Preferential sampling in veterinary parasitological surveillance

Lorenzo Cecconi,[1] Annibale Biggeri,[1,2] Laura Grisotto,[1] Veronica Berrocal,[3] Laura Rinaldi,[4] Vincenzo Musella,[5] Giuseppe Cringoli,[4] Dolores Catelan[1,2]

[1]Department of Statistics, Computer Science, Applications, University of Florence, Florence; [2]Biostatistics Unit, Institute for Cancer Prevention and Research, Florence, Italy; [3]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA; [4]Department of Veterinary Medicine and Animal Productions, University of Naples Federico II, Naples; [5]Department of Health Sciences, University of Catanzaro Magna Graecia, Catanzaro, Italy

Abstract

In parasitological surveillance of livestock, prevalence surveys are conducted on a sample of farms using several sampling designs. For example, opportunistic surveys or informative sampling designs are very common. Preferential sampling refers to any situation in which the spatial process and the sampling locations are not independent. Most examples of preferential sampling in the spatial statistics literature are in environmental statistics with focus on pollutant monitors, and it has been shown that, if preferential sampling is present and is not accounted for in the statistical modelling and data analysis, statistical inference can be misleading. In this paper, working in the context of veterinary parasitology, we propose and use geostatistical models to predict the continuous and spatially-varying risk of a parasite infection. Specifically, breaking with the common practice in veterinary parasitological surveillance to ignore preferential sampling even though informative or opportunistic samples are very common, we specify a two-stage hierarchical Bayesian model that adjusts for preferential sampling and we apply it to data on *Fasciola hepatica* infection in sheep farms in Campania region (Southern Italy) in the years 2013-2014.

Correspondence: Dolores Catelan, Department of Statistics, Computer Science, Applications G. Parenti, University of Florence, viale Giovan Battista Morgagni 59, 50134 Florence, Italy.
Tel: +39.055.275.1500.
E-mail: catelan@disia.unifi.it

Key words: Preferential sampling; Veterinary parasitological surveillance; Livestock.

Acknowledgements: the research leading to these results has received funding from the European Union Seventh Framework Programme FP7-KBBE-2011-5 under grant agreement no 288975 and University of Florence, ex 60%.

Note: see online Appendix for a description of latent factors and covariates.

Received for publication: 22 September 2015.
Revision received: 8 February 206.
Accepted for publication: 19 February 2016.

Geospatial Health 2016; 11:412
doi:10.4081/gh.2016.412

Introduction

In parasitological surveillance of livestock, prevalence surveys are conducted on a sample of farms using several sampling designs. While systematic sampling or spatial sampling designs are optimal, often their requirements in terms of confidentiality or list coverage are not fulfilled. On the other hand, informative sampling strategies, which use information from previous surveys conducted on a regular grid, can offer advantages for specific epidemiologic aims (Musella *et al.*, 2014). In analysing the data arising from such surveys, we must take into account the preferential sampling nature of the data (Diggle *et al.*, 2010). Geostatistics refers to the collection of statistical methods used to model and analyse point-referenced spatial data (Cressie, 1991). Typically, this type of data is obtained by sampling a spatially continuous phenomenon at a discrete set of locations in a region of interest. If the sampling locations are considered fixed by design or if they can be considered to be stochastically independent of the spatial process sampled, traditional geostatistical methods are appropriate. However, these methods are not appropriate if either condition is not satisfied. Preferential sampling refers to any situation in which the spatial process and the sampling locations are not independent. Examples of non-preferential designs include completely random samples, and regular lattice designs, while examples of preferential sampling are opportunistic samples in which the sampling locations (location of pollutant monitors, location of sampled animals or trees, residence of asthma patients, to name a few) are concentrated in sub-regions where the underlying values of the spatial process are larger or smaller than average.

Even though it can arise in several situations, preferential sampling has been ignored for many years in the analysis of spatial data. Diggle *et al.* (2010) showed that if one ignores the fact that the data are preferentially sampled, statistical inference could be misleading. To address this issue, they proposed a model, which postulated a shared spatial latent process driving both the sampling process and the underlying spatial process. For parameter estimation, the authors proposed maximum likelihood estimation using a Monte Carlo approach. Diggle *et al.* (2013) presented a more general way to handle preferential sampling using an inhomogeneous spatial point process and suggested the use of log Gaussian Cox processes to model the spatial point process of the sampling locations.

In our study, we consider a finite number of sampling points - those corresponding to the farm locations in the region of interest – with only a subgroup of this having been sampled. When the inclusion probability – *e.g.* the probability to be sampled – varies among sampling locations, a correction, following the Horvitz-Thompson approach, can be used in the geostatistical model. Specifically, we propose to use the inverse inclusion probabilities as observation weights. If not known by design, a statistical model (possibly spatially structured) can be specified to derive these probabilities. Shaddick and Zidek (2014) used a similar approach in the context of environmental epidemiology.

Differently from these authors, we specify two hierarchical Bayesian models: one for the georeferenced data and one for the selection probabilities. Since we use a *two-step model* a matter of propagation of uncertainty arises. We address this by repeatedly sampling from the posterior distributions of sampling intensities.

