

SPATIAL STATISTICS IN THE PRESENCE OF LOCATION ERROR

John Kornak

Program in Spatial Statistics
and Environmental Sciences

(jk@stat.ohio-state.edu)

Joint research with Noel Cressie and John Gabrosek

Spatial Data

Consider a spatial process with *continuous spatial index*:

$$\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}; \quad |D| > 0,$$

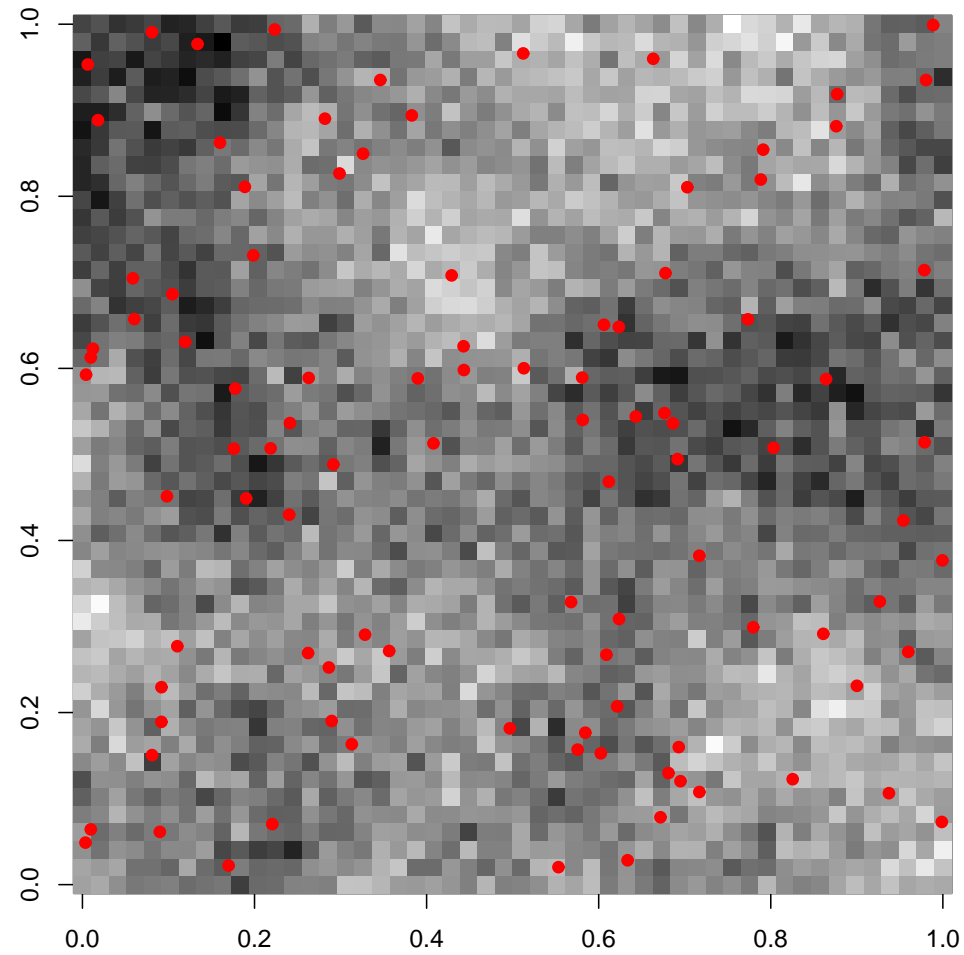
sampled at locations

$$S = \{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subset D.$$

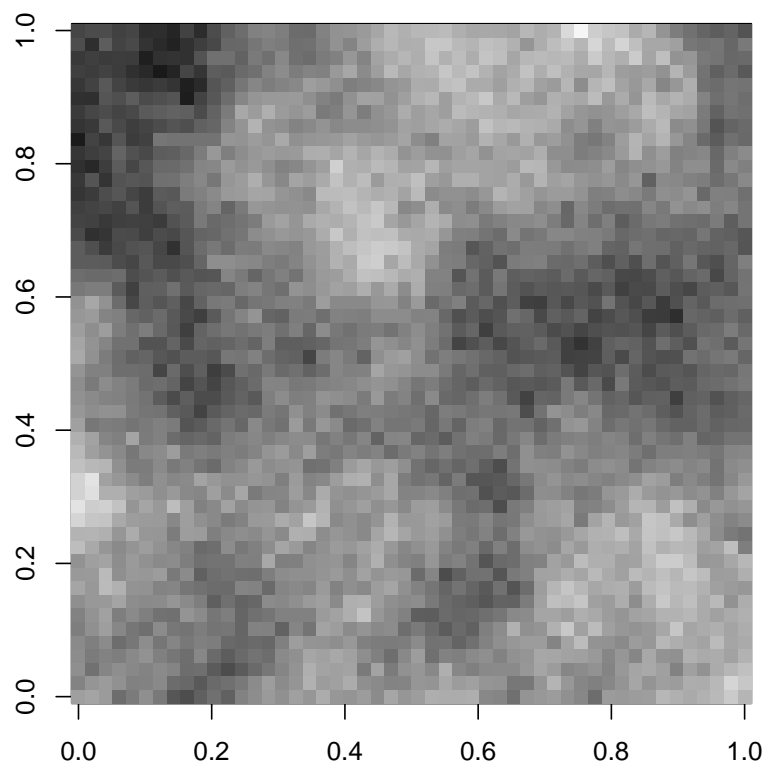
Data are

