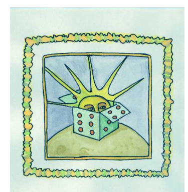
Measuring air quality standards and spatial designs

Douglas Nychka and Eric Gilleland
Geophysical Statistics Project,
National Center for Atmospheric Research
`www.cgd.ucar.edu/stats`

- The problem: A nonlinear, seasonal statistic
- Space filling designs
- A space-time model for ozone
- Thinning the network some examples



The problem

A suggested ozone pollutant standard is based on the fourth highest (max) 8-hour daily average recorded during the year. Compliance is related to a three year average being less than 85 PPB.

Thinning monitoring networks

How should the existing network be reduced but still maintain the best performance?

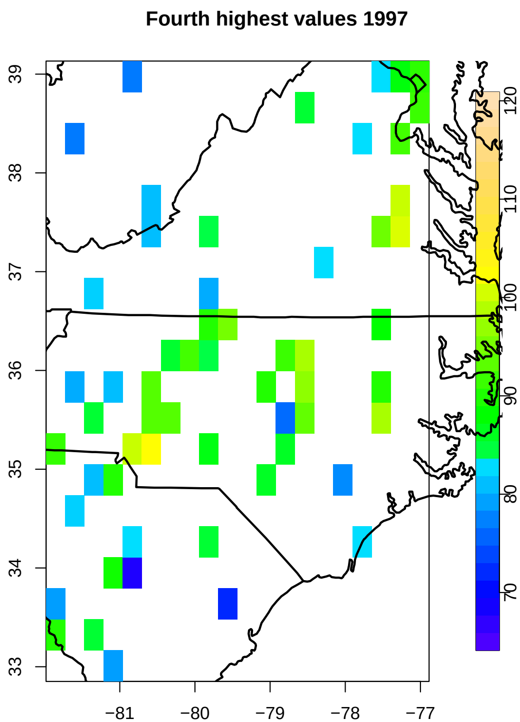
How accurately does a monitoring network measure this standard?

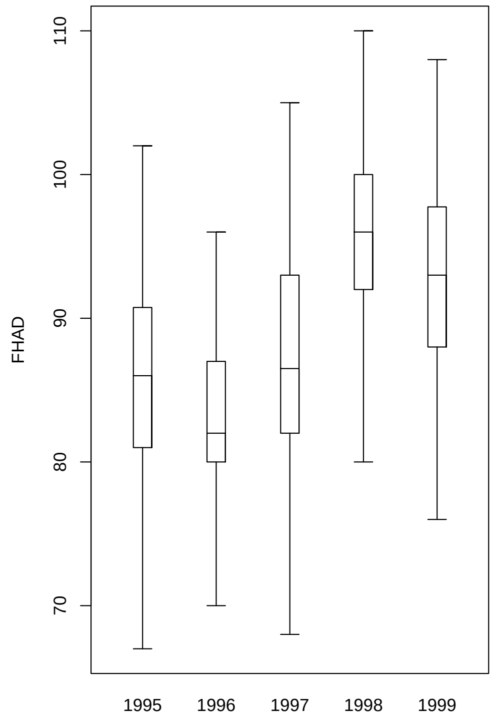
Main idea:

The proposal is to create network designs independently of their evaluation.

An example for North Carolina

Fourth highest daily average (FHDA) values (1997)





Optimal spatial designs: Some formalism

$z(\boldsymbol{x})$ pollutant field at $\boldsymbol{x}$

Observations $\{z_j\}$ at spatial locations $\{\boldsymbol{x}_j\}$ Conditional distribution $[z(\boldsymbol{x})|\boldsymbol{z}]$.

Find $\{\boldsymbol{x}_j\}$ that concentrates distribution at $z(\boldsymbol{x})$.

Kriging

$\hat{z}(\boldsymbol{x}) = E[z(\boldsymbol{x})|\boldsymbol{z}]$ or the BLUE.

We wish to find $\{\boldsymbol{x}_j\}$ that makes MSE of $\hat{z}(\boldsymbol{x})$ small.

