Journal of Data Science **7**(2009), 43-59

Two Approaches to Imputation and Adjustment of Air Quality Data from a Composite Monitoring Network

Alessio Pollice[1] and Giovanna Jona Lasinio[2]
[1] *Università degli Studi di Bari and* [2] *Universita di Roma*

Abstract: An analysis of air quality data is provided for the municipal area of Taranto characterized by high environmental risks, due to the massive presence of industrial sites with elevated environmental impact activities. The present study is focused on particulate matter as measured by PM10 concentrations. Preliminary analysis involved addressing several data problems, mainly: (i) an imputation techniques were considered to cope with the large number of missing data, due to both different working periods for groups of monitoring stations and occasional malfunction of PM10 sensors; (ii) due to the use of different validation techniques for each of the three monitoring networks, a calibration procedure was devised to allow for data comparability. Missing data imputation and calibration were addressed by three alternative procedures sharing a leave-one-out type mechanism and based on *ad hoc* exploratory tools and on the recursive Bayesian estimation and prediction of spatial linear mixed effects models. The three procedures are introduced by motivating issues and compared in terms of performance.

Key words: Air quality, Bayesian Kriging, calibration, missing data.

1. Introduction

This paper is motivated by an analysis of air quality data for the municipal area of Taranto (southern Italy) characterized by high environmental risks as formally decreed by the Italian government in the '90s with two administrative measures. This is due to the massive presence of industrial sites with elevated environmental impact activities along the NW boundary of the city of Taranto conurbation. Such activities include one of the largest iron production plants in Europe, an oil-refinery, cement production, fuel storage, power stations, waste materials management, mining industry and many others. Some other highly environmental impacting activities are more integrated within the urban area and have to do with the presence of a large commercial harbour and quite a few military plants (a NATO base, an old arsenal and fuel and munitions storages).

All the afore mentioned activities have effects on the environment and on public health, as a number of epidemiological researches concerning this area reconfirm. As a consequence Taranto was subject to intensive monitoring of the main pollutants in the last few years, leading to an unusually fine composite network that lends itself to the reconstruction of spatial fields at the city level.

Spatio-temporal modelling of PM concentrations can be useful to understand the process dynamics and in producing exposure variables for ecological risk models, as assessing the association between daily concentrations of particulate matter and adverse health effects was the objective of many studies in recent years (Pope *et al.*, 1995 and Biggeri *et al.*, 2004 among others).

Here our main concern is on two methods for pre-processing data recorded from an air quality monitoring network characterized by missing data and heterogeneity. The primary aim is the proposal of a statistical protocol for missing data imputation and adjustment, in view of the spatio-temporal modelling of air quality data. A recent paper by Fasso *et al.* (2007) partly shares the same objectives and also addresses space-time modelling by a geostatistical dynamical calibration model based on a multivariate state-space formulation, dealing with the high dimension of the state equation by the Empirical Orthogonal Function (EOF) approach. The first to introduce a reduced dimension space-time Kalman filter were Goodall and Mardia (1994) who proposed the Kriged Kalman filter (KKF, Mardia et al, 1998). Parameter estimation for both KKF and the afore mentioned EOF-based approach is carried out in the maximum likelihood framework, while Bayesian versions of space-time Kalman filter models were first introduced by Wickle *et al.* (1998), followed by Sahu and Mardia's Bayesian Kriged Kalman filter (BKKF, 2005) and Xu and Wickle (2007) EOF-based model, among others. In a recent paper (Sahu *et al.*, 2005) a point is given in favour of the use of Bayesian Gaussian random effect models (Bayesian LME's) instead of BKKF when there is a reasonable suspect that the space-time process is separable. The application of a separability test by Fuentes (2006) proved that this was indeed the case for the data at hand. In recent years a number of papers was devoted to spatio-temporal modelling of air quality data by Bayesian LME's (Cocchi *et al.*, 2007; Shaddick and Wakefield, 2002 among others). In the following sections we propose and compare three alternative procedures which are suitable for data imputation and adjustment. These procedures share a leave-one-out type mechanism and are based on *ad hoc* exploratory tools and on the recursive Bayesian estimation and prediction of spatial LME's.

The paper is organized as follows. In section 2 we describe the PM concentration data of the Taranto area. These data come from a composite network since they are collected and validated by the regional and municipal governments with different protocols. Section 3 contains a discussion of the methodology used

for missing data imputation and adjustment. In particular in subsection 3.1 we introduce some *ad hoc* tools based on spatial and dynamic regression within a leave-one-out type mechanism. We consider the latter to be a baseline standard approach, to be improved by the Bayesian model-based ones reported in subsection 3.2. A detailed comparison among the performances of the *ad hoc* tools and two Bayesian model-based methods is given in section 4, while section 5 contains some concluding remarks.

