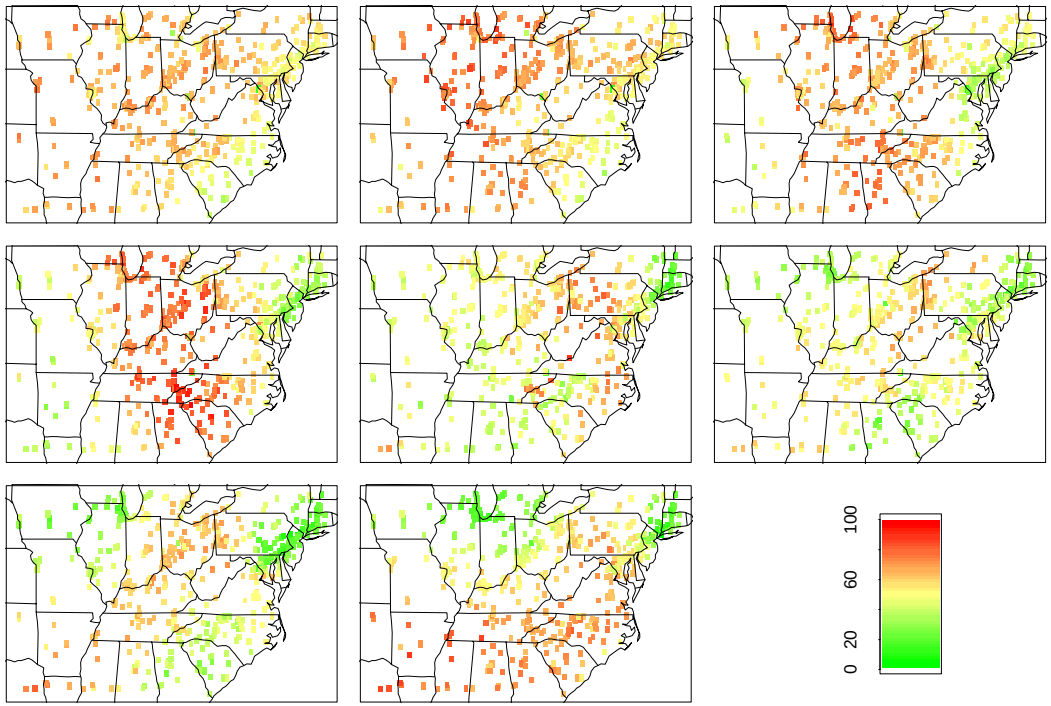
# Optimal Shrinkage of a Monitoring Network

Maria De Iorio

David Higdon Peter Müller

# Overview

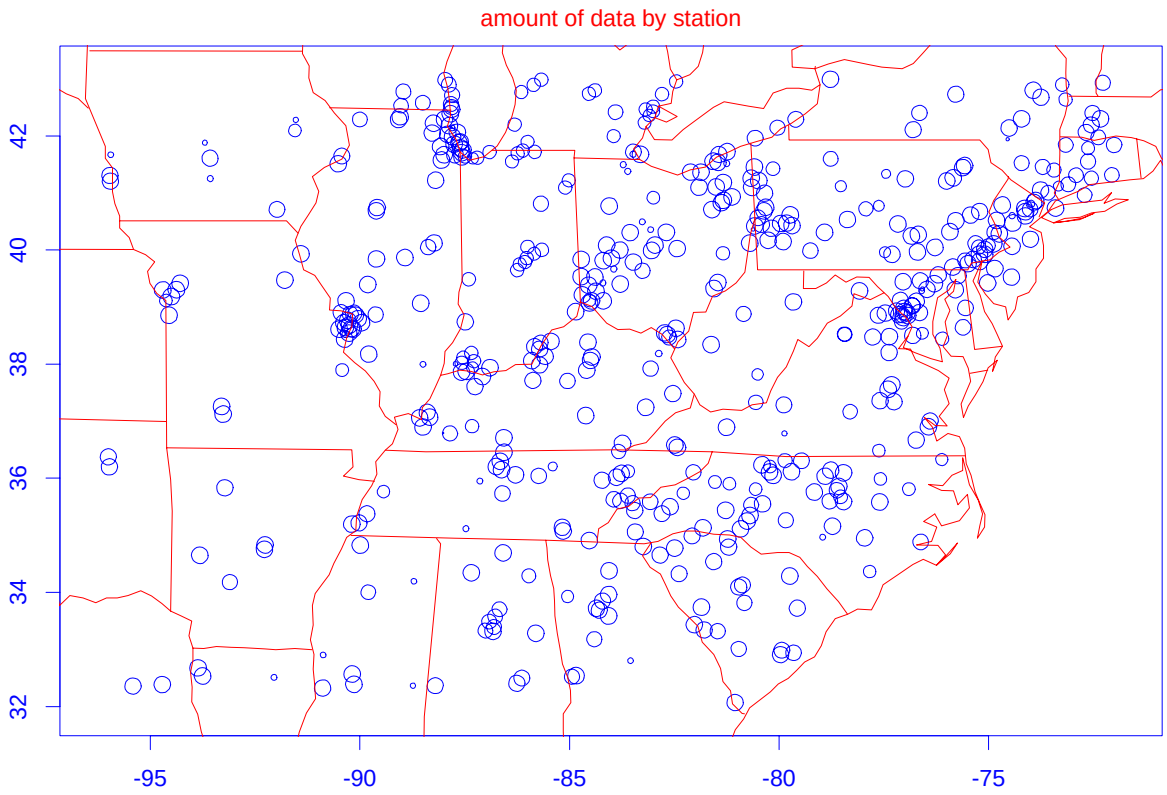
1. Space-time modelling using process convolutions
2. Design problem




