Bayesian Analysis (2015) **10**, Number 3, pp. 741–748

Comment on Article by Ferreira and Gamerman*

Noel Cressie[†§] and Raymond L. Chambers[‡]

Abstract. A utility-function approach to optimal spatial sampling design is a powerful way to quantify what "optimality" means. The emphasis then should be to capture all possible contributions to utility, including scientific impact and the cost of sampling. The resulting sampling plan should contain a component of designed randomness that would allow for a non-parametric design-based analysis if model-based assumptions were in doubt.

Keywords: design-based inference, hierarchical model, informative sampling, preferential sampling, utility function.

1 Introduction

We would like to express our appreciation to Gustavo da Silva Ferreira and Dani Gamerman (hereafter, FG) for their paper on Bayesian preferential spatial sampling (Ferreira and Gamerman, 2015) and to the editor of *Bayesian Analysis* for the opportunity to contribute to the discussion. Building on the papers by Müller (1999) and Diggle et al. (2010), the authors give a Bayesian approach to choosing *new* sampling locations after initial data are assumed to have been obtained under preferential sampling.

2 What is Fixed and What is Random?

Let the initial sample be $\mathbf{y}_{\mathbf{x}}$, obtained at preferential sampling locations $\mathbf{x}$; note that we have emphasised dependence of $\mathbf{y}$ on $\mathbf{x}$ through the notation $\mathbf{y}_{\mathbf{x}}$, but it is exactly the same as what FG notate as $\mathbf{y}$. The observation locations $\mathbf{x}$ and the observations $\mathbf{y}_{\mathbf{x}}$ are known by the statistician designing the next phase of the study, and hence all criteria and inferences should depend on both $\mathbf{x}$ and $\mathbf{y}_{\mathbf{x}}$. One can see this most clearly in FG's definition of the Bayesian design criterion $U(\mathbf{d})$, given at the beginning of FG-Section 4. However, the reader should notice that equations FG-(3) and FG-(4) do not emphasise conditioning on $\mathbf{x}$, along with $\mathbf{y}_{\mathbf{x}}$, something we assume is an oversight on the part of the authors.

It helped us to augment the notation for the utility function from $u(\mathbf{d}, \theta, \mathbf{y}_{\mathbf{d}})$ to $u(\mathbf{d}, \theta, \mathbf{y}_{\mathbf{d}}; \mathbf{x}, \mathbf{y}_{\mathbf{x}})$; and likewise we suggest that the expected utility be notated:

$$U(\mathbf{d}; \mathbf{x}, \mathbf{y}_{\mathbf{x}}) = E(u(\mathbf{d}, \theta, \mathbf{y}_{\mathbf{d}}; \mathbf{x}, \mathbf{y}_{\mathbf{x}}) | \mathbf{x}, \mathbf{y}_{\mathbf{x}}), \tag{1}$$

*Main article DOI: 10.1214/15-BA944.

[†]National Institute of Applied Statistics Research Australia, University of Wollongong, Australia, ncressie@uow.edu.au

[‡]National Institute of Applied Statistics Research Australia, University of Wollongong, Australia

[§]Cressie's research was partially supported by the US National Science Foundation and the US Census Bureau through the NSF-Census Research Network program; and it was partially supported by a 2015-2017 Australian Research Council Discovery Project.

 DOI: 10.1214/15-BA944B

where $\mathbf{d}$ is considered fixed and the expectation is taken over $[\theta, \mathbf{y_d}|\mathbf{x}, \mathbf{y_x}]$. When there are many "stakeholders" (e.g., in an environmental study), each coming with his/her own utility function, how can a single utility function be constructed? Le and Zidek (2006, Chapter 11) opt for one based on entropy. Do the authors have any other suggestions to build "compromise" into a utility function?