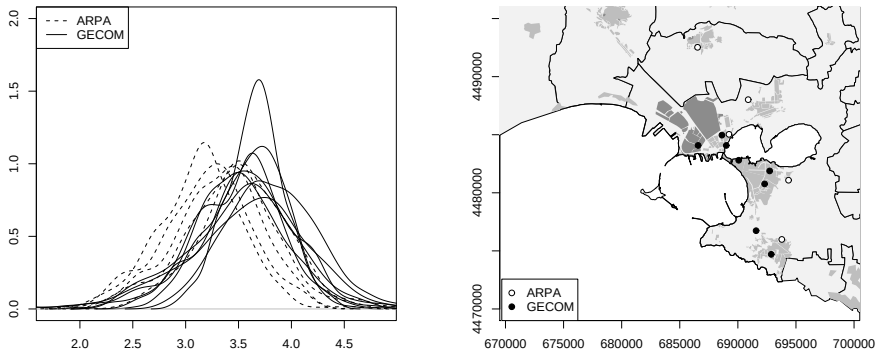

Figure 1: Smooth density estimates of log-average daily PM10 concentrations and spatial distribution of the stations belonging to the ARPA and GECOM monitoring networks (the stations Archimede and Orsini are almost overlapping).

2. The Data

In the context of an agreement between Dipartimento di Scienze Statistiche - Università degli Studi di Bari and the local regional environmental protection agency (ARPA Puglia) air quality data for the municipal area of the city of Taranto were provided belonging to three different monitoring networks pertaining to the regional and municipal government and counting 25 monitoring stations on the whole. Pollutants continuously monitored by some of the stations include sulphur dioxide (SO2), nitrogen oxide (NOx) and dioxide (NO2), carbon monoxide (CO), benzene, PM10 and ozone. At present validated data for the three networks are available for only one common operating period corresponding to year 2005. The present study is focused on particulate matter as measured by PM10 concentrations. All the stations monitoring PM10 are equipped with analogous instruments based on the Beta absorption technology, either reporting hourly, two-hourly or daily measurements.

Log-average daily concentrations were obtained and the 14 stations monitoring PM10 were split into two groups according to the data validation protocol used: the 6 instruments controlled by the regional government (ARPA) were considered to be far more reliable than those managed by the municipal government (GECOM), except for one ARPA sensor which produced sensibly lower values (and was re-calibrated during 2006). The use of the ARPA measurements as reference values lead to the exclusion of that station from the data-base. Among the 13 remaining stations the GECOM ones often appeared to overestimate PM10 concentration levels (Fig. 1 (a)), this behaviour being only partly attributable to the more peripherical location of the ARPA sensors (Fig. 1 (b)). Some adjustment of the GECOM data was thus deemed necessary to allow for data comparability.

A large number of missing data was observed (Tab. 1), due to both different working periods for groups of monitoring stations and occasional malfunction of PM10 sensors. Missing data were thus considered to be missing at random (MAR) wether they occurred during the operating periods of the measuring devices.

Table 1: Taranto PM10 concentration data, year 2005: operating periods starting dates and percentages of missing daily averages (ARPA stations in bold).

Station	Starting date	% missing
Ancona	01/01/2005	1.10
Camuzzi	01/01/2005	2.19
Carcere	01/01/2005	1.64
Gennarini	01/01/2005	9.32
Stadio	01/01/2005	9.59
Talsano	01/01/2005	9.04
Talsano (A)	01/01/2005	2.74
Testa	01/01/2005	2.74
Paolo VI	15/01/2005	9.86
Peripato	15/01/2005	25.75
Orsini	08/02/2005	17.81
Archimede	07/04/2005	29.04
Statte	07/04/2005	34.79

3. Methods

Missing data is a ubiquitous problem in evaluating long-term experimental measurements, such as those associated with air quality monitoring. Spatio-temporal modelling often implies that such gaps in the measured data are filled or imputed. So far, no standardized method has been accepted and the imputation methods used are largely dependent on the researchers' choice.

The objective of the methods to be described in this section is to obtain a "clean" database by imputing missing values and adjusting data recorded at presumably overestimating (GECOM) stations. These tasks are undertaken by three alternative procedures. In the first occasional NA's are imputed by linear spatial regression, then GECOM data are calibrated by dynamic linear models within a leave-one-out scheme (this procedure will be denoted by *ad hoc*). The other two procedures both rely on the Bayesian estimation of daily hyerarchical spatial linear models (Bayesian Kriging) within a leave-one-out scheme for missing data imputation and data adjustment (they will be called `krg` and `s/t krg` respectively). In section 4 the three procedures are compared considering the performance of the first as a baseline standard.

3.1 Ad hoc exploratory tools

First of all to deal with occasional NA's assumed to be MAR (i.e. those occurred during the instruments operating periods), an imputation technique based on linear spatial regression is used (Le and Zidek, 2006). In particular, consider measurement at site i missing at time t:

i. fit a regression model with PM10 log-mean daily concentrations at site i as predictand and the time series of all other sites available at time t as predictors;

ii. obtain the prediction of the fitted model at time t;

iii. impute the missing value by simulating at random from a Normal distribution with mean and variance respectively equal to the fitted value and the estimated residual regression variance.

