Geostatistical integration and uncertainty in pollutant concentration surface under preferential sampling

Laura Grisotto,[1] Dario Consonni,[2] Lorenzo Cecconi,[1] Dolores Catelan,[1,3] Corrado Lagazio,[4] Pier Alberto Bertazzi,[2] Michela Baccini,[1,3] Annibale Biggeri[1,3]

[1]Department of Statistics, Computer Science, Applications, University of Florence; [2]Epidemiology Unit, Department of Preventive Medicine, Ca' Granda Hospital, Milan; [3]Biostatistics Unit, Institute for Cancer Prevention and Research, Tuscany Region, Florence; [4]Department of Economics, University of Genoa, Genoa, Italy

Abstract

In this paper the focus is on environmental statistics, with the aim of estimating the concentration surface and related uncertainty of an air pollutant. We used air quality data recorded by a network of monitoring stations within a Bayesian framework to overcome difficulties in accounting for prediction uncertainty and to integrate information provided by deterministic models based on emissions meteorology and chemico-physical characteristics of the atmosphere. Several authors have proposed such integration, but all the proposed approaches rely on representativeness and completeness of existing air pollution monitoring networks. We considered the situation in which the spatial process of interest and the sampling locations are not independent. This is known in the literature as the preferential sampling problem, which if ignored in the analysis, can bias geostatistical inferences. We developed a Bayesian geostatistical model to account for preferential sampling with the main interest in statistical integration and uncertainty. We used PM_{10} data arising from the air quality network of the Environmental Protection Agency of Lombardy Region (Italy) and numerical outputs from the deterministic model. We specified an inhomogeneous Poisson process for the sampling locations intensities and a shared spatial random component model for the dependence between the spatial location of monitors and the pollution surface. We found greater predicted standard deviation differences in areas not properly covered by the air quality network. In conclusion, in this context inferences on prediction uncertainty may be misleading when geostatistical modelling does not take into account preferential sampling.

Correspondence: Laura Grisotto, Department of Statistics, Computer Science, Applications G. Parenti, University of Florence, viale Morgagni 59, 50134 Florence, Italy.
Tel: +39.0552.751500 – Fax: +39.0554.223560.
E-mail: grisotto@disia.unifi.it

Key words: Preferential sampling; Prediction uncertainty; Bayesian geostatistics; Air pollution; Italy.

Acknowledgements: we thank the Air Quality Unit of the Lombardy Regional Agency for the Environment and the Health Directorate of the Region of Lombardy Government for providing the data that made this work possible. This work was supported by the Regional Government of Lombardy, Milan (grant VIII/10462), and the Ministry of University and Scientific Research, Rome, Italy.

Received for publication: 3 November 2015.
Revision received: 19 January 2016.
Accepted for publication: 22 January 2016.

Geospatial Health 2016; 11:426
doi:10.4081/gh.2016.426

Introduction

Geostatistics refers to statistical methods used with data obtained by sampling a spatially continuous phenomenon at a discrete set of locations in the region of interest (Cressie, 1991). Generally speaking, the interest is in predicting the mean surface for the phenomenon under study. In this paper, we focused on environmental statistics with the aim of estimating the concentration surface and related uncertainty of an air pollutant from air quality data recorded by a network of monitoring stations. We did so within a Bayesian framework to overcome difficulties in measuring prediction uncertainty (Diggle *et al.*, 1998; Pilz and Spöck, 2008; Vicedo-Cabrera *et al.*, 2013; Cecconi *et al.*, 2016), which are usual when land-use regression (Hoek *et al.*, 2008) or ordinary Kriging (Banerjee *et al.*, 2004; Son *et al.*, 2010) are used. However, it is important to acknowledge that, alternatively to statistical approaches, deterministic models based on emissions meteorology and chemico-physical characteristics of the atmosphere are of great value [*e.g.* community multi-scale air quality (CMAQ) (http://www.epa.gov/asmdnerl/CMAQ; Zanini, 2009] and might be preferable to approaches based on observed monitor data. Indeed, the number of monitoring stations in existing air quality networks can be very small and the monitors may be sensible to local disturbances, which affects the validity of the data for interpolation of the concentration levels. Integration of the two approaches – statistical modeling of observed concentrations and deterministic emissions modelling – has been proposed by several authors, *e.g.* Berrocal *et al.* (2010). However, such approaches rely on representativeness and completeness of existing air pollution monitoring networks. Air quality networks may be problematic in this respect, because the geographical location of the monitors may have been deliberately chosen for a number of reasons, including: i) background pollution levels outside urban areas, where

the location is chosen on the basis of a prior expectation of low concentration levels; ii) air quality in residential areas, where the location is chosen on the basis of population density and land use; and iii) pollutant concentrations near important emission sources, where the location is chosen on a prior expectation of high concentration levels (Guttorp and Sampson, 2010).