Data: daily maximum for the eight hour running average of ozone concentration over the Eastern United States.

The data consist of around 500 measurements recorded for 30 days in summer 1999.

Want: Space-time model for ozone concentration.




Problem: missing values

## 1. Basic Model

$z(s)$ : Gaussian process over a spatial region $\mathcal{S}$

$x(s)$ : continuous white noise process, $s \in \mathcal{S}$

$k(s)$ : smoothing kernel

$$\Downarrow$$

We can construct $z(s)$ by convoluting $x(s)$ with $k(s)$

$$z(s) = \int_{\mathcal{S}} k(u-s)x(u)\mathrm{d}u, \qquad s \in \mathcal{S}$$

The covariance function for $z(s)$ depends only on the displacement vector $d = s - s'$

$$c(d) = \mathrm{Cov}(z(s), z(s')) = \int_{\mathcal{S}} k(u-s)k(u-s')\mathrm{d}u = \int_{\mathcal{S}} k(u-d)k(u)\mathrm{d}u$$

If $\mathcal{S} = \mathbb{R}^m$ and $k(s)$ is isotropic $\Longrightarrow$ $z(s)$ is isotropic and $c(d)$ depends only on the magnitude of $d$.

**Generalize:**

- Dimension reduction: restrict the latent model $x(s)$ to locations $\omega_1, \ldots, \omega_m$ (i.e. over a coarse lattice) $\Longrightarrow$ a small number of parameters effectively control the entire spatial process $z(s)$

$$\begin{aligned} z(s) &= \sum_{j=1}^{m} x(\omega_j) k(s - \omega_j) \\ x(\omega_j) &\sim \mathrm{N}(0, \sigma_x^2) \end{aligned}$$

$k(\cdot - \omega_j)$ is a kernel centered at $\omega_j$.

- Space-time models: Augment $\mathcal{S}$ with time $\mathcal{T}$ so that the latent process and the smoothing kernel are defined over $\mathcal{S} \times \mathcal{T}$

$$\begin{aligned} x(s) &\rightarrow x(s,t) \\ k(s) &\rightarrow k(s,t) \end{aligned}$$