Within this procedure no spatial correlation structure is assumed for the imputed data, all the stations being considered equivalent in the linear predictor.

As a second step adjustment of data recorded at the GECOM stations is dealt with. Generally speaking *calibration* is often referred to as the process of adjusting the output of a measurement instrument to agree with the values of some specified standard. In Statistics calibration is a reverse process to regression (Osborne, 1991) and can be summarized as follows:

i. a *dependence model* is estimated between a response variable (the specified standard) and an explanatory variable (the output of the measurement intrument);

ii. it is used to obtain *predictions* of other values of the explanatory variable from new observations of the response variable.

In the Taranto case-study each GECOM monitoring station is adjusted to agree with the values reported by a specified station belonging to the ARPA network (taken as the reference standard). Notice that in order not to generate further missing values in the adjusted GECOM data, only two ARPA stations were available, i.e. those having no missing values after the imputation process (Carcere and Talsano (A)). Then for each GECOM station one of the two ARPA stations was chosen as a reference standard on the ground of spatial proximity (Fig. 1 (b)) and maximum correlation (Tab. 2). In this case no *new* values of the output of the measurement instrument are available to obtain predictions of the reference standard, but rather the same observations of the explanatory variable are used to base model estimation and prediction. This is accomplished within a leave-one-out scheme where each daily observation is deleted in turn and the dependence model is estimated by the remaining 364. The prediction of the log-average daily PM10 concentration at the ARPA station for the deleted day is then considered as the corresponding adjusted measurement for the GECOM station.

Table 2: Correlation matrix between stations belonging to the GECOM (rows) and ARPA (columns) networks

	Archimede	Carcere	Paolo VI	Talsano (A)	Statte
Ancona	0.18	0.69	0.56	0.68	0.65
Camuzzi	0.55	0.53	0.54	0.56	0.49
Gennarini	0.29	0.67	0.69	0.76	0.72
Orsini	0.67	0.30	0.31	0.28	0.11
Peripato	0.68	0.66	0.70	0.76	0.53
Stadio	0.34	0.40	0.39	0.44	0.45
Talsano	0.08	0.43	0.38	0.52	0.47
Testa	0.28	0.50	0.54	0.58	0.59

The form chosen for the dependence model was dynamic linear regression (Pankratz, 1991), where the AR(1) autocorrelation structure was assumed for both the response and the explanatory variable:

$$Y_t = \phi Y_{t-1} + \beta_1 X_t + \beta_2 X_{t-1} + \epsilon_t \tag{3.1}$$

here Y_t and X_t respectively stand for the log-average PM10 concentrations at the ARPA and GECOM stations on day t and ϵ_t is an i.i.d. random term. Model (1) is estimated by OLS and adjusted values of the GECOM output are obtained as $\tilde{X}_t = \hat{Y}_t$ within the afore mentioned leave-one-out scheme.

3.2 Bayesian Kriging

While explicitly taking the temporal correlation structure into account, the methods outlined in the latter section don't include any spatial information. We then consider them to be a baseline standard approach to be improved by a

unique "spatial" procedure, to be used for both missing data imputation and data adjustment. The basic idea is to use daily spatial interpolation models, in order to predict missing and overestimated data. This approach is taken to obtain an efficient tool for data pre-processing, reducing the computational complexity implied by considering a full spatio-temporal model. Alternatively the consideration of a unique marginal spatial model would lead to neglect the predictable changes in the spatial structure of the data along time. Hierarchical Bayesian models embracing properly defined spatial autocorrelation structures can admit any pattern of missing measurements in a partially observed spatial process, as this approach provides a predictive distribution that can be used for imputation.

The usual LME model is chosen as the daily spatial interpolation model (Diggle and Ribeiro, 2007):

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + S(\mathbf{s}) + \epsilon(\mathbf{s})$$

where $Y(\mathbf{s})$ is the observed process at a set of n spatial locations $\mathbf{s}$, $\mu(\mathbf{s})$ is a spatial trend, $S(\mathbf{s})$ is a Gaussian spatial random effect and $\epsilon(\mathbf{s})$ is an independent random error term. More precisely:

- $\mu(\mathbf{s})$ is a function describing the large scale variation of the spatial phenomenon. It can be modeled as a function of covariates (coordinates and/or other spatial information) or set as a constant;
- $S(\mathbf{s})$ is a second order stationary isotropic Gaussian spatial process with null mean and covariance structure depending on the distance between spatial locations through two parameters σ^2 and ϕ, respectively the variogram *sill* and a vector of spatial correlation parameters;
- $\epsilon(\mathbf{s})$ is a vector of i.i.d. Normal random variables with null mean and common variance τ^2 (the variogram *nugget*). $\epsilon(\mathbf{s})$ accounts for measurement error and microscale uncertainty, i.e. the *noise* affecting the readings of the spatial signal.

