Designing and Integrating Composite Networks for Monitoring Multivariate Gaussian Pollution Fields

Author(s): James V. Zidek, Weimin Sun and Nhu D. Le

Source: *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 49, No. 1 (2000), pp. 63-79
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2680861
Accessed: 18-05-2019 10:23 UTC

Appl. Statist. (2000)
49, *Part* 1, *pp.* 63–79

Designing and integrating composite networks for monitoring multivariate Gaussian pollution fields

James V. Zidek,

University of British Columbia, Vancouver, Canada

Weimin Sun

Statistics Canada, Ottawa, Canada

and Nhu D. Le

University of British Columbia, Vancouver, and British Columbia Cancer Agency, Vancouver, Canada

[Received September 1997. Final revision April 1999]

Summary. Networks of ambient monitoring stations are used to monitor environmental pollution fields such as those for acid rain and air pollution. Such stations provide regular measurements of pollutant concentrations. The networks are established for a variety of purposes at various times so often several stations measuring different subsets of pollutant concentrations can be found in compact geographical regions. The problem of statistically combining these disparate information sources into a single 'network' then arises. Capitalizing on the efficiencies so achieved can then lead to the secondary problem of extending this network. The subject of this paper is a set of 31 air pollution monitoring stations in southern Ontario. Each of these regularly measures a particular subset of ionic sulphate, sulphite, nitrite and ozone. However, this subset varies from station to station. For example only two stations measure all four. Some measure just one. We describe a Bayesian framework for integrating the measurements of these stations to yield a spatial predictive distribution for unmonitored sites and unmeasured concentrations at existing stations. Furthermore we show how this network can be extended by using an entropy maximization criterion. The methods assume that the multivariate response field being measured has a joint Gaussian distribution conditional on its mean and covariance function. A conjugate prior is used for these parameters, some of its hyperparameters being fitted empirically.

Keywords: Air pollution; Data missing by design; Entropy; Hierarchical Bayes method; Inverted Wishart distribution; Kriging; Monitoring network; Nonparametric covariance estimation; Optimal design; Ozone; Spatial prediction; Sulphate

1. Introduction

In this paper, a companion to that of Le *et al.* (1997), we show how to apply the Le–Sun–Zidek (LSZ) theory in designing environmental monitoring networks. We describe this theory in Section 2. Then we demonstrate its applicability in Section 3 by extending a composite network of ambient air pollution monitoring stations in southern Ontario.

In general, ambient monitoring networks are set up to provide regular measurements of random environmental fields of societal importance. To fix ideas we consider a comparatively simple network of 20 continuous ambient air quality monitoring stations maintained by the Greater Vancouver Regional District (GVRD); see Greater Vancouver Regional District

Address for correspondence: James V. Zidek, Department of Statistics, University of British Columbia, 6356 Agriculture Road, Vancouver, British Columbia, V6T 1Z2, Canada.
E-mail: jim@stat.ubc.ca

 0035–9254/00/49063

(1996). These stations transmit hourly data to an air quality monitoring system computer database. Local air quality can then be compared against national and provincial guidelines. (In the paper we refer to locations (e.g. building roof-tops) of ambient monitoring stations as 'gauged sites'. Numerous other sites are potentially available for the installation of other stations. We call them 'ungauged' sites.)

Each of these 20 gauged sites in the GVRD network has seven positions at which monitors or 'gauges' are installed, one for each of the seven fields being measured (e.g. micrograms of sulphur dioxide per cubic metre). As a purely conceptual device for explaining our theory we call the positions with monitors 'gauged pseudosites'. However, we do not alter in any substantive way the special relationship that the pseudosites have with their partners at any given site even though they are formally treated as though they were different sites. That relationship is preserved in both the multivariate models we adopt in the next section as well as in the estimated covariances themselves.

At the ungauged sites all seven positions to which monitors could be attached are vacant. We call the vacant positions 'ungauged pseudosites'. Although, in the GVRD network, monitors have been installed at every pseudosite, in other networks like that to be redesigned in this paper some gauged sites have some vacant pseudosites. In other words some gauged sites have both gauged and ungauged pseudosites on them. Then the designer can put gauges at an ungauged site or put additional monitors on ungauged pseudosites at gauged stations.

We now turn to the network to be redesigned in this paper. In reality this network consists of the union of three monitoring networks established at various times for various purposes:

(a) the environment air quality monitoring network;
(b) the air pollution in Ontario study, APIOS;
(c) the Canadian acid and precipitation monitoring network described by Burnett *et al.* (1994), CAPMON.

By generalizing the theory of Brown *et al.* (1994), Le *et al.* (1997) showed how to integrate these three networks statistically.

To understand how the need to integrate networks arises, a brief history is informative. Reflecting concerns of the day, both APIOS and CAPMON were established with the initial purpose of monitoring acidic precipitation (see Ro *et al.* (1988) and Sirois and Fricke (1992) for details).

CAPMON's history is particularly instructive (see Sirois and Fricke (1992)). It began monitoring in 1978 with just three sites in remote areas. In 1983 its size increased when the acid and precipitation network was merged with it; the merged network came to be used for a second purpose, tracing source–receptor relationships; monitoring sites could then be found closer to urban areas. More recently a third purpose for the network has been identified: that of discovering relationships between air pollution and human health (Burnett *et al.*, 1994; Zidek *et al.*, 1998).

The composite network now monitors hourly levels of nitrogen dioxide (in micrograms per cubic metre), ozone (O_3) in parts per billion, sulphur dioxide in micrograms per cubic metre and the sulphate ion (SO_4) (in micrograms per cubic metre). For simplicity, we consider only two pollutants: ozone and sulphate. These are the pollutants that Zidek *et al.* (1998) found to be associated with respiratory morbidity.

Since two responses are being considered we are dealing with a $k = 2$-dimensional random vector field. Our data were obtained by measuring this field from January 1st, 1983, to December 31st, 1988, in southern Ontario and its surrounding areas. However, data quality

considerations led us to use data from just 31 of the 37 stations, but not all had been equipped with monitors for both pollutants.

Abstractly, we may view responses as random column vectors of fixed length (2 in our example). They are indexed by site $s \in \mathcal{S}$ and time $t \in \mathcal{T}$, $\mathcal{S}$ and $\mathcal{T}$ being finite sets. Each co-ordinate of the column vector for site s and time t represents the response at one of the pseudosites. In our example $\mathcal{S}$ has 46 elements; in addition to the 31 gauged or partially gauged sites we posit 15 additional ungauged sites to which one or two monitors could be attached in a redesign.

The idea of the pseudosite simplifies the formulation of our design problem. It enables us to conceive the two positions at any given site as if they were two different sites. As a result we can string out the responses for all 46 sites into a $2 \times 46 = 92$-dimensional vector corresponding to all the pseudosites. By arranging these co-ordinates appropriately we can partition the column vector of responses into subvectors of those that are measured (at the gauged pseudosites) and those that are not (at the ungauged pseudosites).

Responses at the ungauged pseudosites yield no data. When some of the ungauged pseudosites are actually at the gauged sites, Le *et al.* (1997) called such data *missing by design*. In contrast, gauged pseudosites yield time series of measurements. Some of these may be *missing at random* because of equipment failure for example. These missing measurements are imputed in our analysis and ignored in our exposition.

