Bayesian entropy for spatial sampling design of environmental data

Montserrat Fuentes, Arin Chaudhuri and David M. Holland [1]

Institute of Statistics Mimeo Series No. 2571

SUMMARY

We develop spatial statistical methodology to design large-scale air pollution monitoring networks with good predictive capabilities while minimizing the cost of monitoring. The underlying complexity of atmospheric processes and the urgent need to give credible assessments of environmental risk create problems requiring new statistical methodologies to meet these challenges. In this work, we present a new method of ranking various subnetworks taking both the environmental cost and the statistical information into account. A Bayesian algorithm is used to obtain an optimal subnetwork using an entropy framework. The final network and accuracy of the spatial predictions is heavily dependent on the underlying model of spatial correlation. Usually the simplifying assumption of stationarity, in the sense that the spatial dependency structure does not change location, is made for spatial prediction. However, it is not uncommon to find spatial data that shows strong signs of nonstationary behavior. We build upon an existing approach that creates a nonstationary covariance by a mixture of a family of stationary processes, and a propose a Bayesian method of estimating the associated parameters using the technique of Reversible Jump Markov Chain Monte Carlo. We apply these methods for spatial prediction and network design to ambient ozone data from a national monitoring network in the eastern US.

[1]Montserrat Fuentes is an Associate Professor at the Statistics Department, North Carolina State University (NCSU), Raleigh, NC 27695-8203. Tel.:(919) 515-1921, Fax: (919) 515-1169, E-mail: fuentes@stat.ncsu.edu. Arin Chaudhuri is a researcher at SAS, Cary, NC. David M. Holland is a senior statistician at the U.S. Environmental Protection Agency, RTP, NC.
This research was sponsored by a National Science Foundation grant DMS 0353029 and by a US EPA award R-8287801.

Key words: Bayesian inference, Matérn covariance, nonstationarity, simulated annealing, spatial statistics.

1 Introduction

Environmental monitoring agencies around the world maintain large-scale air monitoring networks to assess the efficacy of regulatory controls, determine current levels and trends, and provide air quality inputs to risk assessment and source attribution analyses. However, these networks need to be managed such that changing priorities and needs, both national and local, can be accommodated with the understanding that there could be constraints in future funding for these networks. The proposed reduced network should maintain sufficient spatial information to ensure reasonable statistical inference about air pollution. A major criterion for modifying an existing network is the quality of the spatial predictions, and minimizing the monitoring costs of obtaining such predictions or ensuring that monitoring continues in areas with high pollution levels. In this work, we propose a new method for ranking various subnetworks (most informative subsets) using an entropy measure of the spatial information and giving priority to monitoring sites with high pollution values. Given this optimization criterion, a heuristic algorithm for determining near optimal subnetworks of different sample sizes is described.

The spatial configuration of final subnetworks is heavily dependent on the underlying model for spatial covariance. Complex, atmospherically driven pollutants are not expected to have simple, stationary forms of spatial covariance. We build upon an existing approach for modeling underlying nonstationary covariance, or heterogeneous covariance structure over large spatial ranges, by using a mixture of stationary processes. Properties of this approach are given, along with a method for estimating the covariance parameters using a Reversible Jump Markov Chain Monte Carlo (RJMCMC) approach. These methods are applied to ozone design values for 1997-1999 observed at 513 monitoring sites in the eastern U.S..

The basic problem of air quality monitoring design is selecting the number and spatial configuration of sites to allow, in some quantitative sense, optimal predictions of the air quality field subject to certain constraints on monitoring resources. The design literature contains several different approaches for design. The idea of using entropy in the context of experimental design goes back to at least Lindley (1947) and Bernardo(1979). Caselton and Zidek (1984), Guttorp et al. (1993) and Zidek et al. (2000) developed the maximum entropy design approach by modeling observations at different monitoring locations as a multivariate time series to maximize information "expected" about potential non-monitored sites. Other authors

have considered different approaches for network design. Warrick and Myers (1987) and Müller and Zimmerman (1999) considered design criteria for precise estimation of attributes of the semivariogram that affect kriging. Others have considered design approaches under the assumption that the semivariogram is known (Bras and Rodriguez-Iturbe, 1976; Yfantis, Flatman, and Behar (1987); and Cressie, Gotway, and Grondona, 1990). The design criteria considered by these authors are generally either the average kriging variance or the maximum kriging variance over a region of interest. Wikle and Royle (1999) considered time-varying or dynamic designs to evaluate the efficiency of allowing monitoring locations to change with time. Nychka and Saltzman (1998) considered geometric space-filling designs and Müller (1999) investigated approximate or simulation based approaches for optimal design using utility functions. The approach presented here uses ideas from information and entropy theory integrated into a Bayesian framework for spatial prediction incorporating the uncertainty of the covariance parameters. Given the regulatory need to maintain sites near air quality standards, we give priority to retaining these sites. Our primary objective to downsize the existing network, although these techniques could be applied to augment the network and select the most informative sites for additional monitoring sites.

This paper is organized as follows. In Section 2 we introduce the scientific problem that motivated this research and describe the data. In Section 3 we propose a fully Bayesian framework for selecting optimal reduced networks. In Section 4 we extend this framework for network design to account for constraints, including environmental cost. In Section 5 we discuss the design optimization problem. In Section 6 we introduce our model for spatial nonstationarity. Section 7 presents an application using an air pollution dataset.

1.1 Entropy as a measure of information

If Y is an uncertain, random quantity or vector of such quantities, and $f(y)$ is the probability density function of Y, then uncertainty about Y can be expressed by the entropy of Y's distribution,

$$H(Y) = \int -f(y)log(f(y))dy.$$

For a distribution with low spread and a sharp peak near the mode, the mode provides a good indication of where a "typical" observation might lie. However, for less peaked distributions with larger spreads, the

region where a "typical observation" might lie could be quite large. In general, among a given family of distributions the members with higher spreads have higher entropy. A nice exposition of the statistical significance of entropy is given by Theil and Fiebig (1984). Thus, among all distributions having support in the interval [a,b] we expect the uniform distribution on [a,b] to have the maximum entropy. We illustrate this fact with few examples:

- If $Y \sim N(\mu, \sigma^2)$, then,

$$H(Y) = 0.5(\log(2\pi) - 1) + log(\sigma).$$

  Hence, the entropy is an increasing function of the variance

- If $Y \sim U[a, b]$, then,

$$H(Y) = \log(b - a).$$

  The entropy function is an increasing function of the width of the interval.

2 Data

As national monitoring priorities and funding changes, it has become critical to optimize resources available for national monitoring networks. Thus, there is an urgent challenge to provide credible statistical approaches for reducing or downsizing existing monitoring networks to find the most informative reduced set of monitoring sites that will still meet multiple objectives of major monitoring programs. Here, we consider downsizing the National Air Monitoring Stations/State and Local Air Monitoring Stations (NAMS/SLAMS) (U.S. Environmental Protection Agency, 2003) ozone (O_3) network. NAMS/SLAMS is the major source of O_3 data in the U.S. and monitors O_3 at approximately 800 sites in the conterminous U.S. to determine compliance with the O_3, assess regional transport, and for use in estimating trends in this pollutant. Although most NAMS/SLAMS sites are located in urban and suburban areas where air quality is influenced primarily by local sources, some sites are located in rural areas to characterize regional air quality.

