Bayesian Analysis (2015) **10**, Number 3, pp. 753–758

Rejoinder*

Gustavo da Silva Ferreira† and Dani Gamerman‡

1 Introduction

We would like to start by thanking the Editor of BA for the opportunity to discuss our work and the discussants Peter J. Diggle, Michael G. Chipeta, James V. Zidek, Noel Cressie and Raymond L. Chambers for a very thorough evaluation of our contribution and for their thoughts on the topic. Our rejoinder to the discussion will be presented in the following topics: Preferential Sampling; Auxiliary Information; Models for $[X|S]$; Utility Functions; Sequential Design; and Approximations.

2 Preferential Sampling

Preferential sampling plays an important role in surveys routinely carried out by Official Statistics agencies, as those developed in the Brazilian Institute of Geography and Statistics (IBGE), the institution that one of us is affiliated to. So we are well aware of the relevance of this issue. In addition to the areas of application of preferential sampling cited by the discussants, it is important to mention the very topical area of publication bias (see Bayarri and DeGroot, 1993; Franco et al., 2014).

The link between preferential sampling and the methodology of survey-sampling presented by discussants is also very helpful in clarifying the similarities of approaches. We agree with Cressie and Chambers (hereafter CC) that papers from the latter may bring important aspects of sampling design to the context of Geostatistics. However, an important distinction between the approaches is needed. While in the context of survey methods the population size is generally fixed at a finite N, this feature is not true in the context of Geostatistics. In fact, in Geostatistics a finite value of N may only be associated with the discretization of a continuous process. Thus, part of the similarity between the approaches stems from the current limitation of many approaches to handle inference and prediction in Geostatistics appropriately due to the use of discrete approximations.

We agree with Chipeta and Diggle (hereafter CD) that preferential sampling is a method of adaptive design, which may depend on the previous design without relying on the underlying process S. Similarly, inference may be simplified after assuming that the process X is governed by the values of a spatially distributed covariate W, thus rendering conditional independence between X and S given W. The main difficulty is finding and quantifying such a covariate. We will return to this issue in the next section.

*Main article DOI: 10.1214/15-BA944.

†National School of Statistical Sciences, Brazilian Institute of Geography and Statistics, Rio de Janeiro, Brazil, gustavo.ferreira@ibge.gov.br

‡Department of Statistical Methods, Mathematical Institute, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, dani@im.ufrj.br

 DOI: 10.1214/15-BA944REJ

3 Auxiliary Information

All discussants brought in the issue of replacement of some of the latent and unobserved random processes by observed quantities. These processes are a mere artifact to best represent our lack of knowledge about the true mechanism underlying the observation processes, if such truth exists. If we knew what variables cause our observational processes to behave as they do and were able to measure these quantities, we should use them. Otherwise, models that incorporate relevant features such as smoothness of these processes are a useful alternative. This issue is not only relevant to Spatial Statistics but to any other area of Statistics where absence of complete knowledge of the sources of variation forces the use of qualitative information in the form of latent processes, and Gaussian processes are one of the most adequate choices as a first step.

We agree that covariates could (and should) be part of all model components. Covariates may also be used to construct a deterministic intensity function of the point process X, or as proxy of a random process S in a model with a random intensity function. They may reduce remaining spatial heterogeneity in the mean or covariance structure of S, when available. CC suggested the use of other covariates Z, assumed to be highly correlated with the process S, as a proxy. This is a practical solution that needs to be used carefully. Despite simplifications produced in inference, it is important to bear in mind that this solution may also introduce another source of error in the model.

However, despite being a natural choice, we do not always have available covariates. In other situations, the covariates are available only in part of the region of interest. In these cases, it is often necessary to interpolate values before including them in the model, adding more uncertainty to the results.