The problem that we address is that of deciding which if any unmeasured responses should henceforth be measured. The corresponding pseudosites can be at either gauged or ungauged sites. The latter entail large site preparation costs. We estimate these to be \$40000 (in 1997 Canadian dollars) depending on such factors as location. The monitors cost about \$15000 each. Let us spread these initial costs over 5 years and ignore inflation. Then the site preparation costs will add \$40000/5, i.e. \$8000 to the annual cost. The analogous figure for each monitor would be \$3000.

We estimate the annual operating costs as \$5000 for the first and \$500 for each additional monitor placed at a site. A previously ungauged site to which both O_3 and SO_4 monitors are attached would therefore have an annual cost of about $\$8000 + 2 \times \$3000 + \$5000 + \$500 = \$19500$ per annum.

How should potential gains in information be traded off against the cost of installing gauges at ungauged pseudosites? These questions are answered in this paper through the example described above and the theory described below. The theory is an adaptation for design of the LSZ theory.

2. Information and the network design problem

'Redesigning a network' can mean adding to it or deleting from it gauged pseudosites (Caselton *et al.*, 1992; Haas, 1992; Wu and Zidek, 1992; Guttorp *et al.*, 1993; Le and Zidek, 1994). However, in our application it will mean their addition. The design theory that we use is based on ideas from information and entropy theory integrated into a Bayesian framework that allows 'uncertainty' to be handled in a natural way. In particular that framework allows us to incorporate our uncertainty about model parameters into the process of selecting a design.

Uncertainty is reduced by measurement. It completely eliminates uncertainty about the associated responses at gauged pseudosites (assuming, as we do throughout this paper, negligible measurement errors). Through spatial association or correlation, it reduces uncertainty about responses at ungauged pseudosites and parameters of stochastic response

models. Design theory exploits the benefits of measurement but to be useful these benefits must be given an operational meaning.

The history of CAPMON in Section 1 illustrates the difficulty of attaching such a meaning. Specific design objectives can be quite ephemeral even over relatively short time periods and finding good designs can prove quite challenging since the former cannot be based on narrowly defined objectives. Caselton and Husain (1980) and Caselton and Zidek (1984) proposed a theory based on entropy that avoids the need to specify such objectives like 'parameter estimation' or 'hypothesis testing'.

2.1. Univariate setting

To describe the theory, suppose that $k = 1$ so that we are dealing with a univariate rather than a multivariate field. We are to add u_1 gauged sites from among u candidates to g currently gauged sites. A particular subset, 'add', of ungauged sites is under consideration for possible addition to the network.

Let $\mathbf{X}_f$ be a random vector whose co-ordinates represent future realizations of the field at the gauged and ungauged sites combined. Partition the transpose $\mathbf{X}_f^T$ as $\mathbf{X}_f^T = ((\mathbf{X}_f^u)^T, (\mathbf{X}_f^g)^T)$, $\mathbf{X}_f^g$ being the g-dimensional vector of responses at gauged sites and $\mathbf{X}_f^u$ that at the ungauged sites. By suitably relabelling co-ordinates, $\mathbf{X}_f^u$ can be further partitioned as $(\mathbf{X}_f^u)^T = ((\mathbf{X}_f^{rem})^T, (\mathbf{X}_f^{add})^T)$, $\mathbf{X}_f^{add}$ and $\mathbf{X}_f^{rem}$ being respectively the vector of responses at currently ungauged sites in add and the remainder. How much information would the measurements from the gauged sites in add yield?

The answer can be expressed by the uncertainty still left (i.e. residual uncertainty) after incorporating the information in the data coming from add. To express this uncertainty we need the notion of 'entropy', a quantity associated with a random vector that resembles the variance of a random variable: the larger the entropy, the more uncertain the random vector.

However, unlike the variance whose size depends on the scale on which the variable is measured, the entropy of a random vector can be specified so that it is invariant under diffeomorphic transformations of the random vector. More precisely taking the approach of Jaynes (1963) we can define the entropy relative to a 'reference density' $h(\mathbf{V})$ of a random vector $\mathbf{V}$ with density f as

$$H(\mathbf{V}) = E\left[-\log\left\{ \frac{f(\mathbf{V})}{h(\mathbf{V})} \right\} \right].$$

As noted by Caselton *et al.* (1992), the choice of h (which Jaynes described as a 'measure' representing 'complete ignorance') need not have a finite integral and is in fact somewhat arbitrary. The problem of its choice, which has not been fully resolved, is analogous to selecting a reference prior in conventional Bayesian analysis. So the results of our analysis must therefore be viewed as somewhat tentative. A natural choice on the range of the response field seems to be the uniform density since, after co-ordinates of the response vector have been log-transformed, they have approximately a joint Gaussian distribution on Euclidean space.

Once the entropy has been defined we find that it has various natural properties in addition to invariance. For example, if $\mathbf{W}$ denotes a second random vector we can easily see that (in an obvious notation)

$$H(\mathbf{V}, \mathbf{W}) = H(\mathbf{W}) + H(\mathbf{V}|\mathbf{W}).$$

To compute the entropy we need to specify an appropriate joint distribution. For this we adopt the Gaussian model of Le *et al.* (1997). Let

(a) $\theta = (\Sigma, B)$ denote the first-level parameters in our conditional Gaussian model for the random field and
(b) D be the set of all data currently available from previously gauged sites.

Then given D the conditional entropy above with respect to a constant reference density may be used to represent the total uncertainty about a future realization of the random field X_f and the unknown parameter vector θ: $H(X_f, \theta|D)$. Moreover

$$H(\mathbf{X}_f, \theta|D) = H(U|G, \theta, D) + H(\theta|G, D) + H(G|D) \tag{1}$$

where $G = ((\mathbf{X}_f^{add})^T, (\mathbf{X}_f^{g})^T)^T$ and $U = \mathbf{X}_f^{rem}$.

Adding a site to the current network will not change the total uncertainty $H(\mathbf{X}_f, \theta|D)$. When the response vector at gauged sites and added sites is observed the uncertainty represented by $H(G|D)$ is entirely eliminated. We can see from equation (1) that the total uncertainty decomposes into two parts, the first being reduced to 0 by the process of measuring the future realization of the responses at the add sites whereas the second is unaltered. Therefore selecting add to minimize future uncertainty about responses at the remaining 'rem' sites is equivalent to selecting the add sites about which our uncertainty is maximal Thus our selection problem reduces to selecting the add sites to maximize $H(G|D)$.

To describe how we find $H(G|D)$ under the assumptions made in Le *et al.* (1997) let us assume more simply that $\mathbf{Y}$ is any random vector for which

$$\mathbf{Y}|\Sigma \sim N_g(0, \Sigma) \quad \text{and} \quad \Sigma|\Phi, \delta \sim \mathrm{IW}_g(\delta, \Phi)$$

where $\mathrm{IW}_g(\delta, \Phi)$ denotes the g-dimensional inverted Wishart distribution with hyperparameter scale matrix Φ^{-1} and degrees of freedom δ in the notation of Dawid (1981) as presented by Brown (1993).