Alternatively we can allow the latent process $x(s,t)$ to evolve over time, with the kernel being a function only of $s$
$\Longrightarrow$ a space-time process can be constructed by

$$z(s,t) = \int_{\mathcal{S}} k(u-s)x(u,t)\mathrm{d}u$$

## 2. A space-time model for ozone concentrations

Data: 500 measurements recorded for 30 consecutive days

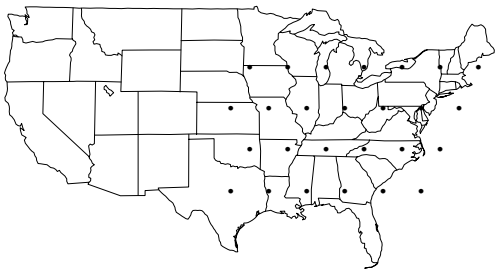
$$y_t = (y_{1t}, \ldots, y_{nt})'$$

Data are recorded at $s_1, \ldots, s_n$ stations.
$t \in \mathcal{T} = \{1, \ldots, 30\}$
$n = 500$ stations
Let

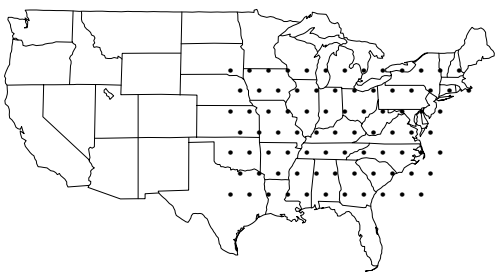
$$\begin{aligned} W^c &= \{\omega_1^c, \ldots, \omega_{27}^c\} \\ W^f &= \left\{\omega_1^f, \ldots, \omega_{83}^f\right\} \end{aligned}$$

spatial supports for 2 independent latent processes $x^c(s,t)$ and $x^f(s,t)$. We account for temporal dependence within $x^c(s,t)$ and within $x^f(s,t)$.

coarse scale: $W^c$



fine scale: $W^f$



The space-time process is then obtained

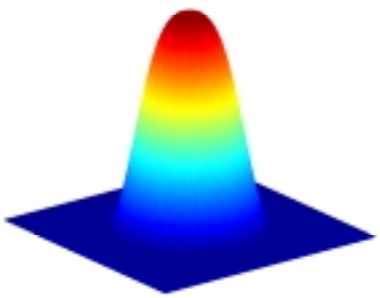
$$
\begin{aligned}
z(s,t) &= \int_{W^c} k(u-s)x^c(u,t)\mathrm{d}u + \int_{W^f} k(u-s)x^f(u,t)\mathrm{d}u \\
&= \sum_j k(\omega_j^c - s)x^c(\omega_j^c, t) + \sum_j k(\omega_j^f - s)x^f(\omega_j^f, t)
\end{aligned}
$$

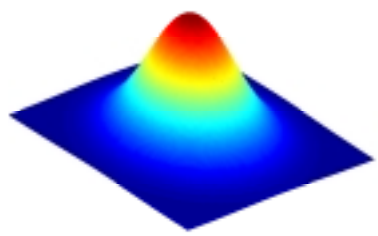
Set

$$
\begin{aligned}
x_t^c &= (x^c(\omega_1^c, t), \ldots, x^c(\omega_{27}^c, t)) \\
x_t^f &= \left(x^c(\omega_1^f, t), \ldots, x^c(\omega_{83}^f, t)\right)
\end{aligned}
$$

for $t = 1, \ldots, 30$.

**Smoothing kernel $k(s)$ is an isotropic two dimensional tricube kernel with a range of 9 degrees.** Hence the induced covariogram dies off at about 15 degrees.





kernel covariance function

The covariogram is fairly flat near the origin and the dependence dies off after a fixed distance