Tropospheric ozone continues to be one of the most significant air pollutant concerns in the United States. Ozone is a photochemical oxidant and a major component of smog. Scientific evidence indicates that high

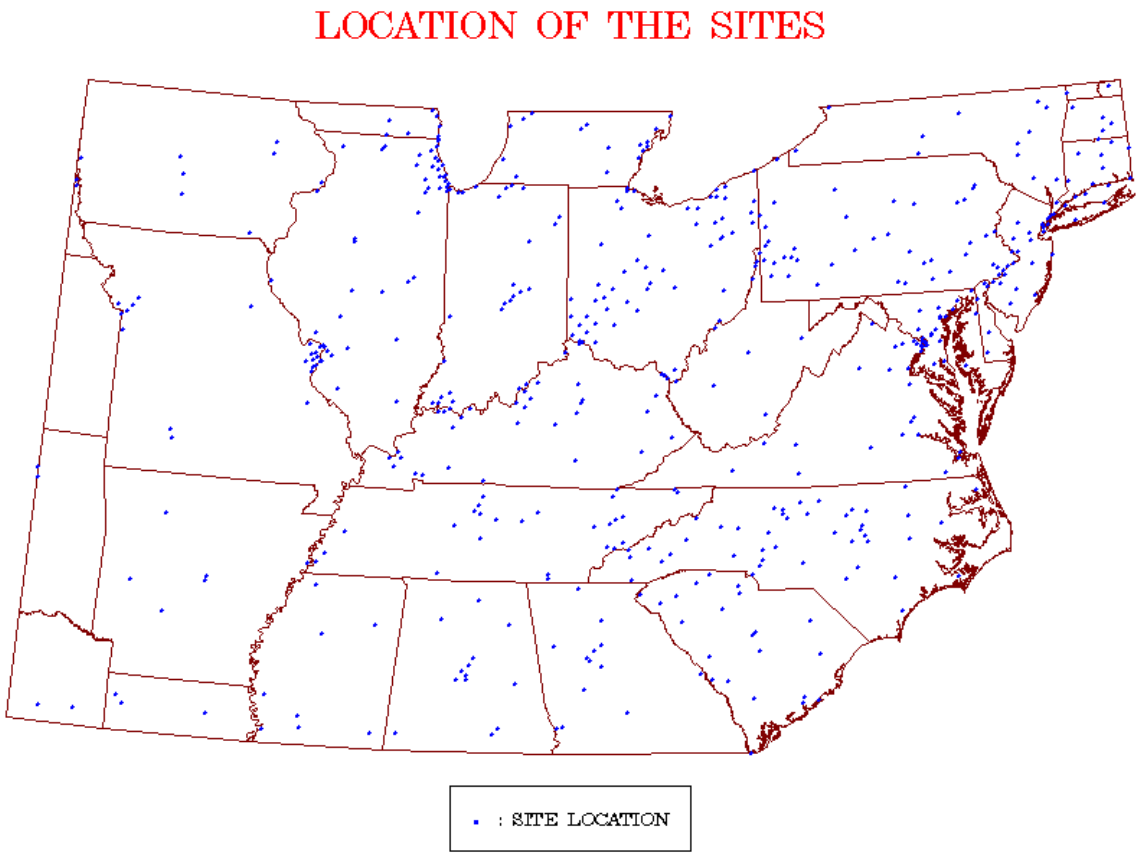

Figure 1: Locations of 513 ambient ozone monitoring sites.

levels of O_3 not only affect people with impaired respiratory systems, but healthy adults and children as well. The U.S. Environmental Protection Agency (EPA) first set ozone National Ambient Air Quality Standards in 1971. These include primary standards to protect human health, and second standards to prevent ecological and agricultural damages. In July 1997, EPA strengthened the O_3 national ambient air quality standard (NAAQS) based on scientific evidence showing adverse health effects from exposures allowed by the existing standard. The new standard was defined in terms of 8-hour averaging times: the 3-year average of the annual fourth-highest daily maximum 8-hour average ozone concentration is less than .08 parts per million (ppm). This 3-year average is usually referred to as the ozone "design value". The maximum daily 8-hour average ozone concentration is the highest of the 17 possible running 8-hour daily average concentrations. We investigate new reduced designs using O_3 design values from 1997 to 1999 for 513 NAMS/SLAMS sites throughout the eastern U.S. (Figure 1).

3 Fully Bayesian approach for network design

We present a fully Bayesian approach for monitoring network design, taking into account the potential lack of stationarity of the environmental process of interest (ozone, in this case), and some monitoring constraints by using a utility function.

Consider a Gaussian spatial process $\{Z(x) : x \in D \subset \mathcal{R}^2\}$, with mean $E[Z(x)] = \mu(x)$. In the application presented in this paper, the mean is a polynomial function on x. The covariance of $Z(\,)$ depends on a parameter θ, $\text{cov}[Z(x)Z(y)|\theta] = C_\theta(x, y)$. We put a prior distribution on θ, $\theta \sim \pi(\,)$. We observe the process at locations $x_1, x_2, \ldots x_N$, that is we have a vector of observations $Z = (Z(x_1), Z(x_2), Z(x_3), \ldots, Z(x_N))$. In the problem of environmental network design we have to choose a subset of $\{x_1, x_2, \ldots x_N\}$ of a given size such that the loss of statistical "information", here defined as entropy-utility, is minimal. For many atmospheric processes, the spatial covariance function is nonstationary, in the sense that the spatial structure changes with location. In Section 6, we present a nonstationary covariance model that is used in the proposed entropy design framework.

3.1 The posterior predictive distribution

Consider the following Bayesian framework

$$Y \sim f(y|\theta), \;\; \theta \sim \pi(\,).$$

That is, we have a random variable Y, whose density is given by a parametric form $f(y|\theta)$, and $\pi(\,)$ is a prior distribution for θ. In the Bayesian setup, the marginal density of Y at the point y^*, is given by,

$$\int f(y^*|\theta)\pi(\theta)d\theta.$$

However, after having an observation of the variable of interest, $Y = y$ we update $\pi(\theta)$ to $\pi(\theta|y)$, and obtain the posterior predictive density. So, if we observe a realization of Y, say y, the posterior predictive density of Y at a point y^*, after having observed y, is defined as,

$$f_P(y^*) = \int f(y^*|\theta)\pi(\theta|y)d\theta.$$

Sampling from the posterior predictive density is easy, if we can sample from the posterior distribution $\pi(\theta|y)$. Then, to generate a sample from the posterior predictive density, we first generate an observation

from the posterior distribution $\pi(\theta|y)$, say θ^*, and then we generate an observation y^*, from $f(y^*|\theta^*)$. Thus, y^* is a sample from the posterior predictive density.

3.2 Fully Bayesian Network design

Our goal is to select an optimal subnetwork i of size $k < N$ under a Bayesian framework by considering all subsets of size k of $(x_1, \dots x_k)$. We calculate the entropy of the posterior predictive density of $S_i = (Z(x_{i_1}), Z(x_{i_2}) \dots Z(x_{i_k}))$ and choose the subnetwork with the maximum posterior predictive entropy. Sites with high uncertainties, which are generally more difficult to predict, are retained. Sites with smaller uncertainties are eliminated from the subnetwork. This has the desirable predictive feature of excluding sites characterized by high entropy values or higher uncertainties.

Let $g()$ be the predictive posterior density of S_i. In calculating the entropy of $g()$, we should point out that if $S_1, S_2, \dots, S_p$ is a sample from $g()$. Then,

$$\frac{1}{p}\sum_{i=1}^{p} -\log(g(S_i))$$

is an unbiased estimator of the entropy of $g()$. Therefore, if we can compute the value of $g(\,)$, and generate a sample from $g(\,)$, we should be able to estimate the value of the entropy of $g()$. Even though, we cannot explicitly compute the value of $g()$ at any point, say s_0, we can still estimate its value using the following approach:

- First, we generate a sample $\theta_1, \dots \theta_k$, from the posterior distribution of θ.
- Then, the posterior predictive density can then be estimated as

$$\hat{g}(s_0) = \frac{1}{k}\sum_{i=1}^{k} f(s_0|\theta_i).$$

We estimate the posterior predictive entropy as

$$\frac{1}{p}\sum_{i=1}^{p} -\log(\hat{g}(S_i)).$$

However, since $\log(\hat{g}(S_i)) \neq \log(g(S_i))$, the above expression may not be unbiased. Then, since

$$\frac{1}{k}\sum_{i=1}^{k} f(s_0|\theta_i)$$

converges almost surely to the value of the posterior predictive density at s_0, as $k \to \infty$ for each s_0, we can then get good estimates by choosing k large enough.

4 Utility of a Design

For a potential design S, we define a utility function $U(S)$. We selected a utility function that gives higher priority or weight to sites with measurements near the NAAQS for ozone. Other utility functions such as minus the monitoring cost could be used. Following Zidek *et al.* (2000), it seems natural to determine an optimal design by maximizing a combined monitoring objective:

$$H(S) + \gamma U(S),$$

where γ is a utility to entropy conversion factor. However, there is no natural way to choose γ, and we decided to pursue modifications to this approach that are detailed in the following Section.