Here we assume the trend to be constant $\mu(\mathbf{s}) = \beta_0$, concentrating our attention on the latent spatial part of the process $S(\mathbf{s})$. Prior specification then concerns parameters β_0, σ^2, τ^2 and ϕ. Diffuse priors are chosen for β_0 and σ^2. A scalar correlation parameter ϕ (the variogram *range*) is considered, measuring how quickly the correlation function decays to a particular reference value when the separation distance between pairs of locations increases. The prior represents a guess about its possible values, varying in the interval $[0, \infty)$. The `krige.bayes` function of the `geoR` R library implements two types of prior distributions for discrete sets of values of ϕ:

1. flat priors: $pr(\phi) \propto 1$,
2. decreasing priors: $pr(\phi) \propto 1/\phi^{\delta}$ or $pr(\phi) \propto \exp(-\delta\phi)$, with $\delta > 0$

The uniform prior (type 1) represents the belief that, a priori, all the values in the specified discrete set are equally plausible. Priors of type 2 allow the user to flexibly choose the shape hyperparameter δ and express a prior belief that small values of ϕ in the discrete set are more likely.

The so called noise-to-signal ratio $\tau^2_{rel} = \tau^2/\sigma^2$ can be treated as a fixed or random quantity. In the second case it is possible to describe the prior knowledge in the same way adopted for the correlation parameter ϕ.

For missing data imputation and adjustment we propose a procedure making use of two daily spatial kinds of models specified as Bayesian LME's, namely *prediction models* and *estimation models*. The structure of the algorithm is iterative and based on a leave-one-out scheme: iterations are necessary to reconstruct the spatial variation, while the by leave-one-out scheme we obtain homogeneous predictions of data to be calibrated, missing or outlying.

Available daily averages are often too few and show little spatial structure: our way to make use of such slight spatial variability to obtain spatially smooth daily series without using any external information consists in iteratively estimating the spatial structure parameters and predicting the observations to be treated. Along the iterations estimates and predictions become more and more stable and the daily spatial structure becomes more apparent. Further insights on the iteration mechanism and its effectivness are given in section 4 (figure 2 (a) and (b) and comments).

Within each iteration we obtain predictions using as many data points as possible: a single daily model predicting all observations to be treated by those not to be treated would be too expensive in terms of degrees of freedom and estimates' variability would change daily as a function of the number of observations to be treated. Then daily spatial prediction models are fitted within a leave-one-out scheme and used to predict each observation to be treated by all the data available on the same day. This procedure is repeated updating observations to be treated untill convergence is reached.

The iterative reconstruction of the daily spatial variation and the prediction of missing and overestimated values are formalized in the following `krg` algorithm. Let x be the vector of daily observations and J the set of indices denoting the monitoring stations to be treated.

step 0.1 Fit the estimation model to vector x where data corresponding to the stations to be treated are omitted. After some sensitivity analysis a discrete uniform prior was chosen for τ^2_{rel} on the interval (0,1) with 0.1 increments.

We allowed ϕ to vary in a discrete sequence between 1 and 7 km with 0.5km incremental value; a type 2 (reciprocal) prior was judged appropriate (faster convergence and less smoothing in the returned values). Obtain posterior estimates of ϕ and τ^2_{rel}.

step 1.1 For $i \in J$ let $x_{(i)}$ be obtained by omitting station i in the vector of daily observations x. Iteratively predict each x_i from $x_{(i)}$ using posterior estimates of ϕ and τ^2_{rel} obtained in the previous step in the prior specification of the prediction models. Store predicted values in vector z and substitute them to corresponding values in x.

step 2.1 Store the current z values in z_{old} and repeat step 1.1 to obtain a new z.

step 3.1 If $|z_{old} - z| < \varepsilon$ ($\varepsilon = 0.0001$) or the iterations number is ≥ 100 stop, otherwise repeat step 2.1 until convergence.

An explicit consideration of the time correlation structure is neglected in the `krg` algorithm. Exploratory data analytic results (Tab. 6) show that, as expected for PM10 concentrations (Cocchi *et al.*, 2007), the time series have an autoregressive order 1 correlation structure. In order to include this time dynamic into the Kriging procedure daily priors are recursively updated at each iteration in algorithm `s/t krg`. More precisely the priors of the prediction models are daily updated by posterior estimates obtained by the estimation model on the previous day. The spatial variation is thus believed to follow a sort of order 1 time dependence, with daily covariance parameter estimates depending stochastically on those of the day before. Then step 0.1 in `krg` takes the following modifyed form, while next three steps remain the same:

step 0.2 For day 1 run step 0.1. For days 2 to 365 fit the estimation model to vector x of the previous day, where data corresponding to the treated stations z are substituted. The same priors as in step 0.1 are used.

