Discussion of the paper by Diggle, Menezes and Su

Richard D Wilkinson, University of Nottingham

October 7, 2009

Preferential sampling occurs when sampling locations X and the process S are dependent. This implies that when choosing X, the survey designer had prior knowledge of S which was used in some manner in the design. For example, the locations in the 1997 sample in the Galicia dataset were chosen to lie in regions where the surveyors believed a priori that there were likely to be large gradients of lead concentrations. This dependence complicates the analysis as we can then no longer factorize the distribution as

$$[S, X, Y] = [Y|S(X)][X][S] \tag{1}$$

My question concerns whether we can bypass the problem of preferential sampling by conditioning on the information that induced the dependency between S and X. Suppose that B is the surveyor's prior belief about S before observing Y and that this information includes the model for how X was chosen. For example, we might hope to elicit the prior expected response surface $\mathbb{E}S(x)$, and perhaps also the variance. This corresponds to specifying a mean and correlation function in *Assumption 1* for the Gaussian process assumed in the paper. Models for how the locations X were chosen, can be decided after discussion with the surveyor. Examples might include choosing X where we expected S to be large (e.g., $X \sim U[x : \mathbb{E}S(x) > \epsilon]$), or where the derivative is large ($X \sim U[x : |\frac{d}{\mathrm{dx}}\mathbb{E}S(x)| > \epsilon]$) etc. The model stated in *Assumption 2* in the paper, namely that X is a Poisson process with rate $\lambda(x) = \exp(\alpha + \beta S(x))$, could never arise in practice as the surveyors do not know S. However, it is feasible that the rate could be $\lambda(x) = \exp(\alpha + \beta\mathbb{E}S(x))$, for example.

For each of these models, S and X are dependent. However, they are conditionally independent given $\mathbb{E}S(x)$. Hence, if B contains the information that induced the dependency between S and X, then $[X \mid S, B] = [X \mid B]$.

Thus, given prior knowledge B, the distributional relationship becomes

$$\begin{aligned}[S, X, Y \mid B] &= [Y \mid S(X), B][X \mid S, B][S \mid B] \\ &= [Y \mid S(X), B][X \mid B][S \mid B] \end{aligned} \tag{2}$$

which returns us to a non-preferential setting (cf. Equation (1)). Typically, $[Y \mid S(X), B] = [Y \mid S(X)]$ if the prior information solely concerns S and X, and the distribution $[X \mid B]$ is the model used by the surveyors to choose X. The final term in Equation (2), $[S \mid B]$, is tractable for many types of prior information if we assume that S is a Gaussian process. In conclusion, by using expert knowledge there may be an alternative approach to dealing with preferential sampling.