In the statistical literature, most examples of preferential sampling are in environmental statistics, and are related to the placement of pollutant monitors (Diggle *et al.*, 2010, 2013; Shaddick and Zidek, 2014). Few examples (Zouré *et al.*, 2014; Rinaldi *et al.*, 2015a) are present in the context of veterinary surveillance, where informative or opportunistic sampling is very common. In this paper, we tackle the problem of preferential sampling in parasitological veterinary surveillance and we specify a two-stage hierarchical Bayesian model to adjust for the preferential sampling nature of the data. We apply this model to data on parasitic infections in sheep farms in the Campania region, located in southern Italy.

The paper is organised as follows: in *Materials and Methods* we introduce the data and the information on the environmental covariates that we use to better explain the spatial pattern of the infection probability. In the same section we also introduce the Bayesian models for the inclusion probabilities and the geostatistical process. We compare the various models and we evaluate the underestimation of uncertainty due to the two-step approach using the Kullback-Leibler measure (McCulloch, 1989) and the variance inflation factor (VIF), respectively. Finally, we present the *Results*, followed by a *Discussion* and *Conclusions*.

Materials and Methods

Data collection

The data arise from two surveys, both instances of preferential sampling: a survey based on an informative sampling design, and an opportunistic survey.

In the first survey – the informative sampling design – we use preliminary data from a survey originally planned to visit 150 farms and sample farms according to the posterior predicted probabilities of infection. Specifically, using the Bayesian geostatistical model proposed by Musella *et al.* (2014) to predict the probability of several parasitic infections for all sheep farms in the Campania region (Italy), we derive an informative sampling design based on the posterior predictive distribution of infection.

The second survey is an opportunistic survey of routine diagnosis performed at the regional center for monitoring of parasitic infections (CREMOPAR, Campania region; Cringoli *et al.*, 2015). This survey consists of local veterinarians or farmers who spontaneously bring samples for parasitological examinations of livestock.

Altogether, the two surveys provide information and data from a total of 89 farms: 50 farms coming from the first aforementioned on-going survey with data collected in 2013, and 39 farms coming from the second survey with data collected in 2014 (Rinaldi *et al.*, 2015a).

Study outcome

Our interest is on *Fasciola hepatica,* partly because of the impact of liver fluke upon animal health, welfare and productivity (Rinaldi *et al.*, 2015b) and partly because of the recent outbreaks of fasciolosis in sheep farms in the Campania region, considered to be a possible consequence of climate change (Bosco *et al.*, 2015).

Pooled faecal samples (Rinaldi *et al.*, 2015a) were collected for biological examinations and faecal egg counts (FEC) were determined using the FLOTAC dual technique (Cringoli *et al.*, 2010; Rinaldi *et al.*, 2015a), which has an analytic sensitivity of 6 eggs per gram of faeces (EPG). To detect and count the number *F. hepatica* eggs, we used a zinc sulphate-based flotation solution (ZnSO4 specific gravity=1.350).

Covariates

Using GIS, we derived 19 bioclimatic layers and 30 Moderate-resolution Imaging Spectroradiometer (MODIS) variables that can be used to explain the spatial pattern of the infection (Rinaldi *et al.*, 2015a). To assign a covariate value to each farm, we overimposed a 10x10 km grid on the Campania region, and we identified buffer zones of 3 km around each sampled farm using the geographic information system (GIS). In the Appendix we report the list of covariates for both types of variables.

Due to the high number of covariates (P=49) compared to the slightly larger number of observations (n=89), we reduce the covariates' dimensionality by performing a factor analysis within the Bayesian geostatistical approach adopted. In particular, we summarise the covariate information with 3 latent factors that are introduced as explanatory variables in the linear predictor term of the Bayesian geostatistical model (see below). We fixed the number of factors *a priori* and set it equal to three and we specified which variable belongs to each factor (see Appendix for the definition of the three latent factors). The first factor summarised those covariates that refer to *position indexes*: mean, maximum and minimum (*e.g* annual mean temperature). The second factor included those covariates that refer to variability: range and coefficient of variation (*e.g.* mean diurnal range), while the third factor includes covariates that represent seasonality: amplitude or phase of cycle (*e.g.* amplitude of annual cycle-middle infra-red). We call these factors *position*, *variability* and *periodicity* respectively.

Statistical modelling

Bayesian modelling of sampling probabilities

Using GIS we were able to georeference all the farms in the Campania region. We overimposed a 10x10 km grid on the region for a total of 184 cells. We obtained the number of sampled farms over the total number of farms contained in each cell (observed sampling fractions). We then specify a Bayesian hierarchical model to the observed sampling fractions with a log linear predictor expressed as a function of both spatially structured and unstructured random terms (Besag *et al.*, 1991).

More specifically, let n_j be the number of farms population belonging to the j-th cell (j=1,…,184) of the grid and k_j be the number of sampled farm in the j-th cell. We assume that:

$$K_j \propto \text{Binomial}(p_j, n_j)$$

$$z_j = \text{logit}\ (p_j) = \kappa + \upsilon_j + \upsilon_j$$

where p_j is the sampling probability for a farm in the j-th cell, κ is the intercept term in the model for the spatially-varying log-odds, υ_j is a spatially unstructured random term provided with as $N(0,0.001)$ prior and υ_j is a spatially structured random term provided with an improper conditional autoregressive (ICAR) prior (Besag *et al.*, 1991). That is, conditionally on $\upsilon_{l\sim j}$ terms, where $_{\sim j}$ indicates cells adjacent to *j*-th one, (*j*=1,…,184), we assume that υ_j is distributed as, $N(\bar{u}_j, \tau_u n_i)$ where $\bar{u}_j = \sum_{l\sim j} \frac{u_l}{n_j}$ and τ_u is the precision. *A priori* we assume that τ_u is distributed as *Gamma(0.5,0.0005)* while for the intercept term κ, we specify a flat prior. Using the model above, we are able to estimate, for each of the 89 georeferenced sampled farms, the posterior probability $\tilde{p}_j$ that a farm in the j-th cell is sampled. We then use the posterior estimates $\tilde{p}_j$ of p_j as covariate in the geostatistical model for the risk of infection.

