EPA Workshop, RTP, NC. December 3, 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. Since there is no much experience or data with these new pollutants, new methods of network design that take advantage of selected **covariates** must be employed.

Slide 2

GOAL:

Help EPA to design monitoring networks with good predictive capability.

In our current project the objectives are:

- To choose a subset of monitoring sites (in this case from the SLAMS/NAMS network) with good predictive capabilities.
- Determine measures of appropriateness of a subset.
- Determine the size of a minimal optimal set.
- Joint analysis of multi-pollutants (e.g. ozone and PM).
- Combine data from different networks, or different sources (i.e. models-3 and ground measurements).

Slide 3

PROBLEM:

The complex spatial-temporal structure of air pollution processes. The challenge of dealing with a large number of pollutants, and data from different sources, i.e. monitoring data and models-3 output.

SOLUTION:

Novel statistical approaches are proposed, exploiting particularly the potential for Bayesian hierarchical models both in handling spatial variation, and for the joint analysis of a large number of pollutant variables.

Slide 4

Two strategies

Mathematical formulations of the network design problem follow generally one of two broad strategies, though there are several more *ad hoc* approaches.

The two broad strategies may be characterized as:

- Maximum Entropy.
- Optimal Design Approach.

Slide 5

Maximum Entropy

Suppose that EPA has the funds for n sites, we distribute these sites in m desirable locations, where m is bigger than n.

The problem: divide the m sites into n "gauged" sites (instrumented sites) and $m-n$ "ungauged" sites.

One possible way to formulate that problem is to choose the *gauged* sites so that predictions of the whole field based on those sites will provide the maximum possible information about the *ungauged* sites.

We define the **information** contained in a variable X with density function f as

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

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

Slide 6

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.

Bernardo (1979), following earlier work by Lindley (1955), proposed the following measure of the information gained about Z_1 as a result of *measuring* Z_2

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

this is called Shannon's information index.

In the Gaussian case the above formula simplifies to

$$-\frac{1}{2}\sum_i \log(1-\rho(i))^2$$

where $\rho(i)$'s are the canonical correlations between Z_1, Z_2.

- We choose Z_1 to maximize:

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

Slide 7

The previous approach might be ignoring the information in the gauged stations themselves.

One can decompose the total entropy

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

Another entropy criterion:

- Minimizing $H(Z_1|Z_2)$, or equivalently to maximizing $H(Z_2)$.

  In the Gaussian case, this means to maximize the covariance of the gauged sites.

Slide 8

Issues in maximization and computation

- A complete solution to the maximization problem involves searching over a prohibitively large set.
- The data may be non-stationary so the covariance should be estimated using a model that allows for nonstationarity.
- The entropy method just depends on the variance matrix between the sites assuming Gaussian behavior

Slide 9

Optimal Design approach

The major alternative theoretical approach to entropy is based on the theory of **optimal experimental design**.

The traditional formulation of optimal design theory is for a *linear regression* problem in which certain variables x_i are chosen by the experimenter and p covariates of interest are known functions of x_i, denoted $f_1(x_i), \ldots, f_p(x_i)$. The i'th data point is,

$$y_i = \sum_{i=1}^{p} f_j(x_i)\beta_j + \epsilon_i$$

where f_j are known function of design points x_i, $\beta_1, \ldots, \beta_p$ are unknown coefficients, and ϵ_i are uncorrelated errors with mean 0 and common variance σ^2.

Slide 10

Typical optimality criteria include:

- *D-optimality* minimizes the volume of a confidence ellipsoid of fixed significance level for β.
- *A-optimality* minimizes the average variance of the parameter estimates.
- *E-optimality* minimizes the variance of the least well estimated contrast subject to a normalizing condition on the contrast.
- *G-optimality* minimizes the variance of the estimated response function.

Slide 11

Limitations of optimal design approaches

D-optimality designs may be directly applicable in a spatial context, but it does not hold when realistic spatial models are considered. The difficulty is that classical optimality criteria tend to produce designs which involve replications at a relatively small number of design points.

Limitations of entropy approaches

Maximum entropy is the most sophisticated *formulation* of optimal design problems. But are some shortcomings in their actual implementation.

- In general it is implemented by adding and dropping stations one at a time, which may *not lead to the optimal* subset.
- It is implemented without fully consideration of the *uncertainty* about the spatial covariance.

In our research we deal with these two issues concerning entropy approaches.

Slide 12

In our next presentation:

- We will present several efficient algorithms to **calculate** the two entropy criteria.
- We will introduce a model to account for spatial **nonstationarity** of air pollutants.
- We will propose a fully Bayesian approach to take into account **uncertainty** about the spatial covariance.
- We will discuss the importance of metereological and geographical **covariates** for network design.
- We will show **results** for the ozone data from the SLAMS/NAMS network, and we will give some recommendations to EPA in terms of the number and locations of the sites that could be eliminated.

Slide 13