4.1 The utility function

Our objective is to choose a final design that gives more weight to sites that are more likely to exceed the NAAQS for ozone. Thus, given a location x_0 we define the utility as

$$\begin{aligned} u(x_0) &= a_0 \exp((z(x_0) - c_1)/h) \text{ if } z(x_0) \leq k_0 \\ &= a_1 \exp((z(x_0) - c_0)/h) \text{ if } z(x_0) > k_0 \end{aligned}$$

Where c_0, c_1, a_0, a_1, k_0 are constants depending on air quality standards and h is a bandwidth parameter. This utility function assigns more weight to sites with observations that are more likely to be out of compliance with AQ standards. We define the utility of a subnetwork as the sum of the utilities of the sites in the subnetwork,

$$U(S) = \sum_{x \in S} u(x).$$

4.2 Preference relationships between designs

We would like to meet the dual objectives of maximizing the posterior predictive entropy while giving priority to sites with high O_3 observations. The *entropy-utility* combination for a design S, $(H(S), U(S))$ is a point in R^2 . However, it is not clear how to simultaneously achieve both purposes. Therefore, we introduce a preference relationship on R^2 to choose between any two designs, S_1 and S_2. Our goal is to select a subnetwork of sites characterized by both high entropy and high utility. If S_1 and S_2 are two designs, an obvious property any such preference relationship (say $>>$) must satisfy is:

- $H(S_1) > H(S_2)$ and $U(S_1) > U(S_2)$ then $S_1 >> S_2$.

If the entropy of one design is higher, but the utility value is lower,

$$H(S_{i_1}) > H(S_{i_2}) \text{ and } U(S_{i_1}) < U(S_{i_1}).$$

then we base our decision on the relative gain in entropy versus the relative loss in utility value. If

$$\frac{(H(S_{i_1}) - H(S_{i_2}))}{(H(S_{i_1}) + H(S_{i_2}))} > \frac{(U(S_{i_2}) - U(S_{i_1}))}{(U(S_{i_1}) + U(S_{i_2}))}.$$

Then,

$$S_{i_1} >> S_{i_2}.$$

If the reverse inequality holds, then we have

$$S_{i_1} << S_{i_2}.$$

When the ratios are equal, then we are indifferent to the choice between the two designs and we consider the two designs equivalent to each other,

$$S_1 \sim S_2,$$

and we pick one design at random.

5 Optimization problem

To this point, we have discussed a network design criterion that can be used to define a useful subnetwork or partition of the original network, but we have not discussed how to quantify an optimal partition. Typ-

ically, these design optimization problems for large sample sizes are highly formidable and pose enormous computational problems. Many previous design efforts have applied simple one-at-a-time addition and deletion procedures, that often lead incorrect solutions. We could consider a sequence of reduced networks, eliminating one station at a time, with subsequent calculation of the entropy associated with each of the resulting networks. Then choose the natural cut-off for the number of sites by inspecting a plot of entropy vs. number of sites. Another method of searching the design space is to sample blocks of design parameters. We will denote with $d = (d_1, \ldots, d_n)$ a vector of indicators specifying which stations remain in the network ($d_i = 1$) and which do not ($d_i = 0$). With N stations we have 2^N possible designs. For instance for $N = 80$ we could partition the design vector d in, eight blocks $d = (d_1, \ldots, d_8)$. We choose a new value of d_j given currently imputed values for d_i, $i \neq j$, the covariance parameters (θ) and future data. Each d_j has $2^{10} = 1024$ possible values, we would need to evaluate the 1024 possible designs using the information-base criteria, or any other approach. This is essentially Gibbs sampling (see Gelfand *et al*, 1990), since we are conditionally updating the design. Ko et al. (1995) discussed an exact algorithm for determining maximum entropy designs based on establishing the upper bound and incorporating this bound in a branch and bound method. This approach can find the exact solution that maximizes entropy for relatively small sample sizes, but it could not be implemented in our current setting, since we have a relatively large network. Thus, here we use a simulated annealing (SA) approach. SA is a heuristic maximization method. It has been inspired by the technique of slowly cooling a liquid to the lowest possible energy state. The different values a function can take are considered possible energy values, and the optimal value is reached by using a random search in an intelligent way. Suppose $f()$ is a function we want to maximize. Let $S = (s_1, \ldots, s_k)$ be the set of points where it attains a maximum. The goal of SA is to construct a non-homogeneous Markov chain that converges towards $\pi_\infty()$, a uniform distribution over S.

5.1 The simulated annealing algorithm

Consider a sequence $\{T_n\}$, called a cooling schedule, converging slowly to 0. We start with an initial subnetwork S_{i_1}. Then at the i^{th} step,

- We check the entropy and utility value of a random design say S_{i_2}.

- If $S_{i_1} << S_{i_2}$, then we update the value of S_{i_1} to S_{i_2}.

- If $S_{i_1} >> S_{i_2}$, then we update S_{i_1} to S_{i_2} with probability

$$\exp\left(-\frac{\frac{H(S_{i_1})-H(S_{i_2})}{H(S_{i_1})+H(S_{i_2})} - \frac{U(S_{i_2})-U(S_{i_1})}{U(S_{i_1})+U(S_{i_2})}}{T_n}\right).$$

Initially, when T_n is large, the probability of jumping to an inferior point is higher, but since $T_n \to 0$ the probability of jumping to an inferior point should become small after enough iterations. Various choices for the cooling schedule have been suggested in the literature, we use here a geometric cooling schedule, $T_n = T_0 c^n$, for $c = 0.8$. To find the initial subnetwork for our SA algorithm, we use the geometric space-filling design approach as described by Nychka and Saltzman (1998).

6 Modeling nonstationarity

In this section, we present a covariance model to characterize the potential lack of stationarity of the environmental spatial process in the entropy network design framework. We assume the space domain of interest D is divided into small subgrids, $R_1, \ldots, R_n$ centered at the nodes $r_1, \ldots, r_n$. The following equation describes the model we use, which is an extension of the model proposed by Fuentes (2001) (see also Fuentes, 2002),

$$Z(x) = Z_0(x) + \sqrt{\alpha}\sum_{i=1}^{n} K(x - r_i)Z_i(x), \tag{1}$$

$Z()$ is the process of interest, and $Z_0(\,), Z_1(\,) \ldots Z_n(\,)$ are underlying unobservable stationary processes, which are mutually independent and Gaussian, and explain the spatial structure of Z in each one of the subregions R_i. $K(x - r_i)$ is a weight function, e.g., the square inverse distance between x and r_i. Here $Z_0(\,)$ is a background stationary process, and

$$\sqrt{\alpha}\sum_{i=1}^{n} K(x - r_i)Z_i(x)$$

is the nonstationary component of our model. We further assume that the stationary processes $Z_0(\,), Z_1(\,), \ldots, Z_n(\,)$ have a Matern covariance (Matérn, 1960), $C_i(x - y)$, with parameter θ_i for each Z_i

$$C_i(x) = \tau_i^2 I_0(x) + \frac{\sigma_i}{2^{\eta_i - 1}\Gamma(\eta_i)}(2\eta_i^{1/2}|x|/\rho_i)^{\eta_i}\mathcal{K}_{\eta_i}(2\eta_i^{1/2}|x|/\rho_i), \tag{2}$$

where I is an indicator function, that takes the value 1 when $x = 0$ and it is zero otherwise, $\mathcal{K}_{\eta_i}$ is a modified Bessel function (Abramowitz and Stegun, 1964, pp. 374-379), $|x| = \sqrt{x_1^2 + x_2^2}$ denotes de modulus of the vector $x = (x_1, x_2)$. The parameter τ_i^2 is called the nugget parameter, and explains the microscale variability and measurement error. The parameter ρ_i measures how the correlation decays with distance; generally this parameter is called the *range*. The parameter σ_i is the variance of the random field, i.e. $\sigma_i = var(Z_i(x))$, and is usually referred to as the *sill*. The parameter η_i measures the degree of smoothness of the process Z_i, which becomes smoother with higher values of η_i. When η_i equals $\frac{1}{2}$, the Matérn model corresponds to the exponential covariance function. The Gaussian model is the limiting case of the Matérn as $\eta_i \to \infty$

$$C_i(x) = \sigma_s e^{-|x|^2/\rho_i^2}.$$

Parameters of the process

We assume, that given n, the n stationary processes $Z_1(\,), Z_2(\,), \ldots, Z_n(\,)$ are Matérn with range parameters $\rho_1, \ldots \rho_n$, partial sill parameters $\sigma_1^2, \ldots \sigma_n^2$, nugget parameters $\tau_1^2, \ldots \tau_n^2$ and smoothness parameters $\eta_1, \ldots \eta_n$ respectively. The locations $r_1, \ldots r_n$ are n points in our domain, that we call the *centers* of the process. The critical parameter α measures the deviation from stationarity – for a nearly stationary distribution we expect α to be small. This parameter was not included in the previous version of this nonstationary model as introduced by Fuentes (2001).

