THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT #268

Unbiasing estimates from preferentially sampled spatial data

BY

GAVIN SHADDICK & JAMES V ZIDEK

September 2012

Unbiasing estimates from preferentially sampled spatial data*

James V Zidek[1] and Gavin Shaddick[2]
[1]Department of Statistics, University of British Columbia.
[2]Department of Mathematical Sciences, University of Bath

Abstract

This paper explores the topic of preferential sampling, specifically the situation where monitoring sites in environmental networks are preferentially located by the designers. This means the data arising from such networks may not accurately characterize the spatio–temporal field they intend to monitor. Approaches developed to mitigate the effects of preferential sampling have taken in two very different contexts and current methods are reviewed. Building on these approaches, a very general framework for dealing with the effects of preferential sampling in environmental monitoring is proposed. This is followed by a development of strategies for its implementation. A practical example of its implementation is given through a case study in which we consider preferential sampling in the long–term development of an air pollution monitoring network in the UK. We consider changes in locations and periods of operation of black smoke monitoring sites over an extended period (1970-1996). We show how the most appropriate of the strategies for this case can be used to adjust the annual averages to compensate for preferential sampling which generally result in reduction in concentrations over time. Even more dramatic results are seen when adjusting estimates of the number of sites out of compliance.

Keywords: preferential sampling; Horwitz - Thompson estimator; response biased sampling; spatio–temporal fields; spatial point process
AMS 2010 classification: 62P12; 62D99

1 Introduction

This paper is a sequel to the important paper of Diggle et al. [2010] (hereafter D10) and addresses the topic of preferential sampling in the selection of the spatial locations at which to collect samples. The topic of this paper relates to the selection of spatial sampling sites

*The work described in this paper was partially supported by a discovery grant from the Natural Sciences and Engineering Research Council of Canada

for environmental risk management. Such sampling sites have been used by various agencies for a variety of purposes. (See Le and Zidek [2006].) Sometimes they are selected to yield a representative sample of an environmental field. But often they are preferentially selected in a way that depends stochastically on the responses being measured or conditionally on the environmental process parameters, leading to what D10 calls "preferential sampling".

Since the measurements of the random responses at such sites may be critical for analysis, management or administrative purposes, the effects of such sampling may be of concern. For example urban air pollution monitoring sites provide the information that may be used to detect noncompliance with air quality standards (EPA [2005]). Then the designer may locate the sites where air pollution levels are believed to be the highest (although reaching that goal presents its own challenges as shown by Chang et al. [2007]). Reaching this goal would mean the measured concentrations would overestimate the levels of the pollutant in that urban area. That could render these data unsuitable for other purposes, e.g. estimating the relative risk parameter in the health effects model for the pollutant. These considerations point to the need for a model to reflect preferential sampling processes but as well a way of compensating for their effects. Those effects will depend on the inferential objectives for which the data are collected.

The approaches taken in this paper build on two approaches described in Section 2. Neither addresses the issue of concern in this paper, the mitigation of the effects of preferential site selection in monitoring networks that adaptively change over time like the one in the case study .presented in Section 5. The first approach, which has its origins in biostatistics, is based on the idea of response biased sampling surveyed in Scott and Wild [2011] (hereafter S11), which extends the work of Lawless et al. [1999]. The inferential purpose of that approach means that spatial dependence can be ignored, a serious limitation from the perspective of the work described here. In contrast, the second approach does make spatial dependence a fundamental modeling component. It assumes a known parametric form for the preferential sampling process and by treating the selected sites as data, likelihood based methods can be used to estimate both process and selection model parameter, thereby compensating for preferentially siting of the monitors. The second approach unlike the first ignores covariates.

This paper continues the exploration begun in D10 but now in a spatio – temporal context. In addition to providing a sampling framework for preferential sampling, these papers explore ways of compensating for its deleterious effects. Section 3 provides a superpopulation framework for building a unified approach to dealing with those effects. Then Section 4 presents several strategies for implementing that framework, including one that resembles the point process modeling approach of D10, but which can be applied in the domain of interest in this paper. The methods all rely on variations of the Horwitz-Thompson approach for unbiasing estimates based on preferentially sampled data, an idea suggested by Rathburn [2010]. Viewed from an administrative perspective, unbiasedness would be important for official estimates published at any given time. The characteristic ensures that these estimates can be combined to form aggregate estimates with improved accuracy as measured by mean squared error.

Section 5 demonstrates use of our methods in a case study concerning the preferential

adaption over time of the UK's black smoke monitoring network, starting in 1970. We show there, that correcting for bias can substantially reduce estimates of the number of sites, monitored and unmonitored that are out of compliance with regulatory standards. Section 6 discusses our findings and gives some concluding remarks.

2 Background

As noted by Cicchitelli and Montanari [2012] whose work is revisited below, spatial design and inference can be addressed within either the design-based or model-based paradigms. They argue that the former is preferable "if inference focuses on global quantities, such as means or totals..." It was in this context that Horwitz–Thompson estimator (HTE; Horvitz and Thompson [1952]) estimator used in the sequel, was created to construct design - unbiased estimators when finite population elements have unequal probabilities of being selected in a design based analysis.

Cicchitelli and Montanari [2012] suggest the model–based approach for "constructing a map", meaning spatial prediction or interpolation. Within the model-based approach two major approaches have emerged and these are reviewed in this section as they might apply in a spatial sampling context. These approaches are associated with two categories of important inferential objectives.

The first category concerns situations where spatial dependence is central to the investigation, It includes geostatistical modeling, for which an extensive theory has been developed, including variograms and their cousins for representing that dependence in random spatial fields. That dependence determines such things as where monitors of that random field should be sited and predictors of the field at unmonitored sties.

In contrast the second category, concerns the relationship between a spatial response Y and a vector of covariates or explanatory factors X indexed by their spatial locations j. Interest then focuses on estimating or testing the coefficients β the determine the relationship between the Y_j given X_j at sites j through the conditional distribution of the first given the second. Confounders not included in X will mean that conditional on X, the $\{\mathrm{Y}_j,\ j = 1, \ldots, \mathrm{N}\}$ are spatially dependent. However approaches in this category are not concerned with that dependence and assume it is nonexistent with an asymptotic justification described below.

The first category is exemplified by the approach in D10, which analyzes the effect of preferential sampling in a model-based geostatistical framework. The approach models the preferential selection process and treats the selected locations as data. The parameters of both the selection and environmental process parameters can then be estimated within a standard inferential paradigm, for example by maximum likelihood. More precisely D10 assumes a latent, unobservable Gaussian field S over a geographical continuum (domain) $\mathcal{D}$. The sites u are selected at random in accordance with an inhomogeneous Poisson spatial process with intensity function $\lambda(u) = \exp\{\alpha + \beta \mathrm{S}(u)\}, u \in \mathcal{D}$. The measurable response Y is also modeled as dependent on S. Both the vectors of measurements, Y, and sites U provide information about the underlying model parameters, including both those

in the spatial mean as well as spatial covariance matrix for S. The required likelihood for interest is obtained by integrating out the infinite dimensional random field S. Furthermore, through S, U and Y are correlated in their joint spatial distribution.