$$\mathbf{Z} \equiv (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$$

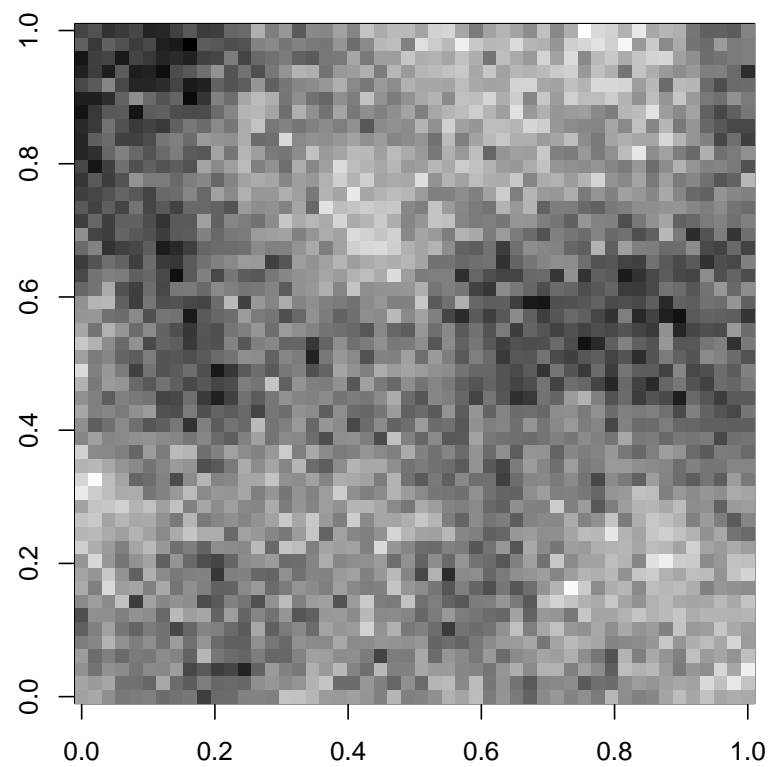
$Z(\cdot)$ and Sample Locations (dots)



$Y(\cdot)$



$Z(\cdot) = Y(\cdot) + \epsilon(\cdot)$



Geostatistical Model:

- $E(Z(\mathbf{s})) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}$
- $cov(Z(\mathbf{s}_1), Z(\mathbf{s}_2)) = C(\mathbf{s}_2 - \mathbf{s}_1; \boldsymbol{\theta})$
- $Y(\cdot)$ is noiseless version of $Z(\cdot)$; i.e., $Z(\mathbf{s}_i) = Y(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i)$

