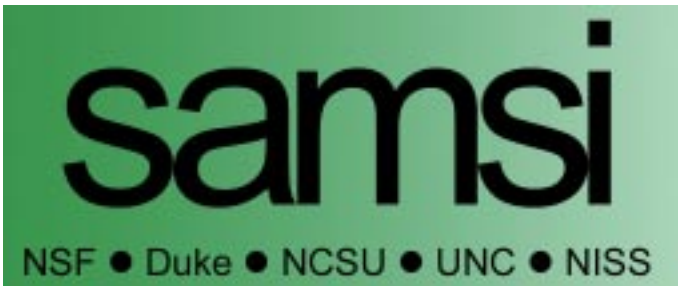

Designing Networks for Monitoring Multivariate Environmental Fields Using Data with Monotone Pattern

Nhu D. Le, Li Sun, James V. Zidek

Technical Report #2003-5
May 23, 2003

This material was based upon work supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

Designing Networks for Monitoring Multivariate Environmental Fields Using Data With A Monotone Pattern

Nhu D Le[1,2], Li Sun[1,2], James V Zidek[2]
[1] Cancer Control Research, BC Cancer Agency
[2] Department of Statistics, University of British Columbia

May 23, 2003

Abstract

This paper presents an approach to network design when the responses at each site are vector rather than scalar valued. The decision problem fits into the general framework of Bayesian ranking and selection theory. We suppose the designer must add sites to or remove sites from an existing environmental monitoring network. Optimality is based on entropy. It yields a natural decision criterion in the usual situation, where there is no single well defined design objective. We assume the responses have (possibly after transformation) a joint multivariate Gaussian distribution. Furthermore, the observed data are from monitored sites where some of the monitors have only recently been installed. Here "recent" means relative to the start-up time of the oldest monitoring site in the ambient monitoring network under consideration. Thus the monitoring data exhibit a monotone pattern, resembling a "staircase" whose highest "step" comes from the oldest monitoring sites. The natural conjugate, generalized inverted Wishart, is adopted as the first stage prior for the Gaussian model parameters. An empirical Bayes strategy is used at the second stage. A reasonably explicit form for the entropy of the resulting predictive distribution is derived. An application of the developed methodology to the hourly PM_{10} concentrations in British Columbia is presented.

Key words: Entropy; Bayesian design; environmental monitoring; network; generalized inverted Wishart; monotone data.

1 Introduction

In this paper, using the approach of Zidek et al (2000) and the predictive distribution of Kibria et al (2002) we develop an optimal method for designing an environmental monitoring network. As in the earlier work, we rely on entropy and information theoretical considerations to define the optimal design. A key feature of this method is its capacity to contend with data that exhibit "monotone pattern" resulting from the monitors having begun operation at different times. Such patterns very commonly occur in practice; our method allows all available data to be used, thereby increasing confidence in the value of the design.

Our approach uses a hierarchical Bayes approach with some estimated components. That framework is natural; designers invariably need prior information. For example, Linthurst and his co–investigators ((Linthurst et al., 1986, p. 4) relied on their "expectations" of low alkalinity to help select their sample of surface water bodies in the United States. After all, the real information only comes from the experiment being designed to produce it!

Optimizing a design requires a design objective. However, defining that objective can be difficult since usually a number of reasonable, often conflicting objectives can be discerned. For example, regulators may wish to build a network that detects non-compliance with proclaimed quality standards. They would prefer to "gauge" sites near anticipated "hot-spots." In contrast, epidemiologists concerned with the health effect of a perceived hazard, would want to split those sites equally between areas of high risk and areas of low risk to maximize contrast and the power of their health effects analyses. Some investigators might be interested in measuring extremes, others trends. Each of these can be measured in a variety of different ways, and those different metrics may well imply different optimal designs. Designing to monitor multivariate response provides even greater challenges since now different levels of importance can attach to the different co-ordinates or to some index computed from them. Also related to specifying the design are factors such as cost as well as levels of temporal and spatial aggregation. In combination, these identified goals and associated factors can lead to a myriad of possible objectives. Although the multi-attribute decision paradigm has a useful role to play here, clearly the combination of terms in the resulting objective may be very large indeed. (For a discussion of the multi-attribute approach in the context of network design, see the document of PD Sampson, P Guttorp and DM Holland, http://www.epa.gov/ttn/amtic/files/ambient/pm25/workshop/spatial/sampsn2.pdf).

Specifying that objective may even seem impossible since many of the future uses of the network simply cannot be foreseen. Zidek et al (2000) give an example of a network comprised of several networks established at different times for different purposes. The original network, established to measure acid deposition, tended to be located in rural areas. Later as the state of knowledge of environmental risk evolved, air pollution came to dominate acid rain as a societal concern and the (by now composite) network tended to be located in urban areas.

Yet the high cost of network construction and maintenance leads to persistent demand for

rational designs that in practice cannot be ignored. This led to a solution that seems to embrace the spirit of all the objectives while not emphasizing any one of them. That solution which uses entropy to define an objective function, was proposed by Caselton and Husain (1984), Caselton and Zidek (1984, hereafter CZ) and again by Shewry and Wynn (1987), and Sebastiani and Wynn (2000). It has also been embraced in the work of Bueso et al (1998, 1999b), Angulo, Bueso, and Alonso (2000) and Angulo and Bueso (2001). GG In fact, the idea of using entropy in experimental design goes back at least to Lindley (1956). There is a substantial body of work on optimal design in the Bayesian context although none as far as we are aware covers the application which is made here. For a review of Bayesian design see Verdinelli (1991).

The entropy approach to design was implemented by Caselton, Kan and Zidek (1990, hereafter, CKZ) to obtain a method of ranking stations for possible elimination from an existing network; refinements are added by Wu and Zidek (1992). Guttorp, Le, Sampson and Zidek (1993) tackle the complementary problem of extending an existing network. Le and Zidek (1994) extend the approach to a multivariate setting. [We rely on the latter for elements of the design review in this paper, provided here for completeness.] Zidek et al (2000) propose a method for incorporating costs. In this paper the design problem for multivariate responses is addressed where the existing monitoring network has stations with different operational periods, resulting in a monotone (staircase) pattern.

The basic idea underlying the just cited work is that all data have a fundamental purpose, that of reducing uncertainty about some aspect of the world. Uncertainty according to the postulates of Bayesian theory is quantifiable in terms of probability distributions. And the postulates of entropy theory, in turn, imply that the uncertainty in any distribution is indexed by its entropy. Ineluctably, an optimal design must minimize residual entropy after data has been collected.

However, before getting to the entropy based approach we give in Section 2, a review of some of the basic approaches that have been taken to network design. From there we turn in 3 to a discussion of the entropy based approach. From there we go in Section 4 to the predictive model that provides the framework in which our entropy approach develops. That leads in Section 5 to the objective function we seek, the entropy based criterion. Section 6 addresses the implementation problem of estimating some of the hyperparameters involved in the predictive distribution. We proceed to an illustrative example in Section 8 and a wrap-up discussion in Section 9.

2 Approaches to Design

Although a sampling domain may seem to offer a continuum of possible monitoring locations (sites), the actual number will typically lie in a small discrete determined by such things as accessibility or availability. Thus, in this paper we posit a discrete population of spatial sites.

Generally designs may be "probability-" or "model-based". The former includes simple random sampling: sites are sampled at random with equal probability (usually without replacement). The measured responses, which may even be a time-series of values, would then be (approximately) independent and their associated inferential theory quite simple. As well, such designs prove quite robust since nothing is assumed about the population of possible responses.

However, these designs can also be very inefficient under the simplest of assumptions about the population. Moreover, sampling sites could end up next to one another by chance, thereby making one of them redundant except in exceptional cases. Thus, samplers commonly rely on population models and achieve potentially dramatic increases in efficiency under these models. For example, they may postulate a population that consists of a union of homogeneous geographical strata. Under that model, only a small number of sites would need to be selected from each stratum. Because of their appeal, such designs have been used in a survey of US lakes (**citation**) and in EMAP (**citation**).

While stratification tends to diversify sampling, adjacent pairs of sites can still obtain within strata and on opposite sites of a common boundary. Moreover, knowledge about environmental fields can well exceed what can be accommodated by the models of probability-based theory. That knowledge can lead to greater gains in efficiency than achievable through probability-based designs and hence model-based designs are commonly used in practice to achieve design optimality.

Broadly speaking, two distinct approaches have emerged for selecting model-based (or optimal) network designs (Federov and Mueller, 1988, 1989), based either on regression models or random field models. Although the latter will be emphasized in this paper, we include a brief review of the former since it has come to be applied to the problem of network design. Furthermore, we will emphasize advantages and disadvantages of that approach.

Regression model (optimal design) theory originally had nothing to do with monitoring networks. It originated in Smith (1918) and was refined by Elfving (1952), Keifer (1959), and others (see Silvey 1980, Fedorov and Hackl 1997, and Müller 2001 for reviews). It concerned continuous sampling domains, $\mathcal{X}$ and optimal designs, ξ, with finite support, $x_1, \ldots, x_m \in \mathcal{X}$ with $\sum_{i=1}^{m} \xi(x_i) = 1$. In all, $n\xi(x_i)$ (suitably rounded) responses would then be measured at x_i for all $i = 1, \ldots, m$ to obtain $y_1, \ldots, y_n$. Key to the method was a regression model, $y(x) = \eta(x, \beta) + \varepsilon(x)$, that related the y's to the selected (and fixed) x's. Also key was the assumption that the ε's were independent from one sample point x to another. Optimality was then defined in terms of the efficiency of estimators of β to obtain an objective function, $\Phi(M(\xi))$, to be optimized where $M(\xi)$ denotes the information matrix and Φ, a positive function that depends on the criterion adopted. For example, in ordinary linear regression, $M(\xi) = \sigma^2[\mathbf{X'X}]^{-1}$. Φ could be any of a number of possibilities including $\Phi(A) = -logdet(A)$ (D-optimality) or $\Phi(A) = Tr(A)$ (A-Optimality). An elegant mathematical theory emerged together with numerical algorithms for computing the optimum design approximately.

To illustrate, suppose that conditional on $x \in [a, b]$, $y(x) = \alpha + \beta x + \varepsilon(x)$ and the ϵ's are

independent of the x's as well as each other. Then to minimize the variance of the the least squares estimator of β, the optimal design would have $x_1 = a, x_2 = b$ while $\xi(x_1) = \xi(x_2) = 1/2$.

Regression based, optimal design theory as described above encounters difficulties in application to network design. There monitors must be located at a subset of available sites and then simultaneously measure the field of interest regularly for an indefinite period. For example, every TEOM particulate air pollution monitor at an urban sampling site yields hourly observations. To measure n responses each time would entail "gauging" n sites, forcing $\xi \equiv 1/n$, that is, completely determined once its support is specified, making the classical theory of design irrelevant.

Nevertheless, a sustained effort has been made to adapt the regression model paradigm to encompass network design. Federov and Müller (1989) cite Gibrik (1976) as an early attempt. However, the major push came later (Fedorov 1987 1989, as well as Fedorov and Müller 1987 1989). The motive may in part have been a unified optimal design theory. However, Fedorov and Müller (1989) give a more pragmatic reason. They argue in their paper that hitherto, only sub-optimal designs could be found, feasible algorithms being limited to adding just one station at a time, albeit optimally. However, algorithms from the regression model theory offered promise (and algorithms!) by which genuinely optimal designs could be computed. (This reason may need be quite as compelling for the maximum entropy designs proposed in the next section where quick algorithms are now available for finding the optimum designs, at least for networks of moderate size.)

To that end, Fedorov and Müller (1988), assumed that at time $t = 1, \ldots, T$, $y_t(x_i) = \eta(x_i, \beta_t) + \varepsilon_t(x_i)$ where once again the ε's are all independent of each other but the β_t's are random and autocorrelated. Moreover, $\eta(x_i, \beta_t) = g^T(x_i)\beta_t$ for a known vector-valued g. Thus, this rather ingenious model captures both temporal and spatial covariance. The latter is not as restricted as it might seem since the co-ordinates of g can be eigenfunctions of the spatial covariance kernel when it is known. That covariance can thus be approximated well if the dimension of g is sufficiently large. But this comes at the expense of fixing the variances of these random effects to be eigenvalues of that kernel. The design objectives embrace the performance of either a linear predictor of a single β_t, say at time t='now' or of the mean of the common β_t distribution. These objectives would not seem compelling when the coefficients are merely artifacts of the eigenvector expansion associated with the covariance kernel rather than quantities of substantive interest such as the slope of a genuine regression model.

The authors recognize the limitation we mention above, that in this context the optimum design must be a subset of the available design set. However, to bring in the classical theory and associated algorithms, they relax that restriction and admit general ξ's, albeit subject to a boundedness requirement, so that established numerical search solutions now obtain. They call this substitution a "continuous approximation" and solve that problem instead of the original. The result will not usually be a feasible solution to the original problem and Fedorov and Müller (1988, 1989) note the challenge of interpreting it, seen variously as a local density, an indicator of

a "hot spot" or a design with more than one monitor at some sites. Further work in this direction described in Müller (2001) may help clarify the nature of this approximation. However, we are unclear about the value of substituting the approximate problem (and big associated toolbox of computational algorithms) for the hard - to - solve exact discrete design problem (and inevitable feasible - to - compute approximations), and issue that seems to need further investigation.