The model proposed in D10 seems a reasonable if not completely general platform on which to examine the effects of preferential sampling. Key to the approach is knowledge of the preferential sampling process. Emphasis is on variogram estimation and spatial prediction and biases induced by preferential sampling, that result from biases in parameter estimation, (except in one simulation study where non–parametric variogram estimators are briefly considered).

This paper diverges from D10 by assuming a finite population of N sampling (i.e. monitoring) sites, $\mathrm{u}_j,\ j = 1, \ldots, \mathrm{N}$. Note that u_j could represent just the label j or more commonly the geographic coordinates of the site, depending on the context. This is not to say that D10's assumption of a continuous domain for site selection is unreasonable in all cases, for example in sampling ocean sediments by grab sampling. However practical and administrative considerations will often restrict S to be a finite dimensional vector–valued process over a discrete domain $\mathcal{D}$. Even D10 needs to discretize S to approximate the marginal likelihood and they do this by replacing the continuous $\mathcal{D}$ by a fairly dense lattice. Sampling points then have to be mapped onto their nearest lattice point neighbours. So populations of potential sampling sites can often be taken as in this paper to be a finite set of possible locations such as the centroids of lakes in a fresh water survey or secluded sites in urban parks for placing an air quality monitor.

This paper also differs from D10 in allowing Y's covariates to be incorporated. Site selection may well depend them, on such things such as an 'urban - rural' classification or the distance of a site from a major roadway. Furthermore the analysis may well be about the significance of the effects of such covariates or design variables on the measured responses. The need to incorporate such covariates has been reflected in the response–biased sample selection approach, the second category referred to above.

Scott and Wild [2011] (hereafter S11), who extend the work of Lawless et al. [1999], survey the literature on that approach, which has origins in case–control observational studies. There the response Y is observed (a "case" or a "control") and then X is observed on samples from the population of cases and controls. Models here assume a finite population of possible sample items, which in our case would be the sample sites u, and also differ from that of D10 in that the covariate– measured response pairs $(\mathrm{X}_u, \mathrm{Y}_u)$ are assumed to be stochastically independent across sites u. S11 cites earlier work that suggests "there are real difficulties when we move away from a diagonal covariance structure." X may well include the geographical site coordinates as in Cicchitelli and Montanari [2012] thereby adding to the challenge of specifying that dependence structure especially when the random spatial field is non-stationary, a common feature of environmental risk analyses.

The assumption of spatial independence could be justified by the well known fact that asymptotically this assumption is not a serious limitation when consistent estimates of model's first order parameters $\boldsymbol{\beta}$, for example in the regression model for the response, are available. More specifically Rao et al. [1998] point out that if the solution to estimating equation (2.1) below is consistent, then the covariance of that solution can also be consis-

tently estimated by the so-called "sandwich estimator". Such parameters would often be the focus of inferential analysis in spatial statistics, for example in universal Kriging. This approach simplifies the analysis while ensuring robustness against misspecification of the spatial dependence structure.

Two approaches for inference for methods in the second category are suggested in S11. The first uses estimating equations in conjunction with a Horwitz – Thompson (HT) approach. To be more precise, define R to be a sampled site indicator so that R_u is 1 or 0 according as site u is selected into the sample or not. Let

$$\pi_{\mathrm{u}} = \pi(y_{\mathrm{u}}, x_{\mathrm{u}}) = P\{R_{\mathrm{u}} = 1 | y_{\mathrm{u}}, x_{\mathrm{u}}\}$$

the selection probability for site u. The HT approach estimates the process parameter vector β by solving the estimating equations

$$\sum_{u} \frac{\mathrm{R}_{\mathrm{u}}}{\pi_{\mathrm{u}}} \frac{\partial \log\left[y_{\mathrm{u}} | x_{\mathrm{u}}, \beta\right]}{\partial \beta} = 0, \tag{2.1}$$

assuming $\pi_{\mathrm{u}} > 0, u \in \mathcal{D}$ are known at the sampled sites.

The second approach involves a partial likelihood (PL) approach where the profile likelihood is found by maximizing out the non parametric marginal distribution. [X]. The corresponding estimating equation for the PL approach is

$$\sum_{\mathrm{u}} \mathrm{R}_{\mathrm{u}} \frac{\partial \log\left[y_{\mathrm{u}} | x_{\mathrm{u}}, \beta, \mathrm{R}_{\mathrm{u}} = 1\right]}{\partial \beta} = 0, \tag{2.2}$$

which depends on the $\{\pi_{\mathrm{u}}\}$. Working under the asymptotic paradigm, the requirement that the selection probabilities be known can dropped in favour of their being consistently estimable.

Spatial statisticians might see the failure of the second approach unlike the first, to reflect spatial pattern as a serious limitation. Moreover spatial location can be a useful explanatory factor. For these reasons, Cicchitelli and Montanari [2012], taking a design–based approach, augment X by constructing quasi–covariates in a matrix Z. More specifically for location u_j, they define $\mathrm{z}_k(\mathrm{u}_j),\ k = 1, \ldots, K$ in terms of $(||\mathrm{u}_j - \kappa_k||)^2 \log(||\mathrm{u}_j - \kappa_k||)$ where the $\{\kappa_k\}$ represent spatial nodes selected by the modeller. With a model assisted approach they assume

$$E_\xi(Y(\mathrm{u}_j) = \beta_0 + \beta_1 x_{j1} + \cdots + \gamma_1 \mathrm{z}_1(\mathrm{u}_j) + \cdots + \gamma_K \mathrm{z}_K(\mathrm{u}_j).$$

The design–based approach is then invoked and HT estimators constructed to fit a regression model of Y on (X,Z). In line with the second approach, the authors assume no spatial dependence in the model assisting their approach. They argue that spatial pattern is provided by the mean function with the augmented covariates. This position can be supported by the well–known duality between first order and second order modeling in geostatistics. Misspecifying the mean function will always lead to bias in the variogram, a second order

feature of the model. Although the paradigm of Cicchitelli and Montanari [2012] is design–based, the idea could really be extended to the model–based context. Moreover instead of splines other bases could well be used to capture spatial pattern. Simpson et al. [2011] also provide some support for that assumption in their paper about computational feasibility on spatial statistics with a title that includes the phrase "we need to forget about the covariance function. " (In reality that paper does rely on a spatial covariance but its role is diminished.)

In spatial sampling, the assumption that the responses Y on which the π_u depend are known seems implausible, unlike say the case of case–controlled studies. More realistically we might assume that these sampling probabilities are uncertain and within a Bayesian context that they depend implicitly on these responses. If that dependence were fixed over time, the probabilities would be learned over time as responses accumulated. In other words they could be consistently estimated.

Although none of the approaches above explicitly refers to data collected over time, they could undoubtedly be adapted to cover such data by allowing the $\{Y_u\}$ to be vectors. Note that extending that in S10 would require an appropriate extension of the point process model since the temporal sampling times would, unlike spatial sampling points, not usually be according to a temporal point process. Thus they provide foundations upon which to build preferential sampling models and ways of adjusting for the effects of such sampling.