Inference on:

- Parameters $\boldsymbol{\beta}, \boldsymbol{\theta}$ – estimate
- $\mathbf{Y}_0 \equiv (Y(\mathbf{s}_{0,1}), \dots, Y(\mathbf{s}_{0,\ell}))'$ – predict

Solutions:

- Two-stage *GLSE* $\Rightarrow \hat{\beta}$
WLS variogram estimation $\Rightarrow \hat{\theta}$
- Gaussian *MLE* $\Rightarrow \hat{\beta}, \hat{\theta}$
- Kriging $\Rightarrow p(\mathbf{Z}; \mathbf{s}_0)$

e.g., $p(\mathbf{Z}; \mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)' \hat{\beta} + \mathbf{c}(\mathbf{s}_0)' \Sigma^{-1} (\mathbf{Z} - X \hat{\beta})$, where
 $\hat{\beta} \equiv (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} \mathbf{Z}$, $X \equiv (\mathbf{x}(\mathbf{s}_1), \dots, \mathbf{x}(\mathbf{s}_n))'$,
 $\mathbf{c}(\mathbf{s}_0) \equiv \text{cov}(\mathbf{Z}, Y(\mathbf{s}_0))$, and $\Sigma \equiv \text{var}(\mathbf{Z})$. (In practice, $\hat{\theta}$ is substituted
into $\mathbf{c}(\mathbf{s}_0)$ and Σ to obtain the predictor $p(\mathbf{Z}; \mathbf{s}_0)$ as a function only of
the data.)

$$E(p(\mathbf{Z}; \mathbf{s}_0) - Y(\mathbf{s}_0))^2 = C(\mathbf{0}) - \mathbf{c}(\mathbf{s}_0)' \Sigma^{-1} \mathbf{c}(\mathbf{s}_0) \\ + (\mathbf{x}(\mathbf{s}_0)' - \mathbf{c}(\mathbf{s}_0)' \Sigma^{-1} X) (X' \Sigma^{-1} X)^{-1} (\mathbf{x}(\mathbf{s}_0) - X' \Sigma^{-1} \mathbf{c}(\mathbf{s}_0))$$