Second step: geostatistical modelling of infection risk

Let Y_i be a binary random variable (1/0) that indicates the presence/absence of a parasitic infection in the *i*-th farm (*i*=1,…,89). We model Y_i as a Bernoulli random variable with parameter π_i denoting the probability of infection. More clearly:

$$Y_j \propto \text{Binomial}(\pi_i)$$

$$Z_i = \text{logit}(\pi_i) = \gamma + s_i + \sum_{l=1}^{3} \beta_l \lambda_{il} + \beta_4 \tilde{p}_i$$

where γ is the intercept term, $\tilde{p}_i$ is the posterior sampling probability of each farm derived in the previous step, λ_{il} are the three latent factors described in *Materials and Methods* related to the GIS covariates. Finally, s_i is the component of a multivariate Gaussian vector with mean zero and variance-covariance matrix Σ at the location of the i-th farm. In turn, the matrix Σ is constructed so that its (i,k)-th element is equal to $\sigma^2 \exp(-\phi\, d_{i,k})$ where σ^2 represents the marginal variance of the multivariate Gaussian vector, $d_{i,k}$ denotes the Euclidean distance between farms *i* and *k*, and ϕ is the parameter that controls the decay of the correlation among farms at distance *d*. The decay parameter ϕ was chosen such that the correlation between two sampled farms is equal to 0.97 at the minimum inter-farm distance (*d*=1.42 km) and equal to 0.017 at the maximum inter-farm distance (*d*=202.23 km) (Banerjee *et al.*, 2014). We complete the specification of the model by providing priors to the model parameters: for the intercept term γ we specify a flat prior, on σ^2 we place a *Gamma(0.05,0.005)* prior distribution, while on the β's coefficients we place weakly informative normal prior distributions *N(0.01,0.001)*. For the factor analysis, we followed the specification of Congdon (2003). Let the matrix X denote the *n* by *p* covariates matrix, we assume:

$$X_{i,k} \propto N\left(\eta_{i,k(l)}, \tau^2_{i,k(l)}\right)$$

where *k*=1,…,*p* indexes the covariates in the *i*-th farm (*i*=1,…,89), and *l*=1,2,3 represents the latent factor. In turn, we assume that the expected value $\eta_{i,k(l)}$ for the *l*-th latent factor is defined as:

$$\eta_{i,k(l)} = \vartheta_{k(l)}\, \lambda_{i,l} + \alpha_{k(l)}$$

where $\vartheta_{k(l)}$ indicates the loading coefficient of covariate k in the factor analysis, $\lambda_{i,l}$ is the common factor to be used as covariate in the full response model and $\alpha_{k(l)}$ is a residual error. The prior distribution of the common factor $\lambda_{i,l}$ is taken to be *N(0,1)*, while we specify non informative N(0,0.00001) prior distributions on $\alpha_{k(l)}$ and $\vartheta_{k(l)}$ (Musella *et al.*, 2011).

We fit the Bayesian factor analysis and the geostatistical model for the presence/absence of infection jointly and we use it to predict the probability of infection in the centroids of the 184 cells in the 10x10 km grid.

Model criticism

To evaluate the effect of not taking into account preferential sampling, we compare the results obtained from fitting a model that accounts for preferential sampling with the results obtained from a geostatistical model for the absence/presence of that does not introduce the sampling weights $\tilde{p}$.

For the comparison between the two models, we consider an influence measure that quantifies locally – at each grid cell – the effect of