But none of these approaches embrace the key feature of spatial sampling addressed in this paper, that of adaption over time. Here the number of sampling sites changes over time (Ainslie et al. [2009]). The case study in Section 5 involves British black smoke (BS) air pollution for which the number of monitoring sites declined from nearly 2000 in 1965 to around 200 at the turn of the last century. So did the overall levels of BS due to improvements in the management of air quality. Shaddick and Zidek [2012] demonstrate conclusively that this decline was made by preferentially removing sampling sites with generally lower concentrations, Thus relative to the decline in overall levels of BS, the data increasingly overestimate BS concentrations. At the other extreme,Le and Zidek [2006] give an example of an network for monitoring ground level ozone concentrations that has been steadily augmented over the last few decades, as recognition of ozone's adverse health effects has been recognized. Here it seems plausible that the addition of sites has been done preferentially to ensure high levels of ozone are detected.

3 A general framework

Section 2 presents a number of paradigms in which to study the issue of preferential spatial sampling within the design– or model–based frameworks. The latter includes the Bayesian approach although it was not explicitly mentioned in that section.

The concept of a superpopulation provides a framework for unifying paradigms for inference and its the one we develop in this section. We suppose a discrete geographical domain $\mathcal{D}$ contains point referenced sites $u_j, j = 1, \ldots, N$. Let T denote the present time and although the spatial locations do not have a natural ordering it is convenient to use

the vector notation $\mathbf{Y}_t : 1 \times \mathrm{N}$ to represent the sequence of responses at those sites at times $t = 1, \ldots, \mathrm{T}$. These sites, which need not be on a lattice, represent a finite population of potential locations at which to site monitors that repeatedly measure at regular times, a random space–time field. The sites could for example lie along a river course. Further let $\mathbf{Y}$ denote the $\mathrm{T} \times \mathrm{N}$ matrix comprised of those row response vectors.

Similarly at time t, let $\mathbf{X}_t$ denote a matrix of covariates or explanatory factors, which we hereafter just call 'covariates' for simplicity. Then $\mathbf{X}$ denotes the corresponding three dimensional covariate array. Following the theory described in S11, we will first develop our theory conditional on $\mathbf{X}$ and base inference on the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ and model parameters $\boldsymbol{\theta}$. Using D10's bracket notation to reduce notational clutter, we denote the conditional distribution of $\mathbf{Y}$, which may be characterized by its probability density, cumulative distribution function and so on, by

$$[\mathbf{y}|\mathbf{x}\ , \boldsymbol{\theta}]. \tag{3.3}$$

If within a specified time period, responses and covariates were observed for every spatial site in our finite population of sites to get a set W, we could proceed in the usual way to make inferences about $\boldsymbol{\theta}$ (and associated constructs). In particular, given $\mathbf{Y} = \mathbf{y}$ and $\mathbf{X} = \mathbf{x}$ the conditional likelihood function would be given by Equation (3.3). The superpopulation's maximum likelihood estimator (MLE) of θ, denoted by $\hat{\theta} = \theta(\mathbf{y}, \mathbf{x})$ would estimate $\boldsymbol{\theta}$, including temporal as well as spatial correlation model parameters along with coefficients in the regression model relating Y and X. If only a random sample of sites were selected from the finite population of sites, and their associated associated response–covariate pairs were observed to yield a set w, then an estimator of $\hat{\boldsymbol{\theta}}$ could be computed.

Alternatively in contexts where official statistics are collected and published or regulatory policy is administered, the target might well be the infinite conceptual superpopulation of Y–X pairs $\mathcal{W}$, from which W and then w have been drawn. Then $\boldsymbol{\theta}$ would be the vector of numerical characteristics of $\mathcal{W}$. Following a standard approach in survey sampling theory (Sarndäl et al. [2003]), we would deem the population parameters $\hat{\theta}$ to be of interest as a representative of θ. (For a discussion of this issue see Pfeffermann [1993].) Thus in summary, two legitimate objects of inferential inference present themselves, θ and $\hat{\theta}$. In either case, we like D10, concern ourselves with the effects of preferential sampling on the estimates derived from the sample θ or $\hat{\theta}$, depending on our perspective. But in either case, we rely on the superpopulation model indexed by the former, which defines the latter, to develop approaches to assess those effects.

With the general notation, $a_{r:s} = (a_r, \ldots, a_s)$ for $r \leq s$ and the null vector if $r > s$ for any object a, and we may express the superpopulation log-likelihood estimating equation for the MLE $\hat{\theta}$ as

$$\sum_{t=1}^{T} \nabla_\theta \ln \left[(\mathbf{y}_t|\mathbf{y}_{1:(t-1)}, \mathbf{x}_{1:t}, \boldsymbol{\theta}\right] = 0. \tag{3.4}$$

As in Section 2, we let $\mathbf{R}$ denote the $\mathrm{T} \times \mathrm{N}$ matrix of indicator random variables where the t^{th}– row of $\mathbf{R}$ consists entirely of zeros except for ones in the columns corresponding to the sites selected for inclusion at time t. Let $\mathbf{Y}(\mathbf{r})$ and $\mathbf{X}(\mathbf{r})$ denote the observed values

of $\mathbf{Y}$ and $\mathbf{X}$ at the design points selected adaptively over time. In other words if $\mathbf{r} = (r_{tj})$, then

$$\mathbf{Y}[\mathbf{r}] = \{Y(u_{tj}) : t, j \text{ for which } r_{it} = 1\}.$$

and so on. Note that in the preferential context of adaptive designs the distribution of $\mathbf{R}$ may stochastically depend on $\mathbf{Y}$, $\mathbf{X}$ and $\boldsymbol{\theta}$. Thus $\pi_{ut} = P(R_{ut} = 1|\mathbf{Y},\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\eta})$, for some parameter matrix $\boldsymbol{\eta}$. Assume the design at time t, R_t, depends only on previously observed values of $\mathbf{Y}$ and $\mathbf{X}$. Under that assumption, we can represent the conditional preferential sampling distribution of $\mathbf{R}$ by

$$[\mathbf{r}|\ \mathbf{y},\ \mathbf{x}, \boldsymbol{\theta}] = \Pi_{t=1}^{T}[\mathbf{r}_t \mid \mathbf{y}(\mathbf{r}_{1:(t-1)}), \mathbf{x}(\mathbf{r}_{1:(t-1)}), \mathbf{r}_{1:(t-1)}, \boldsymbol{\eta}, \boldsymbol{\theta}].$$

Given the data up to time T, how can we compensate for the effects of preferential sampling in estimating either the infinite or finite population parameter vectors, whichever was our objective? Section 2 suggests two approaches to this problem, one like D10 where we model spatial dependence and the other like S11 where we ignore it. Given $\mathbf{X} = \mathbf{x}$, the former approach would lead to the complete likelihood given by

$$L(\boldsymbol{\eta}, \boldsymbol{\theta}) \doteq \Pi_{t=1}^{T}[\mathbf{y}_t(\mathbf{r}_t) \mid \mathbf{y}(\mathbf{r}_{1:(t-1)}), \mathbf{x}(\mathbf{r}_{1:t}), \boldsymbol{\theta}) \times [\mathbf{r}_t \mid \mathbf{y}(\mathbf{r}_{1:(t-1)})], \mathbf{x}(\mathbf{r}_{1:(t-1)}), \mathbf{r}_{1:(t-1)}, \boldsymbol{\eta}, \boldsymbol{\theta}].$$

This likelihood includes the responses and the preferentially sampled spatial sites indices, both of which may provide information about the parameter matrix $\boldsymbol{\theta}$. The sample MLE would in turn provide estimates of the superpopulation MLE, adjusted for sampling preferentially. Like D10, we could carry this one step further and compute the marginal distribution of Y and R after marginalizing out $\boldsymbol{\theta}$ with respect to a prior distribution.