Notation is really important in these complex situations, so in the case of the utility function defined by FG-(4), which involves the latent process (not the observations), we suggest that u be rewritten as:

$$u(\mathbf{d}, \theta, \mathbf{s_d}; \mathbf{x}, \mathbf{y_x});$$

that is, $\mathbf{s_d}$ replaces $\mathbf{y_d}$ in u. Depending on the context, u could be a function of the new observations, $\mathbf{y_d} \equiv (y(d_1), \ldots, y(d_m))'$, or of the corresponding latent process, $\mathbf{s_d} \equiv (s(d_1), \ldots, s(d_m))'$; recall that FG have defined $y(\cdot)$ to be a noisy, shifted version of the mean-zero latent process $s(\cdot)$.

In the rest of our discussion, we follow the authors' lead and use (1), albeit with our modified notation that emphasises dependence on $\mathbf{x}$ and $\mathbf{y_x}$. The utility-function approach to optimal design is attractive, but it will only be truly useful when components that quantify "how much?" and "why?" are specifically included; see Sections 3 and 4 for further discussion.

As FG make clear, the process $s(\cdot)$, the sampling locations $\mathbf{x} \equiv (x_1, \ldots, x_n)'$, and the observations $\mathbf{y_x} \equiv (y(x_1), \ldots, y(x_n))'$ have a possibly complex joint distribution. Following Diggle et al. (2010), the authors put structure on this joint distribution by assuming FG-(1), FG-(2), and a log-Gaussian Cox process. From the point of view of sample survey design, the information in $\mathbf{x}$ and $\mathbf{y_x}$ is comparable to what one would gain from a pilot study, but it requires knowledge of components of θ in order to make the pilot study operational. The following suggestion seems compatible with the authors' approach to optimal design via preferential sampling.

It is hard to design a study if there is no knowledge from which to draw. In the pre-pilot phase, one might choose a simple random sample which, in the spatial context, means that observation locations are sampled uniformly from the spatial region of interest, A. We note that this corresponds to a degenerate case of preferential sampling where β, the coefficient of $s(\cdot)$ in the log-intensity, is equal to 0, and we also note the presence of randomness in this pre-pilot phase.

After gaining knowledge to make the pilot study operational, $\mathbf{x}$ and $\mathbf{y_x}$ are obtained from preferential sampling, and again we note the presence of randomness in choosing $\mathbf{x}$. Given $\mathbf{x}$ and $\mathbf{y_x}$, the next set of locations, $\mathbf{d} \equiv (d_1, \ldots, d_m)'$, need to be chosen, for which there will be a corresponding (based on the latent vector $\mathbf{s_d}$) $\mathbf{y_d}$. This is the problem considered by FG, and their solution follows closely the proposal of Müller (1999). But there is an important difference: Müller considers $\mathbf{d}$ to be a "design parameter" that he clearly treats as non-stochastic (fixed). We would like to ask FG the following question: If $\mathbf{x}$ is considered to be random in the pilot phase, why would $\mathbf{d}$ be treated as fixed in the main phase of the study?

The authors follow Müller's (1999) proposal closely but, in our opinion, they lose an opportunity to build a sequential-sampling-design strategy that updates the posterior distribution for θ and $s(\cdot)$ through *random* sampling from the distribution $[\mathbf{d}|\mathbf{x}, \mathbf{y_x}]$. This can be obtained from $[y(\cdot)|s(\cdot)]$ and $[\mathbf{x}, \mathbf{d}|s(\cdot)] = [\mathbf{d}|\mathbf{x}, s(\cdot)][\mathbf{x}|s(\cdot)]$. The second factor in the product is a log-Gaussian Cox process on A, and the first factor is presumably a log-Gaussian Cox process too, but on $A\backslash\mathbf{x}$. Can the authors comment on our suggestion that designed randomness be used to obtain $\mathbf{d}$?

The following section makes strong links between FG's proposal and the survey-sampling literature. It also reinforces the general desire for a component of randomisation in the design.

3 Spatial Sampling Designs

The development in FG (and in the literature that it refers to) frames the sampling problem as the selection of a set of points on a grid that "covers" the region of interest, A. From this perspective, it is a special case of a finite-population sampling problem, with the N grid points defining the population, and with two random quantities defined on these points. The first is the sample-selection process that results in $\mathbf{x}$ (from the pilot study) and $\mathbf{d}$ (from the main survey). The second is the latent process $s(\cdot)$, from which "noisy" observations $\mathbf{y_x}$ and $\mathbf{y_d}$ are taken at $\mathbf{x}$ and $\mathbf{d}$, respectively.

