A South Boulder guide to spatial statistics

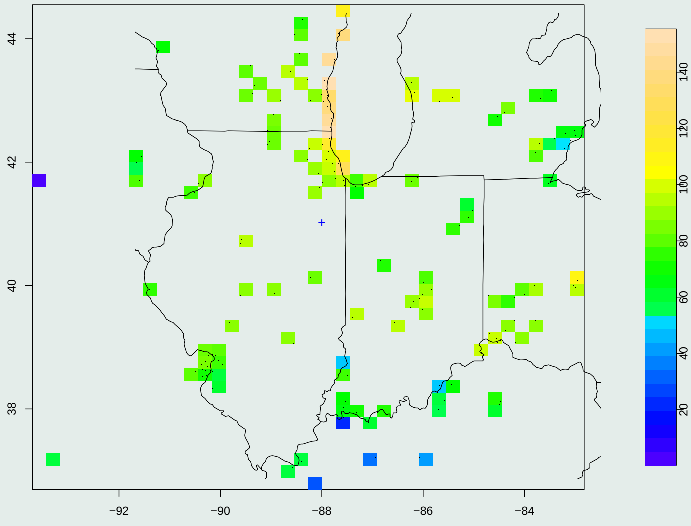
Douglas Nychka Geophysical Statistics Project
National Center for Atmospheric Research

Outline

- Filling in between observations
- A model
- The covariance is everything
- Don't obsess about the covariance
- Things to do.

An example: Daily ozone pollution

Here are the 8-hour average ozone measurements (in PPB) for June 19,1987.



What can we say about ozone at (-88,41)?

Use local information to predict unobserved values

One reasonable method is to predict the location using a linear regression based on close by observations.

$$z_k = \beta_1 + \text{lon}_k\beta_2 + \text{lat}_k\beta_3$$

find $\hat{\boldsymbol{\beta}}$ by least squares

$$\hat{z} = \hat{\beta}_1 + \text{-88}\hat{\beta}_2 + 41\hat{\beta}_3 = w_0 + \sum_{k=1,n} w_k z_k$$

Problems with local regression

How large should the neighborhood be?

What is the uncertainty of the prediction?

How much does the surface depart from a plane?

Spatial models deal with these problems by adding a model for the underlying surface.

A normal world

Suppose $z(\boldsymbol{x})$ is the ozone concentration at location $\boldsymbol{x}$,

We assume that $z(\boldsymbol{x})$ is a Gaussian process, $E(z(\boldsymbol{x})) = 0$

$$k(\boldsymbol{x}, \boldsymbol{x}') = COV(z(\boldsymbol{x}), z(\boldsymbol{x}'))$$

Being a Gaussian process has the practical consequence that *any* discrete subset of the fields has a multivariate normal distribution.

If we know k we know how to make a prediction at $\boldsymbol{x}^*$!

$$\hat{z} = E[z(\boldsymbol{x}^*)|data]$$

i.e. Just use the conditional multivariate normal distribution.

A review of the conditional normal

$$\boldsymbol{z} \sim N(0, \Sigma)$$

$$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11}, \Sigma_{12} \\ \Sigma_{21}, \Sigma_{22} \end{pmatrix}$$

$$[\boldsymbol{z}_2 | \boldsymbol{z}_1] = N(\Sigma_{2,1}\Sigma_{1,1}^{-1}\boldsymbol{z}_1, \;\; \Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2})$$

Thinking of $\boldsymbol{z}_2$ as unobserved locations and $\boldsymbol{z}_1$ as the observations.

The Kriging weights

Conditional distribution of z^* given the data $\boldsymbol{z}$ is Gaussian.

Conditional mean

$$\hat{z}^* = COV(\boldsymbol{z}^*, \boldsymbol{z})\,[COV(\boldsymbol{z}, \boldsymbol{z})]^{-1}\,\boldsymbol{z} = \sum_{k=1,n} \omega_k z_k = \boldsymbol{\omega}^T \boldsymbol{z}$$

$\boldsymbol{\omega}$ are the Kriging weights.