In these cases, the spatial process of interest and the sampling locations are no longer independent. Diggle *et al.* (2010) referred to this problem as *preferential sampling* and showed that geostatistical inferences can be biased if ignored in the analysis. Apart from this reference, several authors have addressed this issue methodologically and provided examples in environmental epidemiology (Gelfand *et al.*, 2012; Diggle *et al.*, 2013; Lee *et al.*, 2015; Shaddick and Zidek, 2015). In a Bayesian contest, Pati *et al.* (2011) proposed a joint modelling of the point process for the sampling locations and the point referenced spatial process for the spatial intensity.

Most of the literature references on preferential sampling focuses on the potential bias in the prediction of pollutant surfaces of geostatistical inferences if preferential sampling is not accounted for. We address here situations in which the accurate estimate of the prediction uncertainty is the main goal. Baccini *et al.* (2015) give an example where statistical integration and uncertainty in pollutant concentration surface is used in health impact assessment of air pollution. However, the geostatistical model used in that paper did not account for preferential sampling, so we developed a Bayesian geostatistical model to account for preferential sampling, when the main interest is in statistical integration and uncertainty. Taking PM_{10} data arising from the air quality network of the Environmental Protection Agency of Lombardy Region in Italy and numerical outputs from deterministic modelling, we focused on the estimate of the prediction uncertainty surface. An inhomogeneous Poisson process for the sampling locations intensities and a shared spatial random model for the dependence between the spatial location of monitors and the pollution surface are specified.

Materials and Methods

Study area

The Lombardy Region has about 10 million inhabitants. The capital city of Milan with its 1.3 million inhabitants is the largest metropolitan area in the region. Part of this territory (40.5%) is mountainous and stretches towards the Alps, 12.5% is located in the declining hills, the Pre-Alps Region, and is occupied by highly industrialized areas. The remainder of the territory (47%) is represented by the plains of the Po River, a predominantly agricultural region (Figure 1). Climatic conditions that are unfavourable to the dispersion of pollutants are present and create a *basin* effect with longstanding thermal inversion periods during winter. Indeed, Lombardy has one of Europe's highest pollution levels (van Donkelaar *et al.*, 2010).

Motivating example

In the paper by Baccini *et al.* (2015), uncertainty arising from different sources was propagated to the impact estimates of PM_{10} on mortality in the Lombardy Region. Annual concentrations of PM_{10} for the calendar year 2007 were available from two sources: a Eulerian photochemical model (Silibello *et al.*, 2008), which was applied on a domain of 244 x 236 km^2 with a resolution of 4 km; and the regional air quality monitoring network of the Regional Environmental Protection Agency (ARPA) of Milan. After controlling for consistency and completeness (Baccini *et al.*, 2011), validated data on PM_{10} concentrations were available for 58 air quality monitoring stations (Figure 2). With the aim of obtaining predictions and related standard deviations on the domain of the Eulerian photochemical model to be used for health impact calculations, Baccini *et al.* (2015) used a Bayesian geostatistical model. Here we evaluate the effect of preferential sampling on prediction uncertainty.

A general preferential sampling model

Following the notation by Pati *et al.* (2011), we are interested in estimating the concentration surface $\mu(s) \in \Re^2$ in the domain $D \subset \Re^2$ accounting for the location sampling process on $s \in D$.

The preferential sampling model is specified as $\mu(s)=\eta(s)+\alpha\xi(s)$ where the first term is a spatially structured term and the second the adjustment for preferential sampling. The term α is a tuning parameter.

In the model proposed by Pati *et al.* (2011) the spatially structured term $\xi(s)$ is the log-intensity of an inhomogeneous Poisson Point process for the monitor locations. Notice that in this modelling specification, the location sampling process is a point process with a continuous spatial intensity on the space, the concentration surface is estimated by a finite number of sampling points and the prediction surface is continuous.