We observe the nonstationary process $Z(\,)$ at m locations $x_1, \ldots x_m$. That is we have observations $Z(x_1), Z(x_2) \ldots Z(x_m)$ where $x_1, x_2 \ldots x_m$ are m points in the plane. From eq. (1) we derive the form of the covariance between $Z(x_j)$ and $Z(x_k)$ where x_j and x_k are any two the points where we observe our process. Note that

$$\begin{aligned} E(Z(x_j)Z(x_k)) =& E(\{Z_0(x_j) + \sqrt{\alpha}\sum_{i=1}^{n} K(x_j - r_i)Z_i(x_j)\} \\ & \times \{Z_0(x_k) + \sqrt{\alpha}\sum_{i=1}^{n} K(x_k - r_i)Z_i(x_k)\}). \end{aligned}$$

Let us denote the stationary covariance of the process $Z_i(\,)$ by $C_i(\,)$. By our assumption of mutual independence of the $Z_i(\,)$'s the covariance between $Z(x_j)$ and $Z(x_k)$ simplifies to

$$C_0(x_j - x_k) + \alpha \sum_{i=1}^{n} K(x_j - r_i)K(x_k - r_i)C_i(x_j - x_k). \tag{3}$$

Hence, the vector $(Z(x_1), \ldots Z(x_m))$ is a multivariate normal distribution with covariance $\Sigma_{m\times m} = (\sigma_{jk})$. From above it follows that

$$\sigma_{jk} = C_0(x_j - x_k) + \alpha \sum_{i=1}^{n} K(x_j - r_i)K(x_k - r_i)C_i(x_j - x_k).$$

6.1 Choice of the Kernel function

For our kernel function K, we choose the Epanechnikov kernel given by [2]

$$K(u) = \frac{2}{\pi}\frac{1}{h^2}(1 - |\frac{u}{h}|^2)^+$$

Here, h is a bandwidth parameter. The choice of the bandwidth is important; large values of the bandwidth lead to oversmoothing, and small values to undersmoothing. We use the criterion developed by Fuentes and Smith (2001) to choose the bandwidth. They propose using $\frac{l}{\sqrt{2}}$ when the process is observed on an uniform grid of width l. However, since our data are not on a grid, we calculate for each point the distance to the nearest neighbor, that is, if we observe our process at locations $x_1, \ldots x_m$, we calculate $l_1, \ldots l_m$, where l_1 is the distance of the point closest to x_1 among $x_2, \ldots x_m$ and so on. We choose $\frac{l'}{\sqrt{2}}$ as our bandwidth, where l' is the median of $l_1, \ldots l_m$. This criterion for choosing the bandwidth coincides with the one proposed by Fuentes and Smith (2001) when the data lie on a grid.

6.2 Covariance Parameter Estimation

6.2.1 Bayesian Framework

We construct a Bayesian framework from our model formulation to estimate the covariance parameters. The parameters of our model are described by the following vector

$$\theta = (n, \alpha, r_1, \ldots, r_n, \rho_0, \ldots \rho_n, \sigma_0^2 \ldots \sigma_n^2, \tau_0^2 \ldots \tau_n^2, \eta_0, \ldots \eta_n)$$

We assume there is a compact rectangle $D \subset \mathcal{R}^2$ which defines our domain of interest and we are not interested in the values of our process outside D. In our applications D is obtained as the bounding rectangle

[2] $x^+ = max\{x, 0\}$

of the points where we observe the process. Note that the dimension of the above vector is $4n + 5$, and it depends on the first coordinate n, which is the number of *centers* of the process. Thus, our parameter vector is of a variable dimension. We choose a Poisson prior for θ.

Prior Distributions:

1. the prior for n, the number of centers, is a Poisson(λ),where λ is given a conjugate Gamma hyperprior with mean a and shape parameter b.

2. Given n, the n centers, $r_1 \dots r_n$ are i.i.d uniformly distributed over the domain D.

3. Given n, the partial sill parameters, $\sigma_0^2, \dots, \sigma_n^2$ are given i.i.d diffuse inverse gamma priors with mean m_s and shape parameter 2.

4. Given n, the nugget parameters, $\tau_0^2, \dots, \tau_n^2$ are given i.i.d diffuse inverse gamma priors with mean m_n and shape parameter 2.

5. Given n, the range parameters, $\rho_0, \dots \rho_n$ are given i.i.d inverse gamma prior with mean m_r and shape parameter 2.

6. Given n, the smoothness parameters, $\eta_0, \dots \eta_n$ are given i.i.d uniform on the set $\{0.5, 1.0, 1.5, 2.0, 2.5\}$

7. α is given a prior distribution that is uniformly distributed on the interval $[0, 1]$

Note that, $n \sim$ Poisson(λ) and $r_1, \dots, r_n$ being distributed uniformly over D, given the value of n, it is equivalent to assume that $r_1, \dots, r_n$ are distributed as a Poisson process over D with a constant rate λ. If we have more information about the number and the spatial distribution of the r_i's one might put a Poisson process prior on $(r_1, \dots r_n)$, with non-constant rate function $\lambda(x, y)$, where the rate function $\lambda(x, y)$ reflects the knowledge about the distribution of the *centers*. The Matérn covariance is very sensitive to changes in the smoothness parameter, a continuous prior for this parameter offers computational challenges. Thus, we work with a discrete prior based on analyses of similar datasets. For the partial sill and range parameters, we use diffuse Inverse Gamma priors (with shape parameter 2). Note that Inverse Gamma distributions with shape parameter 2 have infinite variance. Our inference about θ is based on its posterior distribution, i.e. its distribution conditioned on the observed data. Usually, the inference about the posterior distribution

is made by constructing a Markov Chain which has the posterior distribution as its stationary distribution. However, since our parameter has a variable dimension, the usual Markov Chain Monte Carlo methods do not work. Thus, we use a RJMCMC approach developed by Green (1995) that constructs a Markov chain with a distribution of a specified variable dimension distribution as the stationary distribution. The stages of our approach to estimate the covariance mixture are:

(stage 1) updating the local covariances of the mixture components (θ_i), for a fixed n,

(stage 2) adding or dropping a mixture component,

and we iterate through stage 1 and 2. We judge convergence using the Brooks and Guidici (2000) approach.

7 Application

We apply the entropy-utility design approach to 1997-1999 O_3 design values to find optimal subnetworks of the original network of 513 monitoring sites. This approach involves:

1. modeling the design values as a nonstationary spatial process Z, allowing the covariance parameters to change with location to explain the lack of stationarity. The process Z would have a covariance function C given in (3), that is a mixture of n stationary covariance functions;
2. projecting the site coordinates using the Lambert projection;
3. modeling the mean function of $Z(x)$ as a polynomial on x of degree 3;
4. applying the RJMCMC approach to estimate n, the number of nodes that determine the center of the subregions of stationarity.

In this case the estimated value of n is 3. Figure 2 shows the location of the 3 nodes: in the north-east (node 1); in the southern part of the domain (node 2); and in the Midwest (node 3). This indicates that there are three zones of nonstationarity. The posterior mean of the parameter α was .9. Figures 3 to 5 show the posterior distributions for the partial sill, range and nugget parameters for all nodes, using the priors described in Section 6. The posterior mean for the smoothness parameter was .5 for the 3

subregions of stationarity, so we fixed this parameter at .5 to simplify the entropy computation. The partial sill corresponding to subregion 3 has more uncertainty associated to it (see Figure 3), probably due to the proximity of the edge of our domain. Subregion 1 (North-east) shows larger values for the partial sill. The range of autocorrelation for the subregion of stationarity corresponding to node 1 is significantly smaller than for node 3 (Figure 4), probably due to the spatial heterogeneity in the North-East because of the proximity to the coast, and the presence of big cities. The nugget for subregion 2 is significantly smaller (Figure 5). Figure 6 shows the map of interpolated ozone design values. We interpolated the ozone values using a Bayesian approach for spatial prediction, the values in Figure 6 are the mean of the posterior predictive distribution for the ozone design values. The ozone air quality design values are very high in the north east, in all the geographic areas close to large metropolitan areas. The ozone design values are also high in some areas further south in our domain, e.g. in Atlanta (Georgia), Memphis (Tennessee), and Charlotte (North Carolina). In the Midwest we have some high design values in Dayton and Cleveland (Ohio) and Pittsburgh (Pennsylvania) has high ozone design values.