4 Models for $[X|S]$

We agree that an appropriate model is crucial and robustness considerations are even more important here. Many discussants of Diggle et al. (2010) seemed to agree that, in practical situations, it is unlikely that the design is governed by a log-Gaussian Cox process. This issue is in line with our views expressed in the first paragraph of the previous section. In some cases, it may be enough to assume that the intensity of the point process is somehow proportional to the underlying process at an appropriate scale. In these cases, a log-linear intensity function may be seen as the first option for approximation and may thus be a good starting point to mitigate and understand the consequences of a preferential sample.

In addition to methods for obtaining a robust design against selection bias, alternative specifications for $[X|S]$ may also contribute to a satisfactory model-based inference. All discussants have correctly expressed concern about the sensitivity of the inference to the choice of the model for the intensity function. We concur with their concern and particularly welcome the use of non-parametric specifications to replace the parametric specification of our paper. This is bound to lead to a more flexible and hence robust alternative to the global linearity imposed by the $\alpha+\beta S$ predictor over the entire region of interest.

One approach following this path allows the regression coefficients' processes α and β to vary locally over space. Inference for these models may be performed in a simplified fashion, via discretization of coefficients over sub-regions (Pinto Junior et al., 2015), or exactly, by retaining the continuous variation of the infinite dimensional regression coefficient process (Gonçalves and Gamerman, 2015). Another non-parametric approach was proposed by Kottas and Sansó (2007), based on mixtures of Dirichlet process priors. Their idea could be adapted to a prior for the intensity of a log-Gaussian Cox process that (instead of being concentrated on) is only centered at $\alpha + \beta S$, again allowing more flexibility for the intensity.

Zidek also suggested an interesting situation where the generating process for the locations is based on another process S^*. In this case, it is first necessary to assess the degree of dependence between these two processes, that is, to evaluate if $[S, S^*] = [S][S^*]$. If so, the researcher can consider the points X as ancillary for S and proceed with the non-preferential, standard inference as usual. Otherwise, it may be necessary to establish some form of dependence for $[S|S^*]$ to be able to proceed with inference. A bivariate specification for $[S, S^*]$ may be one way forward (see Gamerman et al., 2007; Crainiceanu et al., 2008).

In conclusion, our model for $[X|S]$ seems to be able to highlight the effects of preferential sampling in the absence of a better understanding of the true causes of variation or of relevant covariates, but more flexible forms are needed.

5 Utility Functions

We agree with CC's suggestions about notation to improve the paper understanding. In particular, we consider pertinent the suggestion of making s_d explicit in the utility function.

The choice of a particular utility function is a crucial step to obtaining the optimal design. For this reason, we completely agree with CC that the questions *how much?* and *why?* cannot be set aside while the researcher is planning a new location sample.

Utility functions based on predictive variance reductions have long been recognized as appropriate to measure improvements in predictive accuracy. Zidek noted that more general evaluation (of predictive/estimation performance) may require the use of other utility functions. We already presented at least one alternative formulation early in the paper to emphasize our adherence to this point and to detach ourselves from the initial approaches based only on a single function, evaluating predictive errors. Our methods apply equally well to any quantifiable utility function.

The combination of different goals — e.g., reducing predictive errors, reducing uncertainty with respect to S, identifying thresholds, reducing estimation errors and evaluating costs involved — allows the researcher to obtain designs in complex situations. Examples of complex utility functions with competing goals can be found in the recent work of Müller et al. (2004), Ruiz-Cárdenas et al. (2012) and Ferreira (2015) to name a few. Note that these goals may be competing, e.g., reduction of the nugget effect estimation error assigns more utility to regions close to locations already sampled

whereas reduction of predictive error assigns more utility for regions far from the points already sampled. The additional challenge in this case is to weigh the various objectives involved.