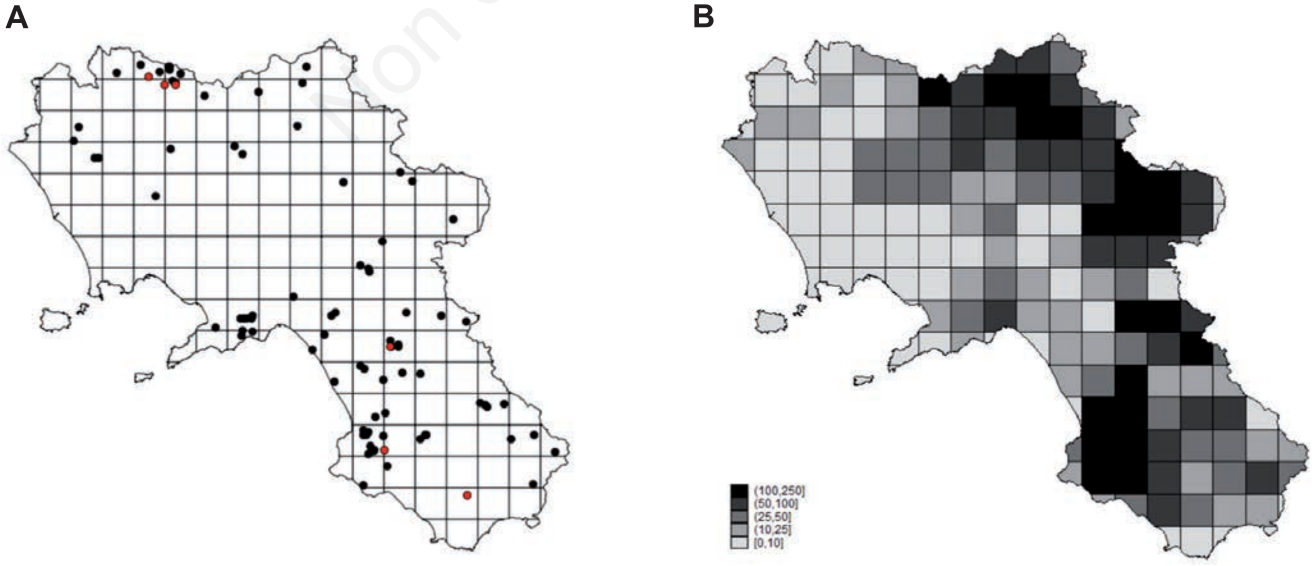

Figure 1. Spatial distribution of presence (black dots: negative farms; red dots: positive farms) of *Fasciola hepatica* observed infection on sampled farms (A); and density – *i.e.*, number of farms in each cell – of farm population (B) in Campania region, Southern Italy (2013-2014).

accounting for preferential sampling. Specifically, as influence measure, we use the calibrated Kullback-Leibler divergence (KL) (McCulloch, 1989), which focuses on how different our inferences would be under alternative models. This discrepancy measure can be expressed as the difference in the expected utilities between the unperturbed and perturbed posteriors, where the model for the actual belief is predefined. In more detail, let M_0 be the model accounting for the sampling intensities – our actual belief – and let M_1 be the alternative model without sampling weights. The loss in utility when preferential sampling is not considered, due to the approximation of model M_0 with model M_1, is expressed by the KL divergence, which we calculate separately for each j-th cell via the following equation:

$$KL_j = \int \log \frac{p(z_j|Y, M_0)}{p(z_j|Y, M_1)} p(z_j|Y, M_0)\, dz_j$$

where $p(z_j|Y,M_0)$ is the posterior marginal distribution of the logit of the probability of infection in the j-th cell given the data and the model. If the posterior distributions of z_j,j=1,…,184, under M_0 and M_1 can be approximated by a Gaussian distributions, then the KL divergence between the distributions resulting from model M_0 and M_1 can be approximated by the KL divergence between two Gaussian distribution with mean m_0 and m_1, respectively, and variances s_0^2 and s_1^2 (the j suffix was omitted for simplicity), respectively. This implies that

$$KL(p(z|Y, M_0), p(z|Y, M_1)) \approx \frac{1}{2}\left[\frac{(m_1 - m_0)^2}{s_1^2} - 1 + \frac{s_0^2}{s_1^2} - \log\left(\frac{s_0^2}{s_1^2}\right)\right]$$

In order to better understand this measure, McCulloch (1989) suggested calibrating Kullback-Leibler divergences in terms of the distance between two Bernoulli distributions. Let c be the distance between two distributions, the calibrated Kullback-Leibler divergence is the value $p(c)$, parameter of a Bernoulli distribution B, such that $KL[B(0.5),B(p(c))]=c$. The value $p(c)$ can be calculated directly as

$p(c) = \left(1 + \sqrt{1 - e^{-2c}}\right)/2$ (McCulloch, 1989).

Uncertainty propagation

One limitation of the two-step model is that it does not allow propagating the uncertainty throughout every step of the model properly. In particular, since in the second step the posterior sampling probabilities estimates p_j are introduced in the model as numerical values, the uncertainty in their estimates is not taken into account in the second step. To quantify the effect of neglecting this source of uncertainty, we sample five random values $p_{1,j}, p_{2,j}, p_{3,j}, p_{4,j}, p_{5,j}$ from the posterior distributions of the sampling probabilities (p_j). For each vector of posterior sampling probabilities, we fit the geostatistical model and obtain the posterior infection probability distributions. In order to evaluate the underestimation of variance due to uncertainty in the sampling probabilities we consider the VIF, defined for each grid cell j, as

$$VIF_j = \frac{\left(1 + \frac{1}{r}\right)\omega_j^2 + \delta_j^2}{\delta_j^2}$$

where ω_j^2 is the square of the range of the r=5 sampled values $p_{1,j}, p_{2,j}, p_{3,j}, p_{4,j}, p_{5,j}$ of the posterior infection probabilities for the j-th cell while δ_j^2 is the variance of the posterior infection probability π_j in the reference model (Little and Rubin, 2014). High values of VIF indicate a substantial underestimation of uncertainty in the predicted probabilities of infection, due to the two-stage approach.

Computational details

Inference for all the models are carried out using Markov Chain Monte Carlo methods that we implement using the WinBugs software (Lunn *et al.*, 2000). For each model, we run two independent chains and we evaluate convergence of the algorithm following the suggestions of Gelman and Rubin (1992). We discard the first 30,000 iterations for burn-in and store and use the following 10,000 iterations for estimation.

Results

Figure 1 (A) shows the spatial distribution of the 89 sampled farms with different colors to indicate presence/absence of *F. hepatica* infection: specifically, blue points represent negative farms while red points are used for positive farms. Overall, the prevalence of the parasite is low [7.9%, 90% confidence interval (CI) 3.7; 14.3] consistently with the climatic characteristics of the region. Indeed, *F. hepatica* intermediate host is a water-snail and therefore wet climate and wetlands are associated with higher prevalence. The density of the farm population is presented in Figure 1 (B). As we have overimposed a 10x10 km grid on the Campania region for a total of 184 cells, Figure 1 shows the number of farms in each 10x10 km cell.