Notice that the leave-one-out structure of both algorithms implies that each observation to be treated is updated by a prediction model, on the basis of the values at the previous iteration. As a by-product estimates of covariance structure parameters τ^2_{rel} and ϕ for each treated observation at each iteration of the two algorithms are obtained as summaries of posterior simulations and are used to assess convergence.

4. Results

The methods outlined in §3 were used to obtain MAR data imputation and adjustment of the GECOM data.

In Fig. 2 the developement of the estimates of τ^2_{rel} and ϕ when passing from one iteration to the next is shown for one day representative of the overall behaviour. It is quite evident that while the covariance structure parameters are updated within `s/t krg`, it is not so in `krg`. This means that in `krg` the provisional a priori spatial structure remains substantially unchanged along the iterations and that treated observations are iteratively adapted to this structure untill convergence. On the contrary `s/t krg` simultaneously adapts the spatial structure and the observations to be treated, so that the final spatial structure is the result of an adaptive iterative process. Given that the spatial structure "moves", the need of a larger number of iterations for `s/t krg` to reach convergence is justified, as shown in Tab. 3. Notice that for both algorithms the maximum iteration number is far below the 100 units limit.

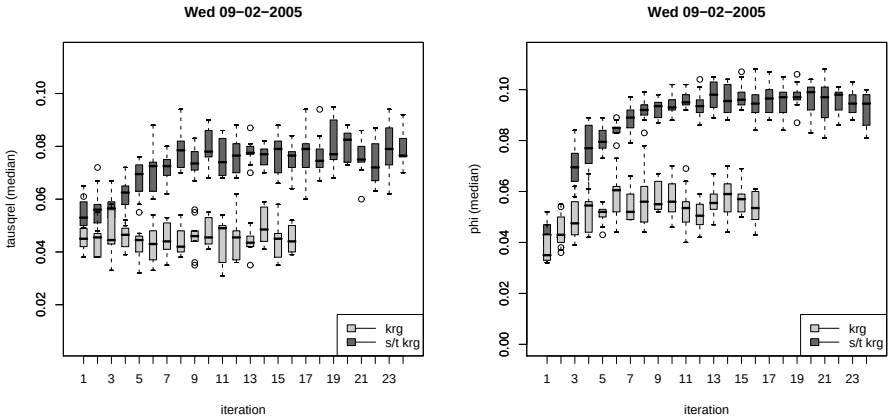

Figure 2: Posterior summaries (medians) of τ^2_{rel} and ϕ by iteration number for day 9/2/2005 (similar results for the 365 days are available from the authors on request).

Table 3: Summary statistics for the number of iterations of the two Bayesian Kriging estimation algorithms

	Minimum	1st Q.	Median	Mean	3rd Q.	Maximum
`krg`	5.00	8.00	9.00	9.81	11.00	25.00
`s/t krg`	6.00	10.00	11.00	13.88	16.00	49.00

For the 8 stations belonging to the GECOM network the outlined procedures produced time series of adjusted data that were compared to the observed ones in terms of root mean squared error. Tab. 4 (left) shows how only for the Gennarini and Stadio stations the *ad hoc* method produces time series closer to those observed. Indeed in these two cases the three predicted time series and the observed one are very close and thus the need for adjustment is questionable.

Table 4: Left: differences between observed and adjusted (GECOM) data (ARPA stations used to obtain *ad hoc* predictions: (*) Carcere, (**) Talsano (A)). Right: differences between observed and calibrated GECOM data and the closest ARPA station (†Carcere, ‡Archimede, §Talsano (A)): root mean squared errors.

	ad hoc	`krg`	`s/t krg`	obs	*ad hoc*	`krg`	`s/t krg`
Ancona (†)	0.48 (*)	0.32	0.42	0.532	0.233	0.277	0.208
Camuzzi (‡)	0.56 (*)	0.38	0.44	0.413	0.567	0.261	0.262
Gennarini (§)	0.28 (**)	0.30	0.39	0.359	0.232	0.154	0.163
Orsini (‡)	0.77 (*)	0.51	0.57	0.416	0.546	0.234	0.236
Peripato (‡)	0.57 (*)	0.33	0.39	0.338	0.535	0.323	0.329
Stadio (†)	0.35 (*)	0.39	0.45	0.556	0.353	0.271	0.204
Talsano (§)	0.46 (*)	0.38	0.43	0.425	0.342	0.141	0.148
Testa (‡)	0.61 (*)	0.47	0.56	0.610	0.592	0.328	0.333

