EPA Workshop, RTP, NC. December 4, 2001.

Entropy approaches for air pollution monitoring network design

Montserrat Fuentes

Statistics Department NCSU

fuentes@stat.ncsu.edu

http://www.stat.ncsu.edu/~fuentes

In collaboration with:

Arin Chaudhuri (NCSU)

Dennis Boos (NCSU)

Dave Holland (EPA)

EPA OAQPS

Slide 1

Motivation

- As the US EPA considers changes in funding for air and deposition monitoring, the Agency might need to **downsize** existing monitoring networks and find the most **informative** set of monitoring sites to achieve similar predictive capabilities of the complete network. EPA also has some national monitoring networks still **under development** for some new pollutants.
- The **Clean Air Act** established national ambient air quality standards for six air pollutants. OAQPS is interested in understanding the spatial behavior of pollutants and determining the spatial areas of non-attainment to judge compliance with ambient air quality standards.

Slide 2

OBJECTIVES:

Our main objective is to help EPA to design monitoring networks with good predictive capability, and estimate the spatial structure of air pollutants. More specifically, we seek to:

- determine a subset of monitoring sites (in this case from the SLAMS/NAMS network) with good predictive capabilities,
- establish measures of appropriateness of a subset,
- determine the size of a minimal optimal set,
- decide where to add sites for a network still under development,
- obtain an optimal network for multi-pollutants (e.g. ozone and PM),
- combine data from different networks, or different sources (i.e. models-3 and ground measurements),
- estimate spatial structure of air pollutants.

Slide 3

OUTLINE

- Entropy approaches to network design.
  - APPLICATION: optimal monitoring designs for SLAMS/NAMS.
- Modeling, prediction and network design for nonstationary air pollution processes.
  - APPLICATION: spatial areas of non-attainment for ozone.
- Fully Bayesian entropy approach to explain uncertainties.
- Adding sites to a new network.
- Combining data to improve air quality prediction and also for network design.
  - APPLICATION: combining Models-3 with CASNet to obtain more reliable maps.

Slide 4

Entropy approaches

DEFINITION: We define the **information** contained in a r.v. X with density f as

$$I(X) = E\{\log f(X)\}$$

The **entropy** is $H(X) = -I(X)$, and explains the uncertainty about X.

THE PROBLEM: Suppose that EPA has the funds for n sites, we want to distribute these sites at m desirable locations, where m is bigger than n.

NOTATION: Z is the vector of observations at all m sites. Z is subdivided into a vector Z_1 at the $m-n$ ungauged sites and Z_2 at the n gauged sites.

Shannon's information index: it is a measure of the **information gained** about Z_1 as a result of

Slide 5

measuring Z_2

$$I(Z_1|Z_2) - I(Z_1)$$

SOLUTION:

- Approach (i).
  We choose the design (Z_1 and Z_2) that maximizes:
  $$I(Z_1|Z_2) - I(Z_1)$$
- Approach (ii):
  We choose the design that minimizes the entropy (uncertainty) in predicting Z_1 given Z_2, i.e. $H(Z_1|Z_2)$. This approach is equivalent to maximize $H(Z_2)$.

Approaches (i) and (ii) are not equivalent.

Slide 6

If we have a Gaussian process:

Suppose we write Z in partitioned form as $\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ and similarly $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, where μ is the mean and Σ the covariance of Z.

Then maximizing Shannon's information is the same as maximizing

$$-\frac{1}{2}\log\prod_{i=1}^{n}(1-\rho_i^2), \qquad (1)$$

where $\rho_1^2, ..., \rho_n^2$ are the eigenvalues of $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$. These are the *canonical correlations* between Z_1 and Z_2. Thus the problem of network design becomes one of minimizing $\prod(1-\rho_i^2)$ where $\rho_1^2, ..., \rho_n^2$ are the squared canonical correlations.

Slide 7

The value of Shannon information will intially increase with the number of network stations but, after reaching a peak value, will subsequently decline as the number of locations in G (gauged sites) begings to dominate the number in U (ungauged sites).

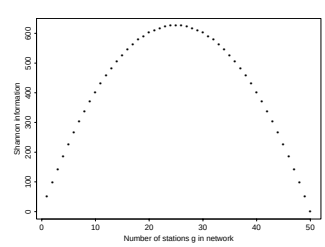


Figure 1: The total number of sites is 50, Shannon information is maximized around 25.