Figure 2 presents the posterior mean of the sampling inclusion probabilities $\tilde{p}_j$. The sampling probabilities range from 0.2 to 19.5% with areas of low sampling coverage being geographically clustered in the mountain areas of the eastern part of the Campania region as well as the wilder southern area of the region.

Given such variability in the sampling probabilities, we expect an

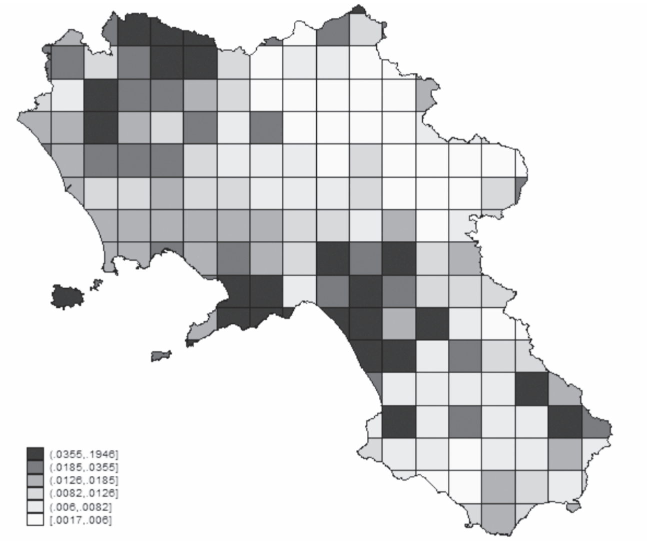

Figure 2. Spatial distribution of posterior estimates of sampling probabilities $\tilde{p}_j$ for Campania region, Southern Italy (2013-2014).

effect on the geostatistical model estimates. However, such effect might be mitigated by the fact that our geostatistical model for the risk infection also account for bioclimatic and remote sensing covariates which can affect the prevalence of *F. Hepatica*. As discussed above, we reduce the dimensionality of the covariates (a total of 30+19) via a factor analysis that we have embedded in the Bayesian geostatistical model. The three latent factors are spatially structured: in particular, the posterior means of the three latent factors highlight a strong spatial pattern with a South East-North West gradient for periodicity and an opposite gradient for position and variability (Figure 3). Seasonality variability is wider in the mountain area, while average temperature and other climatic indexes are higher along the coast. Additionally, the range of these variables is generally greater in the southern areas and along the coast.

F. hepatica infection is rare in the Campania region with a range of posterior predicted probabilities of 0.8-36.0%. The probability of infection at each farm is calculated as the mean of the posterior predictive distribution derived from the model with preferential sampling weights fixed at their posterior mean (*e.g.* Stage 1 model). The geographical distribution of the posterior probabilities of infection and their related standard errors are shown in Figure 4, which shows that the northern and mountain areas of the region appear to be free of risk of infection. While our predictions of infection are generally good as expected, we notice that while one isolated positive farm is positively predicted by our model, two other isolated positive farms in the far south are not captured by the predictions. We believe that this is an effect of the smoothness in the covariates. In fact, *F. hepatica* is a very rare infection in Italy because of Italy's climate characteristics (in contrast for example to Ireland; Rinaldi *et al.*, 2015a) and given the general smoothness of the covariates included in the geostatistical model for risk infection, we expect some smoothing in the predicted infection probabilities.

To study the impact of preferential sampling we evaluate the discrepancies between two alternative models (one that accounts for preferential sampling and a second that does not) by computing for each grid cell the calibrated Kullback-Leibler (KL) divergence p(c). Figure 5 presents the histogram of the KL divergences and their geographical distribution. Overall, there is a strong influence of the preferential sampling adjustment, with p(c) largely above 0.7. High calibrated KL divergence suggest strong difference between predictions derived from simple geostatistical model predictions and predictions from a preferentially adjusted geostatistical model. The areas that emerge most influenced are those with low sampling probabilities (Figure 2).

Finally, Figure 6 shows the spatial distribution of the Variance Inflation Factor across the Campania region. From the Figure we can note that VIF is consistently higher on the north-western areas of the region where the number of farms is smaller.

Discussion

The spread of geospatial tools and the need for large-scale surveillance has made increasingly popular the use of disease mapping methods in veterinary epidemiology. However, it is important to note that in this context, opportunistic samples are very common and the inclusion of a farm in a sample can depend on characteristics which are not under control to the researcher and that are related to the phenomenon under study (infection probability). Thus, consideration of potential bias due to preferential sampling must be addressed in the inferential procedure.

In some instances, inclusion probabilities can be known *a priori* by design. In such case it is simple to use them as sampling weights in the analysis, following a classical Horvitz-Thompson approach. However, informative sampling may be the result of qualitative judgment during the survey and thus no prior weights are defined. In this case, a statistical model to estimate them is needed.

In this paper, we consider also a different mechanism: the situation in which the researcher deliberately chooses informative sampling probabilities. We have discussed this issue in a previous paper as well (Catelan *et al.*, 2012).

In spatial statistics, mostly in applications in environmental sciences and exposure assessment, several approaches have been proposed to address the preferential sampling nature of the data, all of which introduce two processes: the spatial point process for the sampling locations and the spatial point-referenced process for the response under study. The existence and extent of the bias in the statistical inference depends on the degree of overlap between the two underlying spatial processes.