Figure 8 shows the utility function used for each monitoring site. In our entropy framework we give more weight to sites with high utility values, that correspond to sites with greater risk of non-compliance. Accordingly, in Figure 8 ozone design values greater than .08 ppm (80 ppb) have higher utility values since the O_3 NAAQS, is .08 ppm. The bubble graph (Figure 8) shows the estimated probability for an individual site to be included in the final partition or subnetwork. This graph was calculated by simulating samples of covariance parameter values from the posterior distribution of these parameters (using the posterior distributions in Figures 3, 4 and 5), and then computing the best subnetwork for each sample of parameters, using the proposed entropy-utility design approach. Monitoring sites within regions near the industrial areas in the north east are clearly more likely to be part of the final network than sites that have low pollution. Fixing the desired number of monitoring sites to be 252, about one-half of the original network, Figure 9 shows the optimal partition using the Bayesian entropy-utility framework. This subnetwork is chosen as the one with the maximum value (using SA) of the posterior predictive distribution for the entropy (where the covariance parameters are updated using RJMCMC). Our approach can be used to find optimal partitions for any subnetwork size.

We compare our final partition (Figure 9) to a partition obtained using an alternative method proposed by Nychka *et al* (1998) (Figure 10). The final network (of size 252) in Figure 10 has sites very uniformly spaced, this is due to the assumption of stationary in the Nychka *et al*'s approach. Figure 9 shows that the sites in the final partition are much more irregularly spaced, and more dense where the process is more heterogeneous, because it is more difficult to predict ozone values at those locations. Also, due to the utility function used in this application, the final partition in Figure 9 retains most of the sites that have high ozone values, because those sites have higher risk of being out of compliance and our utility function gives them more weight.

8 Conclusions

We propose a new entropy-utility design criterion based on evaluating the posterior predictive entropy constrained by giving higher utility to maintaining sites with measurements near the ozone national ambient air quality standard. Simulated annealing is used to select an optimal subnetwork for a fixed sample size based on a preference relationship between entropy and utility. This approach accounts for the potential lack of spatial stationarity in the underlying pollutant process and implements a Bayesian framework for modeling the uncertainty about the covariance parameters. For the large-scale national ozone monitoring network, this approach shows great promise in efficiently quantifying optimal reduced monitoring networks. Future research will address finding optimal subnetworks relative to monitoring budget limits, evaluate networks where a significant temporal structure might impact the choice of optimal subnetworks, and consider non-Gaussian spatial processes.

Disclaimer

The U. S. Environmental Protection Agency's Office of Research and Development partially collaborated in the research described here. Although it has been reviewed by EPA and approved for publication, it does not necessarily reflect the Agency's policies or views.

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions.* Dover, New York.

Bras, R. L. and Rodriquez-Iturbe, I. (1976). Network design for the estimation of areal mean of rainfall events. *Water Resources Research,* **12**, 1185-1195.

Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics,* **7**, 686-690.

Brooks, SP and Guidici, P. (2000) MCMC Convergence Assessment via Two-Way ANOVA. *Journal of Computational and Graphical Statistics,* **9**, p266-285.

Caselton, W. F., and Zidek, J. V. (1984). Optimal monitoring network designs. *Statistics and Probability Letters,* **2**, 223-227.

Cressie, N., Gotway, C. A., and Grondona, M. O. (1990). Spatial prediction from networks. *Chemometrics and Intelligent Laboratory Systems,* **7**, 251-272.

Fuentes, M. (2001). A new high frequency kriging approach for nonstationary environmental processes. *Envirometrics,* **12**, 469-483.

Fuentes, M. (2002). Modeling and prediction of nonstationary spatial processes. *Statistical Modeling,* **2**, 281-298.

Fuentes, M. and Smith, R. (2001). A new class of nonstationary models. Tech. report at North Carolina State University, Institute of Statistics Mimeo Series #2534.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating margnal densities. *Journal of the American Statistical Association,* **85**, 398-409.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and bayesian model determination, *Biometrika,* **82**, 711-732.

Guttorp, P., Le, N.D., Sampson, P. D., and Zidek, J. V. (1993). Using entropy in the redesign of an environmental monitoring network. Patil, G. P., Rao, C. R. (eds.), *Multivariate Environmental Statistics.* North-Holland, Amsterdam, pp 173-202.

Ko, Chun-Wa, Lee, J., and Queyranane, M. (1995). An exact algrotihm for maximum entropy sampling. *Operations Research,* **43**, 684-691.

Lindley, D.V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27**, 986-1005.

Müller, P. (1999). Simulated-based optimal design. *Bayesian Statistics,* **6**, 459-474.

Müller, W. G. and Zimmerman, D. L. (1999). Optimal designs for variogram estimation. *Environmetrics,* **10**, 23-27.

Nychka, D. and Saltzman, N. (1998). Design of air quality networks. In *Case Studies in Environmental Statistics,* eds. D. Nychka, W. Piegorsch and L.H. Cox, Lecture Notes in Statistics number 132, Apringer Verlag, New York, pp.51-76.

Theil, J. and Fiebig, D. G. (1984). *Exploiting Continuity: Maximum Entropy Estimation of Continuous Distributions.* Cambridge, Massachusetts: Ballinger Publishing Company.

U. S. Environmental Protection Agency (2003). *National Air Quality and Emissions Trends Report, 2003 Special Studies Edition.* U. S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC 27711, EPA 454/R-03-005

Warrick,, A. W. and Myers, D. E. (1987). Optimization of sampling locations for variogram calculations. *Water Resources Research,* **23**, 496-500.

Wikle, C. K. and Royle, J. A. (1999). Space-time dynamic design of environmental monitoring networks. *Journal of Agricultural, Biological, and Environmental Statistics,* **4**, 489-507.

Yfantis, E. A., Flatman, G. T., and Behar, J. V. (1987). Efficiency of kriging estimation for square, triangular, and hexagonal grids. *Mathematical Geology,* **19**, 183-205.

Zidek, J., Sun, W., and Le, N. (2000). Designing and integrating composite networks for monitoring multivariate Gaussian pollution fields. *Applied Statistics,* **49**, 63-79.

Figure 2: Location of the nodes that define the subregions of stationarity. Location of the 3 nodes: in the north-east (node 1); in the southern part of the domain (node 2); and in the Midwest (node 3).