Slide 8

Approach (ii)

One can decompose the total entropy

$$H(Z_1, Z_2) = H(Z_1|Z_2) + H(Z_2)$$

and minimize $H(Z_1|Z_2)$, or equivalently maximize $H(Z_2)$.

In the Gaussian case, this means to maximize the covariance of the gauged sites, i.e. maximize $|\Sigma_{22}|$. The value of $H(G)$ increases as we increase the number of gauged sites.

These two approaches have been particularly developed by Zidek and his co-workers.

Slide 9

We have considered μ, Σ *known*. The simplest case is when we have N observations at time points $t_1, ..., t_N$ (assumed to be independent in time) at all m data locations. Then, we estimate the correlation between two location using the observations over time (sample correlation).

Issues in maximization and computation:

- A complete solution to the maximization problem involves searching over a prohibitively large set.

We implement a computer code called the Simple Genetic Algorithm (SGA) (Goldberg, 1989) to **optimize** the entropy function. Genetic algorithms process populations of strings, a string is a vector of binary integers. In our case the length of each string is m, we use the value 1 for the stations we keep, and 0 for the stations we eliminate.

Slide 10

Approach to initialize the SGA.

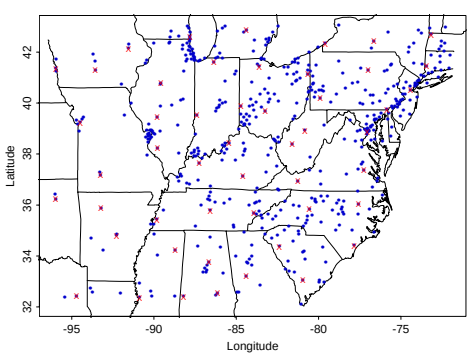


Figure 2: Design obtained by using the routine **cover.design** (Nychka et al). This is a geometric criterion that selects stations evenly distributed over the domain. The total number of sites is 513, and the number of stations in the design is set to be 50. The red cross-points represent the 50 selected stations.

Slide 11

Entropy approach: Application

We have 513 SLAMS/NAMS and CASTNet sites, located irregularly across regions. Hourly ozone concentrations are being measured at each site. We calculated the daily maximum of 8 hour running averages, from May to October of 1995-1999, a total of 920 days.

We present optimal designs using entropy approaches (i) and (ii). The covariance is empirically estimated using the observations over time.

Slide 12

Figure 3: The graph shows the locations of the 513 sites where the ozone ambient concentrations are measured hourly.

Slide 13

Figure 4: Best network design using entropy criterion (ii) (eliminating 30% of sites). The graph shows the locations of the gauged (sites to keep) and ungauged sites (sites to be dropped) using the entropy criterion (ii). The total number of sites is 513 and the number of (gauged) stations in this design in 360.

Slide 14

Figure 5: Best network design of size 252 using entropy criterion (i) (Shannon information). The graph shows the locations of the gauged (sites to keep) and ungauged sites (sites to be dropped) using Shannon information. The number of (gauged) stations in this design in 252.

Slide 15

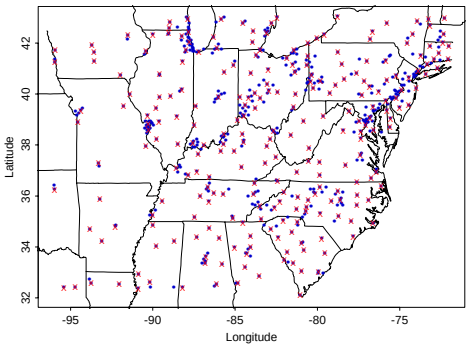


Figure 6: Design obtained by using routine **cover.design**. The number of stations in this design is set to be 252. This is a geometric criterion that selects evenly distributed sites over the domain.The final partition is very different from the resulting one using entropy criteria, because of the **lack of stationarity** of the data. The red cross-points represent the selected stations.