The aim of a spatial sampling design should be to specify a suitable procedure for making a draw from the distribution of $\mathbf{d}$ given $\mathbf{x}$ and $\mathbf{y_x}$; see our discussion in Section 2. What is meant by "suitable" depends crucially on the target of inference for the sampling exercise; FG make their target $s(\cdot)$ and to a lesser extent θ, and they assume that "suitability" can be characterised through a utility function u. Their optimal sample $\mathbf{d}$ is then the set of (presumably so far unsampled) grid points that maximise the expected value of this utility, where recall that the expectation is with respect to the joint distribution of θ and $\mathbf{s_d}$ conditional on $\mathbf{x}$ and $\mathbf{y_x}$.

The authors' optimal-design procedure is explicitly model-based. Furthermore, the fact that selection of $\mathbf{d}$ depends on a log-Gaussian Cox process with intensity function that is a function of $s(\cdot)$ means that the sampling design is informative (Chambers and Clark, 2012, Section 1.4), which Diggle et al. (2010) and FG refer to as preferential sampling. That is, one cannot treat the realised value of $\mathbf{d}$ as ancillary when using the combined pilot-study and main-survey data to make inferences. There is a well developed theory in the sample-survey literature for the analysis of a sample collected via informative sampling; see Chambers et al. (2012). From this perspective, the use of a log-Gaussian Cox process as a model for $\mathbf{x}$ is equivalent to Poisson sampling with inclusion probabilities that depend on the values of $s(\cdot)$ over the grid defining the finite population. We would like to draw attention to an extensive literature on this type of sampling and its implications, including many Bayesian approaches; see Nandram et al. (2013).

More complex informative-sampling methods have also been investigated, principally in the context of sampling spatially clustered populations; see Rapley and Welsh (2008)

for a Bayesian specification and, in the context of sampling on networks, see Thompson and Seber (1996). Awareness of this closely related literature would seem advantageous for further development of the ideas set out in FG's paper.

The main inferential paradigm in survey sampling is design-based inference, which assumes that $y(\cdot)$ is fixed, and all inference is relative to the distribution of $\mathbf{d}$. Moreover, the outcome of the pilot study ($\mathbf{x}$ and $\mathbf{y_x}$) is treated as fixed. Generally, the survey-sampling approach is based on frequentist inference about population summary statistics. The inference uses weights obtained from the randomisation in the design, along with the population values $y(\cdot)$ over the grid. In the simplest case, these weights are defined by the inverses of the inclusion probabilities for each of the elements of $\mathbf{d}$, but more general "calibrated" weights are typically preferred; see Deville and Särndal (1992). When informative (i.e. preferential) sampling is used, design-based inference, although theoretically still applicable, becomes problematic in practice. In order to carry out the survey sampling, one has to have access to the distribution of $\mathbf{d}$, which depends on the latent process $s(\cdot)$. For design-based inference, one might try replacing $s(\cdot)$ in the intensity function of the log-Gaussian Cox process with $z(\cdot)$, a spatial covariate whose value is known for every point on the grid and which is (hopefully) highly correlated with $s(\cdot)$. This is the model underpinning size-biased sampling; see Patil and Rao (1978).

A model-based approach seems therefore necessary under preferential sampling, such as assuming a spatial-statistical model for $s(\cdot)$. However, this does not mean that the basic design-based notions of randomisation, stratification, and clustering cannot be used in a preferential-sampling approach, since they are all useful tools that lead to a better representation of a heterogeneous population. In particular, what happens when the model FG-(1) and FG-(2) does not adequately describe the spatial variability in $y(\cdot)$ and $s(\cdot)$? The optimality of $\mathbf{d}$, and the validity of any consequent inference depends critically on the appropriateness of this model. This is clearly a weakness, should the design be for a highly scrutinised environmental study where scientists are worried not only about the environment but also about the team of lawyers waiting to litigate! Other problems arise when there are relatively few such choices of $\mathbf{d}$, irrespective of the values of $\mathbf{x}$ and $\mathbf{y_x}$, all of whose utilities are comparable.