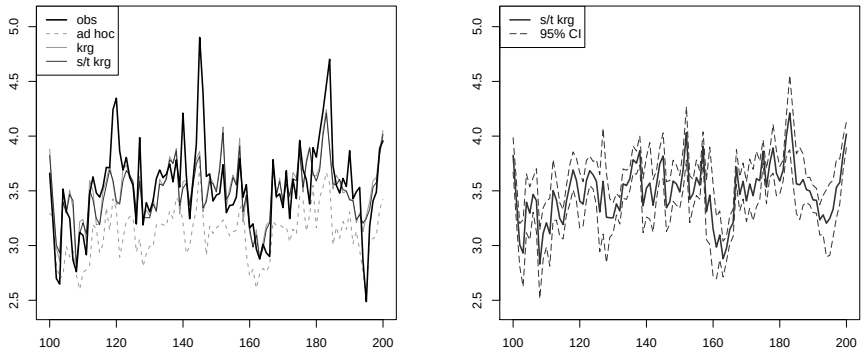

Figure 3: Log-average daily PM10 concentrations for the Camuzzi GECOM monitoring station before and after imputation and calibration (days 100 to 200: 10/4/2005-19/7/2005): (a) comparison among the three methods, (b) `s/t krg` predictions and 95% credibility intervals (similar results for all the 8 GECOM stations are available from the authors on request).

For the remaining six GECOM stations the *ad hoc* predictions are always smaller and far from the Kriging-based ones (as noticed in Fig. 1 (a) the GECOM stations tend to overestimate PM10 concentrations, so adjusted data will be smaller than observed ones), due to the former adjustment method being based on only one reference station rather than on the spatial structure reproduced by all available stations (Fig. 3 (a)). The two Kriging-based methods behave quite similarly, though `krg` predictions denote slightly larger values and variability (less smoothing). Fig. 3 (a) also shows that the shift obtained by the three adjustment procedures does not alter the time dynamics observed in the time series. In figure 3 (b) credibility intervals based on the 2.5% and 97.5% percentiles of the

simulated predictive distribution prove to be quite narrow for the `s/t krg` case. Similar results were obtained for the two Kriging-based methods and for all the eight stations belonging to the GECOM network, the details being available from the authors on request. Notice that the inspection of the graphs didn't produce any evidence for the daily IC's size to vary as a function of the number of stations to be treated.

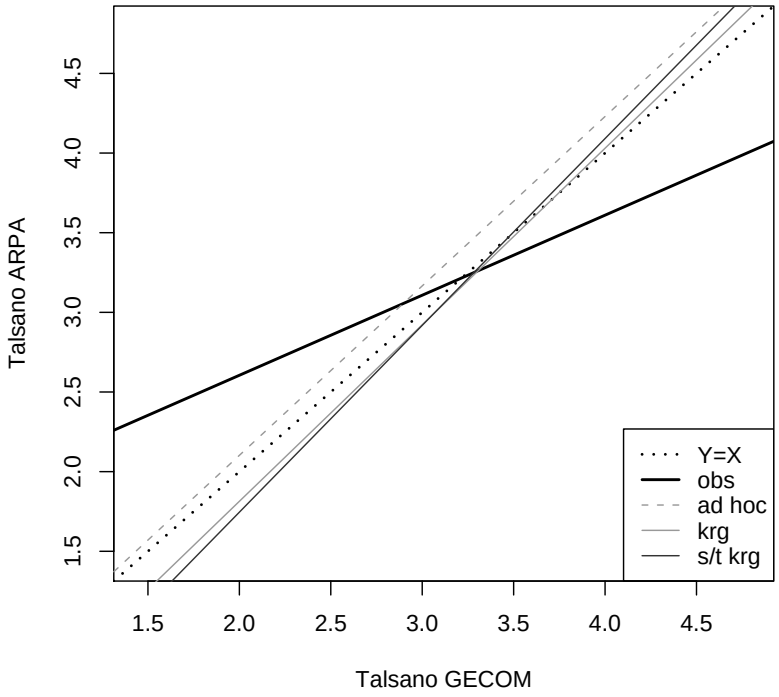

Figure 4: Differences between observed and calibrated data for the GECOM Talsano monitoring station and the closer ARPA Talsano station: regression lines (similar results for all the 8 GECOM stations are available from the authors on request).

For the five ARPA stations missing data imputations obtained by the three methods were quite similar though, as for the GECOM network, more extreme and smaller values were obtained by the `krg` and *ad hoc* methods respectively.

With the aim of obtaining a first assessment of the spatial variability reproduced by the three methods, a second type of diagnostic was produced in order to compare adjusted GECOM data to those observed at the closer ARPA station (considered as a data quality reference standard). For the 8 GECOM monitoring

stations Tab. 4 (right) contains the root mean squared errors corresponding to the regression lines given in Fig. 4 for the Talsano station. It is clear that the *ad hoc* method is the worst among the three and does a poor job in terms of spatial variation, especially when the station used for adjustment does not coincide with the closest one (as in the case of the Ancona and Gennarini stations), due to the presence of missing data. The *ad hoc* method thus results to be very sensible to the choice of the reference station used for adjustment. Algorithm `krg` and `s/t krg` perform quite similarly in reproducing the spatial variation, providing a sharp correspondence between calibrated data for each GECOM station and those observed for the closest ARPA station.