Where the focus of interest is the superpopulation parameters, a Bayesian approach could be used. It would rely on the posterior distribution

$$\pi(\boldsymbol{\eta}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{r}) \propto L(\boldsymbol{\eta}, \boldsymbol{\theta})\pi(\boldsymbol{\eta}, \boldsymbol{\theta}). \tag{3.5}$$

This approach embraces that of S10 since $\boldsymbol{\theta}$ could well include the conceptual device of a latent field that underlies both the measurements as well as the preferential selection process. If $\pi(\boldsymbol{\eta}, \boldsymbol{\theta}) = \pi(\boldsymbol{\eta}, \boldsymbol{\theta}|\nu)$ relies on a hyperparameter vector ν, ν could well be estimated by Type II maximum likelihood to reduce the computational burden that a fully Bayesian approach might entail.

The alternative approaches to the one above, presented in Section 2 ignore spatial correlation and possibly incorporate spatial patterns in an augmented version of $\mathbf{X}$. Suppose $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Psi})$, $\boldsymbol{\Psi} = \boldsymbol{\Psi}_0^{\text{known}}$ representing the case of no spatial correlation. From Equation 3.4, we now get the superpopulation maximum likelihood estimating equation:

$$\begin{aligned} & \sum_{t=1}^{T} \nabla_\beta \ln\left[(\mathbf{y}_t|\mathbf{y}_{1:(t-1)}, \mathbf{x}_{1:t}, \boldsymbol{\beta}, \boldsymbol{\Psi}_0\right] \\ = & \sum_{t=1}^{T}\sum_{j=1}^{N} \nabla_\beta \ln\left[y(u_{tj})|y(u_{1:(t-1),j}), x(u_{1:t,j}), \boldsymbol{\beta}, \boldsymbol{\Psi}_0\right] \\ = & \ 0. \end{aligned} \tag{3.6}$$

Following S11, we now have two approaches for finding calculating the sample–based MLE. The first is provided by the HT approach which yields the following estimating equation

$$\begin{aligned} & \sum_{t=1}^{T}\sum_{j=1}^{N}\frac{r_{tj}}{\pi_{tj}}\nabla_{\beta}\ln\left[y(u_{tj})|y(u_{1:(t-1),j}),x(u_{1:t,j}),\boldsymbol{\beta},\boldsymbol{\Psi}_0\right] \\ = \; & 0. \end{aligned}$$

The conditional maximum likelihood approach for estimating β would rely on the following estimating equation.

$$\begin{aligned} & \sum_{t=1}^{T}\sum_{j=1}^{N}\nabla_{\beta}\ln\left[y(u_{tj})|y(u_{1:(t-1),j}),x(u_{1:t,j}),\boldsymbol{\beta},\boldsymbol{\Psi}_0,R_{tj}=1\right] \\ = \; & 0. \end{aligned} \tag{3.7}$$

In some cases such as that in Shaddick and Zidek [2012] it is reasonable to assume temporal independence of the responses. In that case Equation 3.8 simplifies to

$$\begin{aligned} & \sum_{t=1}^{T}\sum_{j=1}^{N}\frac{r_{tj}}{\pi_{tj}}\nabla_{\beta}\ln\left[y(u_{tj})|x(u_{tj}),\boldsymbol{\beta},\boldsymbol{\Psi}_0\right] \\ = \; & 0. \end{aligned} \tag{3.8}$$

In this way the population parameters are defined.

The maximum likelihood estimating equation is generalized in the estimating estimation described by Godambe and Thompson [1986] which in the case of temporal independence above is for the superpopulation case

$$\sum_{t=1}^{T}\sum_{j=1}^{N}\boldsymbol{\phi}_{tj}(y(u_{tj}),x(u_{tj}),\beta_t) = \; \mathbf{0}. \tag{3.9}$$

Note that under regularity conditions, that gradient has conditional expectation given the superpopulation parameters, equal to zero, a property referred to an "unbiasedness". In fact, Equation (3.9) can be used to define an estimator for any choice of kernel ϕ provided it is unbiased. Thus for example, Binder and Zdenek [1994] formulate a general approach for complex sample surveys based on estimating equations.

Example 1. Suppose for all j and t

$$Y_{tj} = \mu_t + \epsilon_{tj}$$

where the coefficients are fixed but unknown constants while the residual vector is temporally and spatially uncorrelated. In other words,

$$\epsilon_t : N \times 1 \sim N_N(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Then $\theta = (\mu, \sigma^2)$ and the parameters of the population at time t are $\hat{\mu}_t = \bar{Y}_{t\cdot}$ and $\hat{\sigma}^2 = N^{-1} \sum_j (Y_{tj} - \bar{Y}_{t\cdot})^2$. We get these by solving the estimating equations:

$$\sum_j^N \begin{pmatrix} Y_{tj} - \mu_t \\ (Y_{tj} - \mu_t)^2 - \sigma^2 \end{pmatrix} = 0.$$

Applying the HT approach as suggested by Binder and Zdenek [1994], we get a design unbiased version of the two superpopulation estimates. If the superpopulation model were log-normal, the approach above would yield the geometric mean instead of the arithmetic mean with preferential sampling weights entering into the exponents of the factors in that mean.

4 Implementation

Methods are required to implement the theories described in previous sections. Choices will necessarily depend on the context so no single approach can be developed to cover them all. However, the superpopulation model and the finite population parameters it generates as an MLE ($\hat{\theta}$) are the starting points.

4.1 A general result.

A general class of methods have a parameter vector of the form $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T)'$, vector–valued functions $\mathbf{H}$ and $\mathbf{h}_1, \ldots, \mathbf{h}_q$ along with superpopulation estimates defined by solving Equation 3.4, which can be represented in terms of the site specific responses by

$$\hat{\theta}_t = \mathbf{H} \left\{ N^{-1} \sum_j \left(\mathbf{h}_1[y_t(u_j), x_t(u_j)], \ldots, \mathbf{h}_q[\mathbf{y}(u_j), \mathbf{x}(u_j)] \right) \right\}. \tag{4.10}$$

If the $\{\pi_{tj}\}$ are known or well-estimable in such cases, $\hat{\theta}$ can be estimated by

$$\hat{\hat{\theta}}_t = \mathbf{H} \left\{ \sum_j \frac{r_{tj}}{N \pi_{tj}} \left(\mathbf{h}_1[\mathbf{y}(u_j), \mathbf{x}(u_j)], \ldots, \mathbf{h}_q[\mathbf{y}(u_j), \mathbf{x}(u_j)] \right) \right\}. \tag{4.11}$$

Justification for this choice comes from the the unbiasedness of $\{\sum_j \frac{r_{tj}}{N \pi_{tj}} \mathbf{h}_l[\mathbf{y}(u_j), \mathbf{x}(u_j)]\}$ as estimates of their corresponding population averages.