Slide 16

Nonstationary air pollution

We represent the nonstationary process Z observed on a region D as a MIXTURE of orthogonal local stationary processes (Fuentes, *Environmetrics*, 2001):

$$Z(\mathbf{x}) = \sum_{i=1}^{k} Z_i(\mathbf{x}) w_i(\mathbf{x})$$

where $S_1, \ldots, S_k$ are well-defined subregions that cover D, and Z_i is a stationary process with covariance C_i that represents the spatial structure in the subregion S_i, $w_i(\mathbf{x})$ is a positive kernel function. (inverse distance between $\mathbf{x}$ and S_i).

Slide 17

The nonstationary covariance of Z is defined in terms of the stationary covariances of the processes Z_i for $i = 1, \ldots, k$ (Fuentes & Smith, 2001),

$$\mathrm{cov}(Z(\mathbf{x}), Z(\mathbf{y})) = \sum_{i=1}^{k} w_i(\mathbf{x}) w_i(\mathbf{y}) \mathrm{cov}(Z_i(\mathbf{x}), Z_i(\mathbf{y}))$$

this is a valid nonstationary covariance.

Slide 18

We divide the domain D into small subgrids, $S_1, \ldots, S_k$. We represent Z as a weighted average of orthogonal local stationary processes:

$$Z(\mathbf{x}) = \sum_{i=1}^{k} Z_i(\mathbf{x}) w_i(\mathbf{x})$$

Z_i is a local stationary process in the subregion S_i, with stationary covariance $C_{\boldsymbol{\theta}_i}$, and $w_i(x)$ is a weight function.

We start with a small number of subregions (small k), in the next step we increase k by dividing each S_i in half, we iterate this process till a BIC suggests no significant improvement in the estimation of $\boldsymbol{\theta}_i$ for $i = 1, \ldots, k$.

Slide 19

The covariance of Z can be defined in terms of the covariance of the orthogonal local stationary processes Z_i

$$\mathrm{cov}(Z(\mathbf{x}), Z(\mathbf{y})) = \sum_{i=1}^{k} w_i(\mathbf{x}) w_i(\mathbf{y}) \mathrm{cov}(Z_i(\mathbf{x}), Z_i(\mathbf{y}))$$

this is a valid nonstationary covariance. Assuming a Matérn covariance structure $C_{\boldsymbol{\theta}_i}(\mathbf{x} - \mathbf{y})$ with parameter $\boldsymbol{\theta}_i$ for each Z_i we obtain,

$$\mathrm{cov}(Z(\mathbf{x}), Z(\mathbf{y})) = \sum_{i=1}^{k} w_i(\mathbf{x}) w_i(\mathbf{y}) C_{\boldsymbol{\theta}_i}(\mathbf{x} - \mathbf{y}).$$

Slide 20

Estimation of the covariance parameters.

We write $\hat{\rho}_i$ (range), $\hat{\nu}_i$ (smoothness), $\hat{\sigma}_i$ (sill), and $\hat{c}_i$ (nugget), to denote the estimated values of ρ_i, ν_i, $\sigma_i$and c_i that maximize the **likelihood**.

We obtain the estimated covariance, $\hat{C}$,

$$\hat{C}(Z(\mathbf{x}), Z(\mathbf{y})) = \sum_{i=1}^{k} w_i(\mathbf{x}) w_i(\mathbf{y}) C_{\hat{\boldsymbol{\theta}}_i}(\mathbf{x} - \mathbf{y}).$$

where $\hat{\boldsymbol{\theta}}_i = (\hat{\rho}_i, \hat{\alpha}_i, \hat{\sigma}_i, \hat{c}_i)$.

Slide 21

ANOTHER APPROACH:

Bayesian approach for prediction

Using a Bayesian approach to predict/estimate Z at a location of interest, we obtain a predictive distribution at each location, instead of just a predictive value. The predictive distribution can be used for scientific inference.