Then Le and Zidek (1992) showed that

$$\mathbf{Y}|\Phi, \delta \sim t\{0, (\delta - g + 1)\Phi, \delta - g + 1\}.$$

Note that

$$\begin{aligned} H(\mathbf{Y}, \Sigma|\Phi, \delta) &= H(\mathbf{Y}|\Sigma, \Phi, \delta) + H(\Sigma|\Phi, \delta) \\ &= H(\Sigma|\mathbf{Y}, \Phi, \delta) + H(\mathbf{Y}|\Phi, \delta). \end{aligned}$$

To find the entropy for the multivariate T distribution's entropy, $H(\mathbf{Y}|\Phi, \delta)$, we only need to compute $H(\mathbf{Y}|\Phi, \delta, \Sigma)$, $H(\Sigma|\Phi, \delta)$ and $H(\Sigma|\mathbf{Y}, \Phi, \delta)$. Since $\mathbf{Y}|\Sigma, \Phi, \delta$ is multivariate normal, $\Sigma|\Phi, \delta$ and $\Sigma|\mathbf{Y}, \Phi, \delta$ have inverse Wishart distributions. The entropies of the multivariate normal and the inverse Wishart distributions may now be computed by using the results of Caselton *et al.* (1992); see also Guttorp *et al.* (1993) and Le and Zidek (1994). These results prove that for the multivariate T distribution

$$H(\mathbf{Y}|\Phi, \delta) = \tfrac{1}{2}\log|\Phi| + c, \tag{2}$$

where here and in what follows c represents a generic function whose exact form is not relevant to the analysis.

Now, turning to the problem at hand, Le and Zidek (1992) showed that the distributions of $\mathbf{X}_f^{g}|D$ and $\mathbf{X}_f^{add}|\mathbf{X}_f^{g}, D$ are multivariate T distributions and

$$H(G|D) = H(X_f^{g}|D) + H(X_f^{add}|X_f^{g}, D).$$

Applying equation (2), we easily see that

$$H(G|D) = \tfrac{1}{2}\log|\Phi_{\text{add}|g}| + c, \tag{3}$$

where $\Phi_{\text{add}|g}$ is the residual covariance matrix of $X_{\text{f}}^{\text{add}}$ conditional on X_{f}^{g}. More precisely in terms of the hypercovariance matrix kernel we can represent $\Phi_{\text{add}|g}$ as $\Phi_{\text{add}} - \Phi_{\text{add},g}\Phi_g^{-1}\Phi_{g,\text{add}}$. Because only the term $\log|\Phi_{\text{add}|g}|$ is related to the choice of the add sites, maximizing $H(G|D)$ is the same as maximizing $|\Phi_{\text{add}|g}|$.

2.2. The multivariate setting

The results above can now be extended to the multivariate case of k responses. The value of $|\Phi_{\text{add}|g}|$ comes from Φ estimated by the method of Brown *et al.* (1994). To reduce the number of parameters to a manageable level they assumed that $\Phi = \Lambda \otimes \Omega$ where Λ corresponds to the covariation between sites and Ω that between site-specific responses. Letting Φ_{gg} represent the submatrix corresponding to gauged sites, Brown *et al.* (1994) estimated $\Phi_{gg} = \Lambda_{gg} \otimes \Omega$ and δ the prior degrees of freedom by using the EM algorithm of Chen (1979). Then they extended the estimate of Λ_{gg} to an estimate for Λ by using the method of Sampson and Guttorp (1992).

That method adopts a hypothetical Euclidean 'D-plane' with respect to which the co-ordinates of Λ are isotropic, i.e. the covariation between sites is a monotone decreasing function say ζ of their D-plane distances. This function would usually be chosen to have a simple parameter form that is consistent with the scatterplot of empirical correlations at varying intersite distances. For example in the next section we find that a good fit is achieved when it has a negative exponential form and its parameters are estimated by weighted least squares.

The method estimates that monotone function and the D-plane location co-ordinates associated with each of the gauged sites $\mathbf{d}_i = (d_{i1}, d_{i2})$ for site i with geographical co-ordinates $\mathbf{g}_i = (g_{i1}, g_{i2})$. The method relates the $\{\mathbf{d}_i\}$ to the $\{\mathbf{g}_i\}$ through thin plate smoothing splines $d_j = f_j(\mathbf{g})$, $j = 1, 2$. These splines are fitted to the D- and G-plane co-ordinate pairs for the gauged sites, the degree of fit depending on the so-called smoothing parameter. The $\{\mathbf{d}_i\}$ are then replaced by the fits $\{\mathbf{f}(\mathbf{g}_i)\}$, where $\mathbf{f} = (f_1, f_2)$. The parameters would be chosen to ensure a good fit of ζ to that scatterplot of points.

Large values of the smoothing parameter will entail poor G- to D-plane co-ordinate fits. However, those splines will more faithfully maintain the character of the G-plane and lead to a simplicity of interpretation of the results of the analysis. At the other extreme, small values will lead to splines that twist the G-plane into an unrecognizable form while ensuring a good fit to the estimated D-plane co-ordinates.

The choice of this parameter is subjective. 'Small' tends to be better because the co-ordinates of the estimated Λ_{gg} will tend to be more closely isotropic in the **f**-image of the G-plane. However, some smoothing is desirable to achieve a degree of interpretability in the relationship between the resulting plane and its G-counterpart.

Once **f** has been specified, the required extension of Λ_{gg} to Λ can easily be made. Represent the G-plane co-ordinates of sites i and j corresponding to Λ_{ij}, $\mathbf{g}_i$ and $\mathbf{g}_j$ by their **f**-images in the D-plane $\mathbf{d}_i = \mathbf{f}(\mathbf{g}_i)$ and $\mathbf{d}_j = \mathbf{f}(\mathbf{g}_j)$. Finally estimate Λ_{ij} by $\zeta(||\mathbf{d}_i - \mathbf{d}_j||)$.

The LSZ theory enables the multivariate design theory above to apply to the situation where data are missing by design. There Φ and δ can still be estimated by the EM algorithm of Chen (1979). However, changes need to be made in the objective function to account for the ungauged pseudosites at the gauged sites.

To make these changes, we rearrange the response vectors so that the gauged pseudosites are at the 'bottom'. More formally, observe that, given the hyperparameters,

$$\begin{pmatrix} \mathbf{X}_f^u \\ \mathbf{X}_f^{*g} \end{pmatrix} | B = R^* \mathbf{X}_t \sim N(R^* B,\ R^{*\mathrm{T}} \Sigma R^*),$$

$$R^{*\mathrm{T}} \Sigma R^* \sim \mathrm{IW}_{(u+g)k}(\delta,\ \Psi),$$

$$R^* B | \Sigma \sim N(R^* B^0,\ R^{*\mathrm{T}} \Sigma R^* \otimes F^{-1}),$$

where

$$R^* = \begin{pmatrix} I_{uk \times uk} & 0 \\ 0 & R \end{pmatrix},$$

$\Psi = R^{*\mathrm{T}} \Phi R^*$ and R, defined in Le *et al.* (1997), simply permutes the co-ordinates of the response vector in the manner described above. The elements of Ψ are identical with those of Φ except for the diagonal block in the bottom right-hand corner Φ_{gg}, corresponding to the pseudosites at the gauged sites. Here

$$\Psi_{gg} = R^{\mathrm{T}} \Phi_{gg} R = \begin{pmatrix} \Psi_{g_1 g_1} & \Psi_{g_1 g_2} \\ \Psi_{g_2 g_1} & \Psi_{g_2 g_2} \end{pmatrix}$$

where now $\Psi_{g_1 g_2}$ corresponds to the gauged pseudosites and $\Psi_{g_1 g_1}$ corresponds to the ungauged pseudosites at the gauged sites. Consequently an estimate of Ψ_{gg} and hence of Ψ (which is otherwise identical with Φ) can be obtained directly from that of Φ.