Regression of Y on X is an important example where for T dimensional vectors $\mathbf{a}$ and $\mathbf{b}$:

$$H(\mathbf{a}, \mathbf{b}\ ) = \begin{pmatrix} a_1/b_1 \\ \\ a_T/b_T \end{pmatrix},$$

$h_1[y_t(u_j), x_t(u_j)] = y_t(u_j)x_t(u_j)$ and $h_2[y_t(u_j), x_t(u_j)] = x_t(u_j)^2$. This model would allow the relationship between Y and X (e.g. an explanatory factor) to evolve over time.

4.2 Stochastic approaches.

In our context unlike that of survey sampling, site selection probabilities are generally unknown and we now turn to approaches for selecting them. Assume for now that the sample size n_t is fixed for all times t so that $\pi_{t\cdot} = n_t$ or with $\xi = \pi_{tj}/n_t$, that

$$\sum_j \xi_{tj} = 1.$$

Two general approaches now present themselves. The first assumes a purely stochastic relationship between the ξ_{tj} and the $\{\mathbf{y}(\mathbf{r}_{1:(t-1)}), \mathbf{x}(\mathbf{r}_{1:(t-1)}), \mathbf{r}_{1:(t-1)}, \boldsymbol{\theta}\}$. A convenient choice for the required conditional distribution when $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T)'$ has a density proportional to

$$\Pi_j(\xi_{tj})^{\zeta_{tj}-1}$$

where ζ_{tj} depends on $\boldsymbol{\theta}_t$. Although this parameter vector is not known, we assume the experts selecting the design are guided by prior knowledge of its elements. This conditional distribution then gives the marginal result

$$E[\xi_{tj}^{-1}|\theta_t] = (\zeta_{t\cdot}(\theta_t) - 1)/(\zeta_{tj}(\theta_t) - 1)$$

or in other words.

$$E[\pi_{tj}^{-1}|\theta_t] = (\zeta_{t\cdot}(\theta_t) - 1)/(\zeta_{tj}(\theta_t) - 1)n_t.$$

Equation 4.11 then gives the unbiased estimator

$$\hat{\hat{\theta}}_t^*(s_t) = \sum_j \frac{\zeta_{t\cdot}(\theta_t) - 1}{\zeta_{tj}(\theta_t) - 1} \frac{n_t r_{jt}}{N} \mathbf{h}[\mathbf{y}(u_j), \mathbf{x}(u_j)]. \tag{4.12}$$

where $\mathbf{h}[\mathbf{y}(u_j), \mathbf{x}(u_j)] = \mathbf{h}_1[\mathbf{y}(u_j), \mathbf{x}(u_j)], \ldots, \mathbf{h}_q[\mathbf{y}(u_j), \mathbf{x}(u_j)]$.

Estimates of θ_t can then be plugged into this last expression. Alternatively we can let them be learned through a Bayesian updating algorithm if we adopt the model in Equation 3.5.

Example 2: Suppose $Y_{tj} = \beta_{tj} + \beta t + \epsilon_{tj}$. Then $\zeta_{tj}(\theta_t) \propto \beta_{tj}$, $j = 1, \ldots, N$ might be deemed a reasonable choice in contexts like that addressed in D10.

An alternative stochastic approach to characterizing the $\{\pi_{tj}\}$, one that incorporates the problem of estimating the $\{\theta_t\}$, supposes as a purely conceptual rather than literal device, N types of particles, each type arriving during time interval t at its counter in a Poisson process. To describe that process, we first recall that $\mathbf{Y}_t(r_t)$ denotes the measured responses at time t. Next we let

$$\boldsymbol{\Lambda}_t = \{r_{1:t-1}, \mathbf{Y}_1(r_1), \ldots, \mathbf{Y}_1(r_{(t-1)}), \boldsymbol{\theta}_t, \boldsymbol{\eta}_{tj}\}.$$

Now endow the Poisson processes with intensity functions

$$\lambda_{tj} = \lambda_{tj}(\mathbf{\Lambda}_t), \tag{4.13}$$

for particles of type $j = 1, \ldots, N$. The counters are parallelizable, so when a particle of type j arrives at its counter, the counter counts 1 and no further particles after that during the specified time period. The elements of S_t are the labels of the counters that have registered a count during the time interval. In Equation (4.13), the $\{\boldsymbol{\eta}_{tj}\}$ are site specific random effects that can be correlated over space. Note that the intensity depends on the indices of the sites previously selected for monitoring as well as the values measured at those sites. Thus the intensities can be selected to express the goal of retaining previously selected sites and just adding new ones, an increasing staircase of monitors. Alternatively, as in the case study in the next section, they may be made to remove sites according to what was seen at monitoring sites in the past, a descending staircase.

The conditional expected size of S_t as

$$\begin{aligned} E[R_{t\cdot}|\mathbf{\Lambda}_t] &= \sum_j P(R_{tj} = 1|\mathbf{\Lambda}_t) \\ &= \sum_j [1 - \exp(-\lambda_{tj})] \\ &\equiv n_t(\mathbf{\Lambda}_t). \end{aligned}$$

Furthermore, the conditional probability mass function for R_t is

$$\begin{aligned} P(\mathbf{R}_t = \mathbf{r}_t|\mathbf{\Lambda}_t) &= \Pi_j \exp(-\lambda_{tj})^{1-r_{tj}}[1 - \exp(-\lambda_{tj})]^{r_{tj}} \\ &\simeq \Pi_j \lambda_{tj}^{r_{jt}} \exp(-\lambda_{tj}) \\ &= (\lambda_{t\cdot})^{n_t} \exp(-\lambda_{t\cdot})\Pi_j \kappa_{tj}^{r_{tj}} \end{aligned}$$

where

$$\kappa_{tj} = \frac{\lambda_{tj}}{\lambda_{t\cdot}}.$$

To get the component of the likelihood generated by the selection of the sampling sizes, we need to compute the conditional distribution of S_t given just the event $\mathbf{\Lambda}_t^0 = \{\mathbf{r}_{1:(t-1)}, \mathbf{Y}(r_{1:(t-1)}), \boldsymbol{\theta}\}$. The result is

$$P(\mathbf{R}_t = \mathbf{r}_t|\mathbf{\Lambda}_t^0) = E[\Pi_j \exp(-\lambda_{tj})^{1-r_{tj}}[1 - \exp(-\lambda_{tj})]^{r_{tj}}|\mathbf{\Lambda}_t^0].$$

The contribution to the overall likelihood for time period t is therefore

$$f(\mathbf{y}_t(\mathbf{r}_t)|\mathbf{x}_t(\mathbf{r}_t), \boldsymbol{\theta}_t)P(\mathbf{R}_t = \mathbf{r}_t|\mathbf{\Lambda}_t^0) \tag{4.14}$$

and it would be maximized to find an estimate of $\boldsymbol{\theta}_t$ including the non random effects in the intensity function.

4.3 Regression.

Given the requisite temporal data, regression provides a natural approach for estimating the $\{\pi_{tj}\}$. Here we relax the requirement of specified sample sizes and allow these to be random. More specifically, at time t each sight j is included or not according as a biased coin toss yields heads or not. Its probability of heads π_{tj}, depends through a regression relationship, on all the data obtained prior to time t and the sites that were selected in previous years. Fitting the model within a Bayes or non-Bayesian paradigm can be done in the conventional way. The challenge here is in the model selection, in other words, finding the appropriate predictors from the class of all possible metrics that could be computed from previous exposure data. A formal model selection approach will generally be impractical, necessitating reliance on some context specific knowledge to help reduce the class of possibilities. For example, the case study described in the next section involves about 700 initial sites taken as the finite population, from which successively smaller numbers of sites are selected over the years that follow. The large number of spatio-temporal aggregates that could be chosen as possible predictors would force some preliminary reduction in the number of possible aggregates. One such choice is made below in Section 5 for illustrative purposes. The implementation of this approach will rely on Subsection 4.1, notably Equation 4.11.