Apart from the problem of interpreting the optimum, issues of a more technical nature arise. First, suppose a genuine regression model (as opposed to eigenfunction expansion) is used above so that the objective function is substantively meaningful. Then the range of spatial covariance kernels will be restricted unless the ε's are allowed to be spatially correlated. That need is met in the extensions of the model above described in in the reviews of Fedorov (1996) and Müller (2001). However, the resulting design objective function "does not have to much in common with [the original] besides notation" in the words of Fedorov (1996, p524). A new toolbox is needed except in has to be created except in simple cases where an exhaustive search is needed. Back to square one!

While the regression model above does have substantive appeal, its value is uncertain. Environmental space - time fields tend to be so complex that their random response fields are only crudely related to spatial site co-ordinates. Moreover, the shape of that field can vary dramatically over time and season. Thus, a model like that in Example 1 of Fedorov et al (1987) seems quite unrealistic. In other words, finding a meaningful, known vector - valued function g would generally be difficult or impossible.

The alternative, the eigenfunction expansion also presents difficulties according to Fedorov (1996) relating to the problem of accurately approximating the spatial covariance kernel. Complications can arise in particular when the size of the proposed network is large. Moreover, while the eigenfunctions are know to exist under very general conditions, its not clear that actually finding them in usable form will be possible in problems of realistically size.

To summarize, the regression model approach does offer a very highly evolved theory for design, along with a substantial toolbox of algorithms for computing optimal designs, at least approximately. It also offers a broad range of objective functions which formally embraces that which comes out of the maximum entropy approach in the Gaussian case we introduce in the next section. However, forcing the network design problem into the regression model mold proves challenging both in terms of interpretation of the resulting optima as well as satisfying the assumptions underlying that approach.

Perhaps the strongest linkage of the regression modelling and random field approaches can be found in geostatistics. Because until very recently, that subject has concerned spatial fields while we focus on monitoring space - time fields, we will not describe this approach in detail. (Wackernagel 1999 gives a very readable recent account. Myers 2002 addresses space-time processes from a geostatistical modelling perspective.)

Unlike the regression modelling approach above, that emphasizes parameter estimation, geostatistics has tended to focus on the prediction of unmeasured values in a spatial field that para-

doxically, is regarded as random even though its fixed. Two methods are commonly employed, co-kriging and universal kriging. The first concerns the prediction of an unmeasured co-ordinate of the response vector, say $y_1(x_0)$ using an optimal linear predictor based on the observed response vectors at all the sampling sites. The coefficients of that optimal predictor are found by requiring it to be unbiased and to minimize the mean square prediction error. They will depend on the covariances between responses and between the sites, covariances that are unrealistically assumed to be known and later estimated from the data usually without adequately accounting for the additional uncertainty thereby introduced. In contrast to the first, the second relies on a regression model precisely of the form given above $y(x) = g^T(x)\beta + \varepsilon_t(x)$ where the ε's are assumed to have a covariance structure of known form. However, unlike the regression-modelling approach above, there the goal is prediction of the random response (possibly a vector) at a point where it has not been measured. Moreover, g (which can be a matrix in the multivariate case) can represent an observable covariate process. Optimization again relies on selecting coefficients by minimizing mean squared prediction error subject to the requirement of unbiasedness. Designs are commonly found iteratively one future site at a time, by choosing the site x_0 where the prediction error of the optimum predictor proves to be greatest.

Other approaches to model based designs have been proposed. For example, Bueso, Angulo, Quian and Alonso (1999a), offer one based on "stochastic complexity."

That completes our survey of approaches save for the last, the maximum entropy approach described in the next section.

3 Entropy-based Design

As noted in Section 1, entropy can be an appealing design criterion because it sidesteps the problem of specifying a particular design criteria. Moreover, that criterion fits well into the Bayesian framework we adopt for the spatial - temporal stochastic models of interest here.

To describe the approach more precisely, we associate a k dimensional random vector with every site in a spatial random field. The vectors corresponding to the sites in the discrete random field may be "stacked" to obtain a random vector which represents the discrete random field. For simplicity of exposition, we may then assume $k = 1$ in this section. The random vector field is observed at g discrete, "gauged" sites at sampling times, $j = 1, \cdots, n$, to yield $g \times 1$ data vectors, $X_j^{(2)} = (X_j^{(21)}, \cdots, X_j^{(2g)})'$. Of interest is a $u \times 1$ vector, $X_{n+1}^{(1)} = (X_{n+1}^{(11)}, \cdots, X_{n+1}^{(1u)})'$, of unmeasured future values at u "ungauged" sites at time $n+1$. The spatial field is over the $u+g$ discrete sites.

Suppose X_j has the joint probability density function, f_j, for all j. The total uncertainty about X_j may be expressed by the entropy of its distribution, *i.e.* $H_j(X_j) = E[-logf_j(X_j)/h(X_j)]$, where $h(\cdot)$ is a not necessarily integrable reference density (see Jaynes 1963). The inclusion of $h(\cdot)$ in $H_j(\cdot)$ ensures the latter is invariant under one–to–one transformations of the scale of X_j.

Note that the distributions involved in H_j may be conditional on certain covariate vectors, $\{z_j\}$, regarded as fixed.

Given the network's mission to monitor the environment, we regard the next value, X_{n+1}, as being of primary interest. However, X_{n+1}'s probability density function, $f_{(n+1)}(\cdot) = f_{(n+1)}(\cdot \mid \theta)$, depends on a vector of unspecified model parameters, say θ, so cannot be used directly in computing its uncertainty, $H_{n+1}(X_{n+1})$. Uncertainty about θ could be absorbed by averaging $f_{(n+1)}(\cdot \mid \theta)$ with respect to θ's distribution to obtain X_{n+1}'s marginal distribution and hence H_{n+1}. However, θ is of interest in its own right; in our theory θ includes the (spatial) covariance matrix of X_{n+1}, Σ, which has potential use in Kriging. Therefore we, like Caselton, Kan and Zidek (1992), include among the network's objectives, that of reducing the uncertainty about θ. The total entropy is then $H_{n+1}(X, \theta)$ conditional on the data, $D \stackrel{defn}{=} \{X_j^{(2)}, j = 1, \cdots, n\}$.

The two broad design objectives of extending or reducing a given network are well described in Bueso et al (1998). Earlier, CKZ considered the problem of reducing the number of sites in a network which has been providing data for some time. They address the problem of optimally partitioning X so that after appropriately relabeling the coordinates of X, it can be written as $X' = (X^{(1)'}, X^{(2)'})$ where $X^{(1)}$ and $X^{(2)}$ are u and g dimensional vectors respectively, $u + g = p$, of the coordinates of X corresponding to the sites which will be *ungauged* and *gauged* in the future. Wu and Zidek (1992) implement this approach in an analysis of 81 selected sites from the NADP/NTN network, an existing network of wet deposition monitoring stations in the USA, using 48 months of available data.

The objective addressed here is that of augmenting the network by gauging a specified number, u_2, of sites corresponding to coordinates of $X^{(1)}$. That is, we seek an optimal partitioning of $X^{(1)}$ which, after reordering its coordinates, yields $X^{(1)} = (X^{(rem)'}, X^{(add)'})$ where $X^{(rem)'}$ is a u_1-dimensional vector representing the future ungauged sites and $X^{(add)'}$ is a u_2-dimensional vector representing the future gauged sites. The resulting network will consist of the sites corresponding to the coordinates of $(X^{(add)'}, X^{(2)'}) \equiv G$, which is of dimension $(g + u_2)$.

The total a priori uncertainty conditional on D may, for certain purposes, be usefully decomposed as

$$H(X, \theta) = H(X \mid \theta) + H(\theta)$$

where, assuming the reference density to be $h(X, \theta) = h_1(X)h_2(\theta)$,

$$H(X \mid \theta) \quad = \quad E[-\log(f(X \mid \theta, D)/h_1(X)) \mid D]$$

and

$$H(\theta) \quad = \quad E[-\log(f(\theta \mid D)/h_2(\theta)) \mid D].$$

But for our immediate purposes of optimizing design, we need a different decomposition that reflects the partitioning of future observations into ungauged and gauged sites, $X' = (U, G)$

where $U \equiv X^{(rem)'}$. We express the total uncertainty $H(X \mid \theta)$ as

$$TOT = PRED + MODEL + MEAS$$

where, assuming $h_1(X) = h_{11}(U)h_{12}(G)$,

$$\begin{aligned} PRED &= E[-\log(f(U \mid G, \theta, D)/h_{11}(U)) \mid D], \\ MODEL &= E[-\log(f(\theta \mid G, D)/h_2(\theta)) \mid D], \end{aligned}$$

and

$$MEAS = E[-\log(f(G \mid D)/h_{12}(G)) \mid D].$$

If we assume that measurement error is negligible, eliminating all uncertainty about G by observing it will lead to an expected reduction in uncertainty given by $MEAS$. Thus, it will be optimal to augment the present network with sites represented in $X^{(add)'}$ so as to maximize $MEAS$. Since TOT is fixed, it follows that the same selection of additional sites will meet another design objective, that of minimizing $PRED + MODEL$. This latter entropy represents the residual uncertainty about the model parameters and the values of the random field at the ungauged sites, after observing G.

Incidentally, it is easily seen that had we started with $H(X)$ instead of $H(X, \theta)$, and made a decomposition analogous to that given above, we would have arrived at the same optimization criterion: maximization of $MEAS$.

We now move on to more specific issues with the help of a suitable space time model. That model has three components: (1) the measurement model; (2) the process model: (3) the parameter model. Let, $\mathbf{X}_t$ be the $1 \times n_t$ dimensional vector of responses observed at time t. These measured responses are related to $1 \times (u+g)q$ dimensional state vectors, $\mathbf{S}_t$, by the *measurement model*:

$$\mathbf{X}_t = \mathbf{S}_t\mathbf{H}_t + \varepsilon_t, \ t = 1, \ldots, \ (n+1). \tag{1}$$

We may think of this model as a composition of two others. The first is

$$\mathbf{X}_t = (\mathbf{Y}_t + \varepsilon_t^1)\mathbf{F}_t^1, \ t = 1, \ldots, \ (n+1) \tag{2}$$

where $\mathbf{Y}_t$ is of dimension $1 \times (u+g)r$ while $\mathbf{F}_t^1$ is a $(u+g)r \times n_t$ dimensional *design vector* of 1's and zeros that determines which of the responses are measured. In fact, $\mathbf{F}_t^1$ will generally be random, designating the data missing both at random as well as by design. However, we assume the former are missing for reasons ancillary to the process of their generation and measurement. Thus, we can condition on them and treat them as fixed.

The design problem discussed at the beginning of this section, is that of selecting $\mathbf{F}_{n+1}^1$. In other words, we need to find the optimal partition of the vector of all measurements of species

and sites that could be taken at time n + 1, into those that are actually taken and those not. The latter therefore remain uncertain along with all the parameters and latent variables in the process at that time. Following the earlier reasoning in this section, we minimize our residual uncertainty about all these uncertain quantities by selecting $\mathbf{F}_{n+1}^{1} = \mathbf{F}_{n+1}^{1opt}$ so as to maximize MEAS, that is

$$\mathbf{F}_{n+1}^{1opt} = \arg\min_{\mathbf{F}_{n+1}^{1}} H(\mathbf{X}_{n+1}|\mathbf{X}^{n}) \tag{3}$$

superscripts like n denoting here and in the sequel, all items up to and including those up to that specified time. The resulting design will change dynamically as n increases since we are ignoring practical considerations including cost in this section.

The second model needed to reach (5) is

$$\mathbf{Y}_t = \mathbf{S}_t\mathbf{F}_t^2 + \varepsilon_t^2,\ t = 1, \ldots, n \tag{4}$$

$\mathbf{F}_t^2$ relating responses, measured and unmeasured, to the state space vectors, $\mathbf{S}_t$. Generally, the F_t^2, unlike the design matrices, will involve unknowns. Finally, the so - called $((u+g)q \times n_t)$ *output matrix*, $\mathbf{H_t}$ is just the composition of the $((u+g)q \times (u+g)r)$ state transition matrix, $\mathbf{F_t^2}$ with the $((u+g)r \times n_t)$ measured response output matrix $\mathbf{F_t^1}$. The measurement error vectors, $\varepsilon_t = (\varepsilon_t^2 + \varepsilon_t^1)\mathbf{F}_t^1$, resulting from the combination of these two models are assumed to have zero mean, to have covariance matrix, $\mathbf{F}_t^{1\prime}\Sigma\mathbf{F}_t^1$ (assumed known for the purposes of this section), and to be independent of each other as well as other uncertain elements of the process and measurement models. Note that $\Sigma = \Sigma_{\varepsilon_t^2} + \Sigma_{\varepsilon_t^1}$ combines the spatial covariance of the responses with measurement noise.