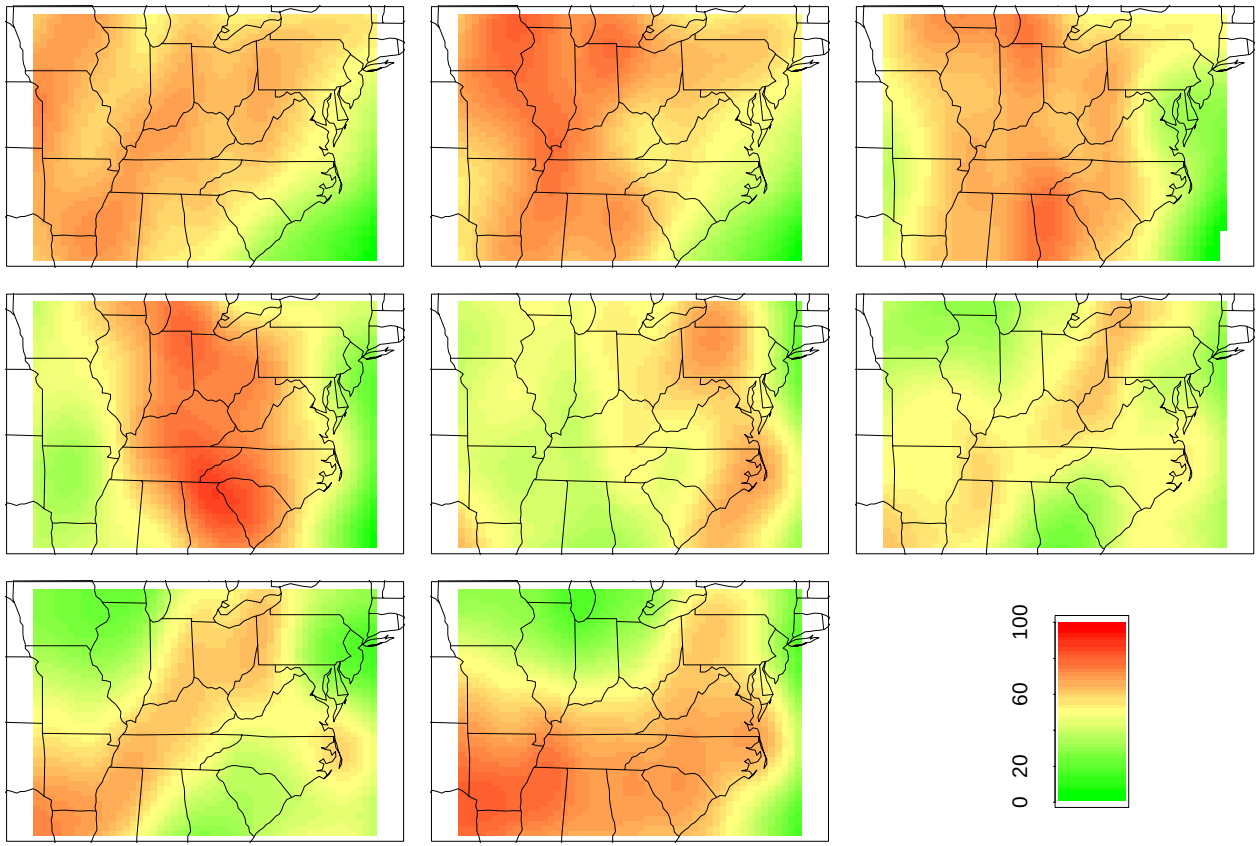
The model can be expressed by 3 evolution equations

$$
\begin{aligned}
y_t &= K'_c x_t^c + K'_f x_t^f + \epsilon_t \\
x_t^c &= x_{t-1}^c + \nu_t \\
x_t^f &= \rho x_{t-1}^f + \eta_t
\end{aligned}
$$