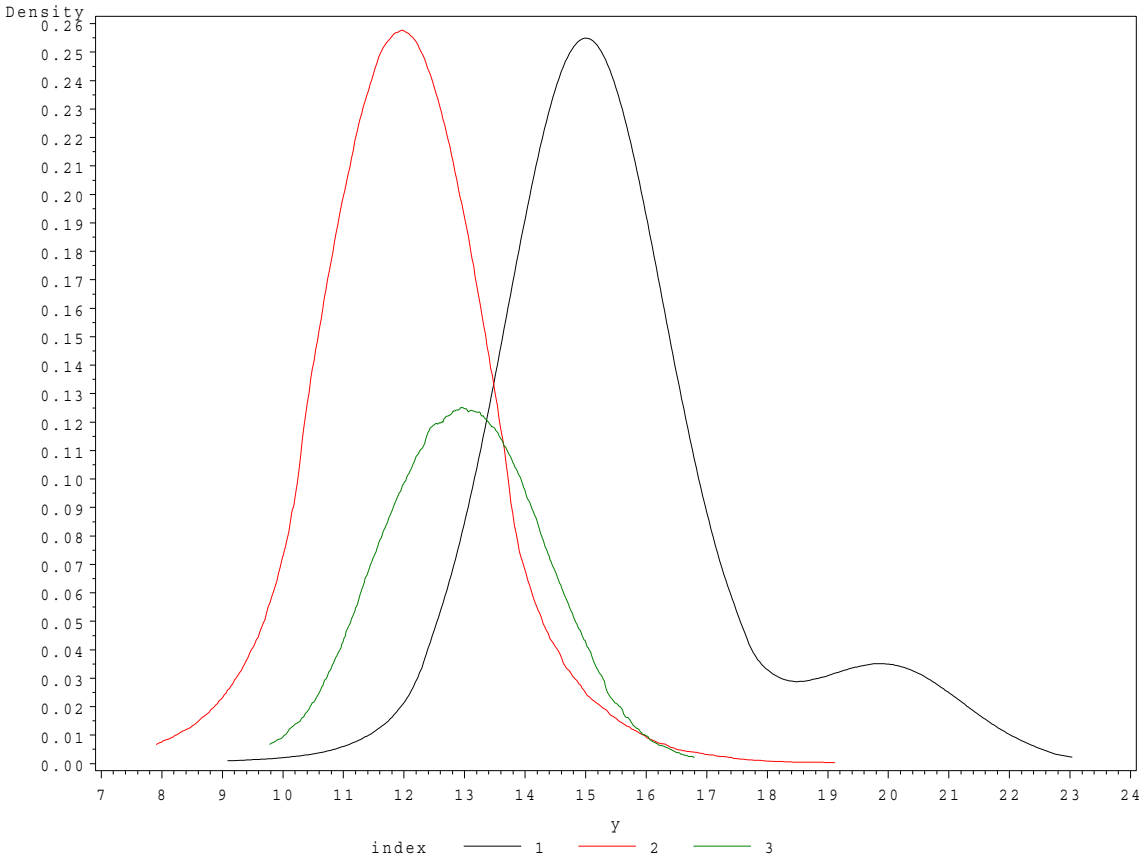

Figure 3: Posterior distribution for the partial sill parameters. The indexes correspond to the three nodes in Figure 2 (node 1 in the North-east, node 2 in the southern part of the domain, and node 3 in the Midwest).

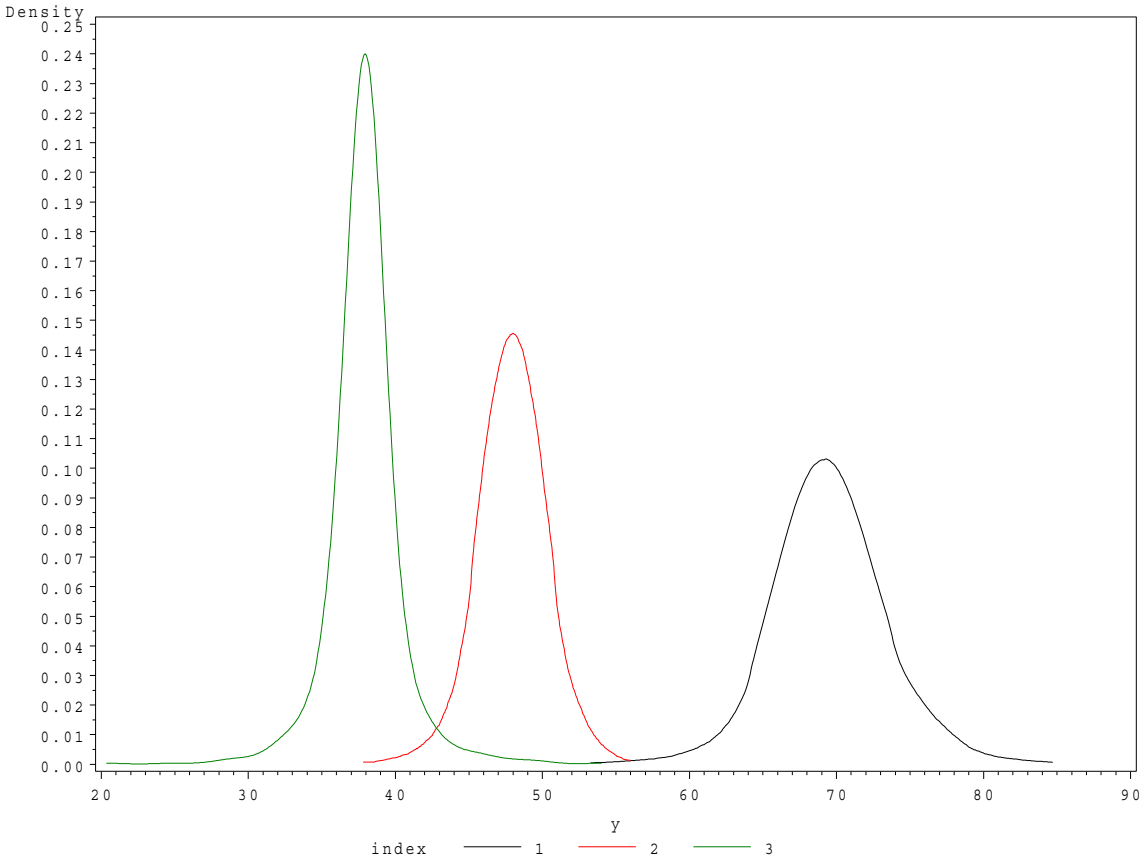

Figure 4: Posterior distribution for the range parameters.

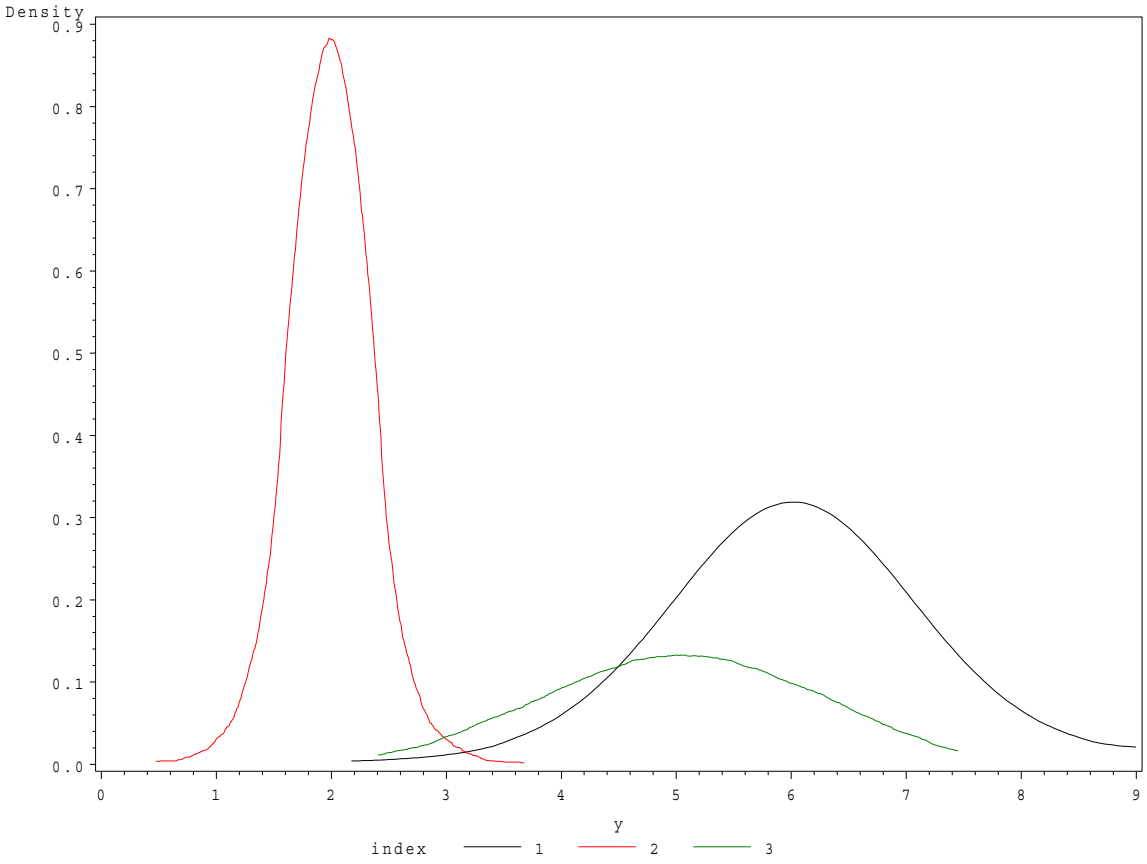

Figure 5: Posterior distribution for the nugget parameters.