Alternatively, the use of entropy in the utility function is usually recommended for situations where multiple goals are involved (Caselton and Zidek, 1984) but this is not an easily implementable solution in practice. We do believe that in practical situations the utility function needs to be specified by whoever is in charge of the analysis, be it a regulatory body, a government institution or a researcher, with the help of a statistician. The responsible entity must be able to value the worth of the information each component of the utility provides. If one can answer, for example, why predictive variance reduction is relevant, one must be able to assign its monetary value; otherwise, one must reassess the relevance of each component included in the utility function. This value is more easily combined with readily quantifiable monetary components such as cost associated with each new location. An economist specialized in the area of the study may be a useful addition to the team setting up this enterprise at this stage.

6 Sequential Design

The approach we used for design (Müller, 1999) allows the planning of m new sampling sites according to the utility function defined by the researcher. This design plan can be sequential or not, while the underlying process S can be static or dynamic in time. In cases where S is a dynamic process, planning of a sequential design scheme is a natural choice. However, in the case where S is fixed in time, one can also plan a sequential design, although it is not possible to ensure that the resulting design will be optimal. Actually, a sequential sample design can be a simple and cheap solution, especially in situations where the costs are not fixed and can increase before obtaining the desired new sample locations.

We anticipate difficulties to assigning a distribution for $[d|S, \mathbf{x}]$ without reducing the flexibility of the model, when a sequential-sampling-design strategy to update θ and S through $[d|\mathbf{x}, \mathbf{y}]$ is built. It would be necessary to assess whether the initial sample is preferential in cases suggested by CC with a pilot study. If it is possible to assume that the pilot phase is non-preferential, then this information will be ancillary to inference. On the other hand, if the pilot sample is preferential, then it is possible to incorporate this information by performing modifications in $[X|S]$. Note that this may remove the Poissonity of the model for $[X|S]$, due to repulsions that may occur around previously selected locations. This will imply more difficulty for likelihood approaches and opens up for interesting research questions.

The case of a dynamic underlying process is more complex. We visualize the following generalization of the 2-stage set-up proposed by CD. In cases where samples are taken at different times in a multi-stage scenario, we may have

$$[Y, X, S] = \prod_t [Y_t|Y_{1:t-1}, X_{1:t}, S_t][X_t|X_{1:t-1}, S_{1:t}][S_t|S_{1:t-1}],$$

where $Z_{a:b} = (Z_a, Z_{a+1}, \ldots, Z_{b-1}, Z_b)$, for $a < b$ integers, and dependence on hyperparameters was removed from the notation as in CD. Conditional independence between observations given the corresponding underlying processes may be assumed in some cases leading to $[Y_t|Y_{1:t-1}, X_{1:t}, S_t] = [Y_t|X_t, S_t]$. Similarly, $[X_t|X_{1:t-1}, S_{1:t}] = [X_t|S_t]$ may be assumed for some sampling schemes, although the general formulation may be required in some cases (see the paragraph above). Further simplification such as those suggested by CD may be assumed, depending on the situation. Ferreira (2015) worked on a similar structure, simplified by his non-preferential sampling scheme.

In the practical situation presented by CD, where the sampling order is a crucial factor, a simple utility function based on predictive variance reductions or exceedances probably would not be enough to produce a satisfactory sample design. In challenging situations like this, it would be necessary to choose more complex forms to reward each sample unit, at each time, in a sequential sampling scheme.

7 Approximations

Questioning the stationarity assumption is a mandatory task in any Spatial Statistics problem. Stationarity can always be seen as an approximation to a more complex underlying dependence. But it turns out that in many practical situations it has proved to be a reasonable, viable solution. Alternatively, approaches based on convolution process (Higdon, 2002) or the use of more flexible (non-parametric) structures to enhance the spatial dependence structure of S can be used. Again, covariates are always relevant options to handle the large-scale heterogeneity considered by CD. Obviously, an increase in the complexity of the model can also complicate the evaluation of the utility function used to obtain the optimal design and parsimony may have to be called into action.