We may now invoke our univariate theory with the pseudosite response vector replacing the site response vector and the estimated Ψ replacing the estimated Φ. The procedure for redesigning a current network now becomes maximize $|\Psi_{\mathrm{add}|g_2}|$, the determinant of the conditional hypercovariance matrix obtained from the estimated matrix Φ, of the future realization at the proposed add pseudosites conditional on the data from the gauged pseudosites.

3. Incorporating cost

In the previous section, we considered the addition of sites or pseudosites to a network purely on the basis of the degree to which the uncertainty (entropy) will be reduced. The theory suggests that we increase the number of pseudosites in the network to the maximum feasible limit since the entropy associated with add will increase monotonically as the number of pseudosites in add equals s, say, increases. However, the monotonic growth rate will begin to tail off at some point and a point of marginal gains in entropy will be seen (Caselton *et al.*, 1992). Let us denote by $E(s)$ the reduction in entropy per time period (which is assumed constant over time periods) to be achieved by adding the pseudosites in s to the network. While $E(s)$ increases with the number of pseudosites in s, the cost of adding the pseudosites in s will increase. When the losses dominate the gains an optimal redesign point will have been reached.

Costs may be classified as

(a) initial preparation and
(b) operating.

A given set s say of proposed pseudosites may be classified as those at formerly gauged and those at formerly ungauged sites. An important component of the initial cost in the latter case will be the site acquisition and preparation costs. These costs may be quite large and could

vary from site to site depending on such things as the ease of access and cost of purchase. At both gauged and ungauged pseudosites in add, monitors must be installed once the site is available. The costs of these monitors can also be pseudosite dependent.

In the application in the next section we spread the initial costs over the realistic lifetime of a site (taken there to be 5 years). In this work we charged out initial costs monthly against entropy 'income'. In a refined approach to our design methodology, we might adjust these monthly costs to account for such things as expected inflation (which would make current dollars larger in the future months over which those costs are being spread). Although we do not introduce such refinements in this paper, our methodology allows for them.

Let us assume that we spread the initial costs as described above and let $C_{\text{prep}}(s)$ represent the resulting cost per monitoring time period. We do this only for proposed new pseudosites in add and ignore the costs of operating the existing network.

The other costs associated with the pseudosites in add would be their operating costs. These could vary from pseudosite to pseudosite according to how the data were extracted. In particular, when analytical methods are needed to analyse material obtained by a monitor, transportation, storage and laboratory costs could be substantial. In contrast, with automated air pollution monitoring equipment for monitors like those considered in the next section, operating costs are quite small. In any case the marginal operating cost of monitoring additional responses will be smaller than that associated with the first pseudosite to have gauges installed at a site. However computed, let us denote by $C_{\text{op}}(s)$ the cost of operating the pseudosites in add for a single time period. Thus the total cost would be $C(s) = C_{\text{prep}}(s) + C_{\text{op}}(s)$ per time period over the projected lifetime of the network.

We now confront competing objective functions $E(s)$ and $C(s)$ representing constant cost entropy reduction and cost per time period associated with the proposed new network pseudosites in s. How should these be integrated into a single analysis?

In principle, this can be done by a multiattribute utility analysis. Such an analysis would in particular measure these two objective criteria on a common scale (of 'utiles') and under general conditions ask us to maximize a linear combination of them on this new scale (Keeney and Raiffa, 1976).

We propose a more direct approach, a variation of that just described. We place both costs on the scale of $E(s)$ and take a linear combination of the result as our single composite objective criterion function: $O(s) = E(s) - \text{DE}\, C(s)$. Here DE denotes the cost-to-entropy conversion factor. In other words, DE gives us the number of entropy units (say 'entiles') that we would have to 'spend' to buy each unit of the currency that we would need to extend the network. A big value for DE would signify that we would have to spend many entiles to build a big network and to make a small network desirable.

In practice this factor would have to be elicited from those in the agency redesigning the network. Those involved in the task would need to have their preferences 'calibrated' perhaps by using existing stations and the value ascribed to them by the agency.

When $\text{DE} = 0$ maximizing $O(s) = E(s)$ amounts to finding the maximal subdeterminant of the hypercovariance matrix described in the previous section. An efficient algorithm for finding an exact optimum subject to linear constraints is available (Ko *et al.*, 1995). However, in general we are forced to find our optimum design by using a 'greedy' algorithm which finds the design sequentially by picking the next best pseudosite to include in the network at each step $S = 1, 2, \ldots$; S represents the number of pseudosites in s. At some step S_{opt}, a maximal value of $O(s)$ and associated optimal s, say s_{opt}, will be found, at which point the algorithm terminates.

4. Application

In this section we apply the theory just described to determine an optimal extension of the composite monitoring network described in Section 1. We demonstrate the use of our theory by an application to monthly pollution fields. This somewhat arbitrary choice reflects an absence of directives from a specific agency. Our choice enabled us to avoid the micro-modelling that was needed to deal with the substantial amount of data missing at random at finer than monthly timescales. Such modelling would also have been necessary to remove autocorrelation from the various series involved and thereby to validate the assumption underlying the linear–Gaussian methodology of Le *et al.* (1997) which assumes independent fields over successive time periods $t = 1, 2, \ldots$. These micromodels would introduce much uncertainty in return for the additional information that we could have extracted from the fine scale measurements. It was not clear whether the potential gains would have outweighed the losses.

Even though we chose to work with monthly air pollution concentration levels we still found that six of the 37 gauged sites had more missing data than we thought on heuristic grounds that we could fill in without biasing the results. Thus we were forced to reject these six on the grounds of their low data quality. That left 31 gauged sites for our study, their latitudes ranging from 42.246° to 49.800° and longitudes from 74.736° to 94.400°.

We also restricted our study to O_3 and SO_4 alone. By doing so we simplified our analysis (although in principle we could have included all four pollutants). More substantively these pollutants, unlike the others monitored by this network, are distinguished in that suitably lagged they seem strongly associated with hospital admissions for respiratory morbidity (see

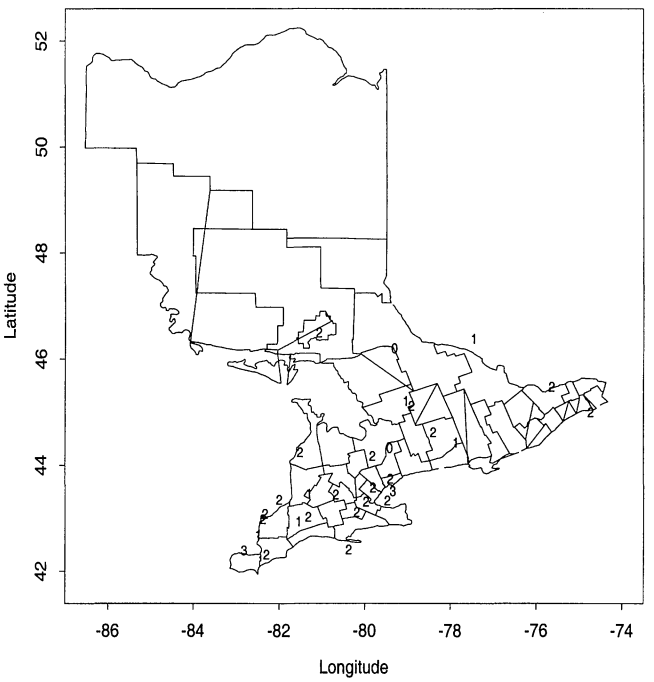

Fig. 1. Locations of 29 of the 31 monitoring sites in the network being redesigned: 0, no ions; 1, SO_4 only; 2, O_3 only; 3, SO_4 and O_3

Zidek *et al.* (1998)). Fig. 1 shows the location of 29 of these 31 sites and the O_3 or SO_4 monitors that they carry (the remaining two, sites 3 and 29, could not be plotted there because they are well off scale).