We adopt the following class of process models:

$$\mathbf{S}_t = \mathbf{S}_{t-1}\theta_t + \nu_t,\ t = 1, \ldots,\ (n+1) \tag{5}$$

where the process noise variables, ν_t, have zero means, covariances Σ_ν and are independent of one another as well as of the other random process vectors. Returning to the general case, we assume known for the purposes of this section, both θ as well as the covariances, Σ_ν and Σ. However, a more realistic approach like that in the next section would add a *parameter model* that specifies prior distributions for these components.

Now assume all measurement and state space processes above, have a multivariate Gaussian distribution (possibly after an appropriate transformation). With these assumptions we can explicitly evaluate the entropy in Equation (3). To that end, let

$$\begin{aligned} \hat{\mathbf{S}}_t &= E(\mathbf{S}_t|\mathbf{X}^t), \\ \hat{\mathbf{P}}_t &= Cov(\mathbf{S}_t|\mathbf{X}^t)) \text{ so that,} \\ \mathbf{S}_t|\mathbf{X}^t &\sim N_{(u+g)q}(\hat{\mathbf{S}}_t,\ \hat{\mathbf{P}}_t) \end{aligned} \tag{6}$$

for all $t = 1, \ldots, (n+1)$. We now find the conditional distribution, $\mathbf{X}_{t+1}|\mathbf{X}^t \sim N_{n_{t+1}}[E(\mathbf{X}_{t+1}|\mathbf{X}^t), Cov(\mathbf{X}_{t+1}|\mathbf{X}^t)]$ needed to compute the entropy. As a first step we find the conditional distribution, $\mathbf{S}_{t+1}|\mathbf{X}^t \sim N_{pq}[E(\mathbf{S}_{t+1}|\mathbf{X}^t), Cov(\mathbf{S}_{t+1}|\mathbf{X}^t)]$. First,

$$\begin{aligned} E(\mathbf{S}_{t+1}|\mathbf{X}^t) &= E(E[\mathbf{S}_{t+1}|\mathbf{S}_t]|\mathbf{X}^t) \\ &= E(\mathbf{S}_t\theta|\mathbf{X}^t) \\ &= \hat{\mathbf{S}}_t\theta. \end{aligned} \tag{7}$$

Similarly,

$$\begin{aligned} Cov(\mathbf{S}_{t+1}|\mathbf{X}^t) &= Cov(E[\mathbf{S}_{t+1}|\mathbf{S}_t]|\mathbf{X}^t) \\ &\quad + E(Cov[\mathbf{S}_{t+1}|\mathbf{S}_t]|\mathbf{X}^t) \\ &= Cov(E[\mathbf{S}_t\theta_t|\mathbf{X}^t) \\ &\quad + E(\Sigma_\nu) \\ &= \theta_t'\hat{\mathbf{P}}_t\theta_t + \Sigma_\nu. \end{aligned} \tag{8}$$

Then

$$\begin{aligned} Cov(\mathbf{X}_{t+1}|\mathbf{X}^t) &= Cov([\mathbf{S}_{t+1}\mathbf{F}^2_{t+1} + \varepsilon^2_{t+1} + \varepsilon^1_{t+1}]\mathbf{F}^1_{t+1}|\mathbf{X}^t) \\ &= \mathbf{H}'_{t+1}Cov(\mathbf{S}_{t+1}|\mathbf{X}^t)\mathbf{H}_{t+1} + \mathbf{F}^{1\prime}_{t+1}\Sigma\mathbf{F}^1_{t+1} \\ &= \mathbf{H}'_{t+1}[\theta_t'\hat{\mathbf{P}}_t\theta_t + \Sigma_\nu]\mathbf{H}_{t+1} + \mathbf{F}^{1\prime}_{t+1}\Sigma\mathbf{F}^1_{t+1}, \end{aligned} \tag{9}$$

by Equation (8). Finally, from standard theory for the Gaussian distribution, it follows that the entropy to be maximized, $H(\mathbf{X_{n+1}}|\mathbf{X^n})$ is, apart from irrelevant constants, the logarithm of the determinant of the (conditional) covariance matrix:

$$|\mathbf{H}'_{\mathbf{n+1}}[\theta'_{\mathbf{n+1}}\hat{\mathbf{P}}_{\mathbf{n}}\theta_{\mathbf{n+1}} + \mathbf{\Sigma}_\nu]\mathbf{H_{n+1}} + \mathbf{F}^{1\prime}_{n+1}\Sigma\mathbf{F}^1_{n+1}|. \tag{10}$$

The logarithm ensures that the optimal design will remain invariant under re-scaling; multiplication of the normal density by the Jacobean of the transformation simply becomes an additive shift. Recalling that $\mathbf{H_{n+1}} = \mathbf{F^2_{n+1}F^1_{n+1}}$, we see that our optimal design matrix is found by finding the maximal $n_{n+1} \times n_{n+1}$ sub-determinant, that is *generalized sub-variance*, of the covariance

$$\mathbf{F}^{2\prime}_{n+1}[\theta'_{n+1}\hat{\mathbf{P}}_n\theta_{n+1} + \Sigma_\nu]\mathbf{F}^2_{n+1} + \Sigma. \tag{11}$$

Example. The apparent simplicity of the state space model above, disguises the difficulty of its formulation in specific cases. Consider the case of Li et al (1999) where an autoregressive model of order three obtains at every site, $j = 1, \ldots, (u+g)$ [and the number of response species is r = 1]. There,

$$\mathbf{Y}_{tj} - \beta_j\mathbf{Z}_t = [\mathbf{Y}_{(t-1)j} - \beta_j\mathbf{Z}_{t-1}]\rho_{1j} + [\mathbf{Y}_{(t-2)j} - \beta_j\mathbf{Z}_{t-2}]\rho_{2j} + [\mathbf{Y}_{(t-3)j} - \beta_j\mathbf{Z}_{t-3}]\rho_{3j} + \varepsilon^2_{tj}$$

where r=1, $\beta_j : 1 \times l$ is a vector of response dependent *trend* coefficients, the $\mathbf{Z}_t : \ l \times r$ are ancillary (and hence fixed) covariates with the same value at all sites j, and $\rho_{ij} : r \times r, \ i = 1, 2, 3$ are the autoregressive coefficient matrices. Equivalently, with an abuse of notation,

$$\mathbf{Y}_{tj} = \mathbf{Y}_{(t-1)j}\rho_{1j} + \mathbf{Y}_{(t-2)j}\rho_{2j} + \mathbf{Y}_{(t-3)j}\rho_{3j} + \beta_j \mathbf{Z}_t + \varepsilon^2_{tj}, \tag{12}$$

where $\mathbf{Z}_t$ now stands for $\mathbf{Z}_t - \mathbf{Z}_{t-1}\rho_{1j} - \mathbf{Z}_{t-2}\rho_{2j} - \mathbf{Z}_{t-3}\rho_{3j}$. A standard reformulation of this model would have $\mathbf{Y}_{tj} = \mathbf{S}_{tj}\mathbf{F}^2_{tj} + \varepsilon^2_{tj}$ where $\mathbf{S}_{tj} : 1 \times (l + 3r) = [\beta_j, \mathbf{Y}_{(t-1)j}, \mathbf{Y}_{(t-2)j}, \mathbf{Y}_{(t-3)j}]$ and

$$\mathbf{F}^2_{tj} : (l + 3r) \times r = \begin{pmatrix} \mathbf{Z}_t \\ \rho_{1j} \\ \rho_{2j} \\ \rho_{3j} \end{pmatrix}. \tag{13}$$

However, this dynamic state - space model fails since space and time are "inseparable" in Le et al (1999) and the residual independence assumption in Equation (5) proves to invalid. That problem is observed by Zidek et al (2002) who adopt a different approach, one that does not model fine - scale auto - covariance structures (and, as a bonus, avoids the risk of mispecifying them). Instead, they adopt the 24 hour site response vector as a basic building block. Since r species are responding each hour at each site (r =1 in the specific case under consideration here), that response vector is $24r \times 1$ dimensional. The resulting vector series at each site turns out to be quite well - modelled by a multivariate AR model of order 1 [MAR(1)] (used below in Section 4). Moreover, in the case of Li et al (1999), the $24r \times 24r$ dimensional autoregression coefficient matrices depend little on j so that $\rho_{ij} = \rho_i$ for all j proves a tenable assumption. This suggests the MAR model,

$$\mathbf{Y}_{tj} = \mathbf{Y}_{(t-1)j}\rho_1 + + \beta_j \mathbf{Z}_t + \varepsilon^2_{tj}. \tag{14}$$

With an appropriate change in dimensions, we can combine Equations (13-17) to get:

$$\mathbf{S}_t : 1 \times (u + g)(l + 24r) = [\mathbf{S}_{t1}, \ldots, \mathbf{S}_{tp}]; \tag{15}$$

$$\mathbf{F}^2_t = \begin{pmatrix} \mathbf{F}^2_{t1} & 0 & \cdots & 0 \\ 0 & \mathbf{F}^2_{t2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \mathbf{F}^2_{tp} \end{pmatrix}. \tag{16}$$

For the autoregressive model in Equation (14) we then obtain,

$$\mathbf{S}_{tj} = \mathbf{S}_{(t-1)j}\theta_t + \nu_{tj},\ j = 1, \ldots, (u+g) \tag{17}$$

where

$$\theta_{tj} = \begin{pmatrix} I_l & \mathbf{Z}_{t-1} & 0 & 0 \\ 0 & \rho_1 & I_{24r} & 0 \end{pmatrix}. \tag{18}$$

We may combine the matrices in Equation (18) to get

$$\theta_t : (u+g)(l+24r) \times (u+g)(l+24r) \quad = \quad \begin{pmatrix} \theta_t & 0 & \cdots & 0 \\ 0 & \theta_t & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \theta_t \end{pmatrix}. \tag{19}$$

Remarks.

3.1 A generalized sub-variance obtained from Equation (11) will tend to be small when either its columns or rows are nearly collinear, i.e. the associated responses are highly associated. That can occur because of strong spatial association between sites, as expressed through the "intrinsic" component of variation, Σ, in that equation. Or it may derive from strong temporal association as expressed through the remaining terms, *i.e.* "extrinsic" component. In any case, sites will tend to be omitted from the network, either because the are predictable from other sites in the present, or from measurements made in the past.

3.2 Typically in applications, data to time n will derive from a permanent set of monitoring sites and the design goal will be judicious augmentation of that network. To address this situation, let us represent the symmetric matrix in Equation (11) more simply as

$$\Xi = \begin{pmatrix} \Xi_{uu} & \Xi_{ug} \\ \Xi_{gu} & \Xi_{gg} \end{pmatrix}, \tag{20}$$

u meaning "ungauged", g, "gauged" while Ξ_{gg} represents covariance at the permanent sites. Then using a familiar identity for matrix determinants, the design optimization problem associated with Equation (11) becomes that of adding a fixed number of sites to the network with maximal generalized sub - variances of $\Xi_{u \cdot g} = \Xi_{uu} - \Xi_{ug}\Xi_{gg}^{-1}\Xi_{gu}$ of appropriate dimension.

3.3 Our formulation of the design problem through F_t^1 allows us to dynamically expand or contract the monitoring network at each successive time, the optimal basis for making alterations being expressed in Remark 3.1. One can conceive of hypothetical cases where

dynamically changing networks might be desirable, for example, in radiation monitoring with mobile monitors following the failure of a nuclear power generator or in military operations following the release of hazardous agents in the battlefield. However, such designs would generally be highly unrealistic for practical reasons involving things like cost and administration.

Through $\hat{\mathbf{P}}_n$, the extrinsic component of the design criterion above is a function of past data. Moreover, that component rather than the intrinsic component may point to the deletion of sites at time t = n+1 whose responses are well predicted from past data including that which they produced. Their deletion will eliminate the very source of information that justified their removable in the first place. Thus in time, the quality of the network in so far as it provides information about non-monitored sites (including some of those that were removed from the network) could degrade.

This suggests a need for a practical compromise and acceptance of a suboptimal permanent design after time t = n. That compromise might be achieved through filtering the data and relying primarily on the intrinsic component of covariance.

> **Example (Continued).** To arrive at an appropriate compromise design criterion we can transform the responses as $\mathbf{Y}^*_{tj} = \mathbf{Y}_{tj} - \mathbf{Y}_{(t-1)j}\rho_1 = \beta_j \mathbf{Z}_t + \varepsilon^2_{tj}$. Then our design at time t = n+1 will not depend on past measurements as predictors of current responses. These transforms having an added benefit in that it eliminates the autoregression matrices from the design criterion, simplifying technical analysis when they are unknown as in the next section. However, in practice this would require that they be well - estimated to enable approximate transformation of the data.