G-optimality

For a design region, $\mathcal{R}$ find $\{\boldsymbol{x}_j\}$ to minimize

$$sup_{\boldsymbol{x} \in \mathcal{R}} E(z(\boldsymbol{x}) - \hat{z}(\boldsymbol{x}))^2$$

Other criteria

Average prediction error, Entropy.

Difficulties

Mean squared error is a highly nonlinear function of the covariance.

Any results will be optimal only for a particular model.

Optimization is difficult ...

Software is usually special purpose and not research-based.

Space filling designs

Determine the design based on geometry.

A coverage criterion

$\mathcal{C}$ candidate set of locations $\mathcal{D}$ subset of design points.

$$max_{c \in \mathcal{C}} \left[min_{d \in \mathcal{D}} || \boldsymbol{x}_c - \boldsymbol{x}_d || \right]$$

Choose the design set,$\mathcal{D}$ to *minimize* this criterion.

- The inner term is the "distance" of a candidate point to the design. The intuition is that this is analogous to the prediction variance for Kriging.
- This criterion can be generalized and made less sensitive by replacing min's and max's with L_p type norms.
- Criterion does not have to tied directly to candidate set.
- Some theory exists. *see:* Johnson, Moore and Ylvisaker

A approximate optimization algorithm

The optimization for finding coverage designs is just as ambiguous as G-optimality.

Consider an approximation to a full optimzation.

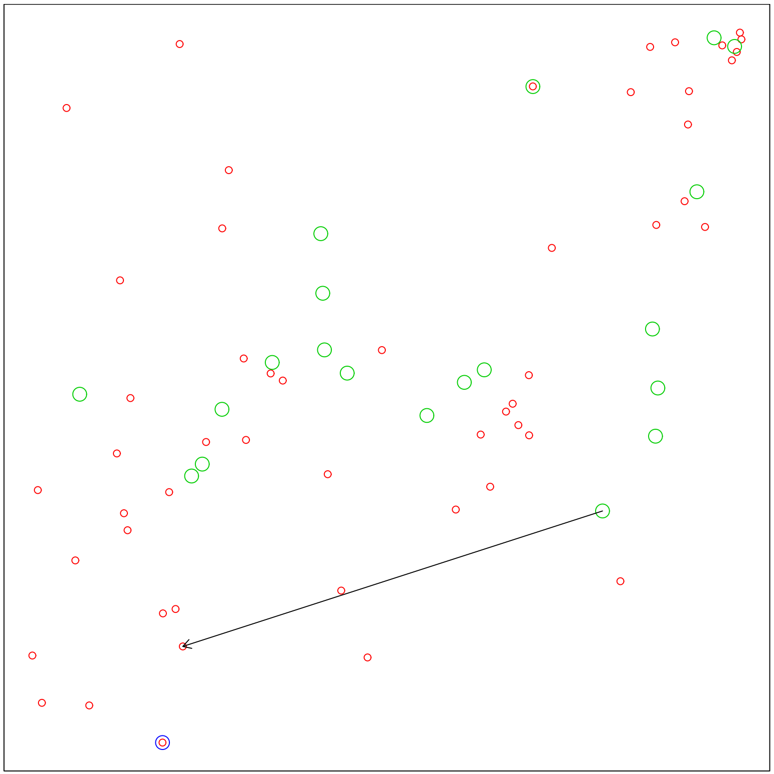
Swapping Algorithm

1. Reduce the coverage criterion by swapping a design point with a candidate point.
2. Repeat swaps until a productive swap can not be made.