Location Error

Co-ordinate-Positioning (CP) Model

Intended locations S

Actual locations R

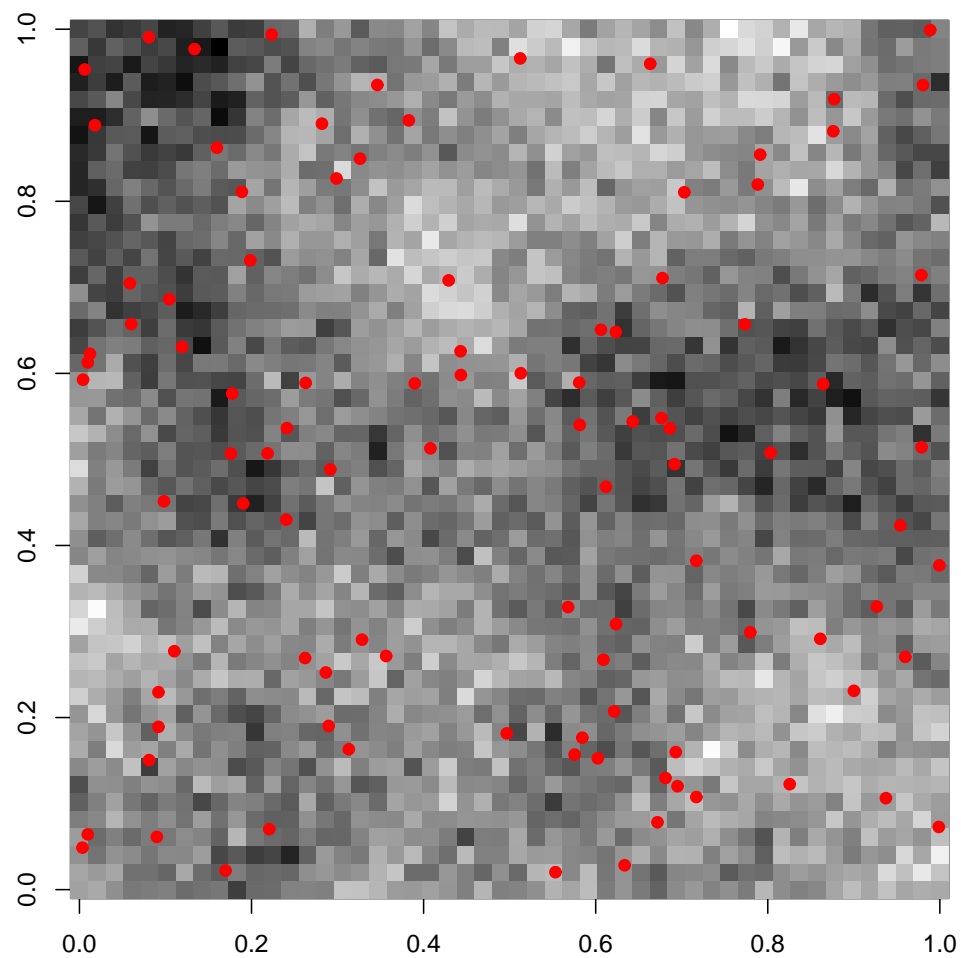
$$R | S \sim (S, \Sigma_g)$$

Notice that S is known, but R is not. Here assume

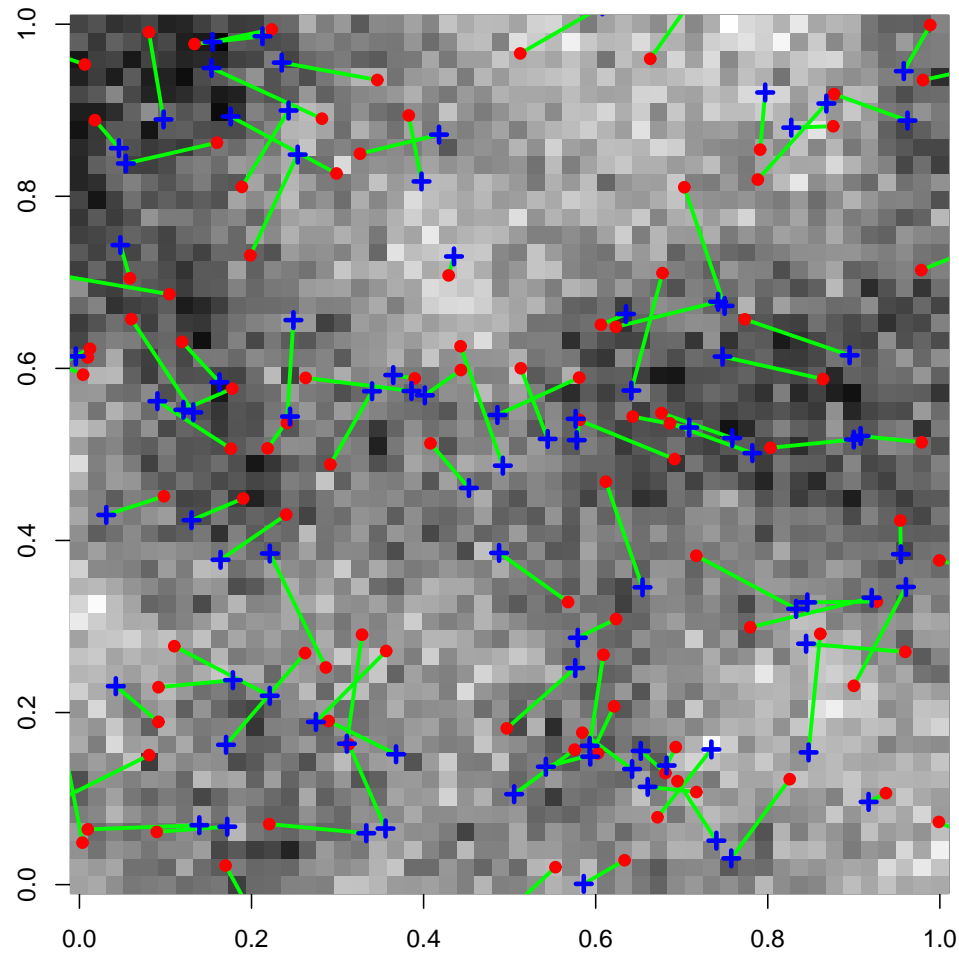
$$\mathbf{r} = \mathbf{s} + \mathbf{p}(\mathbf{s}); \quad \{\mathbf{p}(\mathbf{s})\} \text{ i.i.d. for } \mathbf{s} \in D,$$