Shared models for preferential sampling

We propose a shared spatial random component model (Held *et al.*, 2005) for preferential sampling adjustment. Monitoring stations location and pollutant concentrations are not independent as their correlation depends on an underlying latent spatial process. However, we cannot exclude specific uncorrelated spatial patterns in the two processes, so we specified a model with shared and specific spatial random components.

Let X_s denote the presence of a monitoring station at location s, $\lambda(s)$ the continuous monitoring station spatial intensity, Z(s) the spatial covariates and $v(s)$ the shared spatial process, with a tuning parameter; also let Y(s) denote the pollutant concentration measured by the monitoring station at location $s \in D$, and $\eta(s)$ a specific spatial random process. We assumed an inhomogeneous Poisson point process for monitoring station locations and that the monitoring stations could

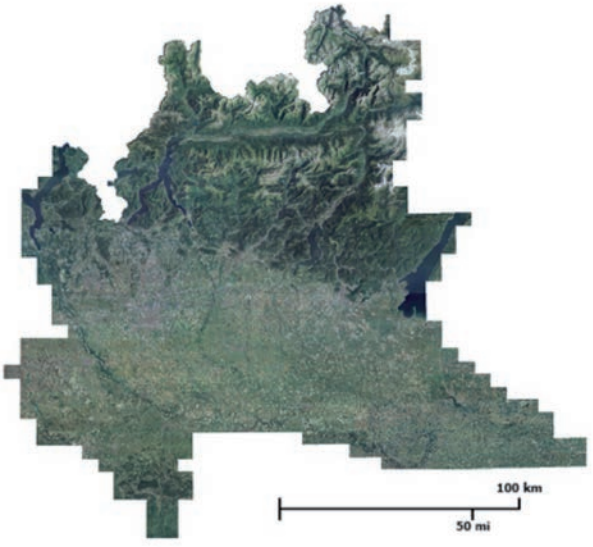

Figure 1. Orthophoto of Lombardy region (Italy) in scale 1:2,000,000 (http://www.geoportale.regione.lombardia.it/)

be virtually located at any point of the region of interest. If geographical barriers or other restrictions were present then a different modelling should be considered. The monitoring spatial intensity was modelled as log-linear function of the covariates and spatial random components.

The pollutant concentrations can be modelled as a Gaussian process with an exponential covariance function. More generally we can specify a measurement process for Y_i (with τ^2 the measurement error variance) and the underlying mean concentration surface as a function of the covariates – the same included in the model for the monitor spatial intensity – a specific spatial random component and a shared spatial component (Diggle *et al.*, 2010). The importance of the shared component in the two joint processes is controlled by the parameter δ. We then arrived at the equation:

$$\begin{aligned} &X_s \propto \text{Poisson PP}\left(\lambda(s)\right) \\ &\log\left(\lambda(s)\right) = \xi(s) = \delta^{-1} v(s) + \beta' Z(s) \\ &Y(s) \propto \mathrm{N}\left[\mu(s), \tau^2\right] \\ &\mu(s) = \eta(s) + \delta\, v(s) + \beta'' Z(s) \end{aligned} \qquad \text{(eq. 1)}$$

Simple algebra gives:

$$\begin{aligned} \mu(s) &= \eta(s) + \gamma\, z'(s) \\ z'(s) &= \delta^{-1} v(s) \\ \gamma &= \delta^2 \end{aligned} \qquad \text{(eq. 2)}$$

The model can be interpreted as a geostatistical model with an unmeasured covariate for which a surrogate (the monitor location intensity) is available. All known covariates Z(s), which may contribute to explain the two joint spatial processes, must be included in both equations. As Pati *et al.* (2011) said: *accounting for informative sampling is only necessary when there is an association between the spatial surface of interest and the sampling density that cannot be explained by the [common] spatial covariates*.

Computational details

In the motivating example described above, we used the output of the Eulerian photochemical model evaluated at the centroids of a 4×4 km grid and matched the 58 PM_{10} monitors locations to the respective 4×4 km grid cell centroids. Then, we specified a Bayesian geostatistical model accounting for preferential sampling and obtained the joint posterior predictive distribution for the annual average concentration of PM_{10} for each cell.

The spatial point process for the monitors location was fitted using the monitor counts on the 4×4 km grid over the region of interest. The same grid was used to predict the continuous pollutant concentration surface by the Bayesian geostatistical model.