We agree with Zidek that other approximation of $V(S|\mathbf{x}, \mathbf{y})$ could be used in this step. A simpler, but expensive, alternative is to generate sub-chains in order to estimate this quantity during MCMC. Alternative analytical approaches may be preferred to avoid the increasing computational cost. It is important to emphasize that

- The approximation was only used to evaluate the utility function, since an analytical expression for $V(S|\mathbf{x}, \mathbf{y})$ is not available; the values of S were always sampled from the correct distribution $[S|\mathbf{y}, \mathbf{x}]$;
- Other utility functions, as those used in our case study, may not require any approximation.

References

Bayarri, M. J. and DeGroot, M. H. (1993). "The analysis of published significant results." In: *Rassegna di Metodi Statistici ed Applicazioni* (W. Racugno, ed.), 19–41. Pitagora. MR1223386. 753

Caselton, W. F. and Zidek, J. V. (1984). "Optimal monitoring network designs." *Statistics & Probability Letters*, 2(4): 223–227. 756

Crainiceanu, C. M., Diggle, P. J., and Rowlingson, B. (2008). "Bivariate binomial spatial modeling of Loa loa prevalence in tropical Africa." *Journal of the American Statistical Association*, 103(481): 21–37. MR2420211. doi: http://dx.doi.org/10.1198/016214507000001409. 755

Diggle, P. J., Menezes, R., and Su, T.-L. (2010). "Geostatistical inference under preferential sampling (with discussion)." *Applied Statistics*, 59(2): 191–232. MR2744471. doi: http://dx.doi.org/10.1111/j.1467-9876.2009.00701.x. 754

Ferreira, M. A. (2015). "Inhomogeneous evolutionary MCMC for Bayesian optimal sequential environmental monitoring." *Environmental and Ecological Statistics*, 1–20. 755, 757

Franco, A., Malhotra, N., and Simonovits, G. (2014). "Publication bias in the social sciences: Unlocking the file drawer." *Science*, 345(6203): 1502–1505. 753

Gamerman, D., Salazar, E., and Reis, E. (2007). "Dynamic Gaussian process priors, with applications to the analysis of space-time data (with discussion)." In: *Bayesian Statistics* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds.), volume 8, 149–174. Oxford University Press. MR2433192. 755

Gonçalves, F. and Gamerman, D. (2015). "Exact Bayesian inference in spatio-temporal Cox processes driven by multivariate Gaussian processes." Technical report, Statistical Methods Laboratory, Federal University of the Rio de Janeiro. 755

Higdon, D. (2002). "Space and space-time modeling using process convolutions." In: *Quantitative Methods for Current Environmental Issues* (C. W. Anderson, V. Barnett, P. C. Chatwin and A. H. El-Shaarawi, eds.), 37–56. Springer. MR2059819. 757

Kottas, A. and Sansó, B. (2007). "Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis." *Journal of Statistical Planning and Inference*, 137(10): 3151–3163. MR2365118. doi: http://dx.doi.org/10.1016/j.jspi.2006.05.022. 755

Müller, P. (1999). "Simulation-based optimal design." In: *Bayesian Statistics* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), volume 6, 459–474. Oxford University Press. MR1723509. 756

Müller, P., Sansó, B., and De Iorio, M. (2004). "Optimal Bayesian design by inhomogeneous Markov chain simulation." *Journal of the American Statistical Association*, 99(467): 788–798. MR2090911. doi: http://dx.doi.org/10.1198/016214504000001123. 755

Pinto Junior, J. A., Gamerman, D., Paez, M. S., and Fonseca Alves, R. H. (2015). "Point pattern analysis with spatially varying covariate effects, applied to the study of cerebrovascular deaths." *Statistics in Medicine*, 34(7): 1214–1226. 755

Ruiz-Cárdenas, R., Ferreira, M. A., and Schmidt, A. M. (2012). "Evolutionary Markov chain Monte Carlo algorithms for optimal monitoring network designs." *Statistical Methodology*, 9(1): 185–194. MR2863607. doi: http://dx.doi.org/10.1016/j.stamet.2011.01.009. 755