By definition we have associated with the 31 gauged sites in our study $2 \times 31 = 62$ pseudosites, specifically those associated with the O_3 and SO_4 co-ordinate responses at the various sites. Of these 62 pseudosites 10 had gauges for SO_4 and 21 for O_3, leaving 31 without gauges.

We should note that although our O_3 data were measured in units of parts per billion we can regard them equivalently as having been measured in the same units as SO_4 in the calculation of our entropy design index, namely in units of micrograms per cubic metre. This is because these two scales differ by a multiplicative factor (of about 2). Under our log-Gaussian model, this multiplicative factor becomes an additive constant and the entropy index is readily shown to be invariant under such changes in location.

For illustration we include an additional 15 ungauged sites, including the six sites rejected earlier on grounds of low data quality. The remaining nine ungauged sites were selected at varying distances from the nearest gauged site so that between them they included both remote and nearby sites. These 15 new sites provide 30 ungauged pseudosites. Fig. 2 depicts the geographical locations of all $31 + 15 = 46$ sites except for the new site 4 as well as sites 3 (here relabelled 18) and 15 (now 44), which were outside the scope of this map.

Using the estimates quoted in Section 1 yields a monthly cost (spread over 60 months) of \$667 (all costs being in 1997 Canadian dollars); that for installing either an O_3 or SO_4 gauge \$250. The monthly operating cost for the first new monitoring device installed is estimated to be \$417 and subsequent monitors at \$42 each. From these estimates, the monthly cost of installing and operating a single gauge at a formerly ungauged site would be \$667 + \$250 + \$417 = \$1334. For an already gauged site, the monthly cost of adding a gauge would be just \$250 + \$42 = \$292. With these estimates the constant monthly cost function $C(s)$ is easily found for any proposed s.

We estimated $E(s)$ by the method described in Section 2.2 (that in Brown *et al.* (1994) and Le *et al.* (1997)). Fig. 3 depicts the result of setting the smoothing parameter to 0. In Fig. 3(b) we see the completely unintelligible character of the G-plane of southern Ontario after transformation by the spline obtained in this case. In particular we see (and again in Fig. 4(b)) how southern Ontario would have to be folded and bent to obtain highly correlated stations (that in some cases are widely separated in the geographic plane) into enough proximity in the D-plane to ensure that the isotropy condition is at least approximately satisfied. Fig. 3 thus shows the extremely non-isotropic character of the pollution field.

In Fig. 3(a), we see the estimated values of the variogram plotted against the estimated D-plane distances between sites. To obtain the value of the variogram associated with sites i and j, we first convert Λ_{gg} as estimated by the EM approach described above to correlation form as $\Lambda_{gg}^{\text{corr}}$ with diagonal elements identically 1. Convert this to variogram form to measure the dispersion rather than correlation between sites: $2(I_{gk\times gk} - \Lambda_{gg}^{\text{corr}})$. The dispersion between i and j is then given by $D_{ij} = 2(1 - \Lambda_{gg:ij}^{\text{corr}})$. In Fig. 3(a) these $\{D_{ij}\}$ have been plotted against their associated $||\mathbf{d}_i - \mathbf{d}_j||$.

Some of the D_{ij} are negative since the associated intersite estimated correlation turned out to be negative. Plausibly their true correlation may be negligible so responses from these sites have little information in them about their counterpart. In contrast some of the D_{ij} are small or near 0; there we are seeing sites which carry a large amount of information about each other.

We obtain the monotone function ζ by fitting the parametric exponential model graphed in Fig. 3(a). The curve cannot exceed 2 since positive intersite correlations are assumed.

In Fig. 4 we see the result of some degree of smoothing. In Fig. 4(b) the surface of the G-plane now becomes at least somewhat apparent although it remains quite contorted near the centre. The high degree of distortion induced by the spline transformation derives from the complex spatial correlation pattern revealed by the data.

By increasing the smoothing parameter to 75 in Fig. 4 we have given up some of the quality of the fit achievable by the exponential variogram model without smoothing. However, the loss seems small in our judgment, whereas we gain a view of the distorted G-plane that gives us some idea of what has been necessary to achieve an approximately isotropic representation of the correlation field.

We experimented with other choices of the smoothing parameter but in the end chose 75 on subjective grounds. The result was quite robust against a variation in that choice. Quite substantial deviations from 75 produced little change in the result.

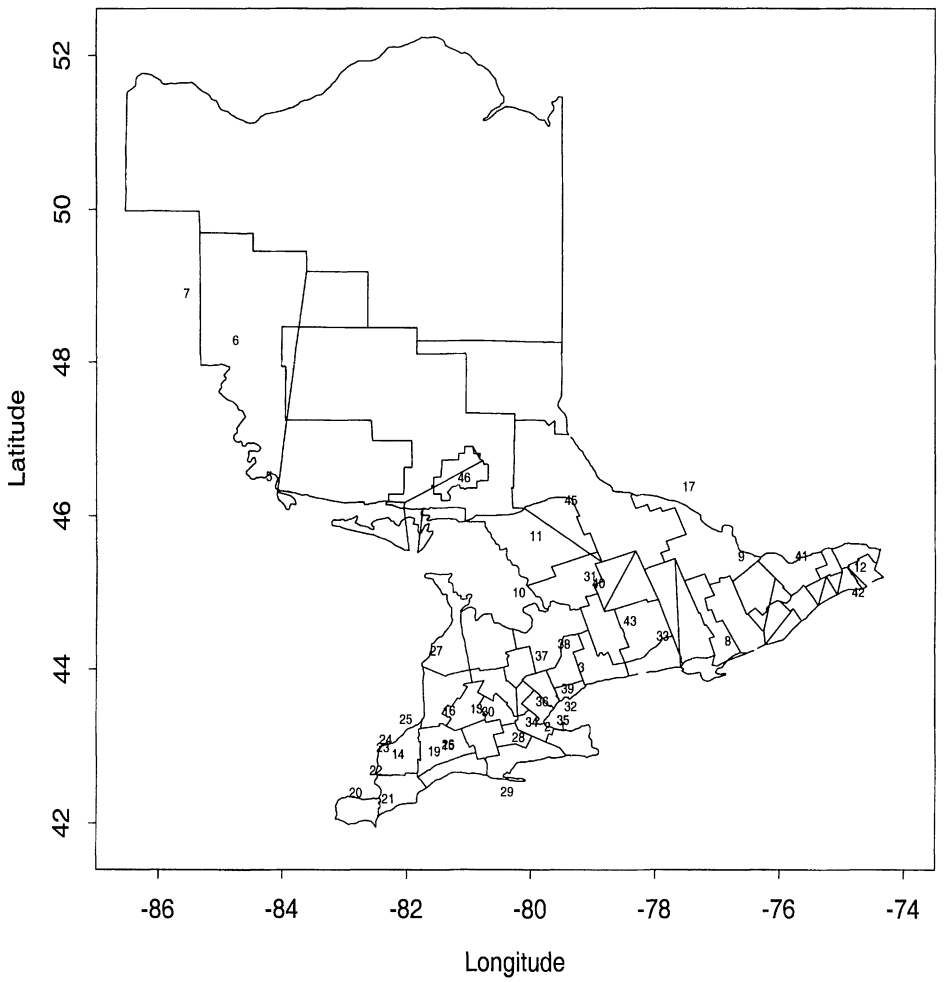

Fig. 2. Locations of 43 of the 46 southern Ontario sites in the study

With this result we could go on to extrapolate from $\Lambda_{gg}^{\text{corr}}$ to Λ^{corr} as described in Section 2.2. To estimate Λ by using the method of Sampson and Guttorp (1992) requires as a further step that we estimate its diagonal elements. This is done by smoothing the diagonal elements of the EM-estimated Λ_{gg}-matrix using a thin plate spline representation of them as a function of the G-plane co-ordinates of the sites to which they correspond. The smoothing spline then enables us to extrapolate to the sites corresponding to the diagonal elements of Λ. In this way we have estimated the site 'variances'. These estimates combine with that of $\Lambda_{gg}^{\text{corr}}$ to give us the required estimate of Λ, Φ and hence Ψ.