Let s_i be the centroid coordinates of the i-th cell over the grid and X_i be the number of monitors in the i-th cell (i=1,…,1679). We assumed that:

$$\begin{aligned} &X_i \propto Poisson\left(\lambda(s_i)\right) \\ &\log\left(\lambda(s_i)\right) = \alpha + \delta^{-1}(u_i + v_i) + \beta' Z(s_i) \\ &\alpha \propto flat(0) \\ &u_i \propto Normal(0, \tau_u) \\ &v_i \propto CAR(\bar{v}_{j\in i}, \tau_v) \end{aligned} \qquad \text{(eq. 3)}$$

where the prior for α is improperly uniform, μ_i and μ_i are the heterogeneity and clustering random terms of the Besag, York and Mollié model (1991). The CAR prior is the normal improper, conditional, autoregressive distribution with a 0,1 adjacency matrix. The two random terms are not identifiable and we considered their sum as the latent shared spatial process. The covariates $Z(s_i)$ are the numerical outputs from Eulerian photochemical model. Note that the monitor counts and the heterogeneity and clustering terms are defined at the grid cell – area level –, while the intensity and the covariate are continuous and identified at the cell centroid coordinates. Let $Y(s_{i(k)})$ be the pollutant concentration observed by the k^{th} monitoring station at location $s_{i(k)}$ k=1,…,58, belonging to i^{th} cell, for some i∈{1,...,1679}. We assumed $Y(s)$ to follow a multivariate Gaussian distribution with a mean vector that depends on covariates $Z(s)$ and a covariance matrix induced by the exponential covariance function. The spatial covariance parameters were chosen in such a way that the rate of correlation by distance produced zero correlation at the maximum distance of 250 km and one at the minimum distance of 3 km (Banerjee *et al.*, 2006). In detail:

$$\begin{aligned} &Y(s_{i(k)})) \propto SpatialExp\left[\mu'(s_{i(k)}); \eta(s_{i(k)}) = \left(f(\varphi,\tau); \sigma\right)\right] \\ &\mu'(s_{i(k)}) = \alpha' + \beta'' Z(s_{i(k)}) + \delta\,(u(s_{i(k)}) + v(s_{i(k)})) \\ &(\varphi,\tau) \propto \text{informative priors} \\ &\alpha' \propto flat(0) \\ &\beta', \beta'' \propto Normal(0, \tau_\beta) \\ &\log(\delta) \propto TN(0, \tau_\delta) \end{aligned} \qquad \text{(eq. 4)}$$

where the term δ is log-normal distributed and allows the shared component to vary by a constant factor. The prior for δ is symmetric around zero on the log-scale; any value is equally likely as the reciprocal values

Figure 2. Spatial distribution of the fifty-eight PM_{10} monitoring stations in Lombardy Region, 2007.

a priori. Having the covariate $Z(s_i)$, the numerical outputs from Eulerian photochemical model and the preferential sampling adjustment ($\mu(s_i)$+ $\upsilon(s_i)$) for all grid cells, predictions and prediction standard deviations were obtained on the set of locations s_i (i=1,…,1679), *i.e*. the centroids of the 4x4 km grid, by the standard Bayesian geostatistical formulation:

$$\left[\tilde{y}\middle|\, y;Z,\upsilon,\nu;\tilde{Z},\tilde{\upsilon},\tilde{\nu}\right]=\int\left[\tilde{y}\middle|\, y,\Omega;Z,\upsilon,\nu;\tilde{Z},\tilde{\upsilon},\tilde{\nu}\right]\left[\Omega\middle|\, y;Z,\upsilon,\nu\right]d\Omega \quad \text{(eq. 5)}$$

where Ω is the set of parameters in the mean and covariance function (α'β"δ;φτσ). The two models (the point process for monitor locations and the geostatistical model for the concentration surface) were run jointly to assure uncertainty propagation in μ, ν and δ. The model was fitted using an MCMC algorithm in WinBUGS (Lunn *et al.*, 2000). We ran two independent chains and checks for achieved convergence of the algorithm following Gelman and Rubin (1992). We decided to run 50,000 iterations and to store the last 20,000 iterations for estimation (Gelman and Rubin, 1992).