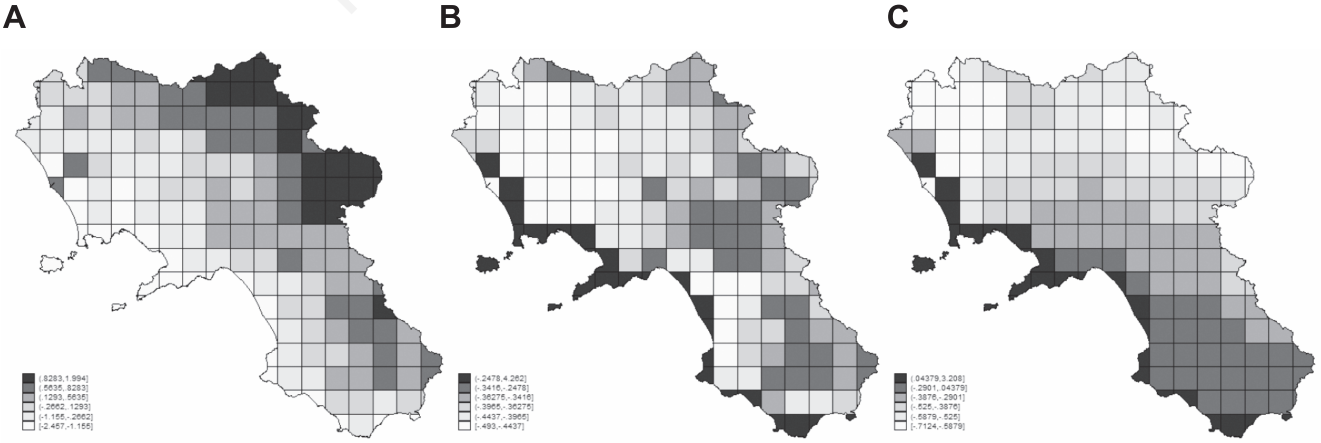

Figure 3. Spatial distribution of posterior means of the three latent factors: periodicity, position index and variability (respectively from A to C). Campania region, Southern Italy (2013-2014).

In this paper, considering as application the prevalence of infection to *Fasciola hepatica* in the Campania region, we propose a two-stage approach. First, we specify an inhomogeneous point process to model the spatial intensity of the sampled farms. The motivating idea behind this modelling choice is that a spatially structured random term (clustering) can capture unknown factors which are spatially structured and that are both related to the sampling probabilities and to the probability of infections. Covariates, if available and pertinent, can easily be introduced. Other specifications are reported in the literature. To model the spatially structured random term, we choose the Besag, York, Mollié (BYM) model because of its flexibility. Although the issue of identifiability of the BYM model components is well know and documented, we are not concerned with it here as this is not relevant in our context: in fact our focus is simply in using the posterior estimates of the sampling

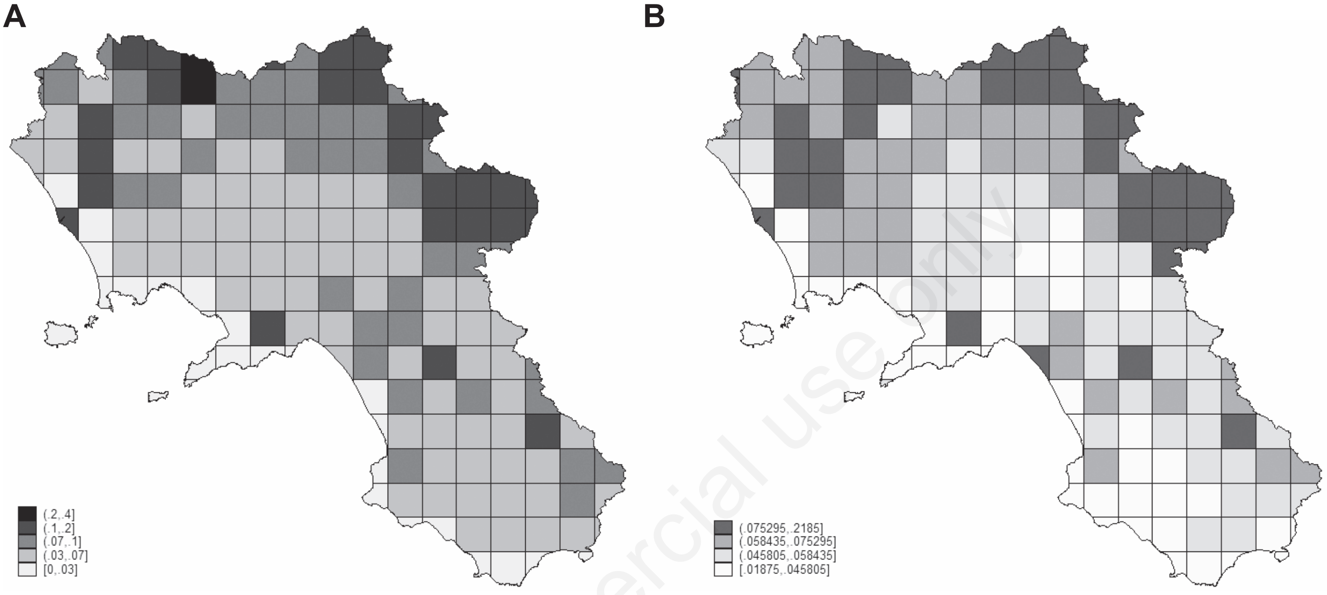

Figure 4. Spatial map of the posterior means (A) and standard deviations (B) of the probabilities of infection of *Fasciola hepatica*. Campania region, Southern Italy (2013-2014).