Applying the method gives us an estimated covariance matrix for the responses corresponding to all 92 pseudosites either with gauges or under consideration. From this covariance matrix the determinant required for $E(s)$ could be found for any proposed subset add equal to s of new pseudosites.

We explored various possible choices for DE in defining $O(s)$ including DE = 0. When currency is expensive in entiles (i.e. DE is too large) ungauged sites will be prohibitively expensive and none will ever be selected for inclusion in the network. But new sites are fitted with gauges from time to time. That behavioural evidence (in the absence of a full utility analysis) leads us to investigate a range of sufficiently small DE-values that at least some

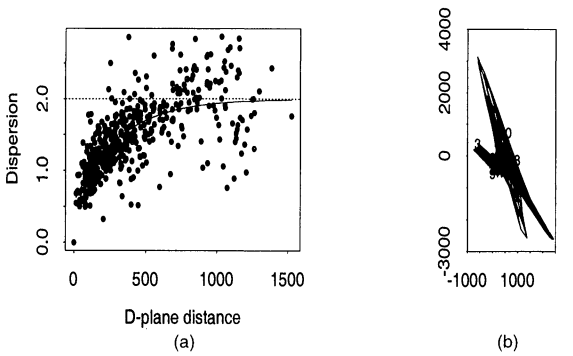

Fig. 3. Fitted Sampson and Guttorp (1992) dispersion matrix with no smoothing: (a) fitted variogram exponential (root-mean-square error 0); (b) D-plane co-ordinates

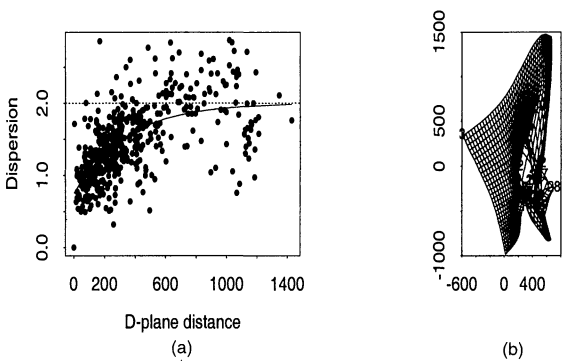

Fig. 4. Fitted Sampson and Guttorp (1992) dispersion matrix with some smoothing (spline smoothing parameter 75): (a) fitted variogram exponential (root-mean-square error 0); (b) D-plane co-ordinates

previously ungauged stations might be admitted. The designer would have to select DE within that range to find the number and identity of ungauged sites to add. Through trial and error our approach leads us to the range [0, 0.015] from which we select for detailed consideration DE = 0, 0.0025, 0.005, 0.015; the last of these values for example says that we would be spending 15 entiles to buy \$1000.

5. Results and discussion

In the accompanying tables and figures we see the results of applying our redesign methodology. Table 1 shows us the best single pseudosite to fit with gauges depending on cost. The entropies appearing there show the increase in $\frac{1}{2}\log|\Phi_{\text{add}|g}|$ of equation (3) as a result of adding the best additional site to the add vector. Cost in Table 1 reflects that addition and depends on whether a new site is fitted with gauges or a monitor is added to a previously gauged site. The last row of Table 1 shows the combined effect of adding the optimum pseudosite.

If DE = 0, the method tells us to install an SO_4 gauge at site 7. From Fig. 2 we find this station to be among the most remote of all the ungauged sites. (From measurements taken at the station, we would gain an entropy reduction of 5.1 entiles.) In fact the first four stations selected when DE = 0 prove formerly not to have gauges, site 4 being so remote that it is off the map and so does not appear in Fig. 2. The first currently gauged station, 45, to be picked by the algorithm is also remote, one of the remotest among the currently gauged stations: it gets fifth place.

Fig. 5(a) shows that the entropy reduction is nearly linear in this range at about 5 units per added pseudosite. As sites are added, the residual gain begins to decline. However, the growth is monotone and when DE = 0 the process of adding pseudosites would never be naturally terminated by the algorithm.

It is worth noticing in Table 2 (where step 1 is continued to select by the greedy algorithm the best 30 pseudosites one at a time) that when DE = 0 only SO_4 gauges are added (until we reach our imposed limit of 50 new gauged pseudosites). This result may be expected; the variation in the SO_4 residual series dominates that in the O_3 residual series as reflected in the estimated variances. However, we would have expected that some new O_3 pseudosites would have been added. Table 2 also shows that when DE = 0 the best four pseudosites are at hitherto ungauged sites whereas the next four tend to be closer to the original network and so on.

Even when DE > 0, SO_4 gauges are favoured over O_3 gauges by the design criterion, again because of the greater uncertainty that is associated with the SO_4 series. However, that criterion now conservatively chooses to add SO_4 monitors to existing sites, thereby avoiding the high initial cost of installation at new sites.

When DE = 0.0025 the criterion first selects a formerly ungauged SO_4 site at step 12 (location 7) and picks several such SO_4 sites among its top 30 (location 5, 6, 4 and 11). As

Table 1. Entropy gains *versus* cost from adding the best single pseudosite

	Results for the following entropy/dollar ratios:			
	0	*0.0025*	*0.005*	*0.015*
Pseudosite	SO_4 (7)	SO_4 (45)	SO_4 (45)	SO_4 (45)
Entropy (entiles)	5.1	4.5	4.5	4.5
Cost (\$)	1333	292	292	292
Combined (entiles)	5.1	3.8	3.8	0.1