Note: $COV(\boldsymbol{z}, \boldsymbol{z})$ is an $N \times N$ matrix, $COV(\boldsymbol{z}^*, \boldsymbol{z})$ an N row vector.

Conditional variance

$$VAR(z^*, z^*) - COV(z^*, \boldsymbol{z})\,[COV(\boldsymbol{z}, \boldsymbol{z})]^{-1}\,COV(\boldsymbol{z}, z^*)$$

My geostatistics/BLUE overhead

For any covariance and any set of weights (not just $\boldsymbol{\omega}$) we can easily derive the prediction variance for z^*.

Minimize

$$E\left[(z^* - \hat{z^*})^2\right] = VAR(z^*, z^*) - 2COV(z^*, \boldsymbol{z})\boldsymbol{w} + \boldsymbol{w}^T COV(\boldsymbol{z}, \boldsymbol{z})\boldsymbol{w}$$

over all $\boldsymbol{w}$.

The answer The Kriging weights ... or what we would do if we used the Gaussian process and the conditional distribution.

Folklore and intuition The spatial estimates are not very sensitive if one uses suboptimal weights, especially if the observations contain some measurement error.
It does matter for finds measures of uncertainty.

Surfaces

The conditional normal tell us how to predict onto an entire grid given the observations. ($\boldsymbol{z}_2$ = grid , $\boldsymbol{z}_1$ = obs.)

Recall:

$$[\boldsymbol{z}_2|\boldsymbol{z}_1] = N(\Sigma_{2,1}\Sigma_{1,1}^{-1}\boldsymbol{z}_1, \;\; \Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2})$$

- The estimated surface has the equivalent form:
  $\hat{z}(x) = \sum_{k=1,n} c_k k(\boldsymbol{x}, \boldsymbol{x}_k)$
  k the covariance kernel and $\boldsymbol{c}$ are estimated from the data.
- We have the full distribution for the surface on the grid and can sample from it. E.g. a realization of the ozone surface given the observed data.
- With some measurement error ($\Sigma_{1,1}$ replaced by $\Sigma_{1,1} + \sigma^2 I$) the conditional mean is a smoother ... but not exactly a kernel estimator or a local linear regression.

Covariance? The variogram.

The preceding discussion is useless without estimating the covariance function (k).

We have to make some assumptions on k to use just one field.
Assume that $z(\boldsymbol{x})$ is stationary and isotropic.

$$k(\boldsymbol{x}, \boldsymbol{x}') = \phi(||\boldsymbol{x} - \boldsymbol{x}'||)$$

$||.||$ great circle distance and we identify ϕ using EDA.

The key is the variogram:

$$E\left[1/2(z(\boldsymbol{x}) - z(\boldsymbol{x}'))^2\right] = \phi(0) - \phi(||\boldsymbol{x} - \boldsymbol{x}'||)$$

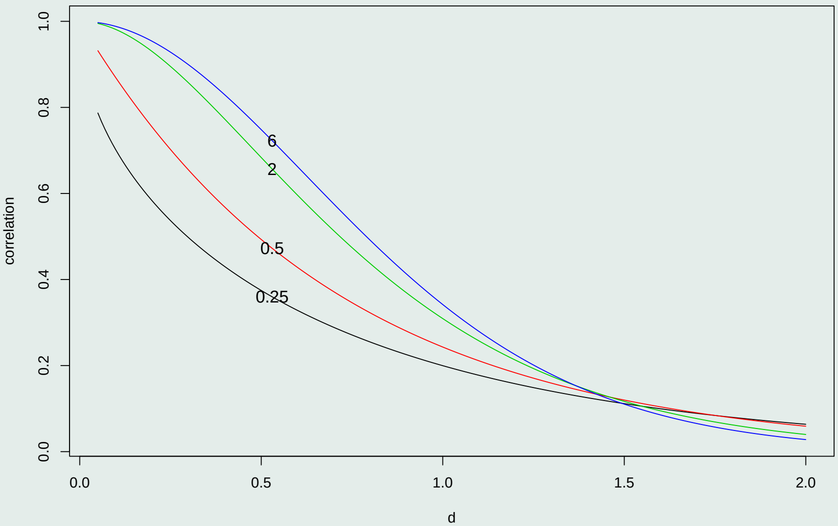
At last! A form we can estimate directly from the observations.