3.4 Another issue we must confront arises from our uncertainty about parameters like the θ's and the covariances that we assumed known. The result of incorporating that additional uncertainty makes the conditional distribution of $\mathbf{X}_{n+1}|\mathbf{X}^n$ non-Gaussian, In fact, that distribution will typically not have a tractable form, making a convenient analytical representation of the entropy impossible. Evaluating that entropy numerically is not a practical option since the combinatorial design optimization problem proves computationally intensive (in fact NP-hard) and finding it, difficult for realistically large values of u+g. Adding the additional burden of numerically evaluating the entropy at each iteration would make the overall burden prohibitively large. Thus, while we allow some of the parameters in the above model to be uncertain (that is, random) in the next section, we eliminate others (as just noted above), to retain the advantage of an explicit design objective function while reducing reliance on extrinsic components of covariance.

3.5 The measurement noise represented by $\Sigma_{\varepsilon^2_t}$ could conceivably vary in magnitude from re-

sponse to response and in extreme cases, could dominate the selection of an optimum design. For this and other reasons, it might be argued that the optimum design should not be selected like that above to include its capacity to reduce uncertainty about measurements that could have been but have not been taken. Instead the goal would be to maximally reduce uncertainty about $\mathbf{Y}_{n+1}$ rather than $\mathbf{X}_{n+1}$ given measurements to time n. These objects would be essentially equal [as assumed in the next section] when measurement noise is negligible. The design objective criterion in that case would obtain from that in Equation (11) after subtracting the measurement noise covariance.

4 Predictive Distribution

In this section, we make some simplifying assumptions while relaxing others imposed in the last section to achieve a more practical design theory. Specifically, we assume that: (1) $\varepsilon_t^1 \equiv 0$ so that the measurements, $\mathbf{X}_{tj} = \mathbf{Y}_{tj}$ for all t and j at which measurements are made; (2) after filtering out any autoregressive effects, $\mathbf{S}_{tj} = \beta_j$ for all t while $\mathbf{F}_{tj}^2 = \mathbf{Z}_t$ for all j. (We are investigating the possibility of relaxing these assumptions in current work. For convenience, we interchange the order of the β's and $\mathbf{Z}$'s in the models below.) Thus, $\nu_t \equiv 0$ and $\theta_{tj} \equiv 1$. However, we will add parameter models below. We then derive the predictive distribution for the unmeasured multivariate responses conditional on the observed data where stations have been added to a monitoring network progressively over time .

Rather than define the data design matrices introduced in Section 3, F_t^1, explicitly, we describe the structure of the data matrix obtained. Specifically, after appropriately reordering the sites, it has a *monotone structure*. That is, when the data are put together in an increasing order of the stations' operational periods, the data matrix has the appearance of a "staircase". Combining active networks with different starting times will yield such a structure.

4.1 Notation

Throughout the rest of the paper we let etr $(\cdot) = \exp[\mathrm{tr}(\cdot)]$; $vec(X)$ will denote a vector obtained by transposing and then stacking as columns the row vectors of a matrix X successively to form one column vector; and $\otimes$ represents the Kronecker product between matrices. In addition, we let

n = number of time points (e.g. number of days);
p = dimension of the multivariate response at each station;
u = number of locations with no monitors - called ungauged sites;
g = number of locations with monitors - called gauged sites.

The g gauged sites are organized into k blocks such that the jth block consists of g_j stations having the same number, m_j, of missing responses and hence $g = g_1 + \cdots + g_k$. These blocks are numbered so that the observed measurements correspond to a monotone data pattern or a staircase structure, that is,

$$m_1 \geq m_2 \geq \cdots \geq m_k \geq 0.$$

If the response values prior to the first monitor in operation are of interest, then m_k is set to be bigger than 0.

The following notation are used to facilitate our presentation. Denote the

- response variables at the gauged and ungauged sites by

$$Y = \left[Y^{[u]}, \begin{pmatrix} Y^{[g_1^m]} \\ Y^{[g_1^o]} \end{pmatrix}, \cdots, \begin{pmatrix} Y^{[g_k^m]} \\ Y^{[g_k^o]} \end{pmatrix} \right],$$

  where:

  - $Y^{[u]}$, $n \times up$ matrix, denotes the responses at ungauged sites;
  - $Y^{[g_j^m]}$, $m_j \times g_j p$ matrix, denotes the missing responses at the gauged sites in the j^{th} block;
  - $Y^{[g_j^o]}$, $(n - m_j) \times g_j p$ matrix, denotes the observed responses at the gauged sites in the j^{th} block.

- observed measurements at the gauged sites by D where

$$D = \left\{ Y^{[g_1^o]}, \cdots, Y^{[g_k^o]} \right\}.$$

- unobserved responses by

$$Y_{unob} = \left\{ Y^{[u]}, Y^{[g_1^m]}, \ldots, Y^{[g_k^m]} \right\}.$$

- unobserved responses in blocks j to k by

$$Y^{[g_j^m, \ldots, g_k^m]} = \left\{ Y^{[g_j^m]}, \ldots, Y^{[g_k^m]} \right\}.$$

- responses from blocks j to k, including both observed and unobserved stacks by

$$Y^{[g_j, \ldots, g_k]} = \left[\begin{pmatrix} Y^{[g_j^m]} \\ Y^{[g_j^o]} \end{pmatrix}, \cdots, \begin{pmatrix} Y^{[g_k^m]} \\ Y^{[g_k^o]} \end{pmatrix} \right].$$

- responses from all gauged sites by

$$Y^{[g]} = Y^{[g_1,\ldots,g_k]}.$$

We postulate l time-varying covariates responses $Z_t = (Z_{t1}, \ldots, Z_{tl})'$, at each time point t being constant across all sites, and

$$Z = \begin{pmatrix} Z_1' \\ \vdots \\ Z_n' \end{pmatrix}.$$

The $l \times (u+g)p$ coefficient matrix $\boldsymbol{\beta}$ corresponding to the l covariates and covariance matrix Σ of dimension $(u+g)p \times (u+g)p$ over gauged and ungauged sites are partitioned conformably:

$$\boldsymbol{\beta} = (\boldsymbol{\beta}^{[u]}, \boldsymbol{\beta}^{[g]}), \quad \Sigma = \begin{pmatrix} \Sigma^{[u]} & \Sigma^{[ug]} \\ \Sigma^{[gu]} & \Sigma^{[g]} \end{pmatrix}.$$

The coefficient matrix $\boldsymbol{\beta}^{[g]}$ and covariance matrix $\Sigma^{[g]}$ on the gauged sites are further partitioned by blocks as

$$\boldsymbol{\beta}^{[g]} = (\boldsymbol{\beta}^{[g_1]}, \ldots, \boldsymbol{\beta}^{[g_k]}), \text{ and } \boldsymbol{\beta}^{[g_j,\ldots,g_k]} = (\boldsymbol{\beta}^{[g_j]}, \ldots, \boldsymbol{\beta}^{[g_k]}).$$

Correspondingly

$$\Sigma^{[g]} = \begin{pmatrix} \Sigma^{[g_1]} & \ldots & \Sigma^{[g_1,g_k]} \\ \ldots & \ldots & \ldots \\ \Sigma^{[g_k,g_1]} & \ldots & \Sigma^{[g_k]} \end{pmatrix}, \quad \Sigma^{[g_j,\ldots,g_k]} = \begin{pmatrix} \Sigma^{[g_j]} & \ldots & \Sigma^{[g_j,g_k]} \\ \ldots & \ldots & \ldots \\ \Sigma^{[g_k,g_j]} & \ldots & \Sigma^{[g_k]} \end{pmatrix}.$$

The following 1-1 transformation (Barlett, 1933) of the matrix Σ is used:

$$\Sigma_{kk} = \Sigma^{[g_k]},$$

$$\Gamma_j = \Sigma^{[g_j]} - \Sigma^{[g_j,(g_{j+1},\ldots,g_k)]}\left(\Sigma^{[g_{j+1},\ldots,g_k]}\right)^{-1}\Sigma^{[(g_{j+1},\ldots,g_k),g_j]},$$

$$\tau_j = \left(\Sigma^{[g_{j+1},\ldots,g_k]}\right)^{-1}\Sigma^{[(g_{j+1},\ldots,g_k),g_j]}$$

where

$$\Sigma^{[(g_{j+1},\ldots,g_k),g_j]} = \begin{pmatrix} \Sigma^{[g_{j+1},g_j]} \\ \vdots \\ \Sigma^{[g_k,g_j]} \end{pmatrix}$$

for $j = 1, \ldots, k-1$. The matrix $\Sigma^{[g]}$ can then be obtained from $\{\Sigma_{kk}, (\Gamma_{k-1}, \tau_{k-1}), \ldots, (\Gamma_1, \tau_1)\}$ by means of this transformation.

4.2 The Model

The response matrix, Y, is assumed to follow a Gaussian-Generalized Inverted Wishart model. Specifically, using the notation described above,

$$\begin{cases} Y \mid \boldsymbol{\beta}, \Sigma \sim N(Z\boldsymbol{\beta}, I_n \otimes \Sigma); \\ \boldsymbol{\beta} \mid \Sigma, \boldsymbol{\beta}_0, F \sim N(\boldsymbol{\beta}_0, F^{-1} \otimes \Sigma); \\ \Sigma \sim GIW(\Psi, \delta), \end{cases} \tag{21}$$

where $N(\cdot, \cdot)$ denotes the Gaussian distribution, $\boldsymbol{\beta}_0$ is the $l \times (g+u)p$ hyperparameter mean matrix of $\boldsymbol{\beta}$, F^{-1} is an $l \times l$ positive definite matrix representing the variance component of $\boldsymbol{\beta}$ between its l rows, and Z is the matrix of covariates. GIW denotes a generalized inverted Wishart distribution of Σ with $\delta = (\delta_0, \delta_1, \ldots, \delta_k)'$ representing degrees of freedom, and Ψ being a collection of hyperparameters defined below. The GIW distribution is defined by

$$\begin{cases} \Sigma^{[g]} \sim GIW(\Psi^{[g]}, \delta^{[g]}); \\ \Gamma^{[u]} \sim IW(\Lambda_0 \otimes \Omega, \delta_0); \\ \tau^{[u]} \mid \Gamma^{[u]} \sim N\left(\tau_{0\,u}, H_0 \otimes \Gamma^{[u]}\right) \end{cases} \tag{22}$$

where $\Gamma^{[u]} = \Sigma^{[u|g]} = \Sigma^{[u]} - \Sigma^{[ug]}(\Sigma^{[g]})^{-1}\Sigma^{[gu]}$; $\tau^{[u]} = (\Sigma^{[g]})^{-1}\Sigma^{[gu]}$. IW denotes the Inverted Wishart with hyperparameters $(\Lambda_0, \Omega, \delta_0)$; the matrix $\tau_{0\,u}$ is the hyperparameter of $\tau^{[u]}$; and the matrix H_0 is the variance component of τ_u between its rows.

Moreover, the above GIW distribution is defined in a stepwise fashion through $\Sigma^{[g]}$ with $\delta^{[g]} = (\delta_1, \ldots, \delta_g)$ and $\Psi^{[g]}$ being another collection of hyperparameters. The distribution of $\{\Sigma_{kk}, (\Gamma_{k-1}, \tau_{k-1}), \ldots, (\Gamma_1, \tau_1)\}$ is defined as follows:

$$\begin{cases} \Sigma_{kk} \sim IW(\Lambda_k \otimes \Omega, \delta_k); \\ \tau_j \mid \Gamma_j \sim N(\tau_{0j}, H_j \otimes \Gamma_j), \quad j = 1, \ldots, k-1; \\ \Gamma_j \sim IW(\Lambda_j \otimes \Omega, \delta_j), \quad j = 1, \ldots, k-1. \end{cases} \tag{23}$$

The hyperparameters involved in our Gaussian-GIW model can be written as

$$\mathcal{H} = \{\boldsymbol{\beta}_0, F, \Psi, \delta\} \tag{24}$$

where

$$\Psi = \{\Psi^{[g]}, \Lambda_0, \Omega, \tau_{0\,u}, H_0\}, \text{ and } \Psi^{[g]} = \{\Lambda_k, \Omega, (\Lambda_j, H_j, \tau_{0\,j}); j = 1, \ldots, k-1\},$$

and

$$\delta = (\delta_0, \delta^{[g]}), \text{ and } \delta^{[g]} = (\delta_1, \ldots, \delta_g).$$

The dimensions of the parameters in Ψ are as follows

$$\Lambda_0 : u \times u, \quad \tau_{0u} : g \times u, \quad H_0 : g \times g, \quad \Lambda_k : g_k \times g_k, \quad \Omega : p \times p,$$

and for $j = 1, \ldots, k-1$,

$$\Lambda_j : g_j \times g_j, \quad \tau_j : (g_{j+1} + \cdots + g_k) \times g_j, \quad H_j : (g_{j+1} + \cdots + g_k) \times (g_{j+1} + \cdots + g_k).$$

In our work, we adopt

$$\tau_{0j} = \Psi^{-1}_{(j+1,j+1)} \Psi_{(j+1),j} \quad \text{and} \quad H_j = \Psi^{-1}_{(j+1,j+1)}. \tag{25}$$

Furthermore, we assume that the degrees of freedom, $\delta_1, \ldots, \delta_k$, follow a gamma prior distribution (Le, Sun and Zidek, 1998) where

$$\pi(\delta) \propto (\delta_1 \cdots \delta_k)^{\alpha - 1} \exp\{-r(\delta_1 + \cdots + \delta_k)\},$$

with α and r specified.