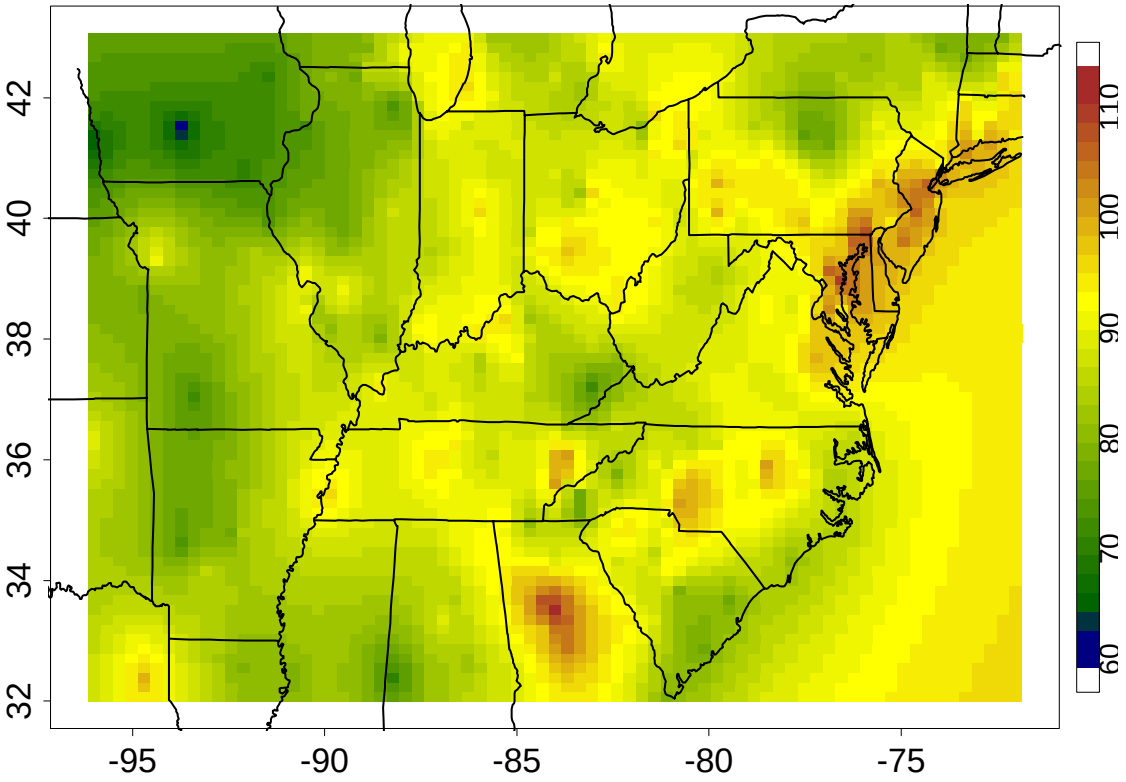

Figure 6: Mean of the posterior predictive distribution for the ozone design values.

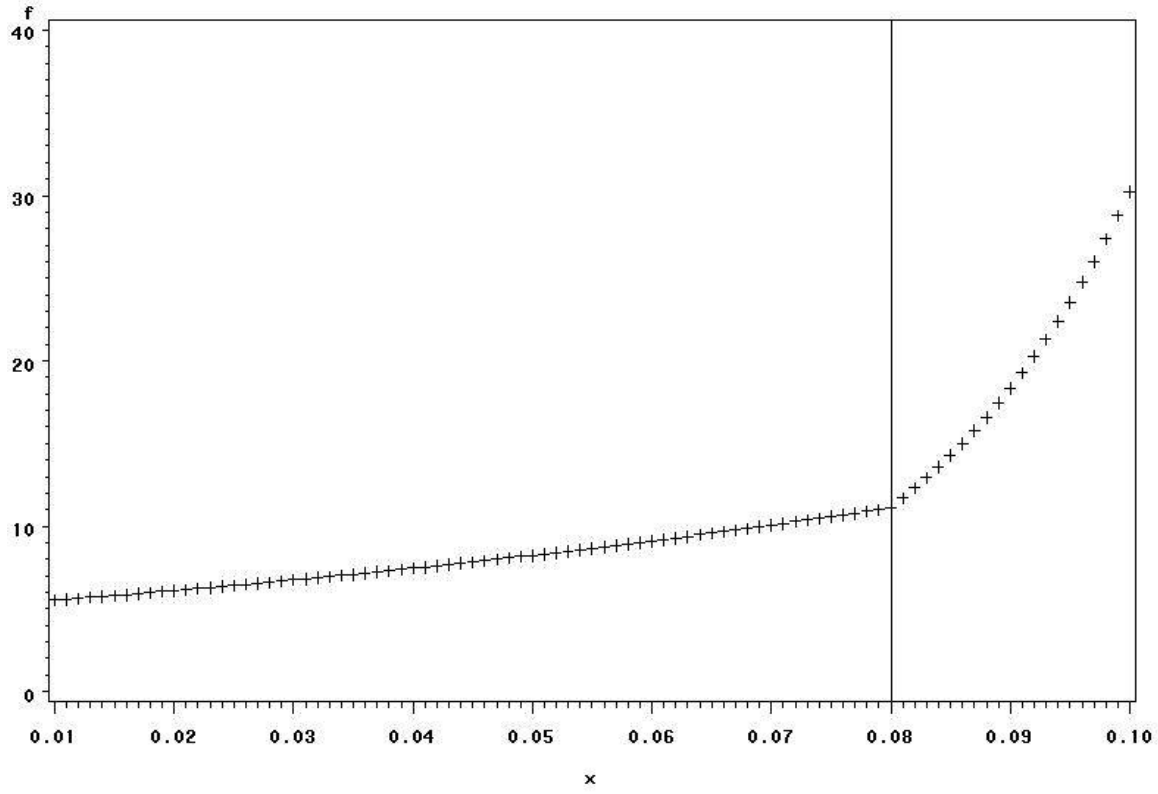

Figure 7: Utility function used for each site. Sites with observations near or above the O_3 air quality standard receive higher priority for inclusion in the final subnetwork. The horizontal axis shows the ozone design values and the vertical axis the corresponding utility values.

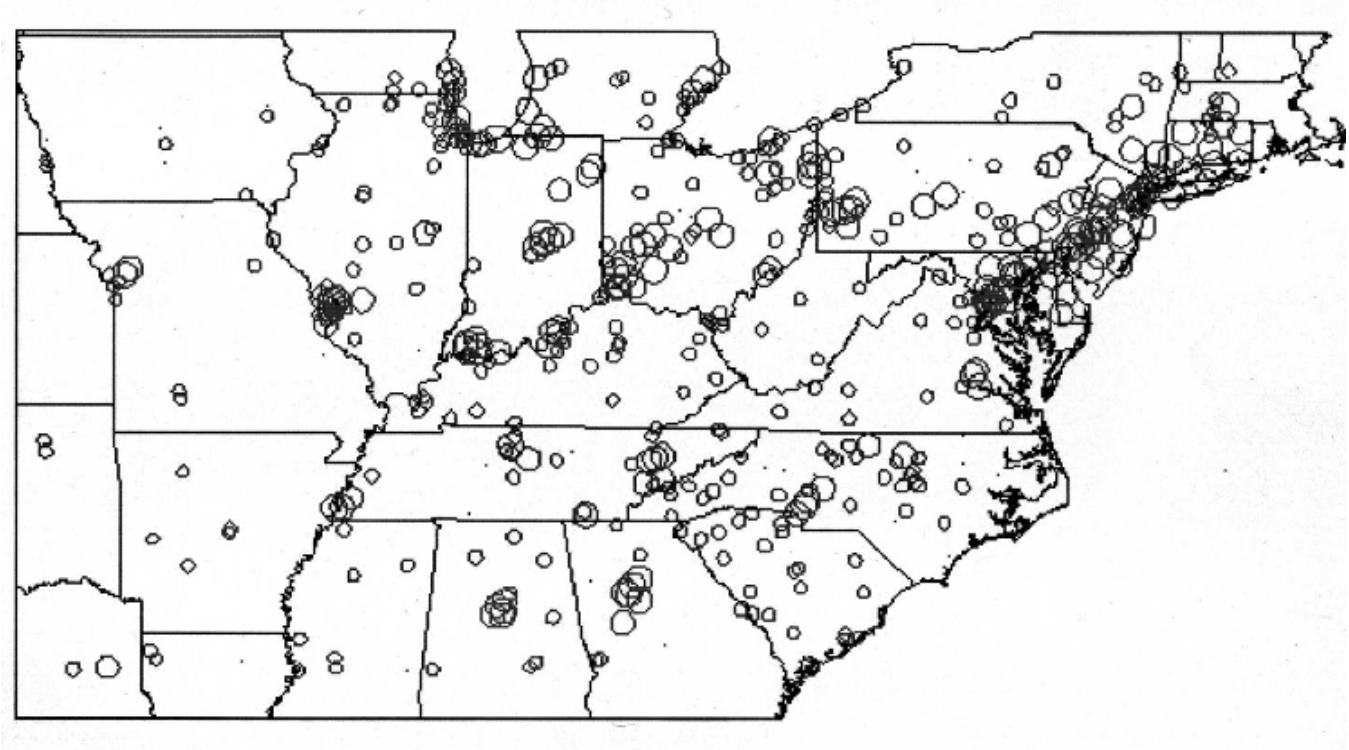

Figure 8: Probabilities of including sites in the final subnetwork.

