Comment on Article by Ferreira and Gamerman*

Michael Chipeta[†‡] and Peter J. Diggle[§]

1 Introduction

This paper is a welcome addition to the growing literature on preferential sampling in a geostatistical setting. Earlier papers cited by the authors have shown that preferential sampling materially affects parameter estimation and prediction. The authors now demonstrate that the same applies to design, or more specifically to the optimal augmentation of an initial set of geostatistical data that has been sampled preferentially. Almost in passing, the paper also sets out an algorithm for Bayesian inference under preferential sampling that is a useful contribution in its own right. Might we look forward to an R package implementation of this?

Our comments fall into two categories: theoretical remarks on what we call *adaptive design*, including an explanation of why this does not necessarily require consideration of preferential sampling issues; practical constraints that may limit the scope for theoretically optimal designs to be used in practice, especially in low-resource settings.

2 Adaptive geostatistical design and preferential sampling

The topic of geostatistical design is multi-faceted. One useful distinction is between adaptive and non-adaptive designs. A *non-adaptive* design is one that is completely determined before any data are collected. An *adaptive* design is one in which an initial design is augmented in a way that depends on the analysis of interim data. We make two theoretical comments that follow from the definition of preferential sampling given in Diggle, Menezes and Su (2010).

Firstly, an adaptive design need not be preferential. To see why, it is sufficient to consider a two-stage adaptive design, $X = (X_0, X_1)$ with associated measurement data (Y_0, Y_1), where subscripts 0 and 1 identify initial and follow-up stages, respectively. Similarly, write $S = (S_0, S_1)$ for the corresponding decomposition of the latent process S. Quite generally, we can factorise the joint distribution of (X, Y, S) as

$$
\begin{aligned}
[X, Y, S] &= [S, X_0, Y_0, X_1, Y_1] \\
&= [S][X_0|S][Y_0|X_0, S][X_1|Y_0, X_0, S][Y_1|X_1, Y_0, X_0, S]. \quad (1)
\end{aligned}
$$

On the right-hand side of (1), if the initial design is non-preferential, $[X_0|S] = [X_0]$, whilst by construction $[X_1|Y_0, X_0, S] = [X_1|Y_0, X_0]$. It then follows that

$$
\begin{aligned}
[X, Y] &= [X_0][X_1|X_0, Y_0] \times \int_S [Y_0|X_0, S][Y_1|X_1, Y_0, X_0, S][S]dS \\
&= [X|Y_0] \times [Y|X]
\end{aligned}
\tag{2}
$$

and the log-likelihood is a sum of two components, $\log[X|Y_0] + \log[Y|X]$. This shows that the conditional likelihood, $[Y|X]$, can legitimately be used for inference although, depending on how $[X|Y_0]$ is specified, to do so may be inefficient. The argument leading to (2) is closely related to the proof that if data are "missing at random" the missingness mechanism can be ignored when using likelihood-based inference (Rubin, 1976), and extends to multi-stage adaptive designs with essentially only notational changes.

Secondly, shared dependence of a design X and the latent process S on observed covariates does not necessarily render X preferential. Specifically, if Z denotes the co-variate process, then $[X, S|Z] = [S|Z][X|S, Z]$. The requirement for the design to be non-preferential is that $[X|S, Z] = [X|Z]$, which in general is a weaker requirement than $[X|S] = [X]$. This illustrates, not for the first time, that spatial statistical inference can be greatly simplified by judicious selection of spatially referenced covariates.

3   Some practical constraints on geostatistical design

The paper makes a number of explicit and implicit assumptions that together provide a very reasonable framework for theoretical analysis, but it is worth bearing in mind that in any particular application, the design problem may be constrained in various ways. These assumptions include the following:

1. *The spatial integral of the predictive variance is an appropriate measure of predictive performance*

   This would not be true if, for example, $S(x)$ represents pollution and the main objective is to monitor compliance with environmental standards; see Fanshawe and Diggle (2013).

2. *Sampling may not be equally costly at every location*

   Put another way, should the design be constrained by the number of locations to be sampled, or by the total sampling effort in the field? An obvious example of this is when travel-time represents a non-negligible proportion of field-effort; see, for example, Figures 2 and 4 of Diggle et al. (2007).

3. *The number of potential sampling points may be finite*

   This applies to disease prevalence surveys when the sampling unit is either a household or a well-defined community. We are currently working on the adaptive design of an ongoing malaria prevalence mapping project around the perimeter of the Majete national park, Malawi, where the first task has been to enumerate and

geo-locate each household in each village within the study-region. In the course of the project, we expect to sample all households, but the order in which they are sampled (in a sequence of monthly field-trips) will be chosen adaptively with the aim of optimising the estimation of the complete spatio-temporal variation in malaria prevalence, which is known to include a strong seasonal component.

None of these these comments are intended to detract from the value of the paper on its own terms. Theoretical studies of this kind help to further our understanding of important, and often subtle, methodological issues around modelling and inference for preferentially sampled geostatistical data.

References

Diggle, P., Thomson, M., Christensen, O., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, H., Boussinesq, M., and Molyneux, D. (2007). "Spatial modelling and prediction of Loa loa risk: decision making under uncertainty." *Annals of Tropical Medicine and Parasitology*, 101: 499–509. 738

Fanshawe, T. and Diggle, P. (2013). "Adaptive sampling design for spatio-temporal prediction." In: Mateu, J. and Müller, W. (eds.), *Spatio-Temporal Design: Advances in Efficient Data Acquisition*, 249–268. Chichester: Wiley. MR3289528. 738

Diggle, P. J., Menezes, R., and Su, T.-L. (2010). "Geostatistical analysis under preferential sampling (with discussion)." *Applied Statistics*, 59: 191–232. MR2744471. doi: http://dx.doi.org/10.1111/j.1467-9876.2009.00701.x. 737

Rubin, D. (1976). "Inference and missing data." *Biometrika*, 63: 581–592. MR0455196. 738