If the goal is to predict Z at a location $\mathbf{x}_0$, the Bayesian solution is the predictive distribution of $Z(\mathbf{x}_0)$ given the observations $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_N))$,

$$p(Z(\mathbf{x}_0)|\mathbf{Z}) \propto \int p(Z(\mathbf{x}_0)|\mathbf{Z}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{Z})\, dh\, d\boldsymbol{\theta}.$$

We simulate m values,$\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{m}$, from the posterior of the parameter $\boldsymbol{\theta}$. Thus, the predictive distribution is approximated by:

$$p(Z(\mathbf{x}_0)|\mathbf{Z}) =$$

$$\frac{1}{m} \sum_{i=1}^{m} p(Z(\mathbf{x}_0)|\mathbf{Z}, \boldsymbol{\theta}^{(i)}).$$

Slide 22

Application to air quality

Our first goal is to understand and quantify the spatial structure of the ozone 8-hour standard using EPA sites, and then interpolate these values to determine the regions of non-attainment.

Useful tools an approaches for computation:

- Hierarchical modeling
- Markov Chain Monte Carlo
- Fast Fourier Transform
- Conjugate gradient algorithm

Slide 23

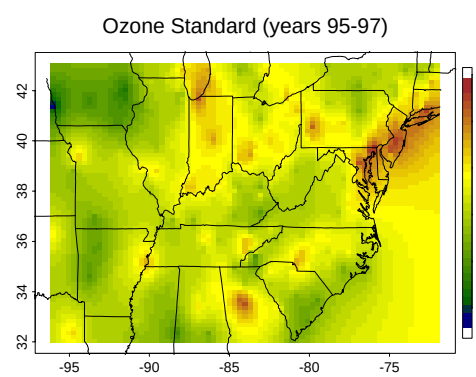


Figure 7: The graph shows the interpolated values (mean of the posterior predictive distribution) of the ozone air quality design values (ppb) using a Bayesian approach. The design values are calculated as the 3-year average of the annual fourth-highest daily maximum 8-hour average ozone concentration at the SLAMS/NAMS sites.

Slide 24

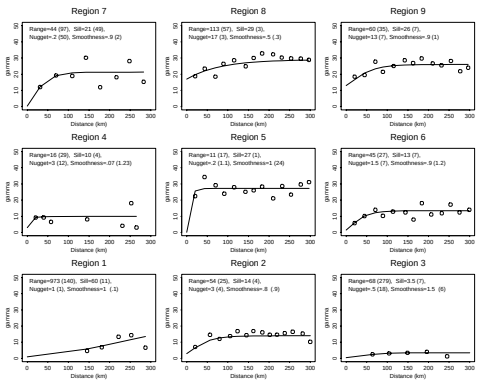


Figure 8: This graph shows the empirical semivariograms for the ozone air quality design values in 9 subregions of equal size that cover the domain. We fitted to the empirical semivariograms the likelihood estimates based on the covariance model, $\mathrm{cov}(Z(\mathbf{x}), Z(\mathbf{y})) = \sum_{i=1}^{9} K(\mathbf{x}-\mathbf{s_i})K(\mathbf{y}-\mathbf{s_i})C_{\boldsymbol{\theta}(\mathbf{s}_i)}(\mathbf{x}-\mathbf{y})$. The semivariograms show how the spatial structure of the ozone design values changes with location.

Slide 25

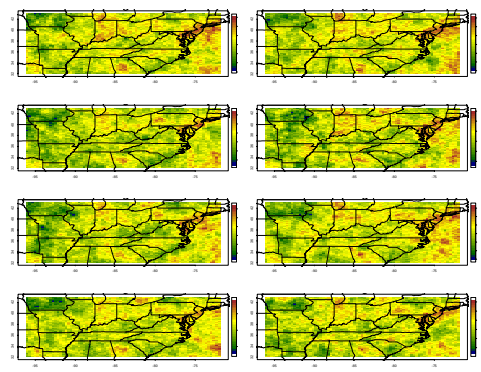


Figure 9: Six simulations from the predictive posterior distribution of the ozone ambient air quality design values (ppb).

Slide 26

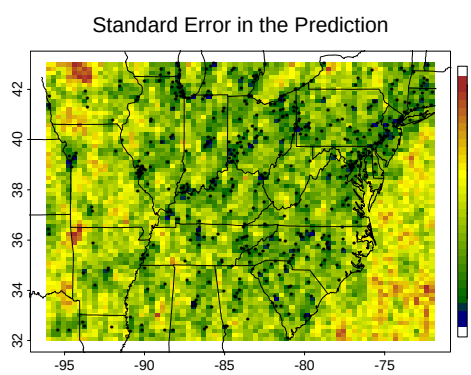


Figure 10: Standard error for the predictive posterior distribution of the ozone ambient air quality design values (ppb). The dots show the locations of the monitoring sites for the ozone.

