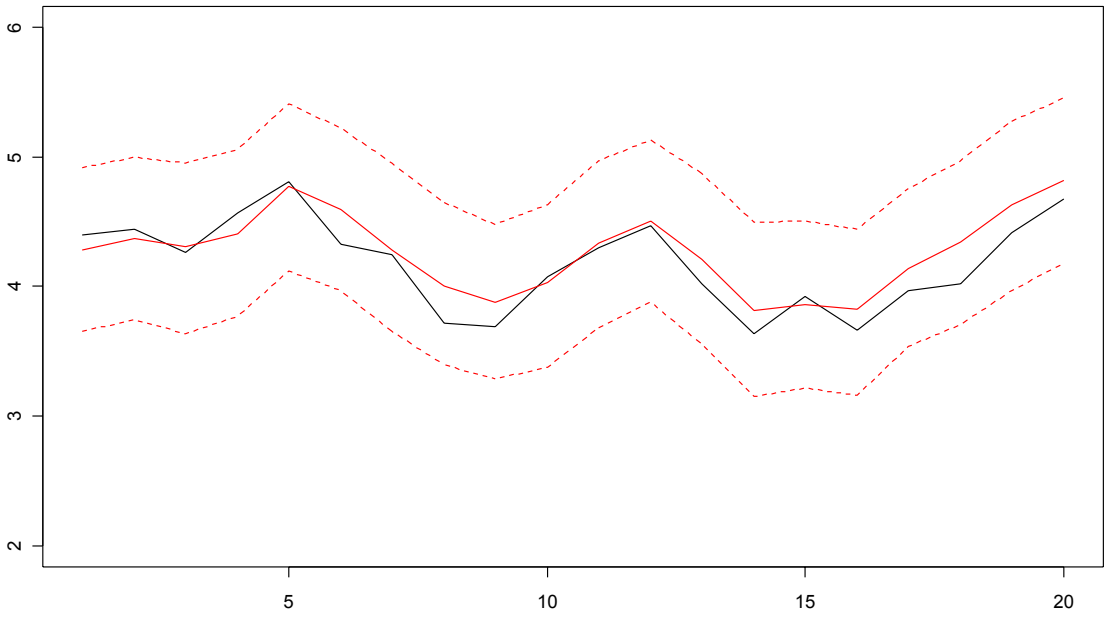


Figure 3: *Missing data imputation: estimated values and 95% credibility interval (red lines); observed data (black line).*

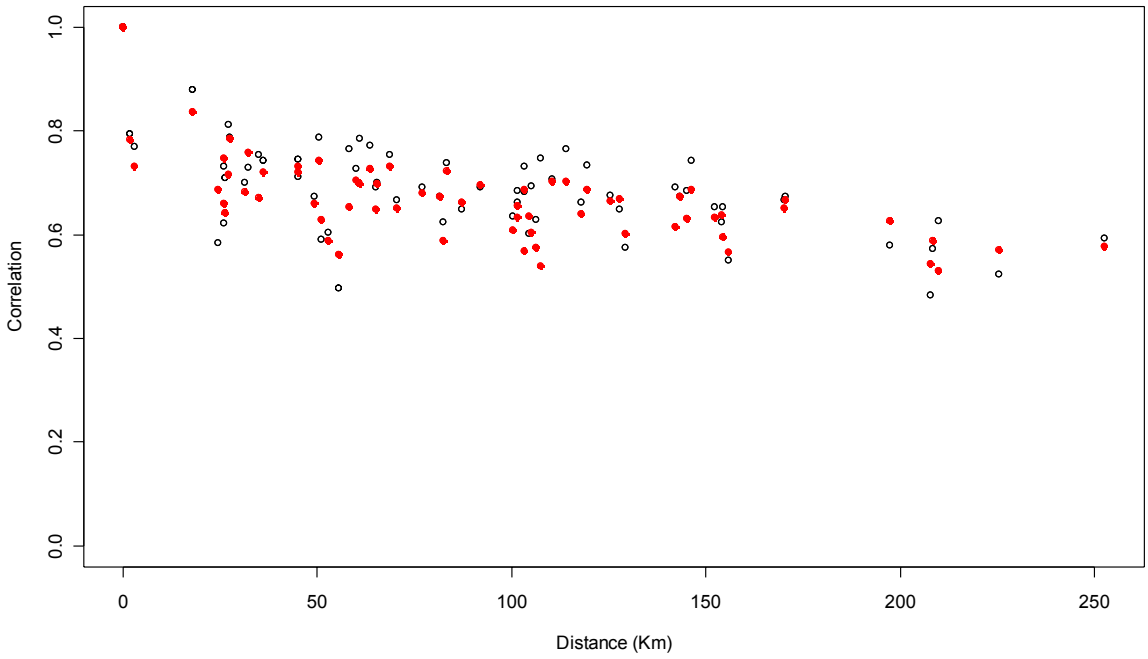


Figure 4: *Observed (black dots) and predicted (red dots) between-sites correlations.*