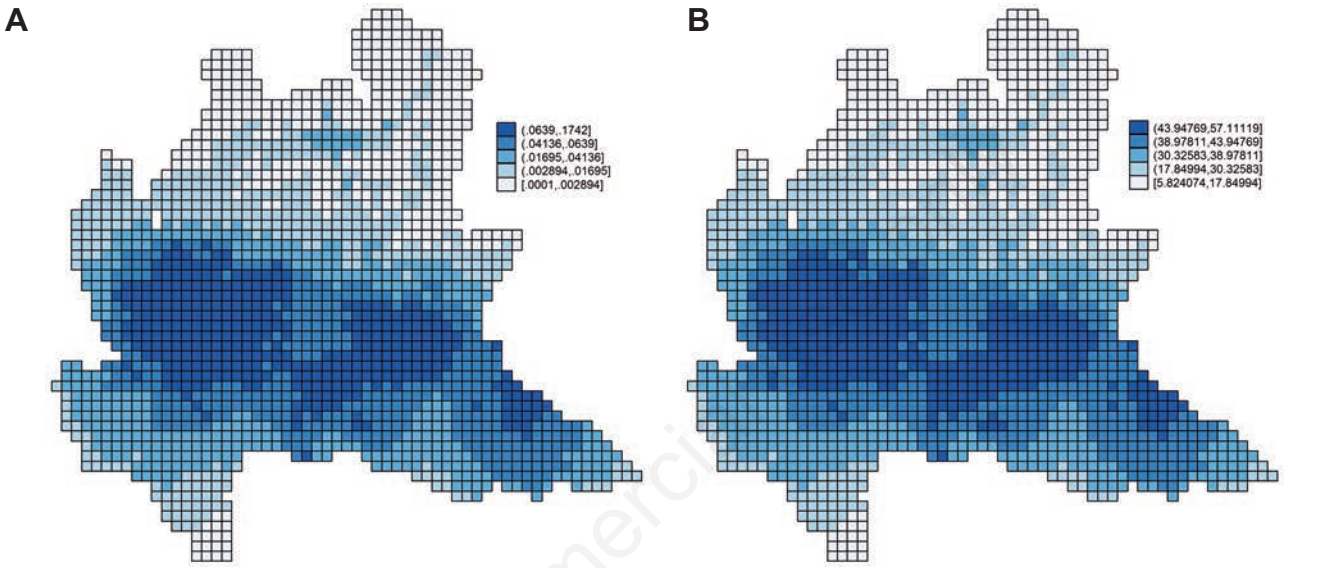

Figure 3. PM_{10} monitor spatial intensity (A) and predicted preferential sampling adjusted PM_{10} concentration by the shared component geostatistical model (B) in Lombardy Region, 2007.

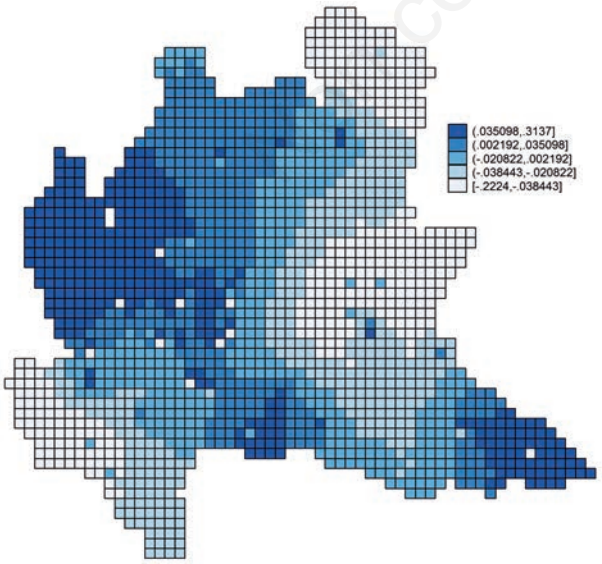

Figure 4. Shared spatially structured component of the PM_{10} preferential sampling adjusted geostatistical model in Lombardy Region, 2007.