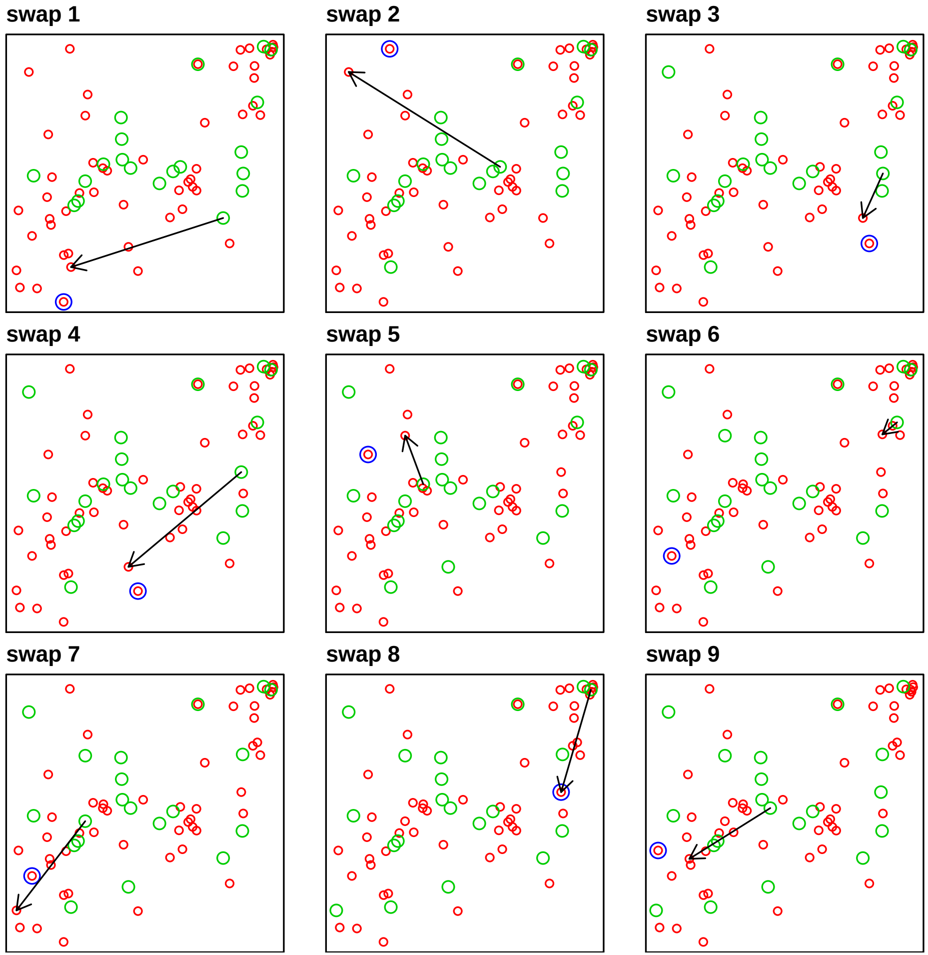
This algorithm will always converge, but may not give the global optimum.
Also, the results can depend on the starting design.