Remarks:

4.1 The GIW distribution, introduced by Brown, Le and Zidek (1994b), generalizes the IW distribution by allowing different degrees of freedom for a random positive definite matrix. A $p \times p$ positive definite matrix S has an IW distribution, denoted by $IW(B, \delta)$, with δ degrees of freedom if its density function is proportional to

$$|S|^{-\frac{1}{2}(\delta + p + 1)} \text{etr} \left\{ -\frac{1}{2} S^{-1} B \right\}.$$

4.2 The GIW distribution is a conjugate prior for a Gaussian distribution. This prior is very flexible and quite natural to deal with the staircase structure of the observed data. For example, different degrees of freedom for the k blocks can be expressed through the hyperparameter vector δ.

4.3 The GIW modeling method also gives us considerable latitude in selecting the numbers of blocks in the GIW structure. For example, we could group all sites that started operation at the same time in one block. Or we could select each site as a block in the stair-case structure.

4.3 The Predictive Distributions

This section presents the joint predictive distributions of all unobserved responses. Their means offer point predictors of those responses while the distribution as a whole allows us to assess the uncertainty of those predictors. Furthermore they allow us to convolve the unknown function with impact distributions so as to incorporate that uncertainty fully in a hierarchical model. The results are the multivariate extension of those derived by Le et al (2001).

To facilitate the presentation of the main results, we introduce the following notation. Let

$$\begin{pmatrix} \mu_{(1)}^{[j]} \\ \mu_{(2)}^{[j]} \end{pmatrix} : \begin{pmatrix} m_j \times g_j p \\ (n-m_j) \times g_j p \end{pmatrix} = Z\boldsymbol{\beta}_0^{[g_j]} + \tilde{\epsilon}^{[g_{j+1},\ldots,g_k]}\tau_{0j};$$

$$\begin{pmatrix} A_{11}^{[j]} & A_{12}^{[j]} \\ A_{21}^{[j]} & A_{22}^{[j]} \end{pmatrix} : \begin{pmatrix} m_j \times m_j & m_j \times (n-m_j) \\ (n-m_j) \times m_j & (n-m_j) \times (n-m_j) \end{pmatrix}$$

$$= I_n + ZF^{-1}Z' + \tilde{\epsilon}^{[g_{j+1},\ldots,g_k]} H_j (\tilde{\epsilon}^{[g_{j+1},\ldots,g_k]})',$$

where

$$\tilde{\epsilon}^{[g_{j+1},\ldots,g_k]} = \begin{cases} Y^{[g_{j+1},\ldots,g_k]} - Z\boldsymbol{\beta}_0^{[g_{j+1},\ldots,g_k]}, \text{ for } j = 1,\ldots,k-1; \\ \\ 0, \text{ for } j = k. \end{cases}$$

Moreover, for $j = 1,\ldots,k$,

$$\mu_{(1|2)}^{[j]} = \mu_{(1)}^{[j]} + A_{12}^{[j]}(A_{22}^{[j]})^{-1}\tilde{e}^{[g_j^o]};$$

$$\Phi_{(1|2)}^{[j]} = \frac{\delta_j - g_j p + 1}{\delta_j - g_j p + n - m_j + 1}\left[A_{11}^{[j]} - A_{12}^{[j]}(A_{22}^{[j]})^{-1}A_{21}^{[j]}\right];$$

$$\Psi_{(1|2)}^{[j]} = \frac{1}{\delta_j - g_j p + 1}\left[\Lambda_j \otimes \Omega + (\tilde{e}^{[g_j^o]})'(A_{22}^{[j]})^{-1}\tilde{e}^{[g_j^o]}\right];$$

$$\delta_{(1|2)}^{[j]} = \delta_j - g_j p + n - m_j + 1;$$

where

$$\tilde{e}^{[g_j^o]} = Y^{[g_j^o]} - \mu_{(2)}^{[j]};$$

and

$$\mu^{[u|g]} = Z\boldsymbol{\beta}_0^{[u]} + \tilde{\epsilon}^{[g]}\tau_{0u};$$

$$\Phi^{[u|g]} = I_n + ZF^{-1}Z' + \tilde{\epsilon}^{[g]}H_0(\tilde{\epsilon}^{[g]})';$$

$$\tilde{\epsilon}^{[g]} = Y^{[g]} - Z\beta_0^{[g]}.$$

A matrix valued random variable $X_{n\times m}$ is said to have a matric t-distribution, *i.e.*

$$X \sim t_{n\times m}(X^{(0)},\ A\otimes B,\ \delta),$$

where A is $n\times n$ and B is $m\times m$, if its density function has the form

$$f(X) \propto |A|^{-m/2}|B|^{-n/2}|I_n + \delta^{-1}[A^{-1}(X-X^{(0)})][(X-X^{(0)})B^{-1}]'|^{-\frac{\delta+n+m-1}{2}}, \tag{26}$$

and the normalizing constant of the density is given by

$$K = (\delta\pi^2)^{-\frac{nm}{2}}\frac{\Gamma_{n+m}[(\delta+n+m-1)/2]}{\Gamma_n[(\delta+n-1)/2]\Gamma_m[(\delta+m-1)/2]},$$

where

$$\Gamma_p(t) = \pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma[t-(i-1)/2] \tag{27}$$

denotes the multivariate Gamma function.

Theorem 1. *The predictive distribution of the unobserved responses conditional on the observed data D and the hyperparameter set $\mathcal{H}$ is given by*

$$(Y_{unob}\mid D,\mathcal{H}) \sim \left(Y^{[u]}\mid Y^{[g_1^m,\ldots,g_k^m]},D,\mathcal{H}\right)\prod_{j=1}^{k-1}\left(Y^{[g_j^m]}\mid Y^{[g_{j+1}^m,\ldots,g_k^m]},D,\mathcal{H}\right)\times\left(Y^{[g_k^m]}\mid D,\mathcal{H}\right), \tag{28}$$

where the three components of the conditional distributions are specified as following.

(i)

$$\left(Y^{[g_k^m]}\mid D,\mathcal{H}\right) \sim t_{m_k\times g_k p}\left(\mu^{[k]}_{(1|2)},\Phi^{[k]}_{(1|2)}\otimes\Psi^{[k]}_{(1|2)},\delta^{[k]}_{(1|2)}\right). \tag{29}$$

(ii)

$$\left(Y^{[g_j^m]}\mid Y^{[g_{j+1}^m,\ldots,g_k^m]},D,\mathcal{H}\right) \sim t_{m_j\times g_j p}\left(\mu^{[j]}_{(1|2)},\Phi^{[j]}_{(1|2)}\otimes\Psi^{[j]}_{(1|2)},\delta^{[j]}_{(1|2)}\right). \tag{30}$$

(iii)

$$\left(Y^{[u]}\mid Y^{[g_1^m,\ldots,g_k^m]},D,\mathcal{H}\right) \sim t_{n\times up}\left(\mu^{[u|g]},(\delta_0-up+1)^{-1}\Phi^{[u|g]}\otimes(\Lambda_0\otimes\Omega),\delta_0-up+1\right). \tag{31}$$

Proof. Following similar arguments as in Le, Sun, Zidek (2001)

5 Entropy Criterion

Following the arguments in Le and Zidek (1994), the entropy of the above predictive distribution can be presented as, conditional on the hyperparameter set $\mathcal{H}$,

$$H\left[Y_{unob} \mid D\right] = H\left[Y^{[u]} \mid Y^{[g_1^m,\ldots,g_k^m]}, D\right] + \sum_{j=1}^{k-1} H\left[Y^{[g_j^m]} \mid Y^{[g_{j+1}^m,\ldots,g_k^m]}, D\right] + H\left[Y^{[g_k^m]} \mid D\right].$$

Here

$$\begin{aligned} H\left[Y^{[u]} \mid Y^{[g_1^m,\ldots,g_k^m]}, D\right] &= \frac{p}{2}\log|\Lambda_0| + \frac{u}{2}\log|\Omega| + \frac{pu}{2}E(\log|\Phi^{[u|g]}|) + c_1 \\ H\left[Y^{[g_j^m]} \mid Y^{[g_{j+1}^m,\ldots,g_k^m]}, D\right] &= \frac{m_j}{2}\log|\Phi^{[j]}_{(1|2)}| + \frac{g_j p}{2}E(\log|\Psi^{[j]}_{(1|2)}|) + c_2 \\ H\left[Y^{[g_k^m]} \mid D\right] &= \frac{m_k}{2}\log|\Phi^{[k]}_{(1|2)}| + \frac{g_k p}{2}E(\log|\Psi^{[k]}_{(1|2)}|) + c_3 \end{aligned}$$

where c_1, c_2, and c_3 are constants depending on hyperparameter estimates. Here Λ_0 denotes the residual hypercovariance matrix among the ungauged stations conditional on the existing sites and Ω represents the covariance matrix among the elements of the multivariate response.

The results can be used to establish the entropy criterion. For augmenting a network, the entropy criterion is to maximize MEAS as described in Section 2 with respect to an "add" subset of ungauged sites among all possible subsets of ungauged sites. This is equivalent to select a subset of "add" sites among the potential sites so that the above $H\left[Y^{[u]} \mid Y^{[g_1^m,\ldots,g_k^m]}, D\right]$ restricted to these add sites would be maximized. Thus, the entropy criterion for augmenting the existing network is to select a subset of "add" sites to

$$\max|(\Lambda_0^{(add)}|. \tag{32}$$

6 Hyperparameter Estimation

The above entropy criterion can be used to revise a monitoring network with multivariate responses having monotone structure. The hyperparameter estimates required can be obtained using the method of moment proposed by Kibria et al (2002). The approach consists of two levels. The first would provide estimates for the hyperparameters associated with the gauged sites. The second is then to use the estimate of spatial covariance between the gauged sites to obtain an estimate for covariance matrix for all sites. That is, to extend the estimated covariance matrix to cover not only the gauged sites but also the ungauged sites. To avoid assuming isotropy and stationarity of the covariance field, we use the nonparametric approach developed

by Sampson and Guttorp (1992) for the covariance extension. More details of these approaches can be found in these references.

7 Computation

The exact optimal design in Equation 32 cannot generally be found in reasonable time since finding it is NP-hard, making suboptimal designs an attractive alternative (Ko et al 1995). Among the alternatives are the "exchange algorithms", in particular, the (DETMAX) procedure of Mitchell (1974a, b) cited by Ko et al (1995). They also cite the "greedy algorithm" of Guttorp et al (1993). At each step, the latter adds (or subtracts if the network is being reduced in size) the station that maximally improves the design objective criterion. Ko et al (1995) introduce a greedy plus exchange algorithm. The former starts with the complete set of all sites, K, and first reduces it to the required number by the greedy algorithm. It then applies an exchange algorithm to the resulting greedy network, S. Specifically, while possible, it successively exchanges site pairs, $i \in S$ and $j \in K \backslash S$ so that the objective function at $(S \backslash i) \cup \{j\}$ exceeds its values at S. Finally, Wu and Zidek (1992) propose the idea of clustering the prospective sites into suitably small subgroups before applying an exact or inexact algorithm so as to suboptimal designs that are good at least within clusters.

Exact algorithms for moderate sized problems are available. The obvious one, complete enumeration, is used in the present paper and in Guttorp et al (1993) in cases when K is not too large. Ko et al (1995) offer a more sophisticated branch and bound approach that we now describe. Using their notation, we let F denote a sub - collection of sites that must be added to the network, K being the collection of all sites. They seek to extend F to some $S \supset F$ of sites that are to be added. Finally, if certain sites, $K \backslash (E \cup F)$, are ineligible, their goal would entail finding

$$\nu(\Lambda_0, F, E, s) := \max_{\substack{S:\#(S)=s \\ F \subseteq S \subseteq E \cup F}} |\Lambda_0[S, S]| \tag{33}$$

and the associated $S = S^{optimal}$ where "$\#(S)$" stands for the number of sites in S and in general, $\Lambda_0[E', F']$ refers to the submatrix of Λ_0 with rows E' and columns F'. The algorithm requires a good initial design, S^*, obtained by the greedy algorithm, for example. This design yields as a target to beat, the initial lower bound, $LB := |\Lambda_0[S^*, S^*]|$. As well, it provides an initial "active subproblems" set $\mathcal{L} = \{\mathrm{L}\}$ consisting of just one element $L := (\Lambda_0, F, E, s)$ as well as a global upper bound $UB := b(\Lambda_0, F, E, s)$. For b, Ko et al (1995) find

$$b(L) := |\Lambda_0[F, F]| \Pi_{i=1}^{s-f} \lambda_i(\Lambda_{0[E \cdot F]}), \tag{34}$$

where $\Lambda_{0[E \cdot F]} = \Lambda_0[E] - \Lambda_0[E, F]\Lambda_0[F, F]^{-1}\Lambda_0[F, E]$ and the $\{\lambda_i\}$ denote the ordered eigenvalues of its matrix argument in decreasing order, $\lambda_1 \geq \cdots \lambda_{s-f}$ with $f = \#(F)$.