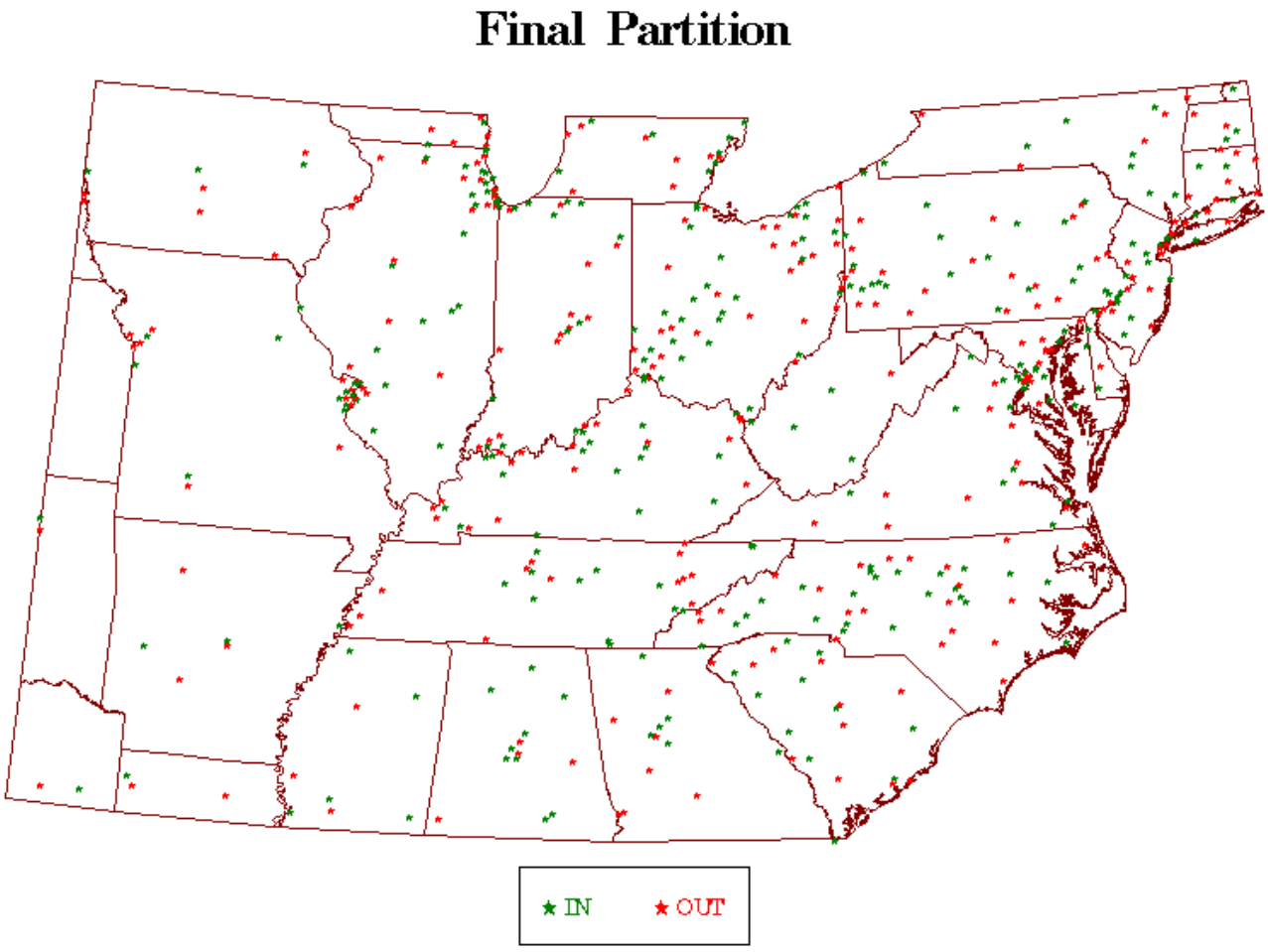

Figure 9: Final partition of size 252 using a fully Bayesian entropy approach.

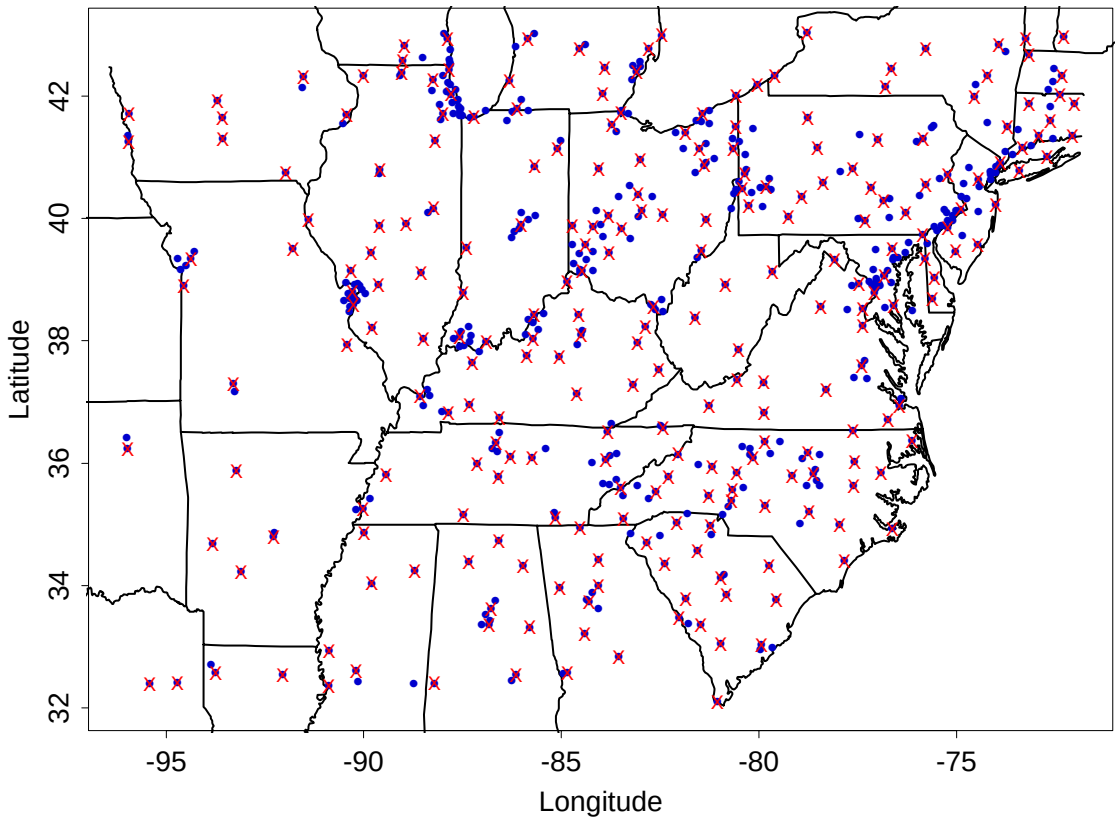

Figure 10: Final partition of size 252 using Nychka and Saltzman (1998) approach. The crosses indicate the sites that we keep in the final partition.