Illustration of swapping for ozone station locations

Starting design and the first swap. Blue point is farthest from $\mathcal{D}$

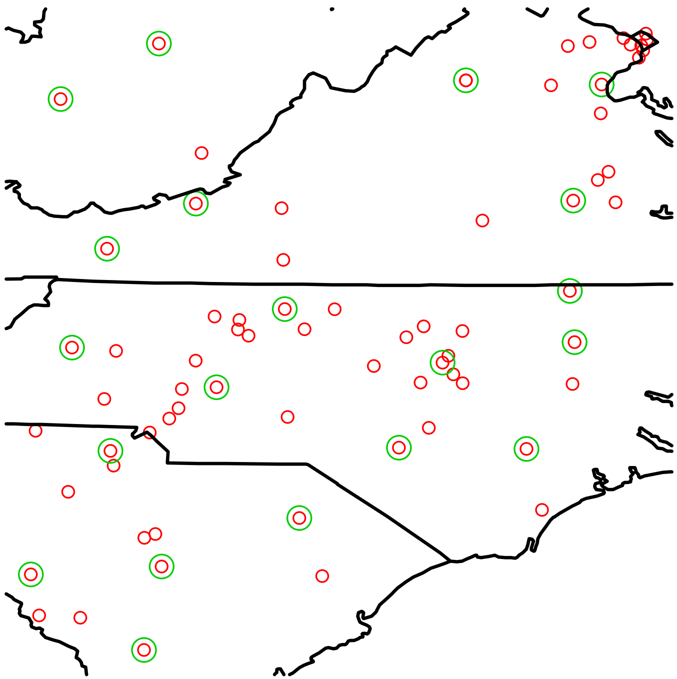


Nine productive swaps



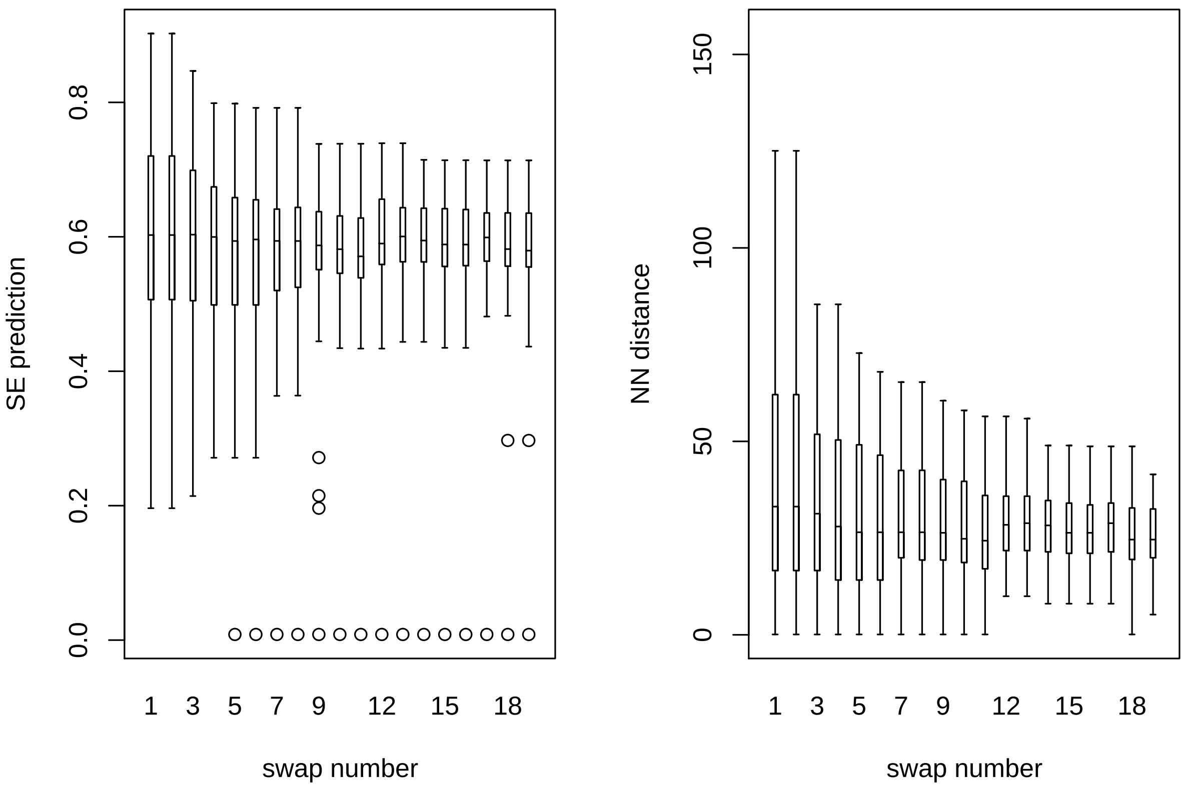
Final coverage design

20 locations out of 72 ozone stations.



How does this compare to Kriging?

Kriging variances and the NN distances for the design complement.



Spatial models for the annual fourth highest

Although it is straight forward to build spatial statistical models for the daily ozone field,

the extension to the fourth highest measurement is difficult ....

- Time dependence
- Nongaussian statistic (extreme value)
- Covariance structure ???

The idea is to determine the distribution of the FHDA from simulating the daily ozone fields.