The matern class of covariances

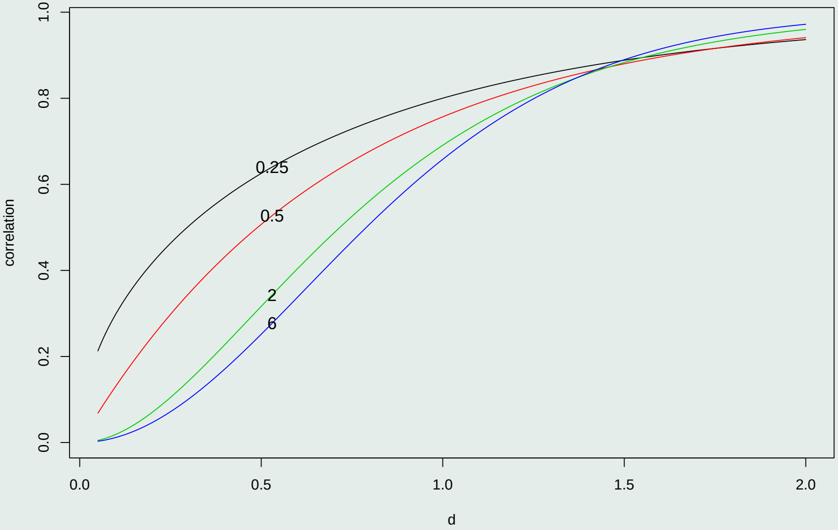
Not any old ϕ will give a valid covariance function. A useful family has four parameters:

$$\phi(d) = \sigma(1 - \alpha * \psi_\nu(d/\theta))$$

ψ_ν is an exponential for $\nu = 1/2$ as $\nu \to \infty$ Gaussian.

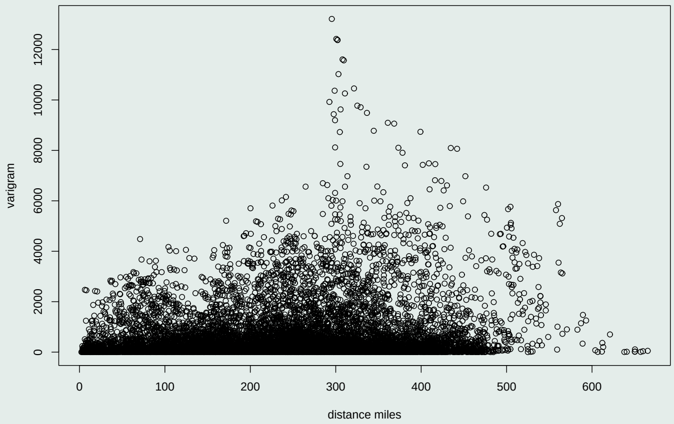


Same models but as variograms



The smoothness properties of the spatial field depend on how smoothly the variogram approaches zero as $d \to 0$.

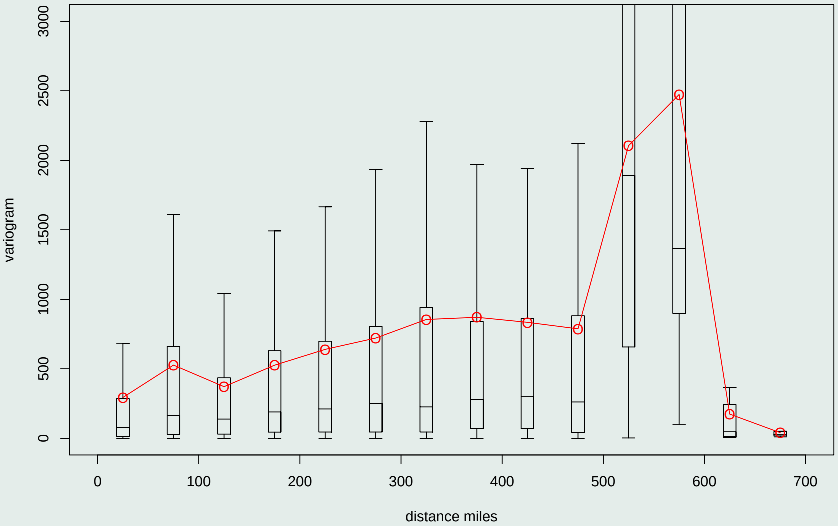
Variogram for ozone data – Day 16



What a mess!