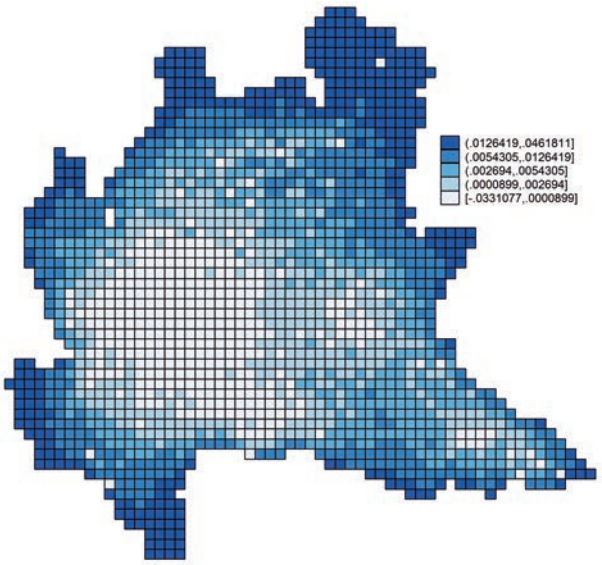

Figure 5. Differences in predictions standard deviations when accounting and not accounting for preferential sampling in the geostatistical model. Positive differences mean that uncertainty is greater when we account for preferential sampling. Data refer to Lombardy Region, 2007.

Results

Monitors locations are shown in Figure 2. The distribution is clearly not homogeneous or regularly spaced over the region. The monitor spatial intensity and predicted PM_{10} concentrations are shown in Figure 3. The two spatial distributions are very similar. Indeed, monitor locations were determined by the Regional Environmental Protection Agency on the basis of the emission inventory and other heterogeneous political considerations, with a preference to monitor highly polluted areas.

The prediction surface was accurately estimated by the Eulerian photochemical model: we did not expect any modification adding the data from the 58 PM_{10} monitors. The Pearson correlation coefficient between Kriging and deterministic model predictions was 0.999 and the Lin correlation coefficient was 0.983. Different consideration would apply with respect to other pollutants, like ozone. Particulate matter is more dependent of local emissions, even in the Lombardy Region context. The shared spatially structured component, which corresponds to the residual spatial variability not explained by the covariates, is shown in Figure 4. It is important to note that preferential sampling is relevant when there is an association of the residual response with the shared spatially structured (residual) component.

The differences between standard deviations accounting *vs* not accounting for preferential sampling are shown in Figure 5. There is an estimated greater uncertainty in the areas not properly covered by the air quality network when accounting for preferential sampling.

Discussion

Deterministic models consider emission sources, photochemical reactions in the atmosphere, meteorology and land use information, resulting in high accurate predictions. Air quality networks are based on too few monitoring stations to produce accurate predictions by geostatistical interpolation. However, monitor networks may provide information on variability of the pollutant concentration measurements, which we used to estimate uncertainty for the predicted concentration surface. Health impact assessment integrates several sources of information, which typically includes baseline occurrence rates, pollutant effect estimates and pollutant concentrations prediction. For all of them, appropriate estimates of uncertainty are needed, unless the calculation is conditional to some observed quantity. In the literature, pollutant spatial predictions and related uncertainty taking advantage of deterministic model outputs are obtained by geostatistical modelling – *e.g.* when misalignment is present by a down-scaler (Berrocal *et al.*, 2010). To the best of our knowledge, preferential sampling has never been addressed in this context. We propose a shared model to account for preferential sampling and discuss the results in term of predicted standard deviation. Our approach combines a Poisson model on spatial data with a Gaussian process on georeferenced data and simplifies calculation using the same fine grid. Diggle *et al.* (2013) discuss the extension of geostatistics to log-Gaussian Cox processes (LGCP). Instead of our modelling choice based on monitors counts on a fine regular grid – through a hierarchical Poisson-Gaussian Markov random field model (Besag *et al.*, 1991) – a LGCP model on the locations can be adopted. This approach leads to spatially smooth maps, the interpretation of which is independent of the particular partition of the region of interest into sub-regions. Illian *et al.* (2012) and Martins *et al.* (2013) discuss the prior choice for log-Gaussian Cox processes and computational details within the integrated nested Laplace approximation framework.

Conclusions

Comparison between predicted surfaces under different preferential sampling processes has been discussed by Gelfand *et al.* (2012). In our case, the interest was in comparing standard deviation surfaces, and we simply report the differences between standard deviations with and without accounting for preferential sampling. We estimated greater differences in the areas not properly covered by the air quality network. Inferences on uncertainty may be misleading when geostatistical modelling does not take preferential sampling into account.

References

Baccini M, Biggeri A, Grillo P, Consonni D, Bertazzi PA, 2011. Health impact assessment of fine particle pollution at the regional level. Am J Epidemiol 67:480-3.

Baccini M, Grisotto L, Catelan D, Consonni D, Bertazzi PA, Biggeri A, 2015. Commuting-adjusted short-term health impact assessment of airborne fine particles with uncertainty quantification via Monte Carlo simulation. Environ Health Persp 123:27-33.