Model components

Transform: $O(\boldsymbol{x}, t) =$ 8-hour ozone at location $\boldsymbol{x}$ and time t.

$$u(\boldsymbol{x}, t) = \frac{O(\boldsymbol{x}, t) - \mu(\boldsymbol{x}, t)}{\sigma(\boldsymbol{x})}$$

Autoregression: $u(\boldsymbol{x}, t) = \rho(\boldsymbol{x})u(\boldsymbol{x}, t) + e(\boldsymbol{x}, t)$

Spatial dependence: $e(\boldsymbol{x}, t)$ uncorrelated over time and stationary over time.

$$COV(e(\boldsymbol{x}, t), e(\boldsymbol{x}', t)) = (1 - \rho(\boldsymbol{x})^2)k(||\boldsymbol{x} - \boldsymbol{x}'||)$$

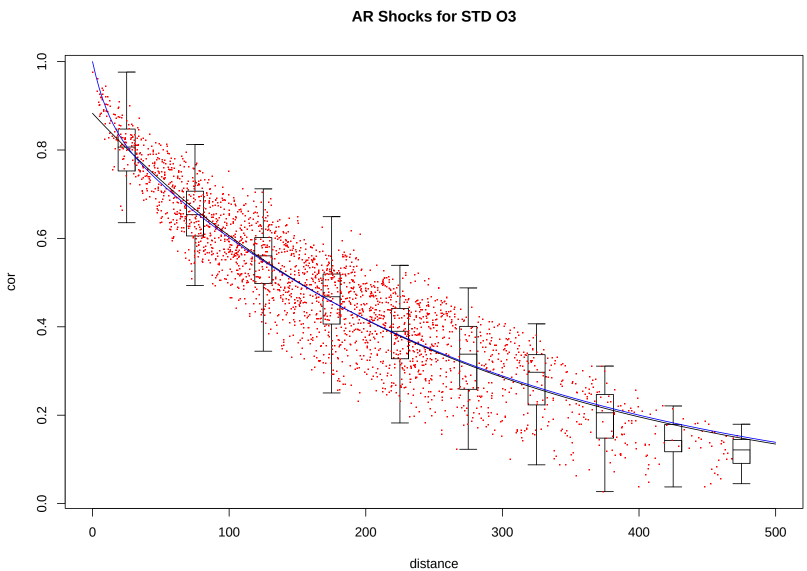
Under the assumption of multivariate normality one can generate fields of daily ozone.

Spatial dependence

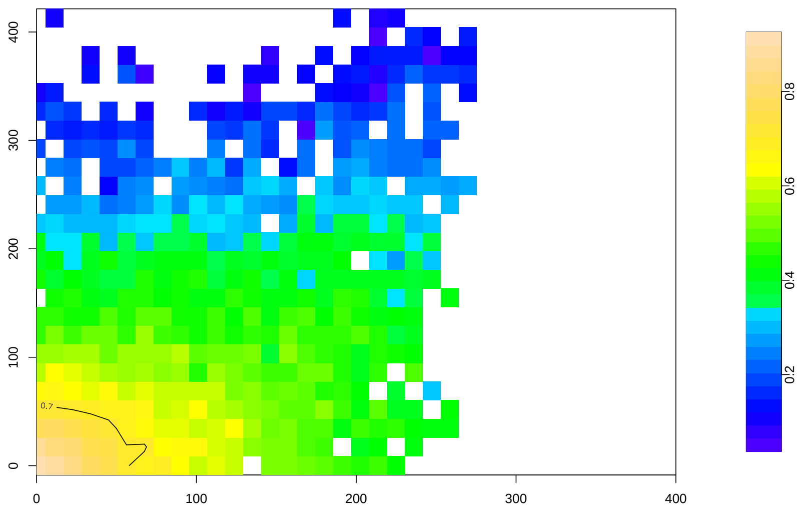
Correlogram of shocks suggests a mixture of exponential covariances

$$k(d) = \alpha e^{-d/\theta_1} + (1-\alpha)e^{-d/\theta_2}$$

with $\alpha = .09$, $\theta_1 = 18$ (miles) and $\theta_2 = 270$ (miles)



Anisotropy?