5 Case study: black smoke over the United Kingdom

The companion paper [Shaddick and Zidek, 2012] clearly shows that sites for monitoring ground level black smoke concentrations in the UK were preferentially removed from 1970 onwards.I this case study we aim for the paper's stated goal of adjusting population level estimates of black smoke levels to make them unbiased. In particular we develop a way of adjusting estimates of overall average annual concentration levels as well as number of sites out of compliance, both those currently monitored as well as those that have preferentially removed over time.

We begin with a summary of the more detailed description given in Shaddick and Zidek [2012] of the monitoring program for black smoke (BS) in Great Britain. There air pollution has been a concern for many centuries and the association between air pollution exposure and mortality or morbidity has been a public health issue for over 700 years. However the subject has only be- come a global health issue in the last fifty years (Firket [1936]; Ciocco and Thompson [1961]; Ministry of Public Health 1954). As a result attempts have been made to measure air pollution concentrations in a regular and systematic way. However, the different directions chosen in making these attempts due to the different jurisdictions in which they were made, led to variations in the pollutants measured, how they are measured, where they are measured and how the results are reported. Moreover over time many of the networks have changed as emphasis changed on pollutants and areas. During much of the twentieth century, concern has focussed on soot (or black smoke) and sulphur dioxide from industry and domestic fires and measurement methods have been used in different areas.

Early air pollution control legislations [Garner and Crow, 1969, Stern et al., 1973] and more recently air quality standards have been related to a list of criteria pollutants (particulate matter (PM); Ozone (O_3); Sulphur dioxide (SO_2); Carbon dioxide (CO); Nitrogen dioxide (NO_2)). Science has supported environmental policy making, in particular that which shows air pollution is harmful to human health and provides quantitative estimates for environmental health risk impact analysis. In 1961 the worlds first co-ordinated national air pollution monitoring network was established in the UK using black smoke and sulphur dioxide monitoring sites at around 1200 sites. The earliest epidemiological studies found that exposures to very high concentrations of particulate matter had adverse effects on human health (Firket [1936]; of Health [1954]). Many focussed on London due to its large population and early availability of data on health outcomes and pollution levels.

The data on annual concentrations of black smoke used in this case study were obtained from the UK National Air Quality Information Archive. The data were collected by a monitoring network established in the early 1960s. By 1971 it included over 2000 sites. As levels of black smoke pollution have declined, the network has been progressively rationalized, reduced, moved, replaced. It currently comprises about 220 sites. Almost certainly these changes were not made purely at random over the monitoring sites locations. Most plausibly those in the most polluted areas had the greatest chance of retention, leading to the possibility that they were in areas of high mortality, and informative sampling. Daily average black smoke has been shown to be a reasonable predictor of PM_{10}. In general, PM_{10} concentrations are usually higher than black smoke except during high episodes, and hence, if smoke exceeds the PM_{10} limit, it is likely that PM_{10} has also done so [Muir and Laxen, 1995]. Black smoke (BS) is one of a number of measures of particulate matter, other examples including the coefficient of haze (CoH), total suspended particulates (TSP), as well as direct measurements of PM_{10}, and $PM_{2.5}$.Each has been associated with adverse health outcomes (for PM10, Samet et al. [2000]; for PM2.5, Goldberg et al. [2001]; for TSP, Lee and Hirose [2010]; for black smoke, Verhoeff et al. [1996]; for CoH, Gwynn et al. [2000]). Attempts have been made to standardize the measures of pollution by converting the measurements into 'equivalent' amounts of PM_{10}, for example $PM_{10} \approx 0.55$ TSP, $PM_{10} \approx$ CoH/0.55, $PM_{10} \approx$ BS and $PM_{10} \approx$ PM2.5/0.6 (Dockery and Pope [1994]).

Our case study is based on the black smoke monitoring network located at a resolution of 10 metres and measuring annual average concentrations of BS (μ g/m3) operating between April 1966 and March 1994 (inclusive). The data come from the Great Britain air quality archive (www.airquality.co.uk). Air pollution statistics for each site are reported in 'pollution years' (April-March) and include mean concentration, standard deviation and number of valid reporting days. Data were obtained for all sites operating within the study period, a total of 3016 sites, and the data filtered to remove all sites which were located in non-residential areas, which fell outside the study area (defined as Great Britain, meaning that Northern Ireland sites were excluded) or for which air pollution data were lacking. Data were obtained for all 2382 sites operating within the study period and sites excluded if located in areas that were not wholly or partly residential.

To unbias estimates of the annual population levels of black smoke in Britain we use the implementation strategy description in Subsection 4.3. For this case study, we consider

the measurements from the network over the period from 1970 to 1996. We use the 624 sites that were operational in 1970 and which had at least 5 measurements in the following 25 years, and these sites define the finite population underlying development of our theory above. In other words, we will think of concentrations measured at these sites as characterizing the BS field over the UK. For each year, t, data are available from n_t sites, $t = 1, ..., 26$. Measurements, Z_{it}, are the log of the annual means of the 24 hour mean concentrations of BS divided by a normalization constant to make them unitless as they must be if their logarithm is to be applied. Over the study period, the number of sites was reduced from $n_1 = 624$ to $n_{26} = 193$ with the yearly means over all sites, $\sum_{i=1}^{n_i} Z_{it}/n_i$, falling from 60.5 to 9.3 μ g/m3 over the same period. A log transformation is appropriate for modelling pollution concentrations, because in addition to the desirable properties of right-skew and non-negativity, there is justification in terms of the physical explanation of atmospheric chemistry [Ott, 1990] and from hereon, $Y_{it} = \log(Z_{it})$ is used.

A logistic model is used to estimate the probability of a monitoring location being present in any particular year. The probability, π_u, of a site u remaining in the network between year $t-1$ and t is modelled as a function of the average concentration at that of inclusion in year t are modelled as a function of the mean concentration at that site over the previous $t-1$ years after adjusting for the trend in concentrations over time by subtracting the yearly means, $\bar{Y}_{.t} = \sum_{i=1}^{n_t} Y_{it}/n_t$.

$$\log(\frac{\pi_j}{1-\pi_j}) = \alpha + \beta(Y_{jt} - \bar{Y}_{.t}) \tag{5.15}$$

Figure 1 (a) shows the results of fitting a series of logistic regression models for $t = 2, ..., 26$. In the period 1971-1981, the values of the slope are close to zero and non-significant indicating that the previous levels of BS at a site had very little effect on its remaining present in the next year. There is a clear decline in the values of the intercept parameters indicating that the overall probability of sites remaining in the network year-on-year is decreasing, reflecting the reduction in sites over this period from 623 in 1971 to 277 in 1981. There was a dramatic reduction, of almost 50%, in the network from 1981 to 1982 with a fall in the number of sites to 140. From 1982 onwards the decrease in the intercept parameters continues, reflecting the continuing decrease in the number of sites. After this point, the effect of the average concentrations over time at a site, represented by the slope parameters, has a much more important effect. This suggests that was preferential sampling in the choice of which sites remained in the network with sites recording high concentrations more likely to stay in the network.