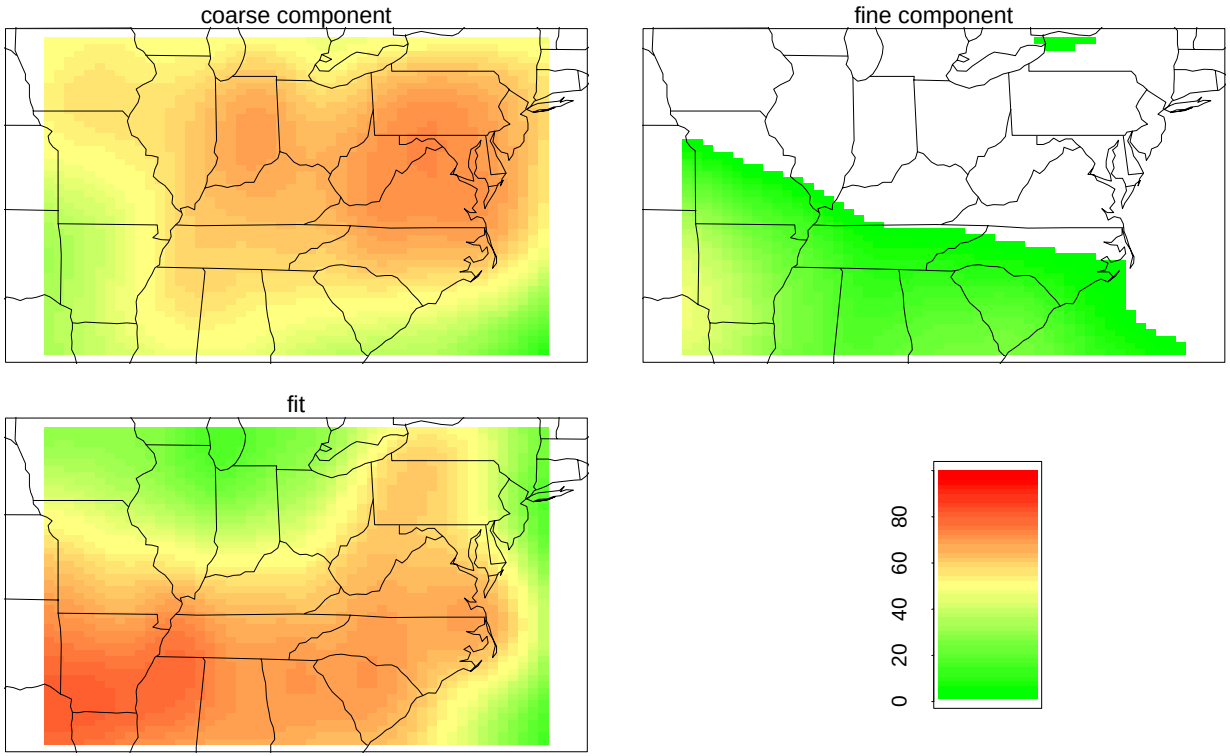
where $K'_c$ is a $n \times 27$ matrix and $K'_f$ is a $n \times 83$

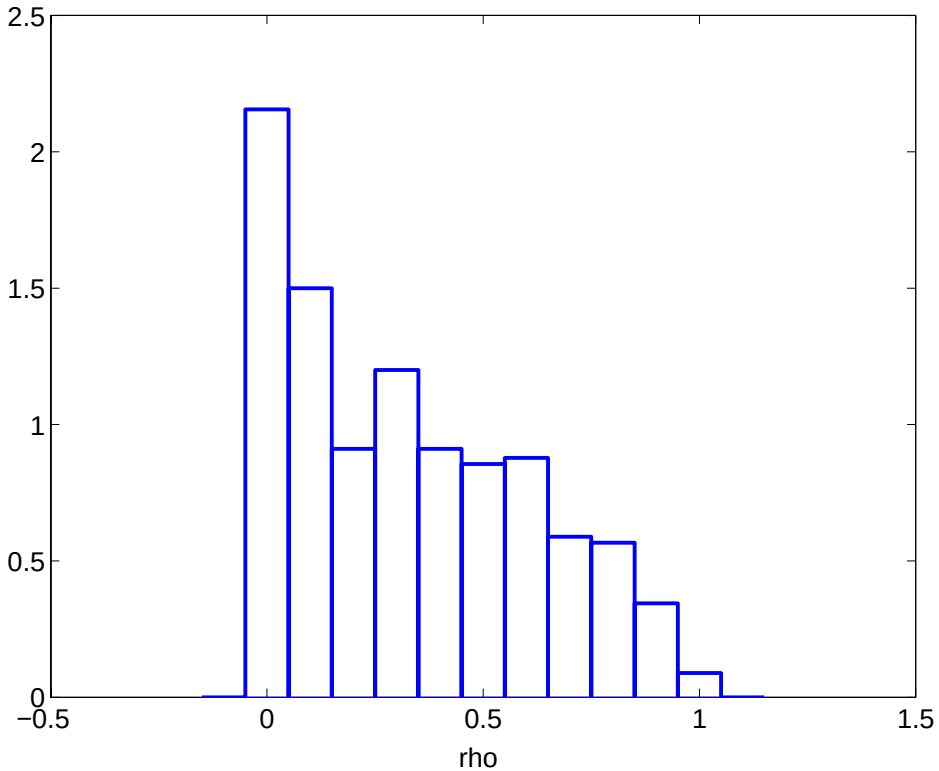
$$
\begin{aligned}
K_{ij}^c &= k(s_{it} - \omega_j^c), && t = 1, \ldots, 30 \\
K_{ij}^f &= k(s_{it} - \omega_j^f), && t = 1, \ldots, 30 \\
\epsilon_t &\overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_\epsilon^2), && t = 1, \ldots, 30 \\
\nu_t &\overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_\nu^2), && t = 2, \ldots, 30 \\
x_1^c &\sim \mathrm{N}(0, \sigma_x^2) \\
\eta_t &\overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_\eta^2), && t = 1, \ldots, 30 \\
\rho &\sim Beta(\alpha, \beta)
\end{aligned}
$$