Banerjee S, Carlin BP, Gelfand AE, 2004. Hierarchical modeling and analysis for spatial data. CRC Press, Boca Raton, FL, USA.

Banerjee S, Carlin BP, Gelfand AE, 2006. Hierarchical modeling and analysis for spatial data. Chapman and Hall, Boca Raton, FL, USA.

Berrocal VJ, Gelfand AE, Holland DM, 2010. A bivariate space-time downscaler under space and time misalignment. Ann Appl Stat 4:1942-75.

Besag J, York J, Mollié A, 1991. Bayesian image restoration, with two applications in spatial statistics. Ann I Stat Math 43:1-59.

Cecconi L, Biggeri A, Grisotto L, Berrocal VJ, Rinaldi L, Musella V, Cringoli G, Catelan D, 2016. Preferential sampling in veterinary parasitological surveillance. Geospat Health 11:412.

Cressie NAC, 1991. Statistics for spatial data. Wiley, New York, NY, USA.

Diggle PJ, Menezes R, Su T, 2010. Geostatistical inference under preferential sampling. J Roy Stat Soc C-App 59:191-232.

Diggle PJ, Moraga P, Rowlingson B, Taylor BM, 2013. Spatial and spatio-temporal Log-Gaussian Cox processes: extending the geostatistical paradigm. Stat Sci 28:542-63.

Diggle PJ, Tawn JA, Moyeed RA, 1998. Model-based geostatistics (with discussion). Appl Stat 47:299-350.

Gelfand AE, Sahu SK, Holland DM, 2012. On the effect of preferential sampling in spatial prediction. Environmetrics 23:565-78.

Gelman A, Rubin DB, 1992. Inference from iterative simulation using multiple sequences. Stat Sci 7:457-72.

Guttorp P, Sampson P, 2010. Discussion of geostatistical inference under preferential sampling by Diggle PJ, Menezes R and Su T. J Roy Stat Soc C-App 59:191-232.

Held L, Natario I, Fenton SE, Rue H, Becker N, 2005. Towards joint disease mapping. Stat Methods Med Res 14:61-82.

Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, Briggs D, 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos Environ 42:7561-78.

Illian JB, Sørbye SH, Rue H, Hendrichsen DK, 2012. Using INLA to fit a complex point process model with temporally varying effects. A case study. J Environ Stat 3:1-29.

Lee A, Szpiro A, Kim SY, Sheppard L, 2015. Impact of preferential sampling on exposure prediction and health effect inference in the

context of air pollution epidemiology. Environmetrics 26:255-67.
Lunn DJ, Thomas A, Best N, Spiegelhalter D, 2000. WinBUGS. A Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10:325-37.
Martins TG, Simpson D, Lindgren F, Rue H, 2013. Bayesian computing with INLA: new features. Comput Stat Data An 67:68-83.
Pati D, Reich BJ, Dunson DB, 2011. Bayesian geostatistical modelling with informative sampling locations. Biometrika 98:35-48.
Pilz J, Spöck G, 2008. Why do we need and how should we implement Bayesian kriging methods. Stoch Env Res Risk A 22:621-32.
Shaddick G, Zidek JV, 2015. Unbiasing estimates from preferentially sampled spatial data. Spatial Stat 9:43.
Silibello C, Calori G, Brusasca G, Giudici A, Angelino E, Fossati G, Peroni E, Buganza E, 2008. Modelling of PM_{10} concentrations over Milano urban area using two aerosol modules. Environ Model Softw 23:333-43.
Son JY, Bell ML, Lee JT, 2010. Individual exposure to air pollution and lung function in Korea: spatial analysis using multiple exposure approaches. Environ Res 110:739-49.
van Donkelaar A, Martin RV, Brauer M, Kahn R, Levy R, Verduzco C, Villeneuve PJ, 2010. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environ Health Persp 118:847-55.
Vicedo-Cabrera AM, Biggeri A, Grisotto L, Barbone F, Catelan D, 2013. A Bayesian kriging model for estimating residential exposure to air pollution of children living in a high-risk area in Italy. Geospat Health 8:87-95.
Zanini G, 2009. Il sistema MINNI, modello integrato nazionale per la valutazione degli effetti dell'inquinamento atmosferico e dell'efficacia delle politiche di riduzione delle emissioni di inquinanti atmosferici. Epidemiol Prev 33:35-42.