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).

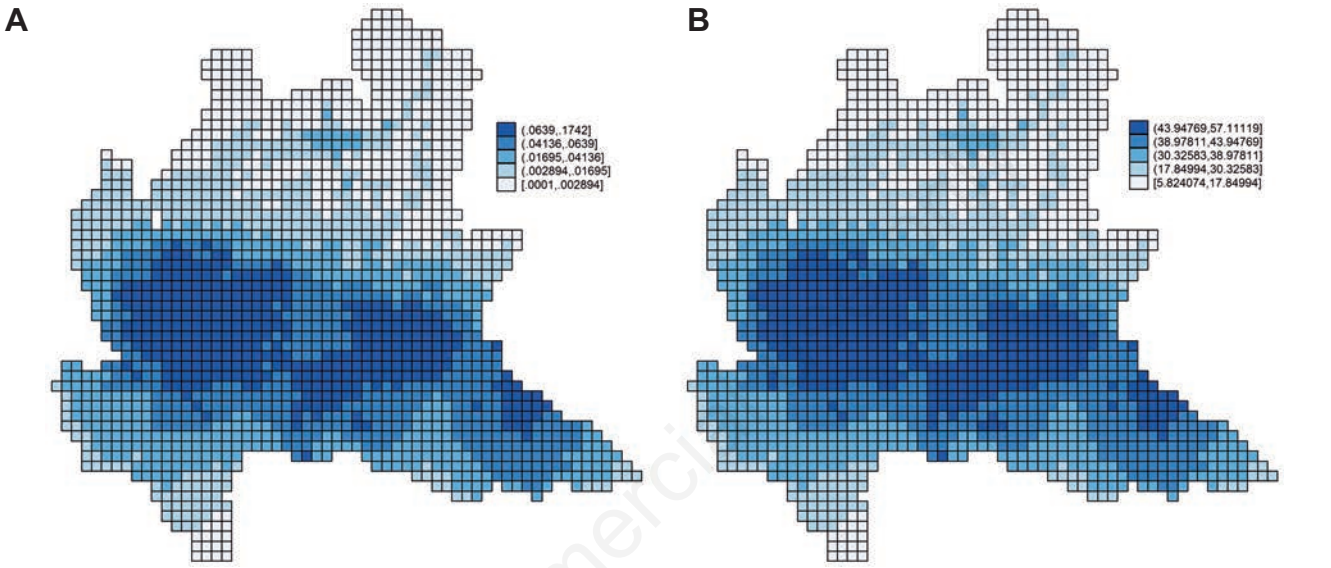

Figure 3. PM_{10} monitor spatial intensity (A) and predicted preferential sampling adjusted PM_{10} concentration by the shared component geostatistical model (B) in Lombardy Region, 2007.

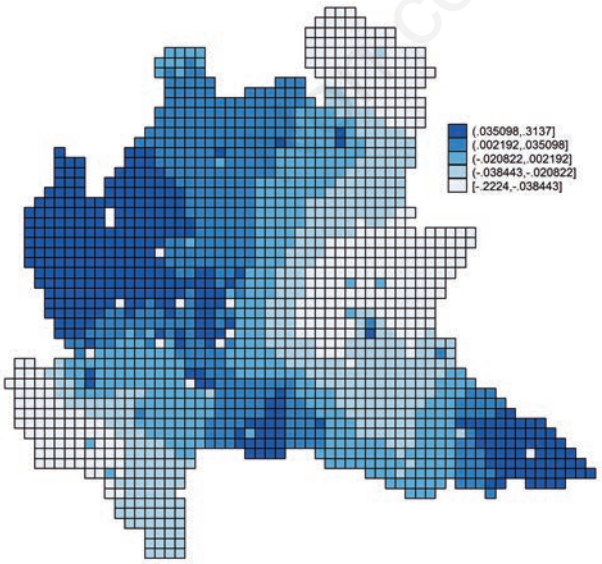

Figure 4. Shared spatially structured component of the PM_{10} preferential sampling adjusted geostatistical model in Lombardy Region, 2007.

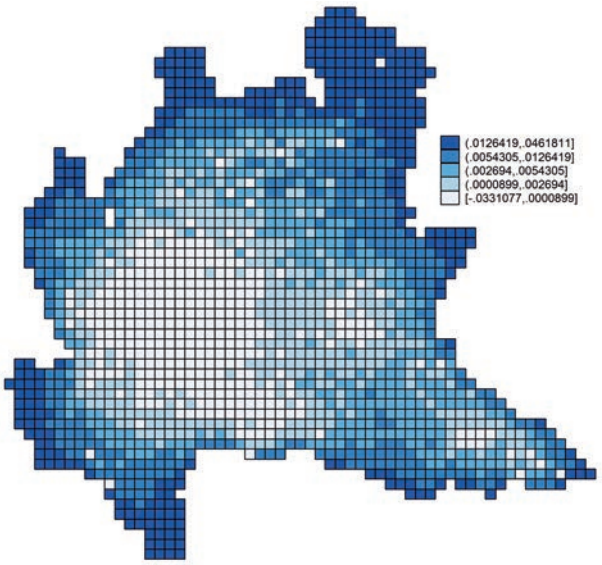

Figure 5. Differences in predictions standard deviations when accounting and not accounting for preferential sampling in the geostatistical model. Positive differences mean that uncertainty is greater when we account for preferential sampling. Data refer to Lombardy Region, 2007.

non-commercial use only

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

Non-commercial use only

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.

Non commercial use only