Bayesian Analysis (2015) **10**, Number 3, pp. 749–752

Comment on Article by Ferreira and Gamerman*

James V. Zidek†

This paper concerns a very topical issue, namely the effect of preferential sampling the locations at which to measure a spatial process. The topic was highlighted at and studied by a research group at the Statistical and Applied Mathematical Sciences Institute (SAMSI) during its 2009–10 thematic year on spatial statistics, and a number papers came out of that initiative.

To put this paper in context, some background seems worthwhile. Selection bias in one form or another has always been an issue in statistical science, and it has been studied since at least the time when Horvitz and Thompson proposed their simple but ingenious approach to unbiasing estimates of finite population averages when sample items are preferentially selected (Horvitz and Thompson, 1952). Survey statisticians have long since recognized the adverse effect of such bias and the need to adjust for it when computing their estimates. Biostatisticians have also been concerned with this issue in the form of response biased sampling in estimating the relationship between a response Y and a covariate vector Z when instead of sites human subjects are the units (Scott and Wild, 2011). There, inter-subject dependence is ignored due to its complexity and the work of Liang and Zeger (1986) which allows that simplification to be made. The responses Y are assumed to be observed (although that assumption can be relaxed by modeling it) and subject selection is biased by these responses. In contrast, the present paper follows Diggle et al. (2010) and assumes instead that the role of Y is implicit and seen through the point process model that "knows" $Y = Y(x)$, or rather the latent process that generates it, through the intensity function $\exp\{\alpha + \beta S(x)\}$, quite a strong assumption. The just cited work in biostatistics would be of potential relevance in spatial regression, an important topic in environmental epidemiology, but the effect of preferential sampling in that domain, especially on the effect on optimal design as seen in this paper, has not been studied as far as we know.

In geostatistics, spatial dependence can often be of central importance especially when spatial prediction is of primary interest. The paper by Diggle et al. (2010) has awakened interest in a topic that has been conveniently ignored even by those charged with setting regulatory standards—where sites may be deliberately sited to detect the non-compliers and placed where response levels are expected to be high (Guttorp and Sampson, 2010).

The present paper shares with Diggle and Ribeiro (2007), Pati et al. (2011) and Gelfand et al. (2012) the goal of determining the effect of preferential sampling on statistical inference, specifically spatial prediction and parameter estimation. The more

*Main article DOI: 10.1214/15-BA944.

†Department of Statistics, University of British Columbia, 2207 Main Mall, Vancouver, BC, Canada V6T 1Z4, jim@stat.ubc.ca

 DOI: 10.1214/15-BA944C

recent work of Shaddick and Zidek (2014) demonstrates fairly conclusively through a case study that administrators, when left to their own devices, will select environmental process monitoring sites preferentially. And the theory in Zidek et al. (2014) shows what to do about it, at least if you are an official statistician reporting annual averages over space of an environmental process field—that theory relies on the Horvitz–Thompson approach albeit with estimated selection probabilities. The effect can be quite dramatic in some years, where the estimated annual averages and numbers of sites out of compliance with regulatory standards is greatly reduced relative to their unadjusted counterparts. Finally, in work about to be submitted, Liu, Shaddick, Zidek, and Cai show that relative risks of respiratory mortality increase once the effect of preferential sampling is accounted for, again as seen in a case study. The last three of the cited works are unlike all the previous papers done in a spatio-temporal context.

The present paper, done in a spatial context, adds to this series of papers by showing yet one more impact of preferential sampling, namely on the all important problem of optimal design. More specifically, the paper shows where to take additional observations based on spatial data already collected at preferentially sampled sites, when the preferential sampling mechanism is known and interest is in a set of future observations to be collected at the new set of monitoring sites. This is done with an optimal design theory based on utilities that measure the benefit of any proposed change to the network, the key element being the inclusion in the posterior distribution of (the informative) existing site locations along with the observations at those sites. Together these two ingredients tell us something about the relationship between site locations X and responses Y.

Although the paper presents a general theory, work is confined to just utilities based on the predictive variance, in keeping with the approach to design developed at least 50 years ago by geostatisticians who added one point at a time to an existing network, placed where predictive variances are large. However, we would argue that this approach is too simplistic. An entropy-based approach in the Gaussian case takes the spatial correlation into consideration so that a site with high posterior predictive variance may not make it into the new network because it can be well predicted from some of its newly added neighbors (Le and Zidek, 2006). That is, in fact, what happens in a real-world application of the entropy-based approach (Ainslie et al., 2009) to the redesign of Vancouver's air quality monitoring network. It would very interesting to see how preferential sampling would affect new site selection in the rainfall example considered in this paper when spatial correlation is included. For recent general discussions of design issues such as these, see Zidek and Zimmerman (2010) and Shaddick and Zidek (2015).