Shortcuts

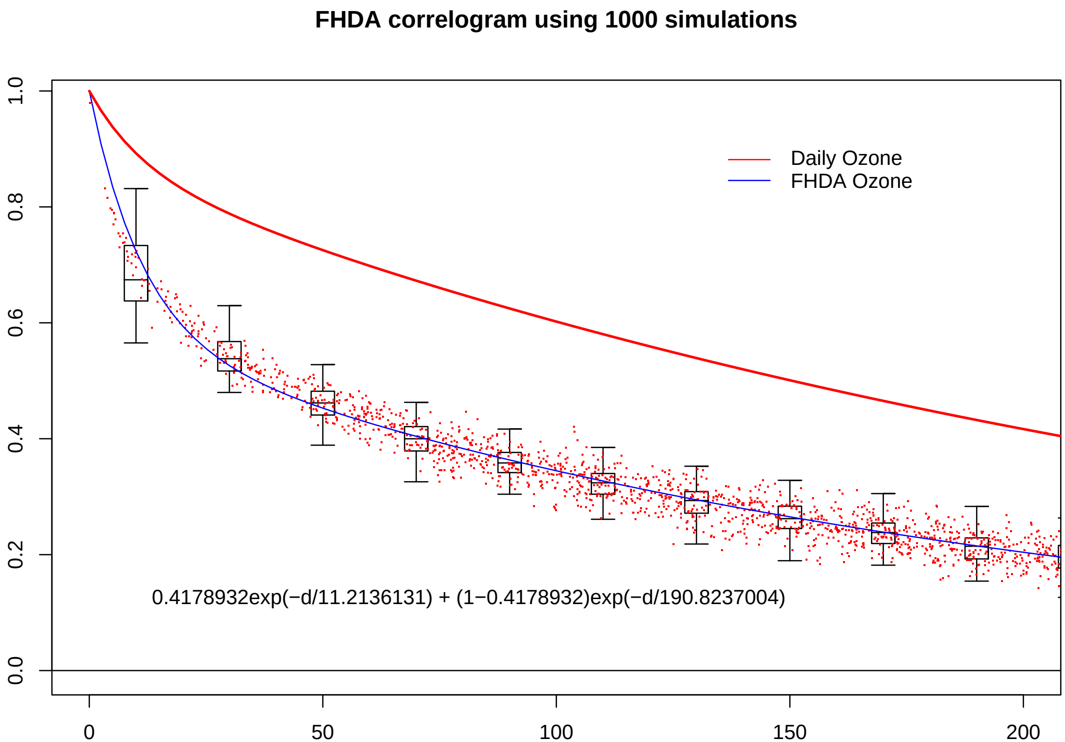
This seems like a lot of work just because we don't know the covariance!

Especially to implement analysis of FHDA in an interactive framework.

What is the joint distribution of FHDA? Is there any hope of being simple?