For PM10 concentration data a strong daily dependence is expected, due to the high atmospheric lifetime of smaller size particles (Cocchi *et al.*, 2007). In figure 5 (a) the boxplots of the partial auto-correlation functions of the 13 monitoring stations are reported (first four lags) showing higher values corresponding to lag 1 for observed and adjusted data. The 13 log-average PM10 concentration time series show a similar AR(1) time-correlation structure and the same conclusion is fostered by the direct inspection of empirical ACF's (not reported).

As a matter of fact adjusting the data according to the three proposed procedures does not alter the AR(1) time-correlation structure. Direct inspection of the 13×4 PACF's shows that they almost invariably fall below significance boundaries for lags greater than 1. Notice that higher values at lag 1 imply that `s/t krg` reproduces a stronger daily dependence with respect to the other two methods and foster the impression of a higher degree of temporal smoothing of `s/t krg` already obtained by the inspection of fig 3 (a) and similar unreported graphs.

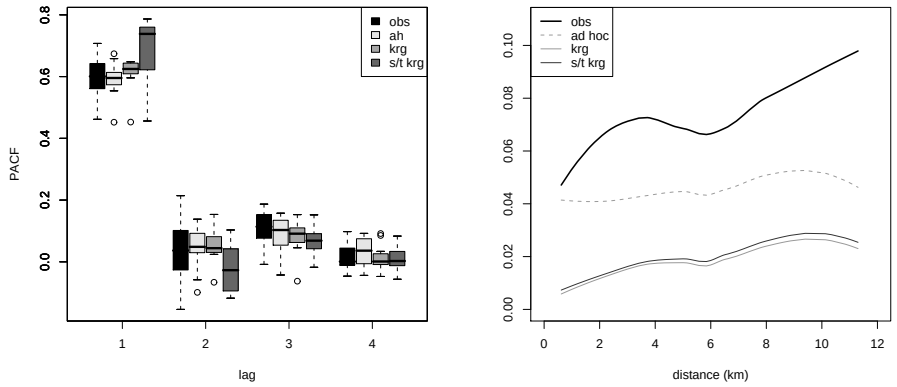

Figure 5: (a) Boxplot of 13 partial auto-correlation functions for observed and adjusted data (first four lags) and (b) smooth loess curves interpolating time de-trended variogram clouds.

Finally empirical variograms of the data were obtained to further investigate their spatial variation before and after missing data imputation and adjustment. To remove the temporal trends the residuals after fitting an AR(1) model to the 13 time series for both the raw PM10 daily log-averages and those after imputation and adjustment were obtained. The estimated autocorrelation functions of the residuals (not shown) confirmed that there were no more temporal effects, then the variation in the resulting data could be expected to have arisen from variation due to space.

Let $w(\mathbf{s}_i, t)$ denote the residuals at location $\mathbf{s}_i$ ($i = 1, \ldots, 13$), time t ($t = 1, \ldots, 365$), assumed to be independent replications at location $\mathbf{s}_i$ since data were time de-trended. We now consider an average variogram estimator (Sahu and Mardia, 2005) defined by

$$\hat{\gamma}(d_{ij}) = \frac{1}{2T} \sum_{t=1}^{T} \{w(\mathbf{s}_i, t) - w(\mathbf{s}_j, t)\}^2$$

where d_{ij} is the distance between the spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$ and $T = 365$. The empirical variogram cloud is obtained by plotting $\hat{\gamma}(d_{ij})$ against d_{ij} for the $13(13-1)/2 = 78$ possible pairs of locations. In figure 5 (b) smooth loess curves (obtained by the R function `loess` with smoothing parameter equal to 0.1) interpolating variogram clouds obtained with the previous method for time de-trended observed and adjusted data are provided.

The *ad hoc* method, which has no consideration for the spatial variation, flattens the variogram observed for the original data producing an almost "pure nugget" spatial random field. On the contrary both Kriging-based methods tend to preserve the spatial variation adjuststing the GECOM data by values perfectly matching the fitted spatial structure and thus lead to a substantial reduction of the nugget effect. This is a desirable feature of the procedure as we expect that a large part of the measurement error is due to the different calibration of the two networks. While sharing the small nugget with `krg`, the smooth variogram of the `s/t krg` method increases more rapidly, implying that the method including the time variability constraint produces more smoothing in the time series and a higher overall spatial variation (sill).