This model can be updated using methods from dynamic linear model theory (West and Harrison, 1997).




Posterior mean surfaces of the 8 hour daily maximum ozone concentrations for eight consecutive days


coarse component
fine component
fit
0
20
40
60
80




Posterior mean: $\mathbb{E}(\rho) = 0.33$

## 3. Design problem

Want: optimal monitoring network

**general design problem**

$$d^\star = \arg\max \underbrace{\int u(d,\theta,y)\mathrm{d}p(\theta,y \mid d)}_{U(d)}$$

$d =$ design

$(\theta, y) =$ random variables (parameters and data)

$y^o =$ possible historical data

$u(\cdot) =$ utility function

$U(d)$ expected utility

$p(\theta, y \mid y^o, d) =$ full probability model

- $n$ = number of stations
- $d = (d_1, \ldots d_n)$, $d_i = 1$ if station $i$ is included in the design, 0 otherwise.
- $T$ = number of days in historical data
- $y^o$ = observed data (30 days)
- $y$ = future data, $y = (y_d, y_{\overline{d}})$
- $y_d$ = future data at selected sites (observable)
- $y_{\overline{d}}$ = future data at NOT selected sites (latent)
- $x$ = parameters in probability model (weights $x$ in the process convolution process

The optimal design is formally defined as the network $d^\star$ s.t.

$$d^\star = \arg\max \int u(d, x, y) p(x \mid y_d) q(y \mid y^o) \mathrm{d}x \mathrm{d}y$$

where $y$ are the future data and $q(y \mid y^o)$ is the design distribution to simulate future data.

## Utility function

payoff for prediction within $\pm\delta$ - sampling cost

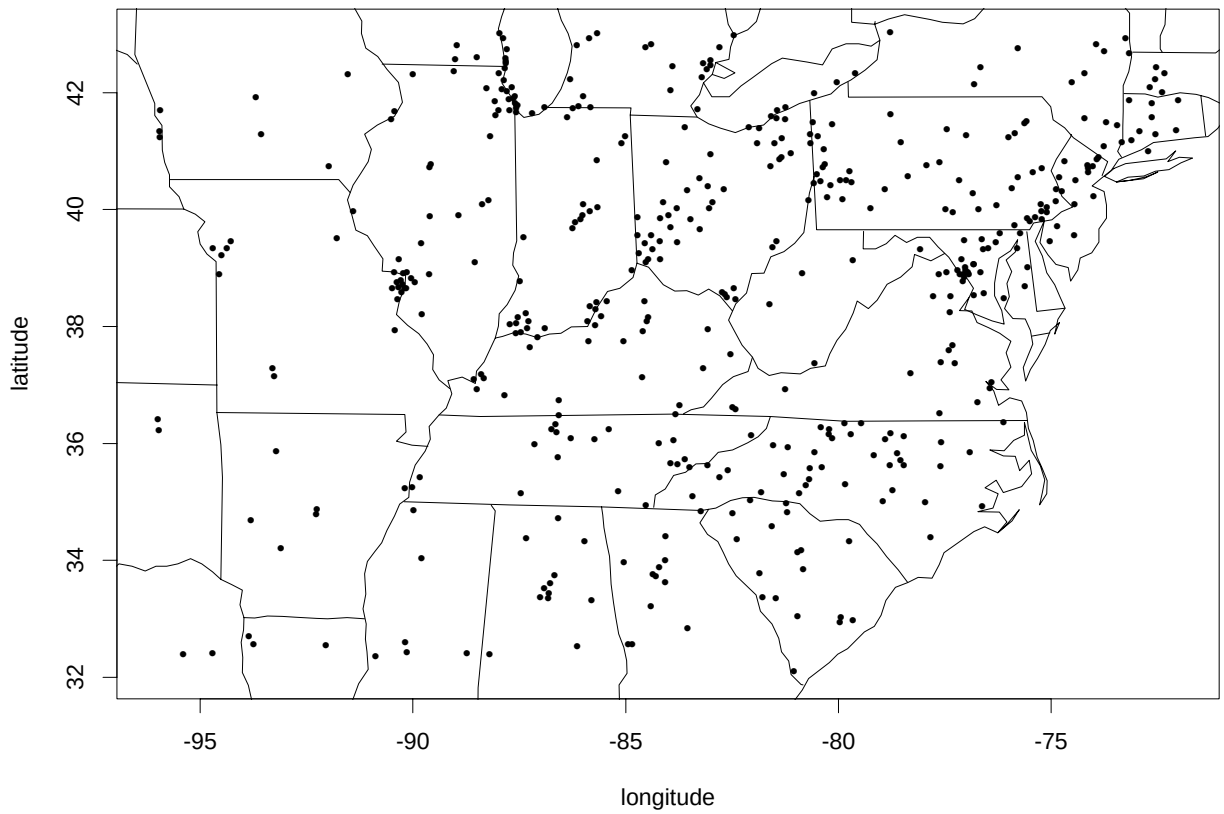
$$u(d, x, y_t) = R \sum_{i:d_i=0} \mathbf{1}(y_{it}^o \in \widehat{y}_{it}(y_d) \pm \delta) - \sum_{i:d_i=1} c_i + C_0$$