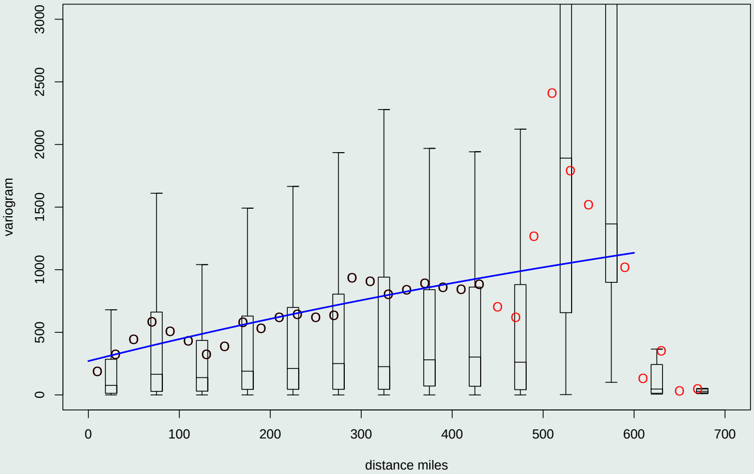
Binning the variogram

Boxplots of squared values in bins with mean added.



Fitting the variogram

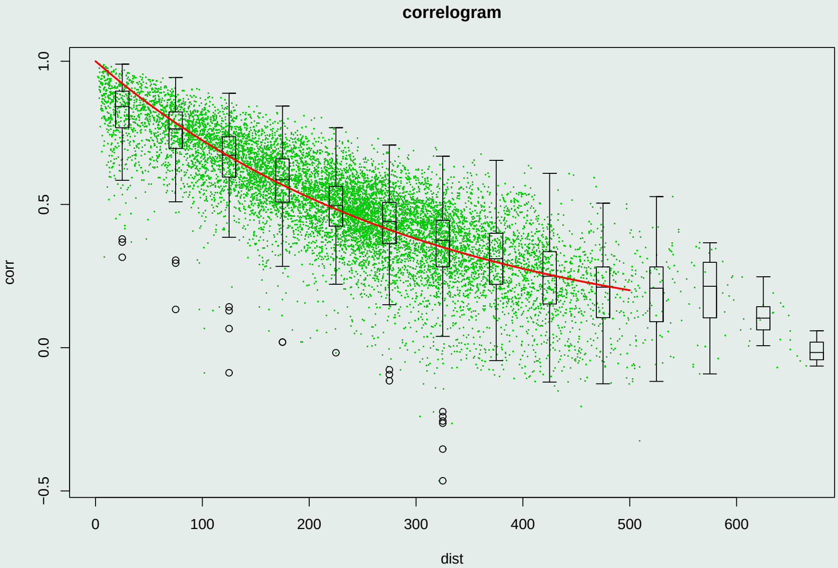
Assume an exponential covariance



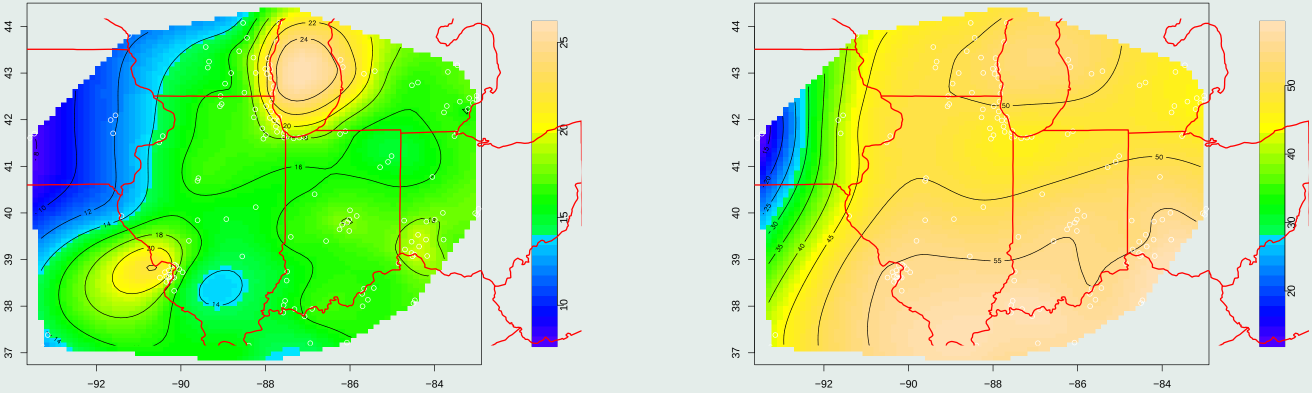
In this case $\theta = 1200$, $\sigma = 49.8$ and $\alpha = .11$ (I don't believe these)

Using the temporal information

In many cases spatial processes also have a temporal component. Here we take the 89 days over the ”ozone season” and just find sample correlations among stations.