where $\mathbf{p}(\cdot)$ is the positioning error with density $g(\cdot)$

$Z(\cdot)$ and Intended Locations (dots)



$Z(\cdot)$, **dots**, and Actual Locations (**pluses**)



CP Model, ctd

Contrast *CP* model with Feature-Positioning (*FP*) model.

Go to features (e.g., trees) in \mathbb{R}^d and observe their locations:

Observed feature locations B

True feature locations A

$$B \mid A \sim (A, \Gamma_g)$$

Notice that B is observed, but A is unknown.

CP Model, ctd

Consider Berkson's (1950) errors-in-variables problem.

$$\left. \begin{array}{l} \text{Observe } z_i \text{ in: } z_i = \alpha + \beta\omega_i + \delta_i \\ \text{Specify } x_i \text{ in: } \omega_i = x_i + p_i \end{array} \right\} \begin{array}{l} \text{Use target } x_i \text{ instead of } \omega_i: \\ z_i = \alpha + \beta x_i + \epsilon_i \end{array}$$

This is analogous to the CP model where:

x_i is analogous to s_i , the intended location (specified controls)

ω_i is analogous to r_i , the actual location (not observed)

KALE analysis adjusts for use of $\{s_i\}$ instead of $\{r_i\}$.

(An alternative model specified by Berkson, 1950, is analogous to the FP model.)

Location-Error Spatial Model

$$\begin{aligned}Z_g(\mathbf{s}) &= Y(\mathbf{r}) + \varepsilon(\mathbf{r}) \\ &= \mathbf{x}(\mathbf{r})'\boldsymbol{\beta} + \nu(\mathbf{r}) + \varepsilon(\mathbf{r}) \\ \mathbf{r} &= \mathbf{s} + \mathbf{p}(\mathbf{s}); \mathbf{s} \in D,\end{aligned}$$

where $\varepsilon(\cdot)$ is the measurement error (mean 0, variance σ^2) and $\mathbf{p}(\cdot)$ is the location error (density $g(\cdot)$). Then decompose:

$$\begin{aligned}Z_g(\mathbf{s}) &= Y(\mathbf{s}) + \epsilon(\mathbf{s}) + \xi_g(\mathbf{s}); \\ \text{that is, } \xi_g(\mathbf{s}) &= Z_g(\mathbf{s}) - Z(\mathbf{s}).\end{aligned}$$

Location-error component of variation:

$$\text{var}(\xi_g(\mathbf{s})) = \text{var}(Z_g(\mathbf{s}) - Z(\mathbf{s}))$$

$$\begin{aligned}
 &= \overbrace{\text{var}(Z_g(\mathbf{s}) - [Y(\mathbf{s}) + \epsilon(\mathbf{s})])}^{\text{zero, if no location error}} \\
 &= \text{var}(Y(\mathbf{r}) - Y(\mathbf{s}) + \epsilon(\mathbf{r}) - \epsilon(\mathbf{s})), \quad \text{where } \mathbf{r} = \mathbf{s} + \mathbf{p}(\mathbf{s}) \\
 &= E\{\text{var}(\dots | \mathbf{p})\} + \text{var}\{E(\dots | \mathbf{p})\} \\
 &= E\{\text{var}(Y(\mathbf{s} + \mathbf{p}) - Y(\mathbf{s}) | \mathbf{p})\} + 2c_{ME} + \text{var}\{\beta' \mathbf{x}(\mathbf{s} + \mathbf{p})\}, \\
 &\hspace{25em} \text{where } \mathbf{p} \text{ has density } g(\cdot)
 \end{aligned}$$