Slide 27

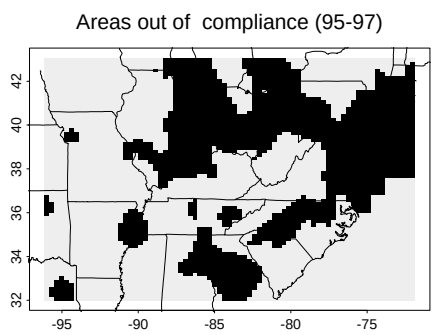


Figure 11: This graph shows non-attainment designations regions (areas in black in the map) based on the ozone 8-hour standard that EPA defined in July 1997. The ozone ambient standards are met when when the ozone design values are less than 85 ppb.

Slide 28

Network Design for Nonstationary Data: A Fully Bayesian Approach

Hierarchical approach:

- Step 1:

$$Z|\beta, \theta \sim \text{Normal}(\mu, \Sigma_\theta)$$

  where Σ_θ is the proposed nonstationary covariance, a mixture of stationary local covariances.

  The mean μ is a function of meteorological and geographic covariates $Y_1, \ldots, Y_k$, with coefficients β.

  We allow for multiple pollutants: $Z(s) = (Ozone(s), PM(s))$.

- Step 2:

$$\beta \sim \pi(\beta)$$
$$\theta \sim \pi(\theta)$$

  we define a prior distribution for β and θ.

Slide 29

We calculate $\hat{H}(G)$, using an iterative simulation Monte Carlo approach to estimate the entropy in G.

Slide 30

Covariates

Weather covariates: average relative humidity, average wind speed, average specific humidity, average temperature, maximum temperature.

Slide 31

This figure shows the loocation of the 513 Ozone Monitoring Sites from SLAMS/NAMS network.

How can we integrate information from both data bases?

Slide 32

- How can we integrate information from both data bases?

  In the next section we will introduce a Bayesian melding approach to combine data from different sources with different spatial locations.

- When are covariates useful?

  Covariates are useful for sparse networks or networks under development.

Slide 33

Adding stations

We choose new sites to maximize $H(G_p)$ where G_p are all the gauged pseudosites after modifying the network.

We decompose $H(G_p)$ into elements representing the existing monitors and the supposed new monitors.

Then, the design criterion becomes choosing the added sites to maximize:

$$H(G_p) = H(G) + H(Add|G) \tag{2}$$

where the notation reflects the subdivision of stations into existing stations G and added stations Add. Since $H(Add|G)$ is the only term in (2) related to the choice of the added sites, maximizing (2) is the same as maximizing:

$$H(Add|G).$$

Slide 34

When the covariance of Add given G is **known**:

$$H(Add|G) = \frac{1}{2}\log|\Sigma_{Add|G}| + c$$

We take into account nonstationarity and uncertainty about the covariance, when it is **unknown**, by using a fully Bayesian approach with the nonstationary covariance model proposed here.

Slide 35

Incorporating Cost:

There is a cost associated with a potential monitoring site s, denoted $C(s)$.

Define $E(s)$ the entropy measure at site s, i.e. the uncertainty associated with adding the site s.

It seems natural to consider a combined objective of maximizing (Zidek, Sun and Le, 2000):

$$E(s) - \gamma C(s)$$

where γ is a cost to entropy conversion factor.

Anoter possibility is to rank potential pseudosites based on the ratio:

$$E(s)/C(s)$$

Slide 36

Combining data

Incorporating all the relevant information can be very difficult for various reasons. For instance, the data could be collected or constructed at different spatial/temporal scales, and the bias and measurement error of the available data might depend on the source of information. We present a method for combining data (Fuentes & Raftery, 2001).

We introduce the methodology with an example in air pollution.

Slide 37

Two sources of information for air fluxes:

I. Point Measures of Pollutant Concentrations

Atmospheric deposition takes place via two pathways: wet deposition and dry deposition. Wet deposition rates of acidic species across the United States have been well documented over the last 10 to 15 years; however, comparable information is *unavailable* for dry deposition rates. Since 1990 EPA operates approximately 50 sites through US to establish spatial patterns of deposition and concentration.