Two characteristics associated with the responses over the years for 1970 and after at all 624 sites seem of natural interest. The first would be the annual averages across these 624 sites as these could be published to show the effect of regulatory policy over time. However, the second characteristic seems of greater operational importance, it being the number of the 624 sites in non–attainment, that is which do not comply with of air quality standards.

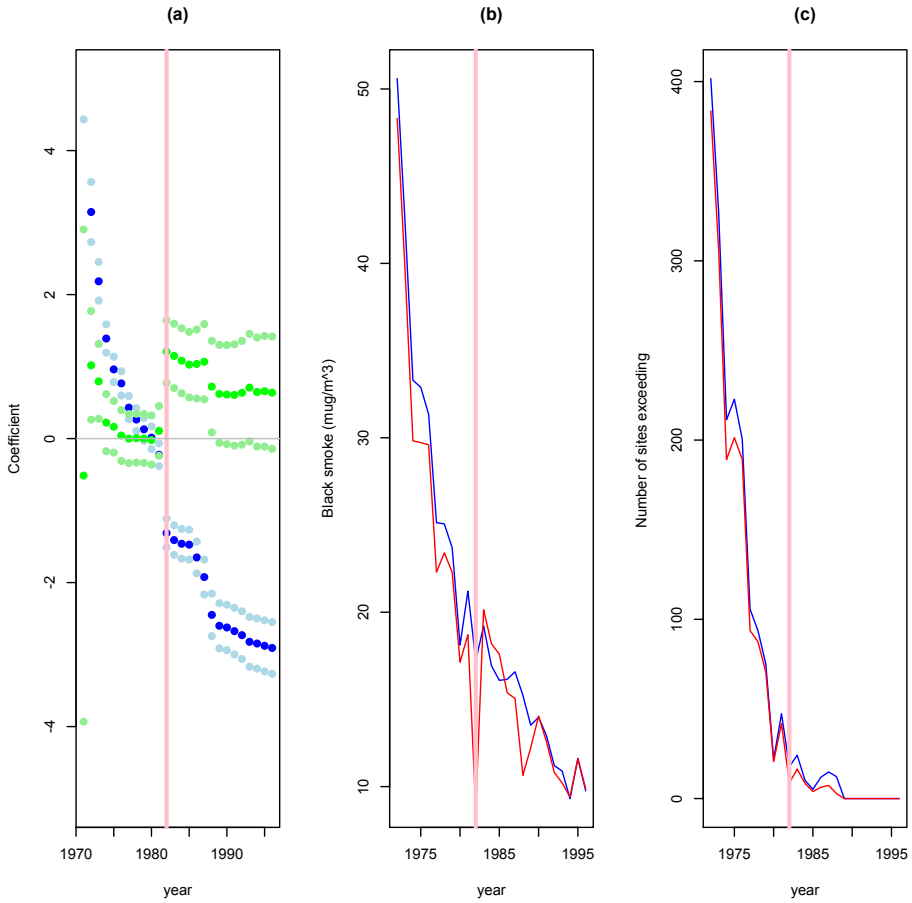

Figure 1: Results from using a logistic regression model to estimate sampling probabilities. Panel (a) shows the coefficients from the logistic model fitted for each year; dark green and dark blue dots denote the intercept and slope coefficients respectively with corresponding light colours showing the limits of 95% confidence intervals. Panel (b) shows the annual means over all sites (blue line) together with the corrected values (red line). Panel (c) shows the number of sites exceeding 34 μg/m-3 (blue line) together with the corrected values (red line). In all three panels, the vertical pink line indicates 1982, the year when the network was dramatically reduced in size.

This number is a surrogate for the cost of mitigation for putting the BS concentrations into compliance. Thus for example, as part of the analysis of the impact of the various ozone standards considered by the EPA's CASAC Ozone Committee in 2006, the EPA Staff predicted the fraction of monitored counties in the United States that would be out of compliance. For the standards that were finally proposed by the Committee, that percent was found to be 86%. Although the US Clean Air Act (epa.gov/oar/caa/title1.html) of 1970 under whose mandate the CASAC was created, rules out economic impact in consideration of standards designed to protect public health, nevertheless policy making cannot ignore the cost of attainment that can be substantial.

Figure 1 (b) shows the yearly averages over time (blue line) together with the corrected versions using Equation 4.11 (red line). It clearly shows the adjustment reduces the estimates of the average levels and the effect of the marked change in the network in 1981 which changed the effects of the logistic regression model, resulting in an over-correction for this year. In practice this could easily be corrected by excluding data from the most recent year from the logistic model. However, this anomaly does serve a useful diagnostic purpose and in general, such marked changes could point to the need for further investigation of underlying changes in the network. Figure 1 (c) shows the number of sites each year that exceed the 1980 EU guide value of 34 μgm/3 [European Commision, 1980]. The blue line is the number of exceeding sites based on the recorded data with the red line the numbers after adjustment for the preferential sampling. Table 1 shows the number of sites exceeding the EU limit of 68 μgm/3 and the guide values of 51 and 34 μgm/3 where a reduction of the number of exceedances can be seen when adjusting for the preferential sampling with the effect being more marked as might be expected for lower values of the cut off point.

6 Discussion and conclusions

A number of methods have been suggested in this paper for modeling the probability of selection in preferential sampling within the general framework we have developed using a superpopulation modeling approach. Taking a public policy perspective, we have emphasized the Horwitz–Thompson (HT) approach to mitigating the effects of preferential sampling in order to get unbiased estimators.

The case study has demonstrated use of the method where the parameters being estimated are numerical features of the finite population of exposures at sites in the UK that in 1970 were measuring black smoke. Subsequently only subsets of those exposures were measured and the case study looks at the use of the methods we have developed, to adjust these estimates for the effects of preferential sample. The results show reductions in the estimates, illustrating how the preferential siting of monitors where exposures are high, gives an exaggerated impression of the level of black smoke and the number of sites that are in noncompliance, using various hypothetical regulatory standards.

The case study does not assume that it is the sites with low concentration levels that were removed in the years since monitoring started. Support for that hypothesis

	Limit 68 μmg-3		**Guide 51 μmg-3**		**Guide 34 μmg-3**	
	Actual	Adjusted	Actual	Adjusted	Actual	Adjusted
1972	129	123	236	225	402	384
1973	73	68	153	143	327	306
1974	31	28	94	84	211	189
1975	21	19	58	52	223	201
1976	19	18	50	47	201	189
1977	7	6	23	20	106	94
1978	7	7	18	17	94	87
1979	8	7	21	19	75	71
1980	0	0	6	6	22	21
1981	2	2	11	10	47	42
1982	0	0	0	0	18	9
1983	0	0	10	5	24	16
1984	0	0	0	0	10	8
1985	0	0	0	0	5	4
1986	0	0	0	0	12	6
1987	0	0	0	0	15	7
1988	0	0	0	0	12	3
1989	0	0	0	0	0	0
1990	0	0	0	0	0	0
1991	0	0	0	0	0	0
1992	0	0	0	0	0	0
1993	0	0	0	0	0	0
1994	0	0	0	0	0	0
1995	0	0	0	0	0	0
1996	0	0	0	0	0	0

Table 1: Number of sites exceeding limits and guide values for black smoke (see text for details). Number of exceedences based on recorded data are presented together with values adjusted for preferential sampling.

emerges instead from the logistic regression analysis, which shows that higher probabilities of inclusion attach to such sites. A particularly dramatic change in that direction occurred in 1981 when the network was reorganized owing to falling urban concentrations and to comply with EC directive 80/779/EEC [Colls, 2002]. Overall the method looks promising where it can be applied.