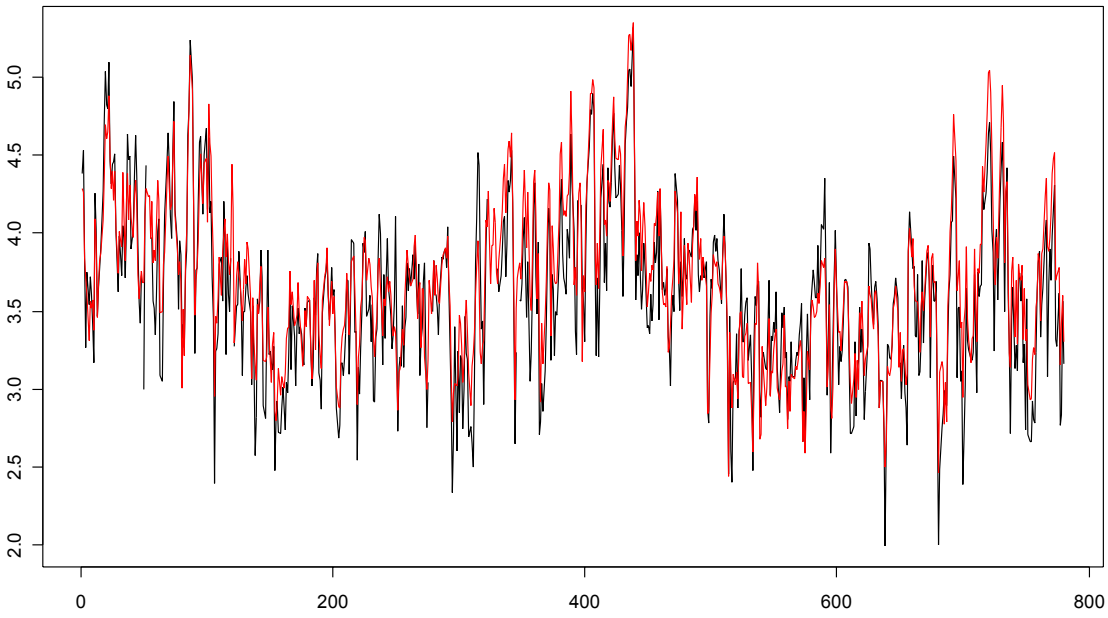


Figure 5: *Observed (black line) and predicted (red line) PM10 levels in the additional monitoring site used as test data.*

	Posterior mean	Posterior median	95% credibility interval			Posterior mean	Posterior median	95% credibility interval	
α	-0.0789	-0.0789	-0.0886	-0.0690	σ_3^2	0.0465	0.0464	0.0343	0.0591
β_1	-0.0013	-0.0013	-0.0016	-0.0011	σ_4^2	0.0917	0.0915	0.0763	0.1083
β_2	-0.0037	-0.0037	-0.0042	-0.0031	σ_5^2	0.0345	0.0344	0.0237	0.0469
β_3	-0.2714	-0.2721	-0.3449	-0.1974	σ_6^2	0.0518	0.0517	0.0381	0.0666
β_4	-0.1659	-0.1650	-0.2356	-0.0973	σ_7^2	0.1014	0.1012	0.0898	0.1138
ϕ	0.0272	0.0271	0.0231	0.0315	σ_8^2	0.0336	0.0336	0.0263	0.0414
σ_ε^2	0.0781	0.0780	0.0702	0.0862	σ_9^2	0.0231	0.0231	0.0156	0.0314
σ_θ^2	0.0553	0.0553	0.0499	0.0613	σ_{10}^2	0.0131	0.0129	0.0084	0.0191
σ_1^2	0.1369	0.1364	0.1138	0.1615	σ_{11}^2	0.0302	0.0299	0.0197	0.0422
σ_2^2	0.0699	0.0698	0.0553	0.0861					

Table 1: *Summaries of model parameters posterior distributions*

Monitoring site	5th percentile	Median	95th percentile
1	**0.992**	0.479	0.679
2	0.387	0.933	0.356
3	0.158	0.141	0.180
4	**0.981**	0.424	0.764
5	**0.998**	0.269	0.720
6	0.770	0.140	0.407
7	**0.996**	0.610	0.784
8	**0.049**	0.554	0.746
9	0.841	0.690	0.883
10	**0.025**	0.296	0.799
11	0.934	0.828	0.770

Table 2: *Bayesian p-values*

Site	1	2	3	4	5	6	7	8	9	10	11
$V(\mu_{\bullet s} \mid Y)$	0.354	0.341	0.343	0.350	0.339	0.350	0.277	0.279	0.294	0.276	0.299
$V(\theta \mid Y)$	0.217	0.217	0.217	0.217	0.217	0.217	0.217	0.217	0.217	0.217	0.217
$V(M_{\bullet s} \mid Y)$	0.019	0.018	0.018	0.018	0.020	0.018	0.017	0.018	0.019	0.018	0.012
$V(\varepsilon_{\bullet s} \mid Y)$	0.076	0.070	0.064	0.057	0.056	0.064	0.043	0.044	0.054	0.053	0.060
$C(\varepsilon_{\bullet s}, \theta \mid Y)$	0.006	0.004	0.008	0.015	0.007	0.012	-0.012	-0.013	-0.011	-0.016	-0.001
$C(\varepsilon_{\bullet s}, M_{\bullet s} \mid Y)$	0.001	0.001	0.002	0.002	0.001	0.001	-0.001	-0.001	-0.001	-0.003	-0.001
$C(\theta, M_{\bullet s} \mid Y)$	0.009	0.008	0.007	0.007	0.009	0.007	0.008	0.009	0.008	0.007	0.005

Table 3: *Explained variances*