We would like to reiterate that some form of randomisation in a design is always a good idea, because it offers protection against a biased (unintentional or intentional) choice of sample sites (e.g. Aldworth and Cressie, 1999). Perhaps more importantly, randomisation ensures that an updated fit of an assumed model for $s(\cdot)$ can be validly assessed from sample data and that replication-based ideas can be used for this purpose. And finally, when the parametric model is in doubt, the presence of randomisation allows the possibility that design-based inference could be used.

As far as we are aware, there has been no work on "robustifying" inference based on data collected via preferential sampling, in order to make it less sensitive to model misspecification. Perhaps FG's paper will stimulate such investigations. Recent research reported in Welsh and Wiens (2013) may provide an indication of how a robust preferential-sampling approach might work, with these authors developing an approach to sampling design that minimises the maximum prediction error in a *neighbourhood* of an assumed model for $s(\cdot)$. A related line of research concerns what could be termed as a

composite approach to preferential sampling, where a proportion of the sampling effort is randomly spread over the spatial grid, with the rest allocated to a more targeted preferential sampling design. An important research question here concerns how this allocation might be determined, based on the information in $\mathbf{x}$ and $\mathbf{y_x}$; this is discussed further in Section 4.

We conclude this section by stating some basic elements of a good sampling design, be it spatial or not. A good design will stratify to ensure sampling over a range of levels of factors or a range of values of covariates. A good design will specify, in advance, inference thresholds and determine the number of observations per stratum needed to achieve those thresholds. Such designs create a rational basis for the inevitable compromise between the cost of the study and the ability to make scientific inferences from incomplete and noisy data (e.g. Cressie, 1998; Zidek et al., 2000). Finally, a good design will involve a component of designed randomness, from which non-parametric, design-based inference is also possible, should the model-based assumptions be in doubt.

4 Utility Functions

The process $s(\cdot)$ and the behaviour of the observations $\mathbf{y_x}$ depend on parameters, which are denoted as θ. If θ were known, then $[\mathbf{x}, \mathbf{y_x}, s(\cdot)|\theta] = [\mathbf{y_x}|\mathbf{x}, s(\cdot), \theta]\,[\mathbf{x}, s(\cdot)|\theta]$, and the predictive distribution is

$$[s(\cdot)|\mathbf{x}, \mathbf{y_x}, \theta] = [\mathbf{y_x}|\mathbf{x}, s(\cdot), \theta]\,[\mathbf{x}, s(\cdot)|\theta] \,/\, [\mathbf{x}, \mathbf{y_x}|\theta]\,. \tag{2}$$

Using the terminology of Cressie and Wikle (2011), an *empirical hierarchical model (EHM)* results if an estimate $\widehat{\theta}$ is used in place of θ in (2), and inference on $s(\cdot)$ is then based on the *empirical* predictive distribution,

$$[s(\cdot)|\mathbf{x}, \mathbf{y_x}, \widehat{\theta}] = [\mathbf{y_x}|\mathbf{x}, s(\cdot), \widehat{\theta}][\mathbf{x}, s(\cdot)|\widehat{\theta}]/[\mathbf{x}, \mathbf{y_x}|\widehat{\theta}]. \tag{3}$$

This EHM set-up is what Diggle et al. (2010) use, and they address the importance of making $\widehat{\theta}$ a function of both $\mathbf{x}$ and $\mathbf{y_x}$.

If there is uncertainty in θ that can be expressed in terms of a prior probability distribution $[\theta]$, then a *Bayesian hierarchical model (BHM)* results. Bayes' Theorem yields the posterior distribution,

$$[s(\cdot), \theta|\mathbf{x}, \mathbf{y_x}] = [\mathbf{y_x}|\mathbf{x}, s(\cdot), \theta]\,[\mathbf{x}, s(\cdot)|\theta]\,[\theta] \,/\, [\mathbf{x}, \mathbf{y_x}]\,. \tag{4}$$

For a BHM, the *Bayesian* predictive distribution is the integral of (4) with respect to θ, namely $\int [s(\cdot), \theta|\mathbf{x}, \mathbf{y_x}]\,\mathrm{d}\theta$.