I have some concern about the accuracy of the approximation given in Appendix C, especially in the particularly important case where $\beta \times S$ is not small. It is difficult to intuit how things might change in that case due to use of the approximation. I wonder if instead, the Laplace approximation could be used as it has some theoretical credentials.

The improved results gained by incorporating preferential sampling into their models relies on knowing the form of model for the preferential sampling effect. In reality, that model would not be anything like the right model–selection is complicated by committees using guidelines, proximity to highways and so on. A starting point in

the exploration of this issue might begin by experimentally adding an S^2 term in the intensity function of the point process.

The work cited above makes the point that covariates need to be included to see if they can explain the site selection before seeing if there is a residual role to be played by preferential sampling, something that is not considered in the theory presented in this paper.

Overall this paper is a welcome addition to the emerging theory of preferentially sampled spatial monitoring things, and it leaves open quite a number of new directions to explore. In particular, it offers a theory for those who must select sites preferentially, for example, to monitor sources of air pollution.

References

Ainslie, B., Reuten, C., Steyn, D., Le, N., and Zidek, J. (2009). "Application of an entropy-based Bayesian optimization technique to the redesign of an existing monitoring network for single air pollutants." *Journal of Environmental Management*, 90(8): 2715–2729. 750

Diggle, P., Menezes, R., and Su, T. (2010). "Geostatistical inference under preferential sampling." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2): 191–232. MR2744471. doi: http://dx.doi.org/10.1111/j.1467-9876.2009.00701.x. 749

Diggle, P. and Ribeiro, P. J. (2007). *Model-based geostatistics*. Springer Science & Business Media. MR2293378. 749

Gelfand, A. E., Sahu, S. K., and Holland, D. M. (2012). "On the effect of preferential sampling in spatial prediction." *Environmetrics*, 23(7): 565–578. MR3020075. doi: http://dx.doi.org/10.1002/env.2169. 749

Guttorp, P. and Sampson, P. (2010). "Discussion of Geostatistical inference under preferential sampling by Diggle, P.J., Menezes, R. and Su, T." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2): 191–232. MR2744471. doi: http://dx.doi.org/10.1111/j.1467-9876.2009.00701.x. 749

Horvitz, D. and Thompson, D. (1952). "A generalization of sampling without replacement from a finite universe." *Journal of American Statistical Association*, 47: 663–685. MR0053460. 749

Le, N. and Zidek, J. (2006). *Statistical analysis of environmental space-time processes*. Springer Verlag. MR2223933. 750

Liang, K.-Y. and Zeger, S. L. (1986). "Longitudinal data analysis using generalized linear models." *Biometrika*, 73(1): 13–22. MR0836430. doi: http://dx.doi.org/10.1093/biomet/73.1.13. 749

Pati, D., Reich, B. J., and Dunson, D. B. (2011). "Bayesian geostatistical modelling with informative sampling locations." *Biometrika*, 98(1): 35–48. MR2804208. doi: http://dx.doi.org/10.1093/biomet/asq067. 749

Scott, A. and Wild, C. (2011). "Fitting binary regression models with response-biased samples." *Canadian Journal of Statistics*, 39(3): 519–536. MR2842429. doi: http://dx.doi.org/10.1002/cjs.10114. 749

Shaddick, G. and Zidek, J. (2014). "A case study in preferential sampling: Long term monitoring of air pollution in the UK." *Spatial Statistics*, 9: 51–65. 750

— (2015). *Spatio—temporal methods in environmental epidemiology*. Chapman and Hall/CRC Press. 750

Zidek, J., Shaddick, G., and Taylor, C. (2014). "Reducing estimation bias in adaptively changing monitoring networks with preferential site selection." *The Annals of Applied Statistics*, 3: 1640–1670. MR3271347. doi: http://dx.doi.org/10.1214/14-AOAS745. 750

Zidek, J. V. and Zimmerman, D. L. (2010). "Monitoring network design." *Handbook of Spatial Statistics*, 131–148. MR2730944. doi: http://dx.doi.org/10.1201/9781420072884-c10. 750