5. Conclusions

In this paper we compare two approaches to the imputation of missing data and calibration of measurements coming from different monitoring networks. More precisely the two methodologies produce adjusted values of the log-average PM10 concentrations for the GECOM network and allow missing data imputation at the ARPA monitoring stations. The first proposed technique (*ad hoc*)

based on linear spatial regression and on a dynamic regression model, does not explicitly account for the presence of spatial variation in the data. As a consequence the initially observed spatial variability is almost eliminated from the final adjusted data set. On the other hand this method preserves the time variability structure quite well. Being based on two different statistical models operating sequentially, the *ad hoc* method does not allow to exactly asses the precision of the final estimates. The latter and the elimination of the spatial variation observed in the data can be considered as serious drawbacks to the adoption of the *ad hoc* approach in practice. The `krg` method, based on Bayesian Kriging, explicitly accounts for spatial variation and its space-time version `s/t krg` for time dynamic as well. Both methods rely on an iterative leave-one-out structure and consider daily spatial models of PM10 concentrations: we avoid the computational complexities implied by considering a full spatio-temporal model and use iterations to reconstruct the spatial variation and the leave-one-out scheme to obtain homogeneous predictions of data to be treated. Indeed temporal and spatial variability prove to be appropriately rebuilt in the final series by these two methods. Furthermore as the imputation/calibration procedure allows to sample from the model predictive distribution, it is possible to build credibility intervals for each treated observation in order to evaluate its precision and that of the overall procedure. Small side effects of the use of Bayesian methods are the computational complexity and time consumption. On the other hand both `krg` and `s/t krg` can be easily implemented in the R environment using library `geoR`. Thus if imputation and adjustment are prerequisites to the reconstruction of spatial fields, the two alternative Kriging-based procedures are suggested. To choose between the two one can consider that the purely spatial `krg` makes use of current day data to set prior distributions and is thus more appropriate when a purely spatial approach to data treatment is recommended.

The space-time Bayesian procedure `s/t krg` revealed to be the most appropriate for the Taranto log-PM10 data. It showed a good capability of spatial variation reconstruction and time dynamic preservation. The slightly higher degree of temporal smoothing together with the larger overall spatial variability (variogram sill) imply a preference of this algorithm with respect to `krg` from the information conservation point of view. `s/t krg` is fairly computationally efficient, due to its iterative nature it provides stable a posteriori estimates/predictions, and enables to assess estimates precision. The exploitation of both temporal and spatial structures to impute missing data, adjust observations and treat outliers revelead to be the best strategy.

References

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

Cocchi, D., Greco, F., Trivisano, C. (2007). Hierarchical space-time modelling of PM10 pollution. *Atmospheric Environment* **41**, 532-542.

Diggle, P. J., Ribeiro, P. J. (2007) *Model-based Geostatistics*. Springer.

Fassò, A., Cameletti, M., Nicolis, O. (2007). Air quality monitoring using heterogeneous networks. *Environmetrics* **18**, 245-264.

Fuentes, M. (2006). Testing for separability of spatio-temporal covariance functions. *Journal of statistical planning and inference* **136**, 447-466.

Goodall, C., Mardia, K. V. (1994). Challanges in multivariate spatial modelling. *Prooceedings of the XVIIth International Biometric Conference*, Hamilton, Ontario, Canada, 8-12.

Le, N. Z., Zidek, J. V. (2006). *Statistical Analysis of Environmental Space-Time Processes*. Springer.

Little, R. J. A., Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley.

Mardia, K.V., Goodall, C., Redfern, E., Alonso, F. (1998) The Kriged Kalman filter. *Test*, **7**, 217-285.

Osborne, C. (1991). Statistical calibration: a review. *International Statistical Review* **59**, 309-336.

Pankratz, A. (1991). *Forecasting with Dynamic Regression Models*. Wiley.

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

Shafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman& Hall.

Sahu, S. K., Jona Lasinio, G., Orasi, A., Mardia, K. V. (2005). A comparison of spatio-temporal Bayesian models for reconstruction of rainfall fields in a cloud seeding experiment. *Journal of Mathematics and Statistics* **1**, 4, 273-281.

Sahu, S. K., Mardia, K. V. (2005). A Bayesian Kriged Kalman model for short-term forecasting of air pollution levels. *Applied Statistics* **54**, 223-244.

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

Wikle, C. K., Berliner, L. M., Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics* **5**, 117-154.

Xu, K., Wikle, C. K. (2007). Estimation of parametrized spatio-temporal dynamic models. *Journal of Statistical Planning and Inference* **137**, 567-588.

Received July 29, 2007; accepted September 26, 2007.

Alessio Pollice
Dipartimento di Scienze Statistiche "Carlo Cecchi"
Università degli Studi di Bari
Via C. Rosalba n. 53, 70124 Bari, ITALY
apollice@dss.uniba.it

Giovanna Jona Lasinio
Dipartimento di Statistica, Probabilità e Statistiche applicate
Universita di Roma "La Sapienza"
P.le Aldo Moro n. 5, 00185 Roma, ITALY
giovanna.jonalasinio@uniroma1.it