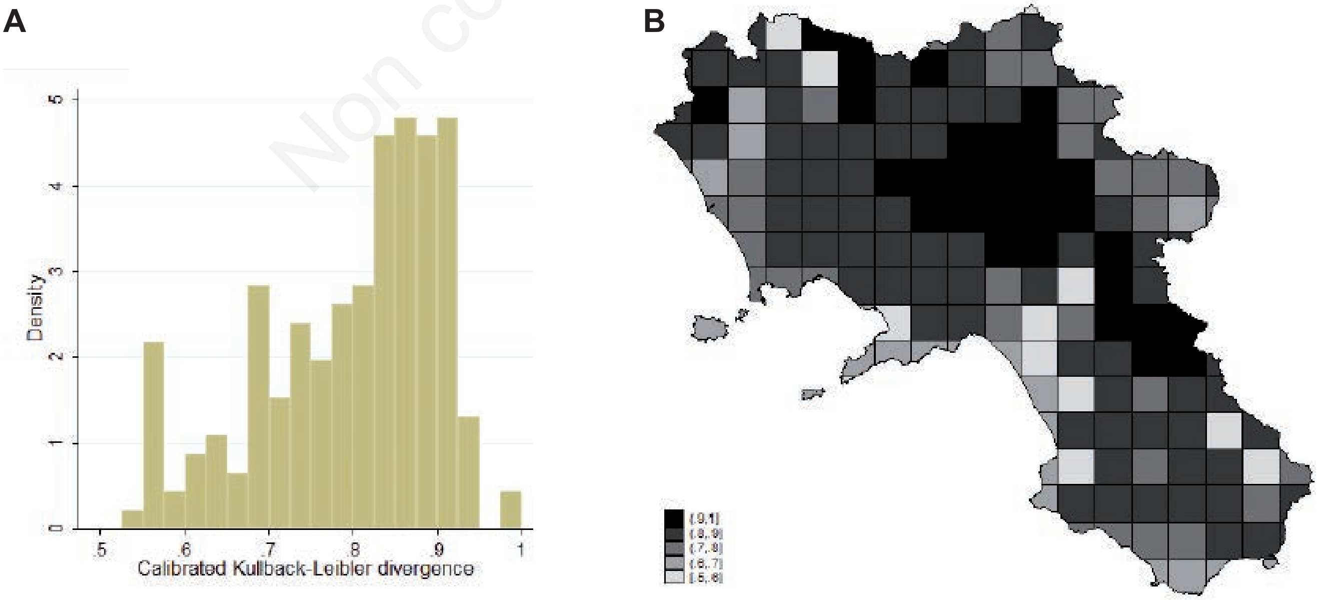

Figure 5. Calibrated Kullback-Leibler divergence between a model that explicitly accounts for preferential sampling and a model without such adjustment. A) histogram of the calibrated Kullback-Leibler divergences; B) spatial distribution of the calibrated Kullback-Leibler divergences. Campania region, Southern Italy (2013-2014).

probabilities and not the estimates of the single components.

From the first step, we estimate posterior sampling probabilities for each grid cell and we use them as weights in a hierarchical Bayesian geostatistical model for the infection probability. The adjustment for preferential sampling is carried out by introducing the estimates of the sampling probabilities as weights in the linear predictor of the geostatistical model. This is similar to other approaches proposed in the literature (Pati *et al.*, 2011; Diggle *et al.*, 2013). Robustness and scale appropriateness in the linear predictor are two aspects, which merit further investigation. Since the mean function is not the identity, the inclusion of the preferential sampling correction in the linear predictor is not straightforward. Several other functions could be specified, analogously to the freedom of choice one has in defining the offset in a generalised linear model. Robustness to different specifications is also an important aspect to consider and investigate.

In our study, we overimposed a fixed grid over the study region. The cell size is important and should depend on the phenomenon under study. Information from previous surveys can be used to this purpose.

Fitting the two models in the two steps jointly can be very cumbersome, especially given the latent factor analysis step. For this reason, we choose to fit the two models separately. It can be argued that a question of uncertainty propagation arises: to address this issue we have decided to follow a simple strategy common in multiple imputation data analysis. More in detail, we sample five values from the joint posterior of the sampling probabilities and we run the model in the second step five times on the five pseudodata. A variance inflation factor is then obtained for each posterior predicted infection probabilities obtained from the geostatistical model. Our results show an underestimation of uncertainty using the two-stage approach thus the VIF should be used to correct the map of posterior standard errors of infection probabilities.

Our data is clearly affected by preferential sampling: we expected this *a priori* and we proposed the statistical model presented in this paper to address this issue. Not surprisingly the calibrated Kullback-Leibler divergence displays high values, greater than 0.70, for almost all grid cells. Even more interesting, some of the cells with higher calibrated Kullback-Leibler divergence are in the highest risk area located in the northern part of the region.