The BHM is coherent in the sense that all inferences emanate from a well defined joint probability distribution. On the other hand, it requires specification of a prior $[\theta]$, and it often consumes a large amount of computing resources. The EHM represents a compromise that may achieve computational efficiency.

Diggle et al. (2010) do *not* address optimal spatial design in the way that FG do. If one sets about doing it, analogous to FG's approach but within Diggle et al.'s EHM

framework, one would modify (1) so that the right-hand side would be the expectation taken over $[\mathbf{y_d}|\mathbf{x}, \mathbf{y_x}, \theta]$, and hence one would write the expected utility as $U(\mathbf{d}, \theta; \mathbf{x}, \mathbf{y_x})$. Then the *empirical* utility is $U(\mathbf{d}, \widehat{\theta}, \mathbf{x}, \mathbf{y_x})$ and, analogous to FG's approach, one would find the EHM-optimal $\mathbf{d}$ by maximising $U(\mathbf{d}, \widehat{\theta}; \mathbf{x}, \mathbf{y_x})$ with respect to $\mathbf{d}$. Is there a more principled way to account for θ (which is considered fixed but unknown) in the EHM framework?

Let $W \equiv \{\mathbf{d}, \theta, \mathbf{y_d}\}$ denote all the unknowns in FG's model. Let $\widehat{W}(\mathbf{x}, \mathbf{y_x})$ be one of many possible decisions about W based on $\mathbf{x}$ and $\mathbf{y_x}$. Some decisions are better than others, which can be quantified through a very general utility function that is bounded above, and which we denote as $\mathcal{U}(W, \widehat{W}(\mathbf{x}, \mathbf{y_x}))$; the utility function, u, used by FG represents a particular form of the more general $\mathcal{U}$ considered here. Note that $\mathcal{U}$ should account for "how much?" and "why?" and could be negative. Obviously, large utilities are preferred, and it is a consequence of decision theory (e.g. Berger 1985) that the optimal decision is:

$$W^*(\mathbf{x}, \mathbf{y_x}) = \arg \sup_{\widehat{W}(\mathbf{x}, \mathbf{y_x})} \left\{ \mathrm{E}(\mathcal{U}(W, \widehat{W}(\mathbf{x}, \mathbf{y_x}))|\mathbf{x}, \mathbf{y_x}) \right\}. \tag{5}$$

Now suppose that the goal is inference on $g(W)$, where $g(\cdot)$ is a known, scientifically interpretable, possibly multivariate function of W. The answer to this inference problem is found in the predictive distribution, $[g(W)|\mathbf{x}, \mathbf{y_x}]$. Let $\widehat{g}$ denote a generic predictor of $g(W)$. The *mean* of the predictive distribution of $g(W)$, namely $\mathrm{E}(g(W)|\mathbf{x}, \mathbf{y_x})$, is a commonly used predictor, but this is just one of many possibly summaries of $[g(W)|\mathbf{x}, \mathbf{y_x}]$.

Why use the mean? Because it is straightforward to show that $\mathrm{E}(g(W)|\mathbf{x}, \mathbf{y_x})$ solves (5) when the utility function is "negative squared-error," $-(\widehat{g} - g(W))'(\widehat{g} - g(W))$. However, a negative squared-error utility assumes equal consequences for under-estimation as for over-estimation, which is not appropriate when $g(W)$ represents extreme events, such as crop failure due to drought.

Notice that we have written the utility as a function of all the unknowns, W, and a decision about all the unknowns, $\widehat{W}$. This gives us the opportunity to design for making inference on $\mathbf{d}$ simultaneously with making inference on θ, for example. Recall from Section 3 our discussion of the composite approach to optimal design. One of the components of θ might be the derivative of the variogram of $s(\cdot)$ at the origin (a critical parameter for kriging), which we simultaneously want to infer along with predicting the hidden spatial process $s(\cdot)$. Laslett and McBratney (1990) give a composite spatial design that distributes sampling locations regularly over A (for inference on $s(\cdot)$) and, around some of those locations, further locations are chosen very close together (for inference on θ). Do FG have any suggestions as to how an *optimal* composite spatial design might be obtained under their utility-function approach?