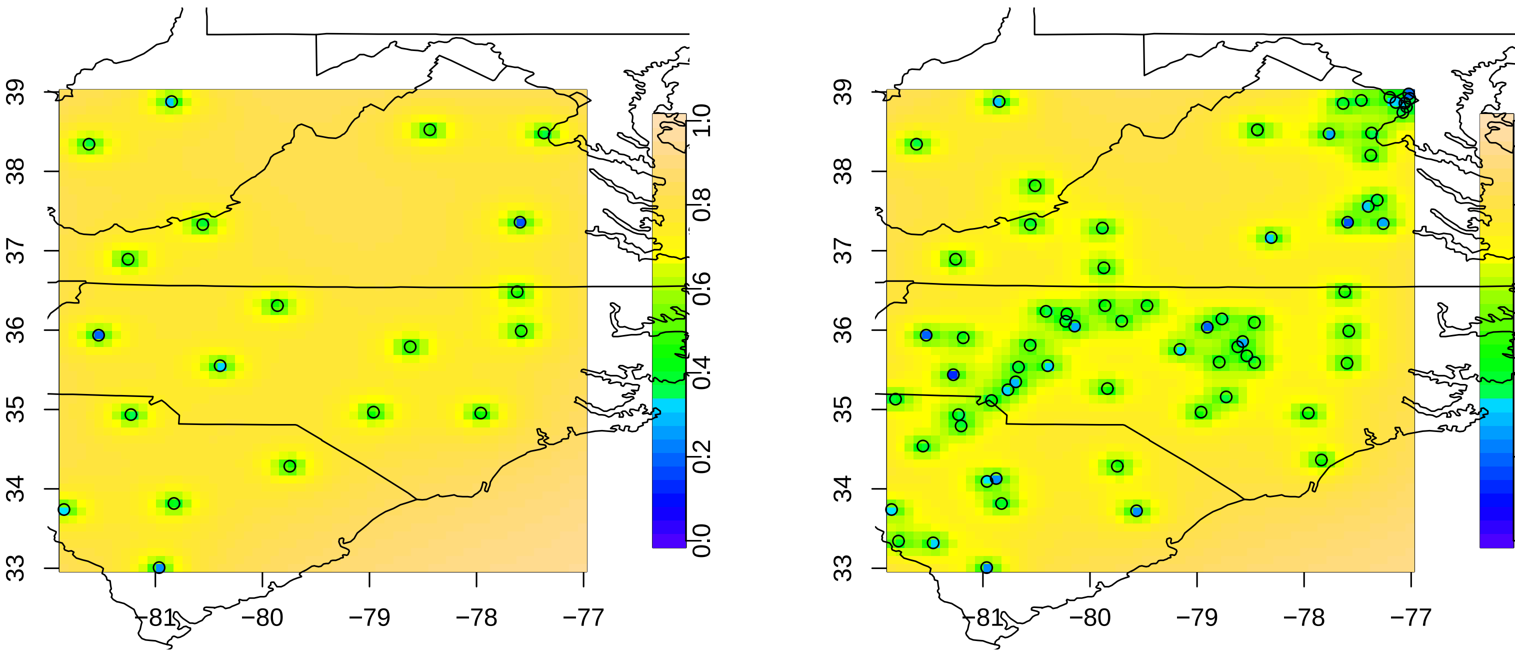
Estimated correlations for FHDA

Relationship among correlations



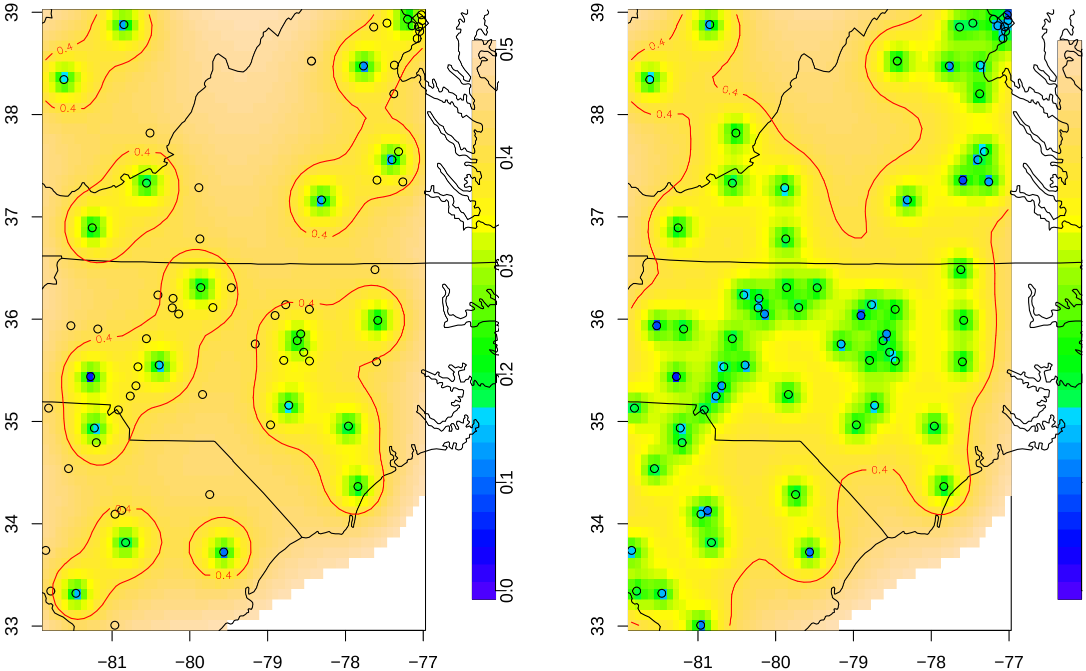
Spatial prediction using FHDA covariance