where $\widehat{y}_{it}(y_d)$ denote the prediction for station $i$ at time $t$ based on the observed stations $y_d$.

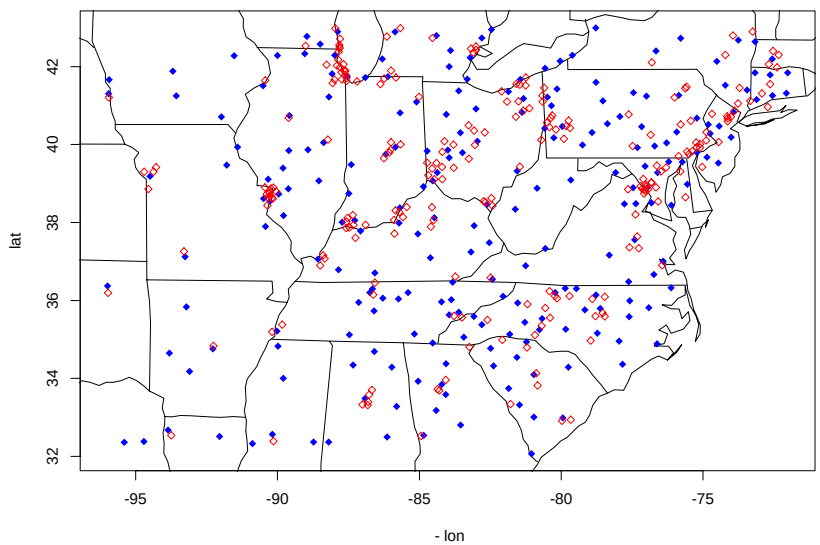
We have set

$$
\begin{aligned}
c_i &= c, \qquad \forall i \\
\frac{R}{c} &= 2 \\
\delta &= 5
\end{aligned}
$$