In conclusion, we thank the authors for their stimulating paper, and we can see a number of very interesting research problems waiting to be solved.

References

Aldworth, J. and Cressie, N. (1999). "Sampling designs and prediction methods for Gaussian spatial processes." In: Ghosh, S. (ed.), *Multivariate Design and Sampling* , 1–54. New York, NY: Marcel Dekker. MR1719054. 744

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis: Second Edition*. New York, NY: Springer-Verlag. MR0804611. doi: http://dx.doi.org/10.1007/978-1-4757-4286-2. 746

Chambers, R. L. and Clark, R. G. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford, UK: Oxford University Press. MR3186498. doi: http://dx.doi.org/10.1093/acprof:oso/9780198566625.001.0001. 743

Chambers, R. L., Steel, D. G., Wang, S., and Welsh, A. H. (2012). *Maximum Likelihood Estimation for Sample Surveys*. Boca Raton, FL: CRC Press. MR2963765. 743

Cressie, N. (1998). "Transect-spacing design of ice cores on the Antarctic continent." *Canadian Journal of Statistics*, 26: 405–418. 745

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley. MR2848400. 745

Deville, J. C. and Särndal, C. E. (1992). "Calibration estimators in survey sampling." *Journal of the American Statistical Association*, 87: 376–382. MR1173804. 744

Diggle, P., Menezes, R., and Su, T. (2010). "Geostatistical inference under preferential sampling." *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 59: 191–232. MR2744471. doi: http://dx.doi.org/10.1111/j.1467-9876.2009.00701.x. 741, 742, 743, 745

Ferreira, G. S. and Gamerman, D. (2015). "Optimal design in geostatistics under preferential sampling." *Bayesian Analysis*, 10: in this issue. 741, 742, 743, 744, 745, 746

Laslett, G. M. and McBratney, A. B. (1990). "Further comparison of spatial methods for predicting soil pH." *Soil Science Society of America Journal*, 54: 1553–1558. 746

Le, N. D. and Zidek, J. V. (2006). *Statistical Analysis of Environmental Space-Time Processes*. New York, NY: Springer. MR2223933. 742

Müller, P. (1999). "Simulation based optimal design." In: Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 6*, 459–474. Oxford, UK: Oxford University Press. MR1723509. 741, 742, 743

Nandram, B., Bhatta, D., Bhadra, D., and Shen, G. (2013). "Bayesian predictive inference of a finite population proportion under selection bias." *Statistical Methodology*, 11: 1–21. MR3000912. doi: http://dx.doi.org/10.1016/j.stamet.2012.08.003. 743

Patil, G. P. and Rao, C. R. (1978). "Weighted distributions and size-biased sampling with applications to wildlife populations and human families." *Biometrics*, 34: 179–189. MR0507202. doi: http://dx.doi.org/10.2307/2530008. 744

Rapley, V. E. and Welsh, A. H. (2008). "Model-based inferences from adaptive cluster sampling." *Bayesian Analysis*, 3: 717–736. MR2469797. doi: http://dx.doi.org/10.1214/08-BA327. 743

Thompson, S. and Seber, G. (1996). *Adaptive Sampling*. New York, NY: Wiley. MR1390995. 744

Welsh, A. H. and Wiens, D. P. (2013). "Robust model-based sampling designs." *Statistics and Computing*, 23: 689–701. MR3247826. doi: http://dx.doi.org/10.1007/s11222-012-9339-3. 744

Zidek, J. V., Sun, W., and Le, N. D. (2000). "Designing and integrating composite networks for monitored multivariate Gaussian pollution fields." *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 49: 63–79. MR1817875. doi: http://dx.doi.org/10.1111/1467-9876.00179. 745