II. Regional Models (Models-3) Estimated Concentrations

The present generation of regional scale air quality models can consider land cover, plant growth rate, topography, and other factors in estimating pollutant concentrations and fluxes in a grid.

Slide 38

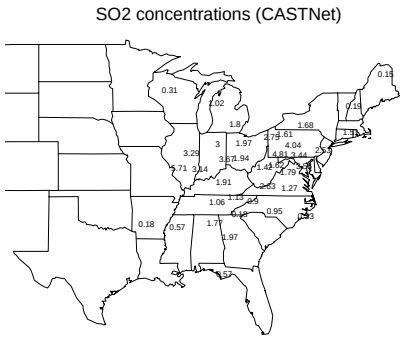


Figure 12: CASTNet weekly concentrations of SO_2.

Slide 39

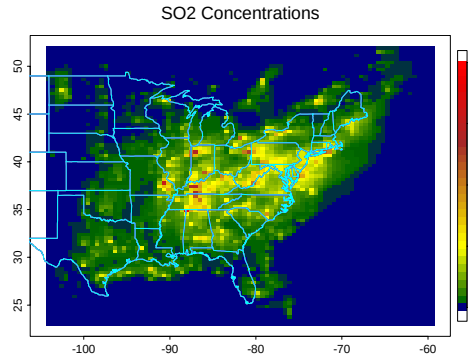
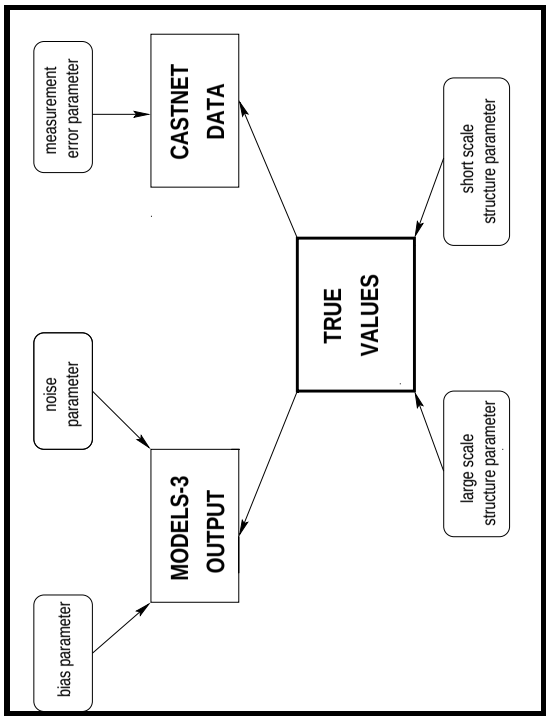


Figure 13: Output of Models-3, weekly average of SO_2 concentrations (ppb), for the week of July 11, 1995. The resolution is 36 km^2.

Slide 40



Slide 41

We should not treat CASTNet ($\hat{Z}$) measurements as the "ground truth". We assume there is some smooth underlying (but unobserved) field $Z(\mathbf{s})$, where $Z(\mathbf{s})$ measures the "true" concentration of the pollutant at location $\mathbf{s}$. We write

$$\hat{Z}(\mathbf{s}) = Z(\mathbf{s}) + e(\mathbf{s})$$

where $e(\mathbf{s}) \sim N(0, \sigma_e^2)$ represents the measurement error (nugget) at location $\mathbf{s}$.

Since the output of Models-3 ($\tilde{Z}$) are not point measurements but areal estimations in subregions $B_1, \ldots, B_N$ that cover the domain, D, we have:

$$\tilde{Z}(B_1) = a(B_1) + b \int_{B_1} Z(\mathbf{s}) d\mathbf{s} + \delta(B_1)$$

Slide 42

The true underlying process Z is a spatial process with a nonstationary covariance,

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s})$$

where $Z(\mathbf{s})$ has a spatial trend, $\mu(\mathbf{s})$, that is a function of some metereological and geographic covariates $f_1, \ldots, f_p$ that are know functions at some locations $\mathbf{s}$, with unknown coefficients $\boldsymbol{\beta}$:

$$\mu(\mathbf{s}) = \sum \beta_i f_i(\mathbf{s})$$

We assume $Z(\mathbf{s})$ has zero-mean correlated errors $\epsilon(\mathbf{s})$. The process $\epsilon(\mathbf{s})$ has a nonstationary covariance with parameter vector $\boldsymbol{\theta}$ that might change with location.