At a general step in the execution of the algorithm, the LB would correspond to the best design S^* obtained to that step. At the same time, $\mathcal{L}$ would have a multiplicity of elements and the required global upper bound would be $UB := \max_{\mathrm{L}\in\mathcal{L}} b(\mathrm{L})$. $UB > LB$ suggests that $S^{optimal}$ has not been reached and new branches need to be explored in search of the optimum, *i.e.* new active subproblems need to be added to the $\mathcal{L}$. We do this by first deleting an active subprogram (Λ_0, F', E', s) from $\mathcal{L}$ and then selecting a "branching index", $i \in E'$. Four (non-distinct) cases obtain and determine which subproblems to add. First, one of Cases, (i)] $\#(F') + \#(E') - 1 > s$ or (ii) $\#(F') + \#(E') - 1 = s$ obtains. If (i),add $(\Lambda_0, F', E'\backslash i, s)$ to $\mathcal{L}$ and compute $b(\Lambda_0, F', E'\backslash i, s)$ (needed to find the new UB). If (ii) $S := F' \cup E'\backslash i$ is the only feasible solution. If $\Lambda_0[S, S] < LB$, S supplants the current S^* and LB moves up to $LB := \Lambda_0[S, S]$. Next, one of Cases, (iii) $F' + 1 < s$ or (iv) $F' + 1 = s$ prevails. If (iii) add $(\Lambda_0, F' \cup \{i\}, E'\backslash i, s)$ to $\mathcal{L}$ and compute $b((\Lambda_0, C(\Lambda_0, F' \cup \{i\}, E'\backslash i, s), s)$. If (iv), $S = F' \cup \{i\}$, the only available feasible solution can supplant the current S^* and move LB (computed as above) even higher. Finally, recompute the UB and determine whether or not the program has terminated with $UB \leq LB$. If not, we would delete another active subproblem, create new branches and carry on as long as possible.

Ko et al (1995) show their algorithm to be much quicker than complete enumeration. Jon Lee suggests (personal communication) that problems with site totals of about 80 can be routinely tackled. No doubt, by improving UB's and methods of selecting the active problems for deletion, further increases in the algorithm's domain are possible. Nevertheless, for realistic continent wide redesign problems having 100's or even 1000's of prospective sites, exact optimization seems out of the question. Therefore, we are encouraged by the finding of Ko et al (1995) that the greedy/swap algorithm described above often produced the exact optimum, where the latter is computable.

The branch and bound algorithm can be extended in various ways. Bueso et al (1998) extend it to the case where observations are made with error and the goal is the prediction not only of responses at ungauged sites but those at the gauges sites as well. Lee (1998) extends it to incorporate linear constraints (*e.g.* limiting cost. His approach defers from the approach of Zidek et al (2000) where cost is also incorporated.

8 Illustrative Example

This section demonstrates use of the above design theory, by redesigning Greater Vancouver's PM_{10} network. When our analysis was done, that network had 10 stations measuring hourly PM_{10} levels with different start dates, resulting in a staircase data pattern. Each step of the staircase consists of stations having the same starting time. Figure 1 shows the names of the stations along with their start dates and the boxplots of the hourly PM_{10} measurements.

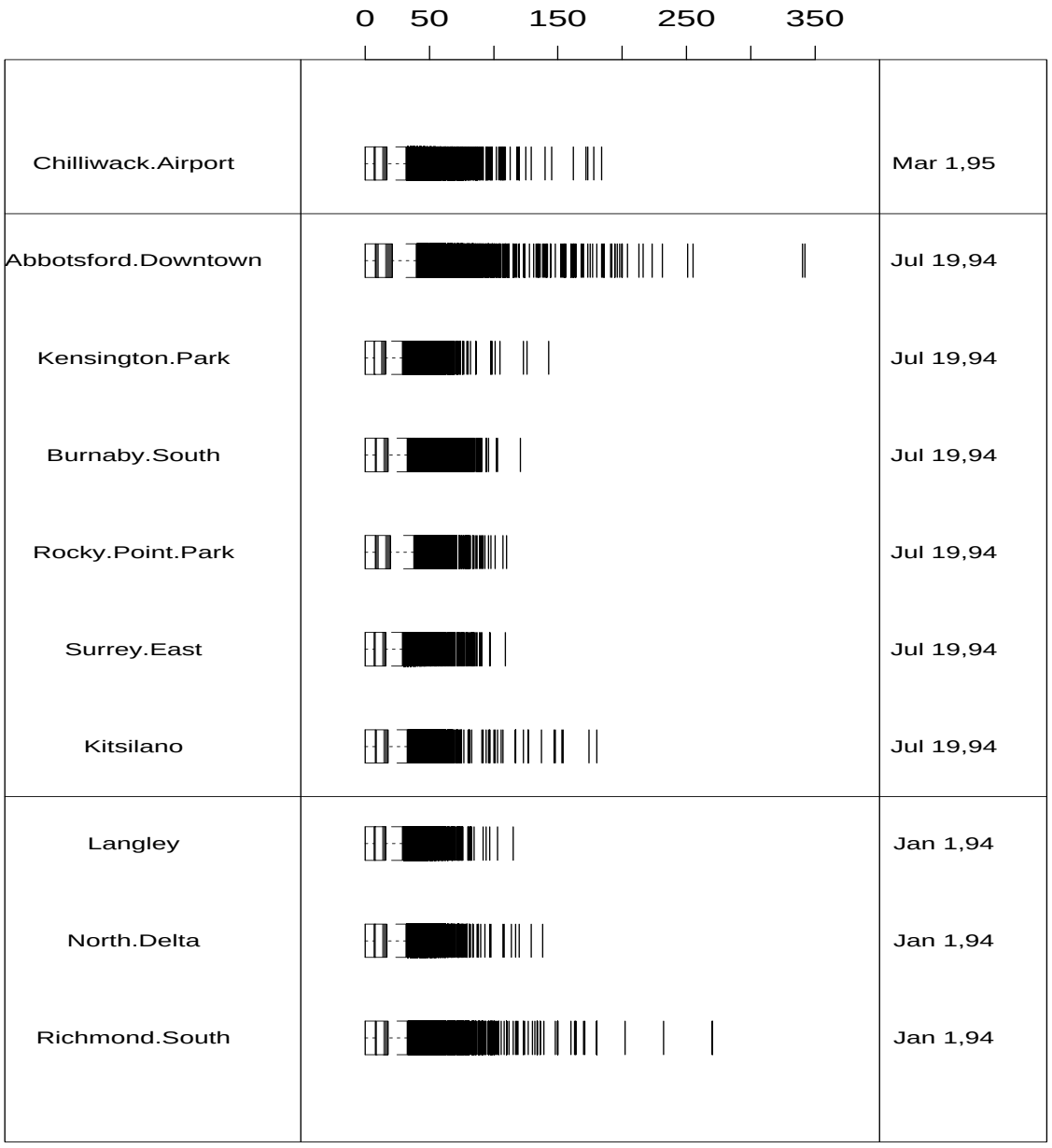
0
50
150
250
350
Chilliwack.Airport
Mar 1,95
Abbotsford.Downtown
Jul 19,94
Kensington.Park
Jul 19,94
Burnaby.South
Jul 19,94
Rocky.Point.Park
Jul 19,94
Surrey.East
Jul 19,94
Kitsilano
Jul 19,94
Langley
Jan 1,94
North.Delta
Jan 1,94
Richmond.South
Jan 1,94

Figure 1: Boxplot of hourly PM_{10} levels ($\mu g/m^3$) at 10 monitor sites in Greater Vancouver and their start-up times.

In this example, the objective is to augment the existing network with an optimal subset of 6 stations among 20 potential sites. Locations of the existing stations and potential sites are displayed in Figure 2.

The trend for the log-transformed hourly PM_{10} levels is modelled with seasonal components, hourly and daily effects, and meteorological covariates. The seasonal components are captured by the sine and cosine functions for monthly, semi-annual, and annual cycles. The meteorological data used include visibility index, sea level pressure, dew point temperature, wind speed, rain and relative humidity.

The 24-dimensional vectors of hourly PM_{10} measurements for each day obtain from the detrended series. The day-to-day autocorrelation are filtered out by fitting a multivariate AR(1) model. The resulting residuals form the multivariate responses used in Model (1) of Section 3.2.

The hyperparameters are estimated using the method moment proposed by Kibria et al (2002). Table 1 shows the estimated hypercovariance matrix between gauged stations, Λ_g is displayed.

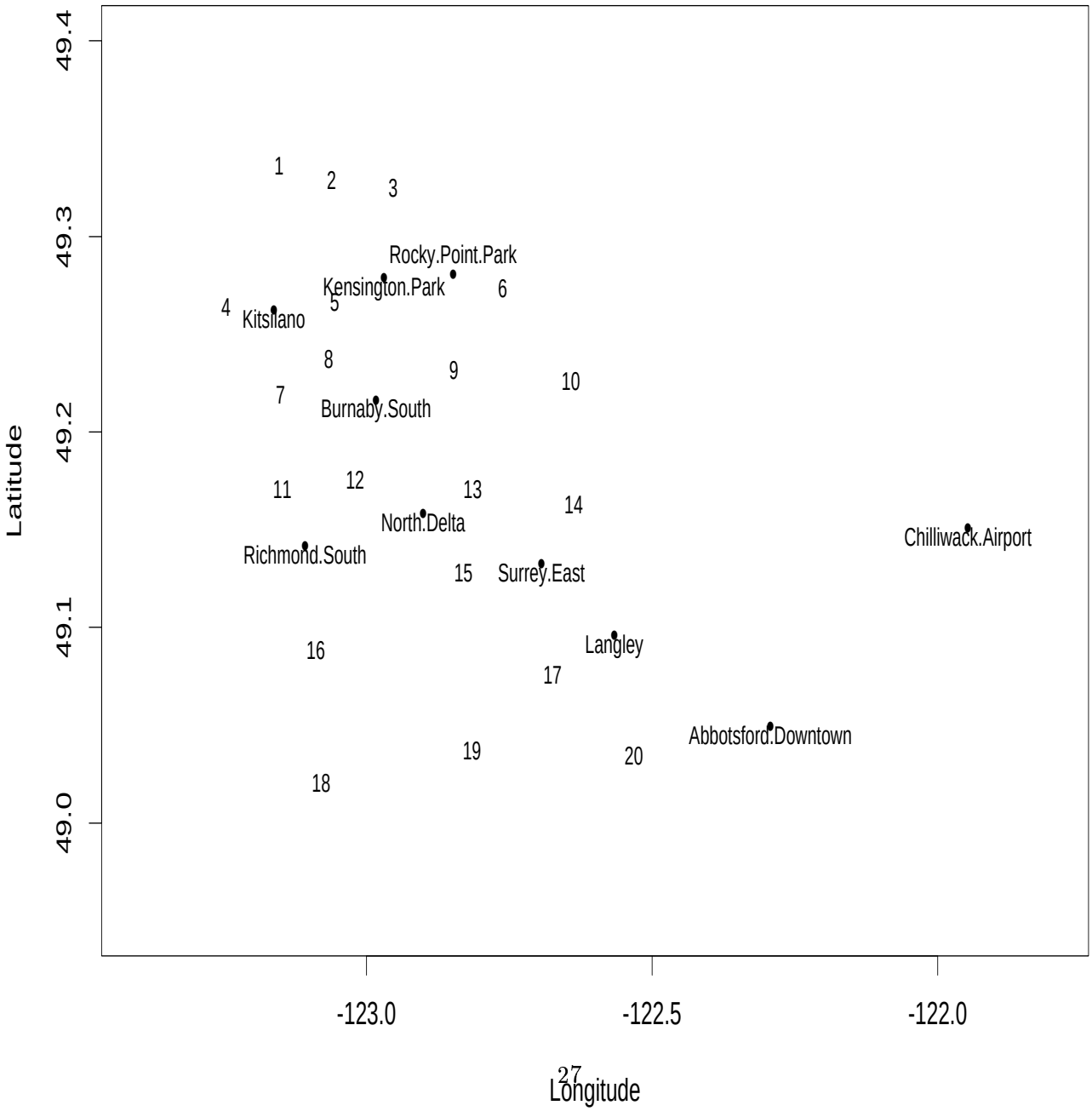
Latitude
49.4
49.3
49.2
49.1
49.0
-123.0
-122.5
-122.0
Longitude
1
2
3
Rocky.Point.Park
Kensington.Park
4
5
6
Kitsilano
8
9
10
7
Burnaby.South
11
12
13
14
North.Delta
Chilliwack.Airport
Richmond.South
15
Surrey.East
16
Langley
17
Abbotsford.Downtown
19
20
18

Figure 2: PM_{10} Monitor Stations and Potential Sites.