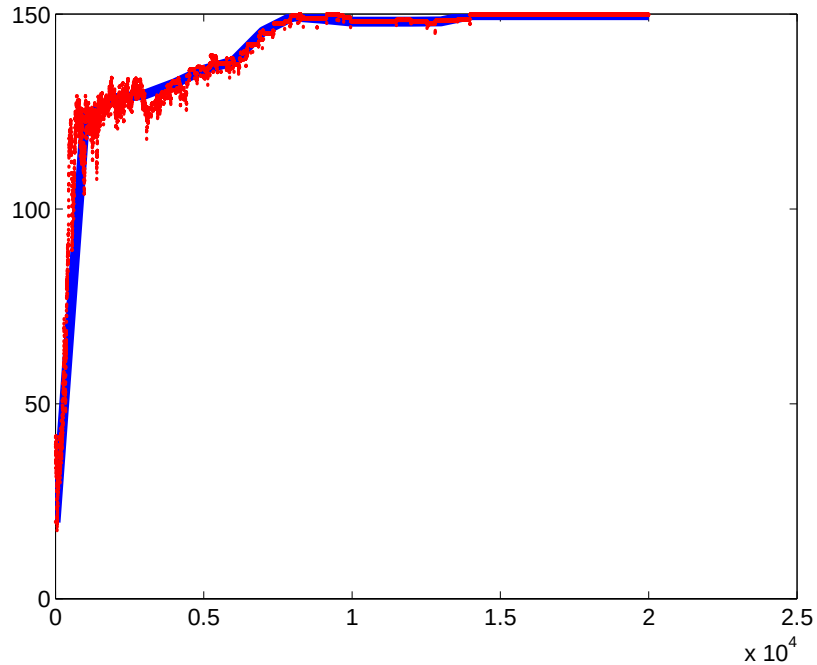
We have forced the number of stations to be between 181 and 281. To find the optimal design we have used simulated annealing technique.




Location of all the stations

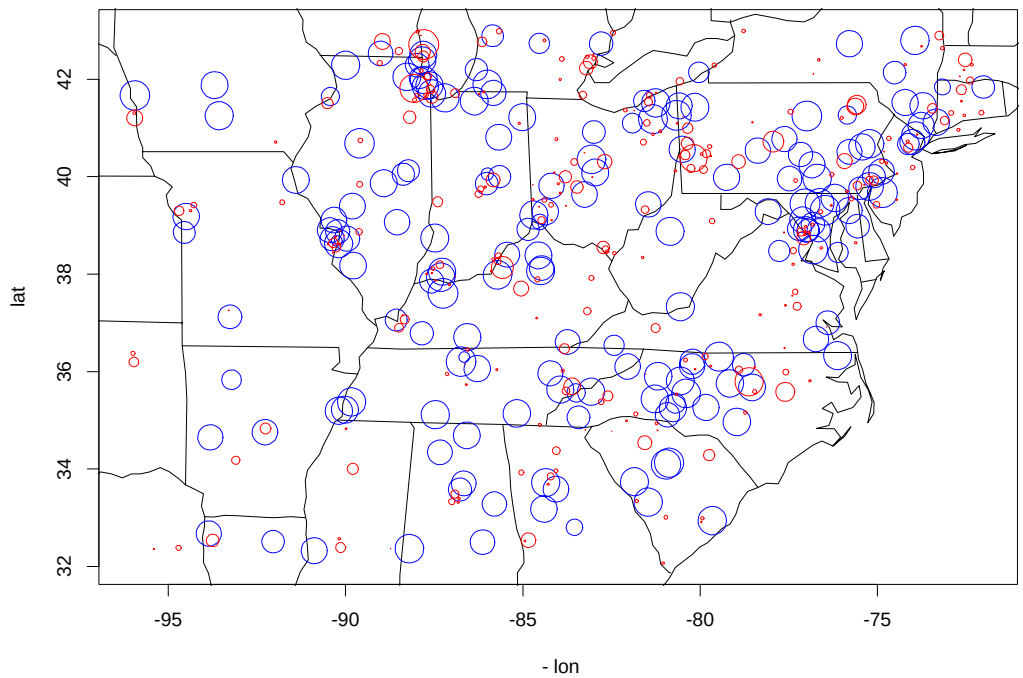



Initial design obtained by using routine **cover.design** (FIELDS by D. Nychka, W. Meiring, J.A. Royle, M. Fuentes, E. Gilleland) → *finds the set of points on a discrete grid (candidate set) which minimises a geometric space-filling criterion.* This set is evenly distributed over the observed domain of data. We have set the number of stations in the design to be equal to 231. The filled blue diamonds represent the selected stations.




Marginal utility $U(d)$.

# Optimal design




There are 181 stations in the design. The blue circles represent the selected stations. The area of the circle is proportional to the marginal frequency of each station.