Slide 43

Combining CASTNet and MODELS-3

The goal is to predict the value of Z at location $\mathbf{x}_0$ given ALL the data (CASTNet and Models-3), thus we need the predictive distribution of $Z(\mathbf{x}_0)$ given the observations ($\hat{Z}$ and $\tilde{Z}$)

Slide 44

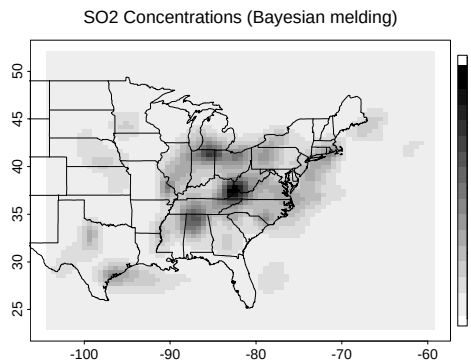


Figure 14: Predicted SO_2 concentrations via a Bayesian melding approach to combine CASTNet and Models-3 data. This graph shows the mean of the posterior predictive distribution for the underlying process Z given CASTNet and Models-3.

Slide 45

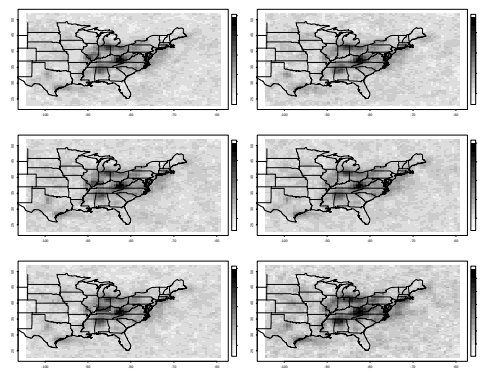


Figure 15: Simulated SO_2 concentrations from the predictive distribution for the SO_2, using a Bayesian melding approach for prediction to combine CASTNet and Models-3 data. The simulated values are from the posterior predictive distribution for the underlying process Z given CASTNet and Models-3

Slide 46

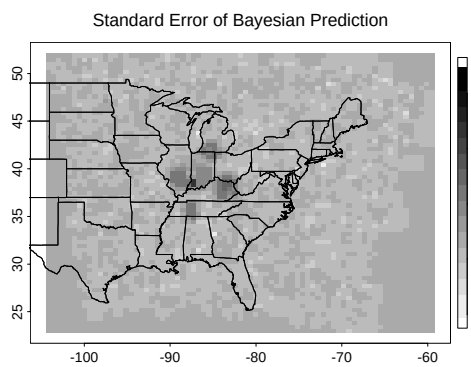


Figure 16: Standard error of the posterior predictive distribution for the SO_2 concentrations of Z, using a Bayesian melding approach for prediction to combine CASTNet and Models-3 data.

Slide 47

Network design with combined spatial data

The conditional distribution of Z given ALL the data $\mathbf{Z}$ (CASTNet and Models-3) is Gaussian with covariance Σ.

We denote the dependence on the design D by writing Σ_D in place of Σ. Then, the design problem becomes to choose D to maximize $\Phi(\Sigma)$ for some suitably chosen design criterion Φ. For example,

$$\Phi(\Sigma) = |\Sigma|,$$

or an entropy criterion to take into account the effect of estimating the model.

Slide 48

Conclusions

- We present a general framework (entropy) for network design, taking into account lack of stationary, different sources of uncertainties, multipollutants, covariates and costs. The entropy approach can be also used to decide where to add sites.
- We propose an approach to combine data from different sources, i.e. different networks.
- The approach discussed here can be used not only for network design but also to produce more reliable maps of air pollution, and to study and estimate the spatial structure of the data.
- The computer implementation is currently being done using genetic algorithms. We are on the process of comparing different algorithms for optimization, and also different approaches for network design.