Chilliwack airport	75	46	38	40	34	38	27	40	36	29
Abbotford downtown	46	79	36	46	33	41	29	40	38	31
Kensington Park	38	36	70	55	47	42	38	38	46	34
Burnaby South	40	41	55	77	48	49	43	43	55	43
Rocky Point Park	34	33	47	48	64	39	36	34	41	33
Surrey East	38	41	42	49	39	64	34	45	47	36
Kitsilano	27	29	38	43	36	34	59	30	37	41
Langley	40	40	38	43	34	45	30	64	41	31
North Delta	36	38	46	55	41	47	37	41	72	39
Richmond South	29	31	34	43	33	36	41	31	39	65

Table 1: **Estimated hypercovariance matrix at existing stations after multiplying entries by 100.**

The residual hypercovariance matrix between potential (ungauged) sites conditional on the existing sites, Λ_0, is estimated using the Sampson and Guttorp (SG) method (Sampson & Guttorp 1992) and based on the estimated Λ_g. That method bypasses the usual requirement of covariance spatial isotropy by smoothly mapping location coordinates in "geographic space" (G - space) into coordinates of another called "dispersion space" (D-space). The mapping is constructed so that the covariance field is isotropic over D - space even if not over G - space. Construction of the mapping starts with the observed spatial covariance matrix for existing monitoring sites. A fitted variogram, or equivalently covariance function, in D-space combined with the smooth mapping can then be used to obtain spatial covariances between all sites in G - space, including those at which no measurements were previously made.

Figure 3 demonstrates the actions of the SG method in this application. The right panel shows the corresponding D-space coordinates, resulting from applying the mapping function to a biorthogonal grid in G-space. The left panel shows the fitted variogram in D-space. The figure shows a good fit for the variogram model using this mapping function with spline smoothing parameter of 2. Users specify this built-in map smoothing parameter that controls the distortion between the G-space and the D-space. This feature ensures that the grid is not folded in the D-space and hence maintains the spatial interpretability of the correlations; that is, correlations are reflected in inter-site distances in dispersion space. The deformation on the right panel indicates the non-stationarity of the field. Failure to capture non-stationarity results in a suboptimal design as illustrated in the analysis below.

1	39	18	9	5	5	1	2	5			1	2								
2	18	40	16	2	8	2	3	7	-1	-1	1	3	-1	-1						
3	9	16	35	1	6	5	2	4				2								
4	5	2	1	36			2				5					2		1		
5	5	8	6	37		6	21	-1	-1	2	10	-1	-1	2						
6	1	2	5			27			4	6			1	1						
7	2	3	2	2	6		36	10	-1	-1	9	11	-1	-1	3	1	1	1		
8	5	7	4	21		10	40	-1	-1	3	16	-1	-1	3		1		1		
9		-1			-1	4	-1	-1	38	17		-1	9	10	-1		-1			
10		-1			-1	6	-1	-1	17	52		-1	7	13	-1		-1		-1	
11	1	1		5	2		9	3			32	5			2	2	1	1	2	
12	2	3	2		10		11	16	-1	-1	5	42	-1	-1	9	1	3	1	3	1
13		-1			-1	1	-1	-1	9	7		-1	35	13			-1		-1	
14		-1			-1	1	-1	-1	10	13		-1	13	44	-1		-1		-1	-1
15					2		3	3	-1	-1	2	9		-1	35	2	8	2	6	3
16				2			1				2	1			2	47	3	17	9	2
17							1	1	-1	-1	1	3	-1	-1	8	3	44	3	14	13
18				1							1	1			2	17	3	60	13	3
19							1	1		-1	2	3-1	-1	6	9	14	13	60	13	
20												1		-1	3	2	13	3	13	58

Table 2: The estimated residual hypercovariance between ungauged sites conditional on existing stations. Entries have been multiplied by 100 and rounded to integers and 0's are replaced by blanks.

The panels in Figure 3 can be used to estimate spatial correlations between any points in the G-space, e.g. by first identifying the points in D-space using the grid, then measuring the distance in D-space between them, and finally applying the fitted variogram to the distance to estimate their spatial correlations. The residual hypercovariance matrix among the ungauged site conditional on existing stations , Λ_0, is estimated accordingly and displayed in Table 2.

Applying the entropy criterion given in (32) to the estimated Λ_0 yields the optimal subset of 6 sites, { Sites: 10, 12, 16, 18, 19, 20 } among the 20 potential sites, to augment the existing network. The locations of the selected sites are depicted in Figure 4 along with the locations

of existing stations and potential sites, the latter accompanied by their ranking based on their estimated hyper-variances (ie. the diagonal element of Λ_0).

The optimum solution seems sensible in that five of the 6 sites, {Sites: 10, 16, 18, 19, 20} have 5 largest variances and are generally far away from existing stations. However, note that the sixth selected site, Site 12, has smaller estimated hyper-variance than two unselected ones (Sites: 14 and 17). The trade-off between variance and correlation with nearby stations is demonstrated here. Sites 14 is not selected in spite of its having a large estimated variance because it is closer to existing stations than Site 12. Furthermore, Site 14 is located in a region of stronger spatial correlation than that of Site 12, as indicated by the stretching in the region containing Site 12 on Figure 3's right panel . The non-stationarity of this field also plays an important role in the selection of Site 12 over Site 17. The two sites are roughly the same distance from existing stations, Site 17 having larger estimated hyper-variance; however, the spatial correlation is weaker at Site 12 than at Site 17. This fact can be clearly seen in the right panel of Figure 3 where the region containing Site 17 does not show any stretching, in comparison with the region containing Site 12 showing more stretching and hence less spatial correlation for the same distance in G-space.

9 Discussion

In this paper we have provided a hierarchical Bayesian framework for enlarging an existing monitoring network whose current stations have been added over time, creating a staircase data pattern. The context is that in which there is no well defined design objective. This is done by adopting the generic objective of minimizing the entropy of the posterior probability distribution of the quantities of interest. Roughly speaking the new network stations are those whose response vector would be the most highly unpredictable either because it is not well correlated with the remaining stations, or because of its high intrinsic variability. The results seem to indicate that the proposed theory works well and is able to capture the non-stationary feature of the spatial field.

We should emphasize that as a compromise, our entropy based approach will not yield an optimal design for specific objectives. In fact, in a paper currently in preparation, we will show limitations of the entropy design when we are interested in monitoring the extreme values of the time series of responses at the spatial sites. However, it needs to be said that the latter is a very challenging problem and we are not aware of any entirely satisfactory where "extreme extremes" are of concern.

The proposed design methodology uses a posterior distribution with its entropy for a multivariate random field. The posterior distribution is obtained from a multivariate Gaussian

model for the field, with a conjugate prior distribution for its parameters, including the generalized inverted Wishart distribution for the spatial covariance matrix, capable of coping with the staircase data pattern. At level two, a Kronecker product is adopted as a covariance model to capture the between station (spatial) and within station (temporal) correlations. That product will clearly not be appropriate in certain applications, but to date, no suitable alternative has been found.

There are a number of practical issues which need to be addressed in implementing our method. It must be possible to transform the data so that the data distribution is approximately Gaussian. Adjustment with this model must yield an approximately uncorrelated data series. Computation is a major practical consideration. Here we have deliberately restricted the number of potential sites involved in our analysis so that complete combinatorial optimization is feasible. But the numerical problem of combinatorial optimization rapidly becomes overwhelming as the size of the existing network increases in size.

This relates directly to an issue raised in Section 3. There we noted that our uncertainty about ungauged sites can be reduced not only by borrowing information from current measurements at gauged sites but to a lesser extent from previous such measurements as well, at least when the autocorrelation in the individual series is sufficiently strongly. However, incorporating that component of the model seems to lead to an intractable entropy calculation and in turn to the computational problem indicated in the last paragraph. This issue is being addressed in current work.

Another issue raised in that section concerns data missing at random. Our space time model there allows for such data in the past. However, the design refers to future values of the response field, and a satisfactory design approach needs to model the random data selection process that determines which measurements will actually be obtained when the design is used.

Finally, we recognize that in practice, "good" rather than "optimal" designs are needed and optimal designs like those in this paper must be considered as tentative proposals susceptible to modification depending on the circumstances prevailing in the context of their implementation. These "optimal" designs may well be valuable starting points, however, since they can be explicated in terms of their axiomatic underpinnings and proposed changes to these optimal designs can be interpreted in terms of the axioms. This can provide a degree of confidence and clarity in the typically complex situation confronting a designer. The redesigns are founded on a coherent theory and we believe that the resulting designs will therefore be defensible in an operational context.

Acknowledgements. This material was based upon work partially supported by the Natural Science and Engineering Research Council of Canada and partially by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

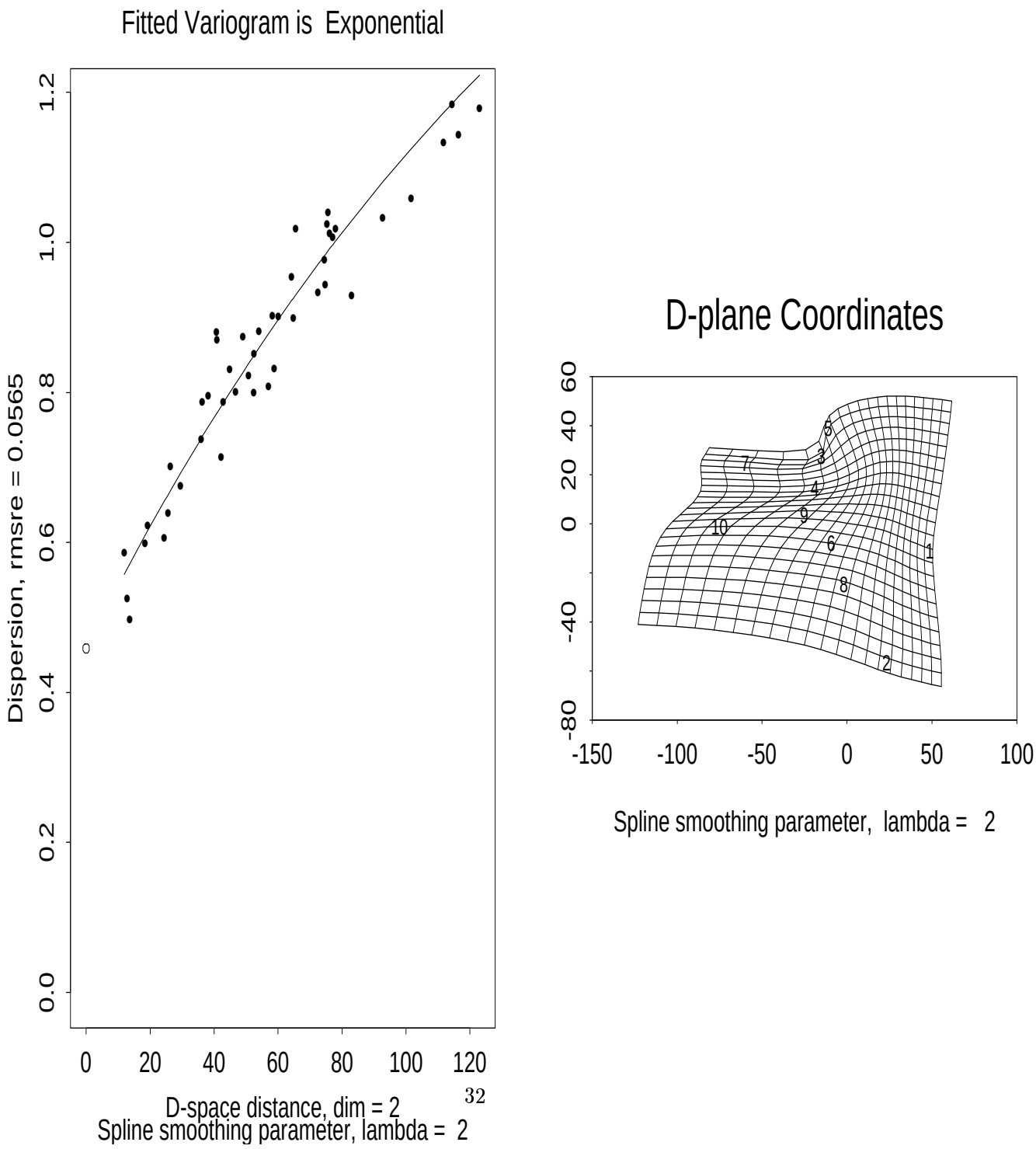

Figure 3: Transforming the geographic plane to dispersion space (right panel) and fitting a variogram to the empirical variogram over dispersion space for Vancouver's hourly PM_{10} field.