20 station thinned network, full set of 72

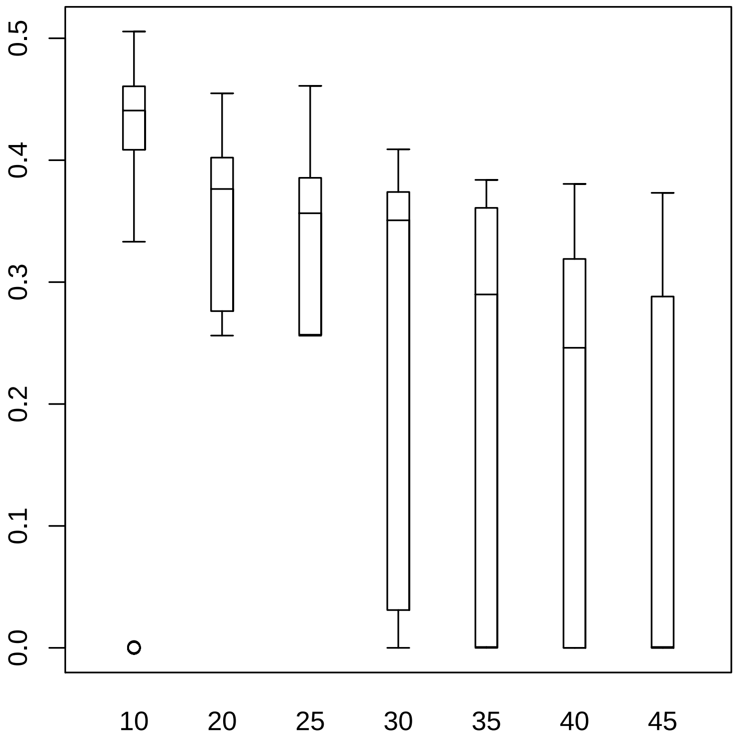


Spatial prediction for just daily ozone

20 station thinned network, full set of 72

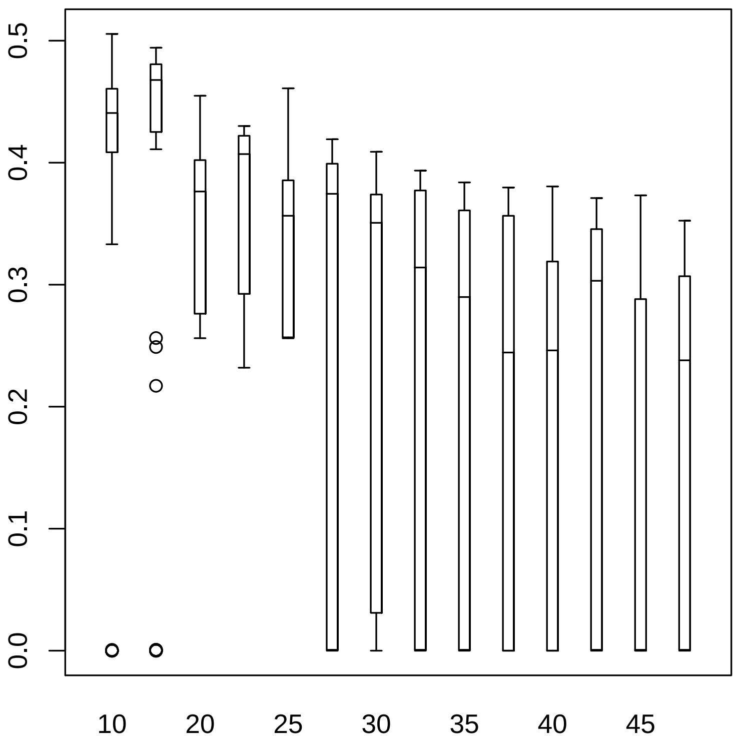


Prediction errors as a function of network size



Comparsion to G-optimal type designs

Swapping algorithm is applied to the max of prediction variances.



Discussion

- Coverage criterion and the swapping algorithm are a flexible tool for generating designs
- FHDA may not be estimated well for a thinned network.
- Blend some covariance information with basic geometric criteria.