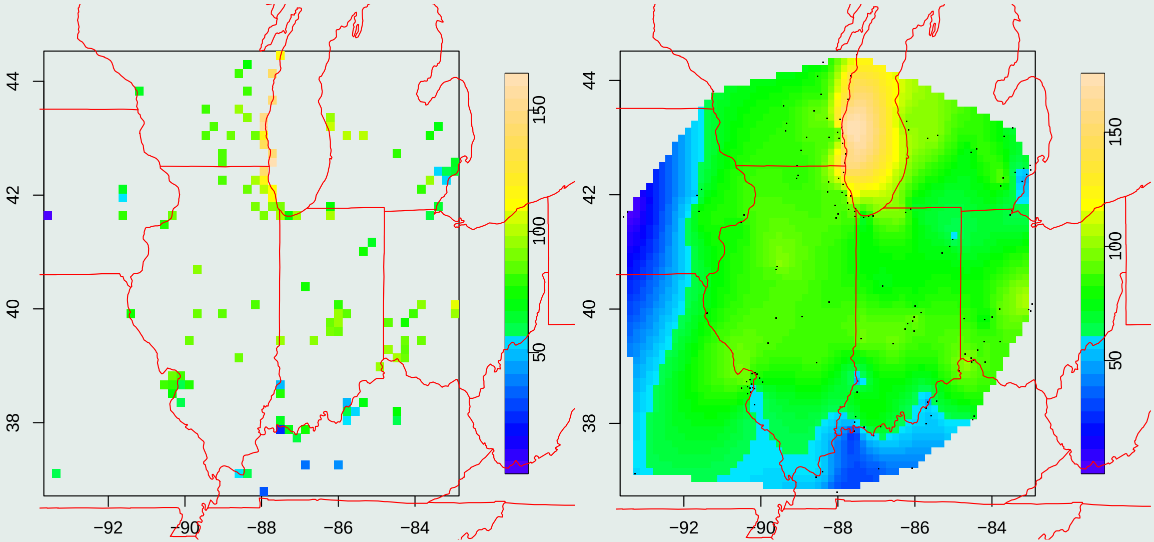
Mean and SD surfaces



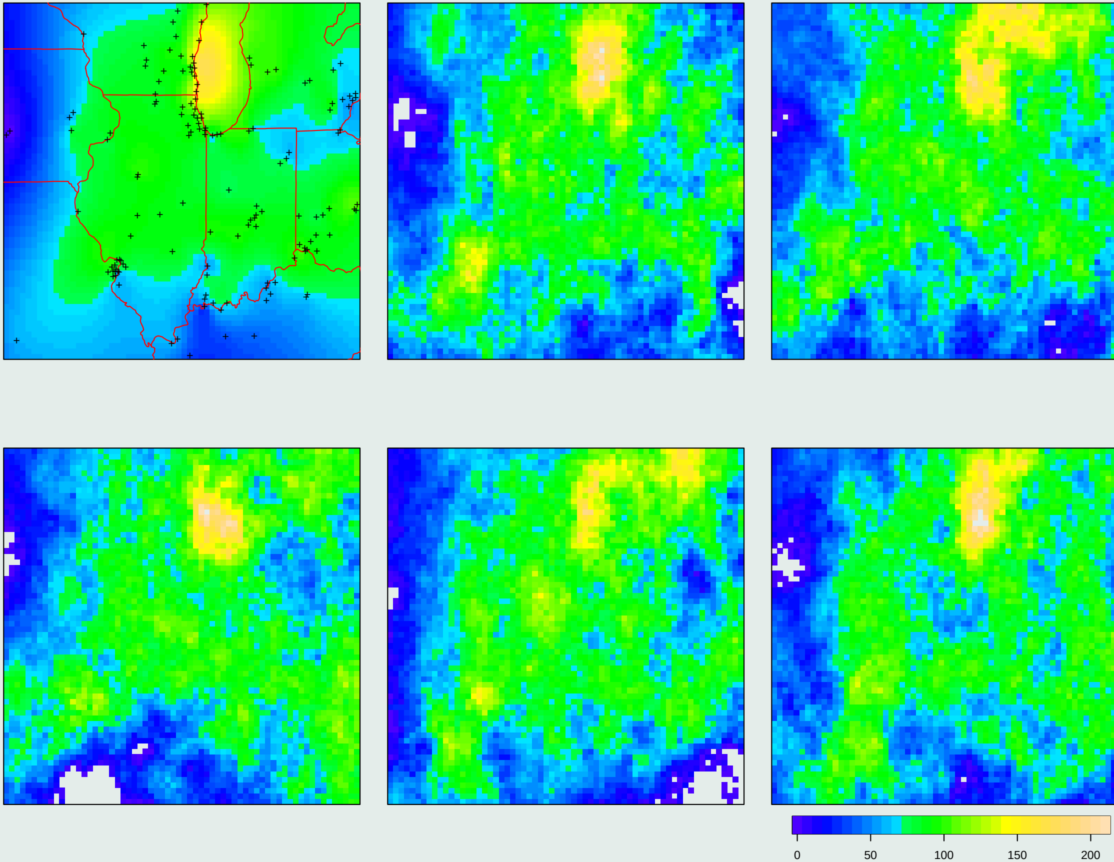
Covariance model: $k(\boldsymbol{x}, \boldsymbol{x}') = \sigma(\boldsymbol{x})\sigma(\boldsymbol{x}')exp(-||\boldsymbol{x} - \boldsymbol{x}'||/\theta)$

Mean model: $E(z(\boldsymbol{x})) = \mu(\boldsymbol{x})$
where μ is also a Gaussian spatial process.

The data for day 16 and the conditional mean surface



Five samples from the posterior



Beyond the covariance

The covariance is rarely of interest on its own.

Some other issues related to finding reasonable posterior distributions of the field

- Handle large numbers of observations
- Nongaussian distributions, robust methods.
- Include temporal as well as spatial dependence.
- Propagate uncertainty in *all* components of the model to uncertainty in field

Examples of useful directions

Dependence over time: $z(\boldsymbol{x}, t) = \rho(\boldsymbol{x}) z(\boldsymbol{x}, t) + u(\boldsymbol{x}, t)$

Where $u(\boldsymbol{x}, t)$ are spatial processes uncorrelated in time.

Design: If the EPA had to reduce the ozone monitoring network by half how should the stations be thinned?

Conclusions

A primary activity in spatial statistics is to develop a (stochastic) model for the unknown surface.

Inferring covariance models from data can be difficult especially when only a single field is available.

The covariance function is an important part of the model but usually not an end in itself.