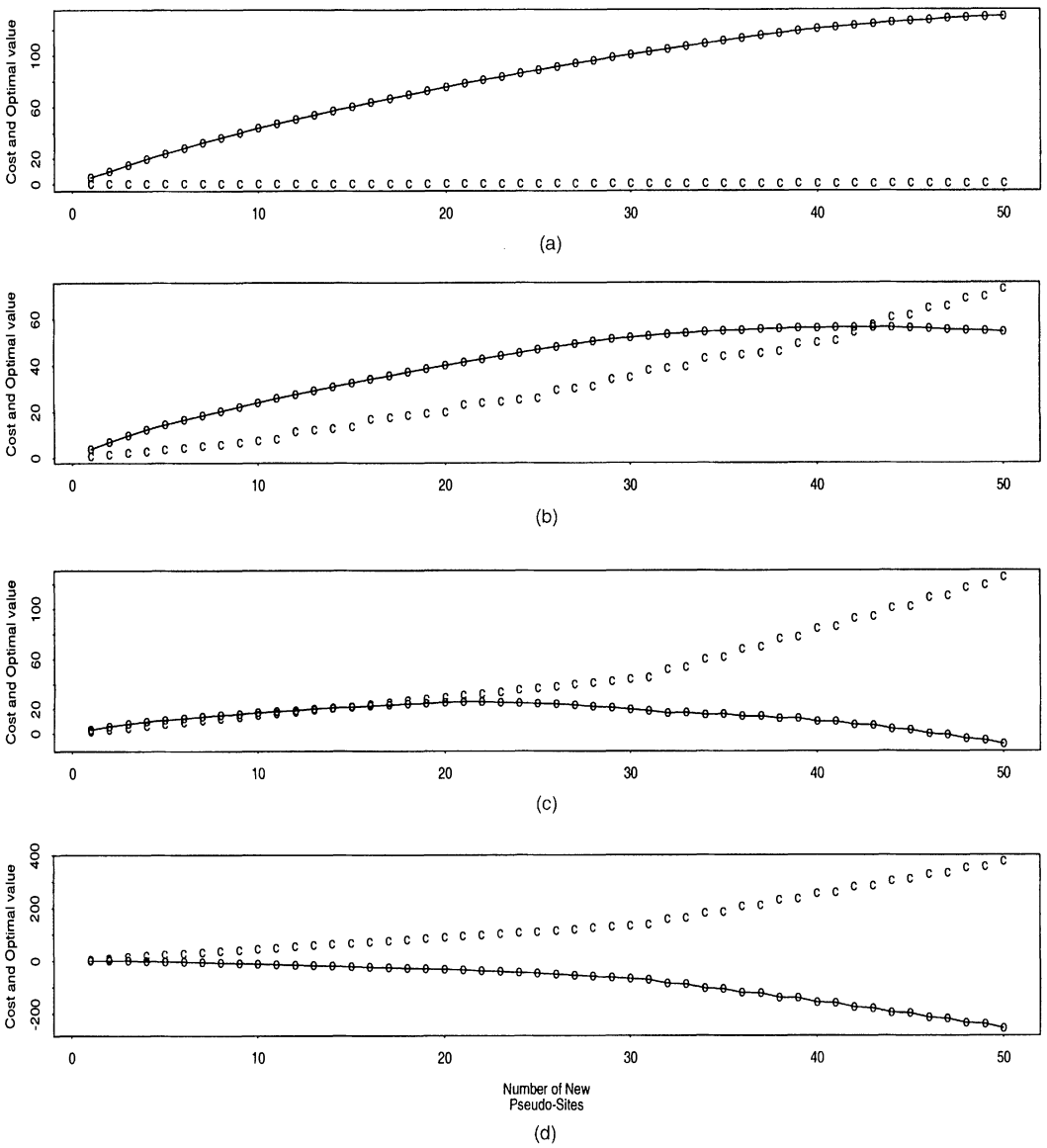

Fig. 5. Optimal network cost (c) and value (○) (entropy units) for various DE ratios: (a) DE = 0; (b) DE = 0.0025; (c) DE = 0.005; (d) DE = 0.0015

might be expected these five new SO_4 sites have the same relative ranking as they had when selected while DE = 0. However, before location 4 receives its monitor, an SO_4 gauge, at step 26 the criterion elects to fit (at step 25) a second O_3 monitor to site 7 gauged with a sulphate monitor at step 12. This is the first O_3 monitor that it installs. The second monitor at site 7 would provide more information than a first SO_4 monitor at either of the sites 4 and 11 in relation to the initial costs (in entiles) of creating a new site.

We were surprised to discover that when DE is at least as great as 0.005 the same top 30 pseudosites are selected and in the same order no matter what the cost. Moreover, the cost of

Table 2. First 30 pseudosites added sequentially to optimize the combined entropy-to-cost function†

Step	*Results for the following entropy/dollar ratios:*			
	0	*0.0025*	*0.005*	*0.015*
1	SO_4 (7)	SO_4 (45)	SO_4 (45)	SO_4 (45)
2	SO_4 (5)	SO_4 (46)	SO_4 (46)	SO_4 (46)
3	SO_4 (6)	SO_4 (38)	SO_4 (38)	SO_4 (38)
4	SO_4 (4)	SO_4 (43)	SO_4 (43)	SO_4 (43)
5	SO_4 (45)	SO_4 (41)	SO_4 (41)	SO_4 (41)
6	SO_4 (11)	SO_4 (27)	SO_4 (27)	SO_4 (27)
7	SO_4 (8)	SO_4 (40)	SO_4 (40)	SO_4 (40)
8	SO_4 (46)	SO_4 (23)	SO_4 (23)	SO_4 (23)
9	SO_4 (10)	SO_4 (21)	SO_4 (21)	SO_4 (21)
10	SO_4 (9)	SO_4 (39)	SO_4 (39)	SO_4 (39)
11	SO_4 (3)	SO_4 (42)	SO_4 (42)	SO_4 (42)
12	SO_4 (38)	SO_4 (7)	SO_4 (37)	SO_4 (37)
13	SO_4 (12)	SO_4 (37)	SO_4 (24)	SO_4 (24)
14	SO_4 (14)	SO_4 (24)	SO_4 (29)	SO_4 (29)
15	SO_4 (13)	SO_4 (29)	SO_4 (25)	SO_4 (25)
16	SO_4 (1)	SO_4 (5)	SO_4 (34)	SO_4 (34)
17	SO_4 (43)	SO_4 (25)	SO_4 (30)	SO_4 (30)
18	SO_4 (15)	SO_4 (34)	SO_4 (36)	SO_4 (36)
19	SO_4 (2)	SO_4 (30)	SO_4 (28)	SO_4 (28)
20	SO_4 (27)	SO_4 (36)	SO_4 (26)	SO_4 (26)
21	SO_4 (40)	SO_4 (6)	SO_4 (35)	SO_4 (35)
22	SO_4 (23)	SO_4 (28)	O_3 (45)	O_3 (45)
23	SO_4 (41)	SO_4 (26)	O_3 (44)	O_3 (44)
24	SO_4 (21)	SO_4 (35)	O_3 (33)	O_3 (33)
25	SO_4 (39)	O_3 (7)	O_3 (17)	O_3 (17)
26	SO_4 (42)	SO_4 (4)	O_3 (18)	O_3 (18)
27	SO_4 (37)	O_3 (5)	O_3 (31)	O_3 (31)
28	SO_4 (24)	O_3 (6)	O_3 (38)	O_3 (38)
29	SO_4 (29)	SO_4 (11)	O_3 (22)	O_3 (22)
30	SO_4 (25)	O_3 (4)	O_3 (16)	O_3 (16)
Entropy (entiles)	100	87	63	63
Cost ($)	24375	13958	8750	8750
Combined (entiles)	100	52	19	−68

†The site number is in parentheses with 1–15 denoting new locations and 16–46 denoting currently gauged locations.

installing gauges at new sites proves too high and no new sites are ranked among the list of the top 30. In other words, the ranked list of the top 30 is invariant under changing DE ratios once DE = 0.005 has been attained.

What does change is the optimal number of pseudosites to fit gauges at as DE varies. Fig. 5 shows the diminishing 'information returns' from adding additional pseudosites as the number of such sites increases. Eventually, except where the cost is zero, an optimum number of pseudosites is reached. In fact, when the new sites are very expensive (relative to the value of information) and DE = 0.15, we find that the optimal decision is to add no new pseudo-stations to the network.