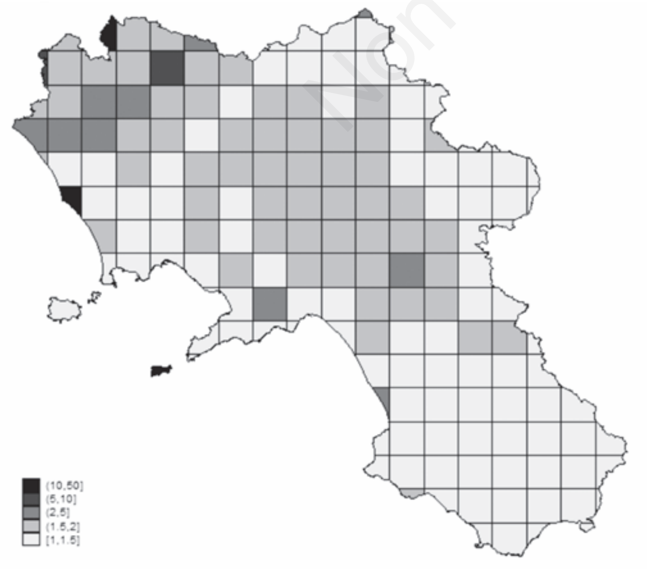

Figure 6. Spatial distribution of the variance inflation factor. Campania region, southern Italy (2013-2014).

Conclusions

In veterinary surveillance the use of remote sensing covariates is becoming fairly popular, mainly because some of these covariates can accurately predict intermediate host ecology. We choose to simplify this aspect via a dimension-reduction factor analysis embedded in our Bayesian geostatistical model. We chose *a priori* the covariates to include in each latent factors thanks to the vast documentation available on the lifecycle and the biology of the *F. hepatica*. On the contrary, if the covariates would have to be chosen arbitrarily, without any support from literature and expert opinion, this approach may distort the analysis and the results could be biased.

Adjusting for preferential sampling is very important to produce reliable prediction maps. The two-step procedure can comply easily in contexts in which sampling probabilities are estimated separately from the disease mapping modelling task.

References

Banerjee S, Carlin BP, Gelfand AE, 2014. Hierarchical modeling and analysis for spatial data. CRC Press, Boca Raton, FL, USA.

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

Bosco A, Rinaldi L, Musella V, Amadesi A, Cringoli G, 2015. Outbreak of acute fasciolosis in sheep farms in a Mediterranean area arising as a possible consequence of climate change. Geospat Health 9:354.

Catelan D, Rinaldi L, Musella V, Cringoli G, Biggeri A, 2012. Statistical approaches for farm and parasitic risk profiling in geographical veterinary epidemiology. Stat Methods Med Res 21:531-43.

Congdon P, 2003. Applied bayesian modelling. Wiley, New York, NY, USA.

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

Cringoli G, Bosco A, Rinaldi L, 2015. CREMOPAR: new strategies for monitoring and control of parasitic diseases of ruminants in the Mediterranean area. In: Proceedings of the XXII International Congress of Mediterranean Federation for Health and Production of Ruminants, 2015 Jun 17-20, Sassari, Italy.

Cringoli G, Rinaldi L, Maurelli MP, Utzinger J, 2010. FLOTAC: new multivalent techniques for qualitative and quantitative copromicroscopic diagnosis of parasites in animals and humans. Nat Protoc 5:503-15.

Diggle PJ, Menezes R, Su T, 2010. Geostatistical inference under preferential sampling. J Roy Stat Soc C 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.

Gelman A, Rubin DR, 1992. Inference from iterative simulation using multiple sequences (with discussion). Stat Sci 7:457-511.

Little RJA, Rubin DB, 2014. Statistical analysis with missing data. John Wiley & Sons, New York, NY, USA.

Lunn DJ, Thomas A, Best N, Spiegelhalter D, 2000. WinBUGS: a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10:325-37.

McCulloch RE, 1989. Local model influence. J Am Stat Assoc 84:473-8.

Musella V, Catelan D, Rinaldi L, Lagazio C, Cringoli G, Biggeri A, 2011. Covariate selection in multivariate spatial analysis of ovine parasitic infection. Prev Vet Med 99:69-77.

Musella V, Rinaldi L, Lagazio C, Cringoli G, Biggeri A, Catelan D, 2014. On the use of posterior predictive probabilities and prediction uncertainty to tailor informative sampling for parasitological surveillance in livestock. Vet Parasitol 205:158-68.

Pati D, Reich BJ, Dunson DB, 2011. Bayesian geostatistical modelling with informative sampling locations. Biometrika 98:35-48.

Rinaldi L, Biggeri A, Musella V, De Waal T, Hertzberg H, Mavrot F, Torgerson PR, Selemetas N, Coll T, Bosco A, Grisotto L, Cringoli G, Catelan D, 2015a. Sheep and *Fasciola hepatica* in Europe: the GLOWORM experience. Geospat Health 9:353.

Rinaldi L, Hendrickx G, Cringoli G, Biggeri A, Ducheyne E, Catelan D, Morgan E, Williams D, Charlier J, Von Samson-Himmelstjerna G, Vercruysse J, 2015b. Mapping and modelling helminth infections in ruminants in Europe: experience from GLOWORM. Geospat Health 9:347.

Shaddick G, Zidek JV, 2014. A case study in preferential sampling: long term monitoring of air pollution in the UK. Spat Stat 9:51-65.

Zouré HG, Noma M, Tekle AH, Amazigo UV, Diggle PJ, Giorgi E, Remme JH, 2014. The geographic distribution of onchocerciasis in the 20 participating countries of the African Programme for Onchocerciasis Control:(2) pre-control endemicity levels and estimated number infected. Parasite Vector 7:326.