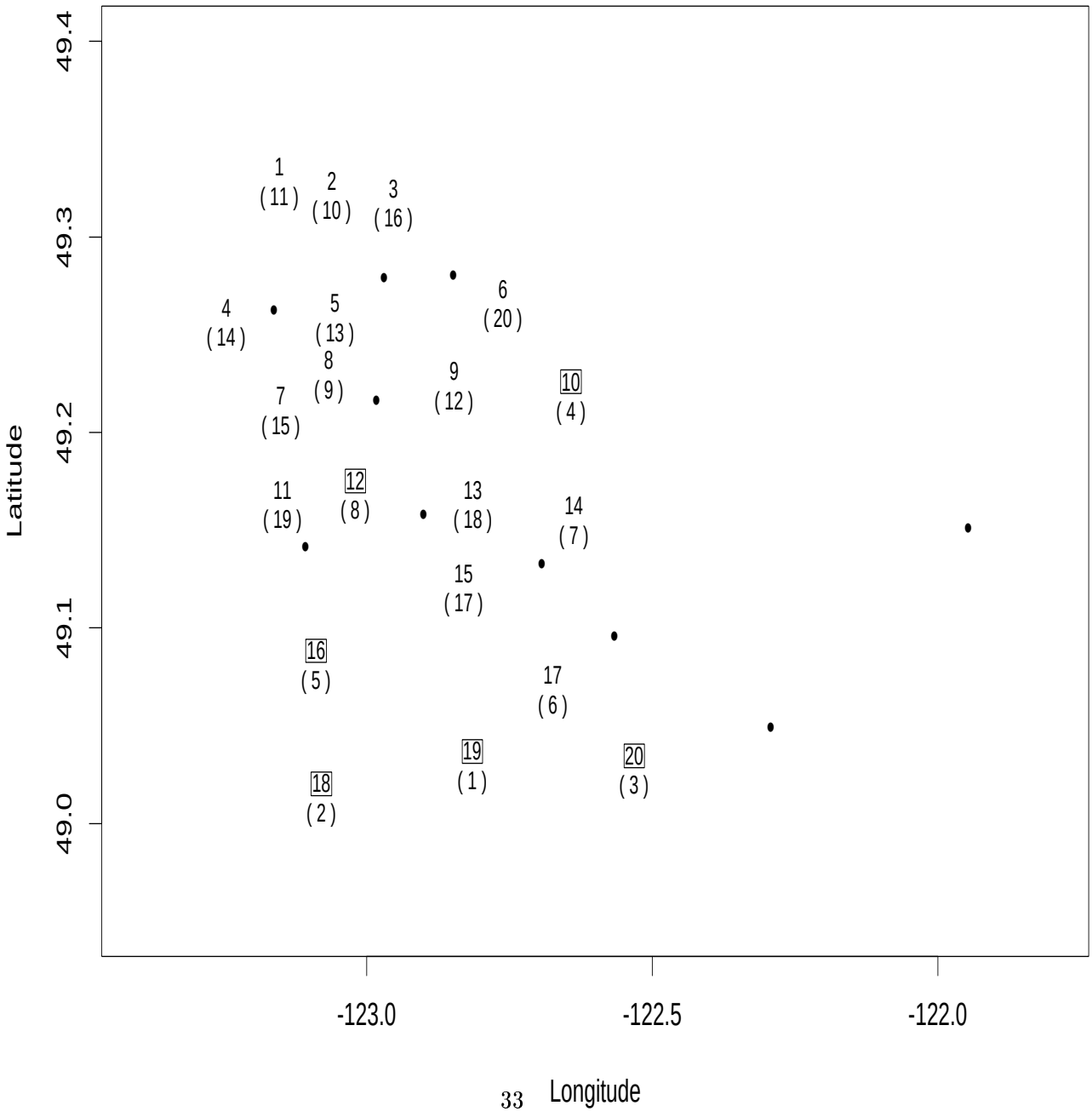

Figure 4: Locations of existing network and potential sites with their rank based on estimated variance (in brackets); selected sites are marked with square.

References

[1] Anderson, TW (1984). *An Introduction To Multivariate Statistical Analysis.* (2nd Ed.) Wiley.

[2] Bartlett, MS (1933). On the theory of statistical regression. *Proc. Roy. Soc. Edenburgh* 53:260-283

[3] Brown, PJ (1993). *Measurement, Regression, and Calibration.* Oxford: Clarendon.

[4] Brown, PJ, Le, ND and Zidek, JV (1994a). Multivariate spatial interpolation and exposure to air pollutants *Can Jour Statist* 22:489-510.

[5] Brown, PJ, Le, ND and Zidek, JV (1994b). Inference for a covariance matrix. In Smith, AFM and Freeman, PR eds. *Aspects of Uncertainty: A Tribute to D.V. Lindley.* Wiley.

[6] Carrol, RJ, Ruppert, D and Stefanski, D (1995). *Measurement Error in Nonlinear Models.* London: Chapman and Hall.

[7] Caselton, WF and Husain, T (1980). "Hydrologic networks: information transmission." *Water Resources Planning and Management Division, A.S.C.E*, 106(WR2), 503–520.

[8] Caselton, WF and Zidek, JV (1984). "Optimal monitoring network designs." *Statist. Prob. Letters,* 2:223–227.

[9] Caselton, WF, Kan, L and Zidek, JV (1992). " Quality Data Networks that Minimize Entropy". In *Statistics in the Environmental and Earth Sciences,* eds. P. Guttorp and A. Walden. London: Griffin.

[10] Chen, CF (1979). Bayesian Inference for a Normal Dispersion Matrix and its Application to Stochastic Multiple Regression Analysis. *Journal Royal Statistical Society, Series B* 41:235-48.

[11] Cressie, NC (1991). *Statistics For Spatial Data.* Wiley.

[12] De Oliveira, V, Benjamin, K and Short DA (1997). Bayesian prediction of transformed Gaussian random fields. *Jour Amer Statist Assoc*, 92, 1422-1433.

[13] Dempster, AP, Laird, NM and Rubin, DB (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal Royal Statistical Society, Series B.* 39:1-38.

[14] Duddek, C, Le, ND, Zidek, JV and Burnett, RT (1995). Multivariate imputation in cross sectional analysis of health effects associated with air pollution (with discussions). *Journal of Environmental and Ecological Statistics*. 2:191-212.

[15] Elfving, G (1952). "Optimum allocation in linear regression theory." *Ann. Math. Statist.*, 23, 255–262.

[16] Federov, V and Muller, W (1988). "Two approaches in optimization of observing networks." In *Optimal Design and Analysis of Experiments* , Y.Dodge, V.V. Fedorov, and H.P.Wynn (Eds.), Elsevier Science Publishers B.V..

[17] Federov, V and Muller, W (1989). "Comparison of two approaches in the optimal design of an observation network". *Statistics*, **3**, 339-351.

[18] Gaudard, M, Karson, M, Linder, E and Sinha, D. (1999). Bayesian spatial prediction (with discussion). *Ecological and Environmental Statistics*.

[19] Guttorp, P, Le, ND, Sampson, PD and Zidek, JV (1993). "Using entropy in the redesign of an environmental monitoring network". In *Multivariate Environmental Statistics* edited by G.P Patil, C.R. Rao and N.P. Ross. North Holland/Elsevier Science, New York:175-202.

[20] Jaynes, ET (1963). "Information theory and statistical mechanics", *Statistical Physics* , 3, Ford, K.W. (ed.), Benjamin, New York, 102–218.

[21] Handcock, MS and Stein, ML (1993). A Bayesian analysis of kriging. *Technometrics*, 35, 403-410.

[22] Handcock, MS and Wallis, J (with discussion) (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, 89:267-378; rejoinder:388-390.

[23] Haas TC (1996). Multivariate spatial prediction in the presence of nonlinear trend and covariance nonstationary. *Environmetrics*, 7:145-165.

[24] Keifer, J (1959). "Optimum experimental design." *J.R. Statist. Soc. B.*, 21, 272–319.

[25] Kibria GBM, Sun L, Zidek V, Le, ND. (2002). Bayesian spatial prediction of random space-time fields with application to mapping $PM_{2.5}$ exposure. J Amer Stat Assoc, 457(97):101-112.

[26] Le, ND and Zidek, JV (1992). Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. *Journal of Multivariate Analysis* 43:351-74.

[27] Le, ND and Zidek, JV (1994). Network designs for monitoring multivariate random spatial fields. in *Recent Adv in Stat and Prob*; edited by JP Vilaplana and ML Puri 191-206.

[28] Le, ND, Sun, W and Zidek, JV (1997). Bayesian multivariate spatial interpolation with data missing-by-design. *Jour Roy Statist Soc, Series B* 59:501-510.

[29] Le, ND, Sun, L and Zidek, JV (1998). A note on the existence of maximum likelihood estimates for Gaussian-Inverted Wishart models. *Stat & Prob Letters* 40:133-137.

[30] Le, ND, Sun, L and Zidek, JV (2001). Spatial prediction and temporal backcasting for environmental fields having monotone data patterns. *Canadian Journal of Statistics* 29(4):516-529.

[31] Li, KH, Le, ND, Sun, L and Zidek, JV (1999). Spatial-Temporal Models for Ambient Hourly PM_{10} in Vancouver. *Environmetrics*, 10:321-338.

[32] Lindley, DV (1956). "On the measure of the information provided by an experiment." *Ann. Math. Statist.* , 27, 968–1005.

[33] Linthurst, RA, Landers, DH, Eilers, JM, Brakke, DF, Overton, WS, Meier, EP and Crowe, RE (1986). *Characteristics of Lakes in the Eastern United States. Volume 1. Population Descriptions and pHysico–Chemical Relationships.* EPA/600/4-86/007a, U.S. Environmental Protection Agency, Washington, DC, 1986, 136 pp.

[34] Little, RJA and Rubin, DB (1987). *Statistical Analysis With Missing Data*, New York: John Wiley.

[35] Liu C (1996). Bayesian robust multivariate linear regression with incomplete data. *Jour Amer Statist Assoc*, 91:1219-1227.

[36] Liu C (1999). Efficient ML estimation of the multivariate normal distribution from incomplete data. *Journal of Multivariate Analysis*, 69:206-217.

[37] Loader, C and Switzer, P (1989). "Spatial Covariance Estimation For Monitoring Data". SIMS Technical Report 133. Stanford University.

[38] Matheron, G (1971). *The theory of regionalized variables and its applications"*. Centre de Geostatistique, Ecole des Mines and Paris, 212pp.

[39] Omre, H (1987). "Bayesian Kriging–merging observations and qualified guesses in Kriging". *Math. Geology*, **19**, 25–39.

[40] Omre, H and Halvorsen, KB (1989a). "The Bayesian bridge between simple and universal Kriging". *Math. Geology*, **21**, 767–786.

[41] Omre, H, Halvorsen, KB and Bertig, V (1989b). "A Bayesian approach to Kriging". In *Geostatistics, 1*, Ed. M. Armstrong, Klumer Academic Publishers

[42] Pierce, DA, Stram, DO, Vaeth, M and Schafer, DW (1992). The errors in variables problem: considerations provided by radiation dose-response analyses of A-bomb survivor data. *Journal of the American Statistical Association* 87:351-359.

[43] Press, JS (1972). *Applied Multivariate Analysis.* Holt, Rinehart and Winston, Inc.

[44] Rubin, DB and Shaffer JL (1990). Efficiently creating multiple imputations for incomplete multivariate normal data, in *Proceedings of the Statistical Computing Section, American Statistical Association*, 83-88.

[45] Sampson, PD and Guttorp, P (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association* 87:108-19.

[46] Sebastiani, P, and Wynn, HP (2000). Maximum entropy sampling and optimal Bayesian experimental design. *J.R. Statist. Soc. B,* 62:145-157.

[47] Shewry, M, and Wynn, H (1987). "Maximum entropy sampling." *Jour. Applied Statist.*, 14, 165–207.

[48] Silvey, SD (1980). *Optimal Design* . London: Chapman and Hall.

[49] Smith, K (1918). "On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and guidance they give towards a proper choice of the distribution of observations." *Biometrika* , 12, 1–85.

[50] Sun, W (1998). Comparison of a co-Kriging method with a Bayesian alternative. *Environmetrics*: 9:445-457.

[51] Sun, W, Le, ND, Zidek JV, and Burnett, R (1998). Assessment of a Bayesian multivariate interpolation approach for health impact studies. *Environmetrics*: 9:565-586.

[52] Verdinelli, I (1991). "Advances in bayesian experimental design." Date+ pages??

[53] Wu, S and Zidek, JV (1992). "An entropy based review of selected NADP/NTN network sites for 1983–86". *Atmospheric Environment.* Date+ pages??

[54] Zidek, JV (1997). Interpolating air pollution for health impact assessment. *Statistics for the Environment 3: Pollution Assessment and Control.* Edited by V Barnett and K Feridun Turkman. New York: Wiley 251-268.

[55] Zidek, JV, Sun, W and Le, ND. (2000). Designing and integrating composite networks for monitoring multivariate gaussian pollution fields. *Appl. Statist.*, 49:63-79.