6. Concluding remarks

The theory developed in this paper as well as that in Caselton *et al.* (1992), Guttorp *et al.*

(1993) and Le and Zidek (1994) differs from but is related to the modern theory of optimal experimental design. These relationships are discussed in the works just cited.

Specifying the DE ratio is of fundamental importance in our approach based on the scale of entiles. We recognize that its specification will prove challenging in practice.

A more refined approach would be based on the measurement of the utilities involved. Processes of conducting such measurements have been developed within the framework of decision analysis. That approach would potentially solve another significant problem (noted by a referee) that arises when the pollution species have differing societal significances which are not adequately captured by the differences in correlations and variances embraced by the multivariate entropy criterion. With the alternative approach, a multiattribute utility analysis would be needed to account for these differences. In particular, in our illustrative example the entropies for O_3 and SO_4 pseudosite selections would have to be computed separately and latterly combined with associated monetary costs, linearly in an overall objective function. This approach would also enable the methodology to cope with a variety of responses (e.g. 'temperature' and 'O_3') that do not, as in our example, share the same scale of measurement. (Nevertheless as shown by Sun *et al.* (1998) even with this new approach the requisite marginal distributions for the responses involved would best be found by marginalizing their joint multivariate distribution rather than conducting separate univariate analyses.)

We have not attempted such a refinement here, partly because of the simple conceptual appeal of the entile scale. And multiattribute utility theory suggests that the utility-based optimization criterion will resemble that used in our analysis, a linear combination of those associated with the entropy and cost functions.

Dr Lawrence Phillips suggested (in a personal communication) an interesting alternative to our optimization approach. He suggested ranking a prospective site on the basis of the ratio of the entropy of the network resulting from the addition of that site to its cost.

These ratios have intuitive appeal. However, we prefer another practical alternative that also bypasses the DE ratio. Our approach would optimize the entropy subject to a fixed upper bound on the cost. By varying the upper bound the best network could be found relative to budget limits.

We prefer our approach because we believe that it will enable us to incorporate a further refinement to our theory. That refinement would replace the greedy one at a time algorithm with an exact combinatorial optimization algorithm proposed by Anstreicher *et al.* (1996) where entropy has the form of a determinant of a covariance function.

We are implementing this algorithm to learn about its value in design problems of the size of that considered in this paper. However, our greedy algorithm would have to be used for problems of large size. Moreover, it will prove a valuable if suboptimal method for use with arbitrary objective functions.

Acknowledgements

We are indebted to Dr R. T. Burnett for providing the data on which the application of this paper is based and Mr Ernie Tradewell from the British Columbia Ministry of Environment for the cost estimates used in our analysis. Thanks also go to the referee and to Professor Constance van Eeden for very thoughtful comments on the work.

This work was supported by grants from the Engineering and Physical Sciences Research Council of the UK and the Natural Sciences and Engineering Research Council of Canada.

References

Anstreicher, K. M., Fampa, M., Lee, J. and Williams, J. (1996) Using continuous nonlinear relaxations to solve constrained maximum-entropy sampling problems. Unpublished.

Brown, P. J. (1993) *Measurement, Regression, and Calibration*. Oxford: Clarendon.

Brown, P. J., Le, N. D. and Zidek, J. V. (1994) Multivariate spatial interpolation and exposure to air pollutants. *Can. J. Statist.*, **22**, 489–509.

Burnett, R. T., Dales, R. E., Raizenne, M. R., Krewski, D., Summers, P. W., Roberts, G. R., Raad-Young, M., Dann, T. and Brook, J. (1994) Effects of low ambient levels of ozone and sulphates on the frequency of respiratory admissions to Ontario hospitals. *Environ. Res.*, **65**, 172–194.

Caselton, W. F. and Husain, T. (1980) Hydrologic networks: information transmission. *Proc. Am. Soc. Civ. Engrs*, **106**, 503–520.

Caselton, W. F., Kan, L. and Zidek, J. V. (1992) Quality data networks that minimize entropy. In *Statistics in the Environmental and Earth Sciences* (eds P. Guttorp and A. T. Walden). London: Griffin.

Caselton, W. F. and Zidek, J. V. (1984) Optimal monitoring network designs. *Statist. Probab. Lett.*, **2**, 223–227.

Chen, C. F. (1979) Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. *J. R. Statist. Soc.* B, **41**, 235–248.

Dawid, A. P. (1981) Some matrix–variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, **68**, 265–274.

Greater Vancouver Regional District (1996) *Ambient Air Quality Annual Report*. Vancouver: Greater Vancouver Regional District. (Available at `http://www.gvrd.bc.ca/air/bro/aqanrep.html`.)

Guttorp, P., Le, N. D., Sampson, P. D. and Zidek, J. V. (1993) Using entropy in the redesign of an environmental monitoring network. In *Multivariate Environmental Statistics* (eds G. P. Patil and C. R. Rao). New York: North-Holland.

Haas, T. C. (1992) Redesigning continental-scale monitoring networks. *Atmos. Environ.* A, **26**, 3323–3333.

Jaynes, E. T. (1963) Information theory and statistical mechanics. In *Statistical Physics* (ed. K. W. Ford), vol. 3, pp. 102–218. New York: Benjamin.

Keeney, R. L. and Raiffa, H. (1976) *Decisions with Multiple Objectives*. New York: Wiley.

Ko, C.-W., Lee, J. and Queyranne, M. (1995) An exact algorithm for maximum entropy sampling. *Oper. Res.*, **43**, 684–691.

Le, N. D., Sun, W. and Zidek, J. V. (1997) Bayesian multivariate spatial interpolation with data missing by design. *J. R. Statist. Soc.* B, **59**, 501–510.

Le, N. D. and Zidek, J. V. (1992) Interpolation with uncertain spatial covariance: a Bayesian alternative to kriging. *J. Multiv. Anal.*, **43**, 351–374.

——— (1994) Network designs for monitoring multivariate random spatial fields. In *Recent Advances in Statistics and Probability* (eds J. P. Vilaplana and M. L. Puri), pp. 191–206. Zeist: VSP.

Ro, C. U., Tang, A. J. S. and Chan, W. H. (1988) Wet and dry deposition of sulphur and nitrogen compounds in Ontario. *Atmos. Environ.*, **22**, 2763–2771.

Sampson, P. D. and Guttorp, P. (1992) Nonparametric estimation of nonstationary spatial covariance structure. *J. Am. Statist. Ass.*, **87**, 108–119.

Sirois, A. and Fricke, W. (1992) Regionally representative daily air concentrations of acid-related substances in Canada, 1983–1987. *Atmos. Environ.* A, **26**, 593–604.

Sun, W., Le, N. D., Zidek, J. V. and Burnett, R. (1998) Assessment of a Bayesian multivariate spatial interpolation approach for health impact studies. *Environmetrics*, **9**, 565–586.

Wu, S. and Zidek, J. V. (1992) An entropy based review of selected NADP/NTN network sites for 1983–1986. *Atmos. Environ.* A, **26**, 2089–2103.

Zidek, J. V., White, R., Le, N. D., Sun, W. and Burnett, R. T. (1998) Imputing unmeasured explanatory variables in environmental epidemiology with application to health impact analysis of air pollution. *Ecol. Environ. Statist.*, **5**, 99–115.