But the HT approach will not always be appropriate and then a likelihood based one may be feasible provided that the preferential sampling can be modelled. That could be the case in the context of air pollution and health in epidemiological analyses, for as Guttorp and Sampson [2010] point out the air pollution monitoring sites may be intentionally located for a reasons such as the need to measure: (i) background levels outside of urban areas; (ii) levels in residential areas; and (iii) levels near pollutant sources. Then the method in D10 or the alternative in Subsection 4.2 may be useful. An important case is that in which the network is augmented in successive time periods rather than removed as in our case study [Le and Zidek, 2006].

The approach taken in the case study worked since representatives of the unmonitored sites remained in sample so that the selection weights would appropriately compensate for their under representation in the computation of population statistics. In the absence of such representation or good background knowledge of how the biased selection was made, there would seem to be no alternative but to augment the network with some, possibly temporary monitors. That leads to a design problem about the optimal selection of those sites. In this case, then we would be unbiasing the design rather than unbiasing the estimates, and that would need a different approach than that in this paper.

We conclude by emphasizing the importance of the issue studied in this paper as well as D10, and the need for much more work on it, given the large amount of environmental monitoring being done today and the need for good data in environmental risk analysis and management.

Acknowledgement. The genesis of the work described in this paper was the 2009-10 thematic program on spatial analysis for environmental mapping, epidemiology and climate change developed by the Statistical and Applied Mathematics Institute on spatial statistics. We are grateful to members of its Working Group on preferential sampling, of which the first author was a member, for discussions that stimulated our interest in that topic.

References

B. Ainslie, C. Reuten, DG Steyn, N.D. Le, and J.V. Zidek. 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, 2009.

D.A. Binder and P. Zdenek. Use of estimating functions for estimation from complex surveys. *J Amer Statist Assoc*, 89:1035, 1043 1994.

H. Chang, A.Q. Fu, N.D. Le, and J.V. Zidek. Designing environmental monitoring networks to measure extremes. *Environmental and Ecological Statistics*, 14(3):301–321, 2007.

G. Cicchitelli and G.E. Montanari. Model-assisted estimation of a spatial population mean. *International Statistical Review / Revue Internationale de Statistique*, pages Online: DOI: 10.1111/j.1751–5823.2011.00164.x, 2012.

A. Ciocco and D.J. Thompson. A follow-up of donora ten years after: methodology and findings. *Am J Public Health Nations Health*, 51:155–164, 1961.

J. Colls. *Air pollution, modelling, and mitigation.* Routledge, Abingdon, Oxford, 2002.

P.J. Diggle, R. Menezes, and T. Su. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232, 2010.

D. Dockery and CA III Pope. Acute respiratory effects of particulate air pollution. *Annu. Rev. Public Health*, 15:107–132, 1994.

EPA. Air quality criteria for ozone and related photochemical oxidants. Technical report, http://oaspub.epa.gov/eims/eimsapi.dispdetail?deid=149923, 2005.

European Commision. Council directive 80/779/eec of 15 july 1980 on air quality limit values and guide values for sulphur dioxide and suspended particulates. 1980.

J. Firket. Fog along the Meuse valley. *Trans Faraday Soc*, 32:1191–1194, 1936.

J.F. Garner and R.K. Crow. *Clean Air-Law and Practice.* Shaw and Sons Ltd., 1969.

VP Godambe and M.E. Thompson. Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review/Revue Internationale de Statistique*, pages 127–138, 1986.

M.S. Goldberg, R.T. Burnett, J. C. Bailar, 3rd, R. Tamblyn, P. Ernst, J. Flegel, K. Brook, Y. Bonvalot, R. Singh, M. F. Valois, and R. Vincent. Identification of persons with cardiorespiratory conditions who are at risk of dying from the acute effects of ambient air particles. *Environ Health Perspect*, 109(Suppl 4):487–494, 2001.

P. Guttorp and P. Sampson. 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, 2010.

R.C. Gwynn, R. T. Burnett, and G.D. Thurston. A time-series analysis of acidic particulate matter and daily mortality and morbidity in the buffalo, new york, region. *Environ Health Perspect*, 108:125–133, 2000.

D.G. Horvitz and D.J Thompson. A generalization of sampling without replacement from a finite universe. *J Amer Statist Assoc*, 47:663–685, 1952.

JF Lawless, JD Kalbfleisch, and CJ Wild. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):413–438, 1999.

N.D. Le and J.V. Zidek. *Statistical analysis of environmental space-time processes*. Springer Verlag, 2006.

A. Lee and Y. Hirose. Semi-parametric efficiency bounds for regression models under response-selective sampling: the profile likelihood approach. *Annals of the Institute of Statistical Mathematics*, 62(6):1023–1052, 2010.

D. Muir and D.P.H. Laxen. Black smoke as a surrogate for pm¡ sub¿ 10¡/sub¿ in health studies? *Atmospheric Environment*, 29(8):959–962, 1995.

Ministry of Health. Mortality and morbidity during the london fog of december, 1962. *H.M.S.O. London*, 1954.

W Ott. A Physical Explanation of the Lognormality of Pollutant Concentrations. *Journal of the Air Waste Management Association*, 40:1378–1383, 1990.

D. Pfeffermann. The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, pages 317–337, 1993.

JNK Rao, AJ Scott, and C.J. Skinner. Quasi-score tests with survey data. *Statistica Sinica*, 8:1059–1070, 1998.

S. Rathburn. 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, 2010.

J.M. Samet, F. Dominici, F.C. Curriero, I. Coursac, and S. L. Zeger. Fine particulate air pollution and mortality in 20 u.s. cities, 1987–1994. *N Engl J Med*, 343:1742–1749, 2000.

C.-E. Sarndäl, B. Swensson, and J. Wretman. *Model assisted survey sampling*. Statisties. Springer New York, 2003.

A.J. Scott and C.J. Wild. Fitting binary regression models with response-biased samples. *Canadian Journal of Statistics*, 39(3):519–536, 2011.

G. Shaddick and J.V. Zidek. Preferential sampling in long term monitoring of air pollution: a case study. Technical Report 267, Department of Statistics, University of British Columbia, 2012.

D. Simpson, F. Lindgren, and H. Rue. In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, page DOI: 10.1002/env.1137, 2011.

A.C. Stern, H.C. Wohlers, R.W. Boubel, and W.P. Lowry. *Fundamentals of Air Pollution.* Academic Press, 1973.

A. P. Verhoeff, G. Hoek, and J.H. Schwartz, J.; van Wijnen. Air pollution and daily mortality in amsterdam. *Epidemiology*, 7:225–230, 1996.