$$\begin{aligned}
 &= 2 \int (C_Y(\mathbf{0}) - C_Y(\mathbf{u})) g(\mathbf{u}) d\mathbf{u} + 2c_{ME} + \\
 &\quad \beta' \left[\int \mathbf{x}(\mathbf{s} + \mathbf{u}) \mathbf{x}(\mathbf{s} + \mathbf{u})' g(\mathbf{u}) d\mathbf{u} - \mathbf{x}_g(\mathbf{s}) \mathbf{x}_g(\mathbf{s})' \right] \beta,
 \end{aligned}$$

$$\text{where } \mathbf{x}_g(\mathbf{s}) = \int \mathbf{x}(\mathbf{s} + \mathbf{u}) g(\mathbf{u}) d\mathbf{u}$$

$$= \text{SPAT. DEP. TERM} + \text{M.E. TERM} + \text{TREND TERM}$$

Moments

$$\mu_g(\mathbf{s}) \equiv E(Z(\mathbf{s})) = \left(\int \mathbf{x}(\mathbf{s} + \mathbf{u})g(\mathbf{u})d\mathbf{u} \right)' \boldsymbol{\beta} = \mathbf{x}_g(\mathbf{s})' \boldsymbol{\beta}$$

$$(\mathbf{x}(\mathbf{s})' \boldsymbol{\beta} = \beta_0 \Rightarrow \mu_g(\mathbf{s}) = \beta_0)$$

$$C_g(\mathbf{h}) \equiv \text{cov}(Z_g(\mathbf{s}), Z_g(\mathbf{s} + \mathbf{h}))$$

$$= \int C_Y(\mathbf{h} + \mathbf{v} - \mathbf{u})g(\mathbf{u})g(\mathbf{v})d\mathbf{u}d\mathbf{v}; \quad \mathbf{h} \neq \mathbf{0}$$

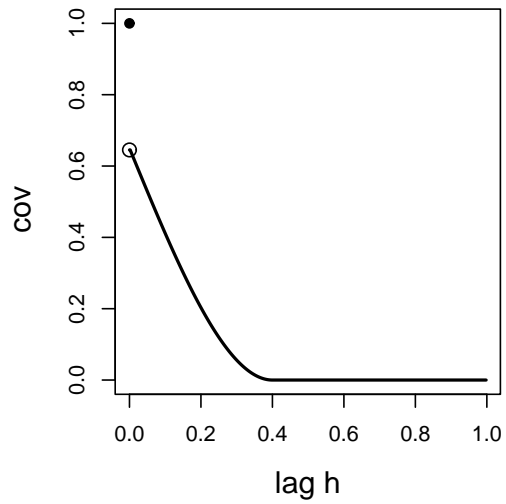
$$v_g(\mathbf{s}) \equiv \text{var}(Z_g(\mathbf{s}))$$

$$= C_Y(\mathbf{0}) + c_{ME}$$

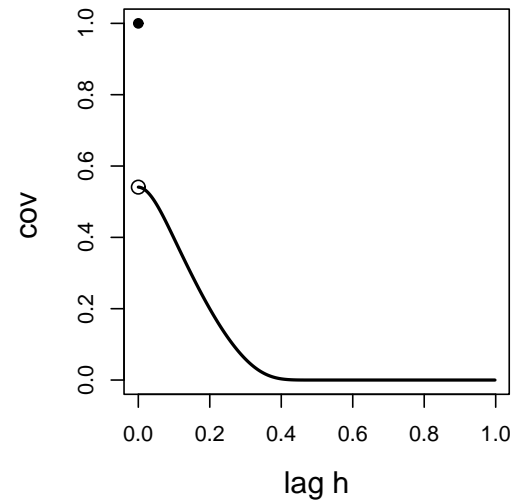
$$+ \boldsymbol{\beta}' \left\{ \int \mathbf{x}(\mathbf{s} + \mathbf{u})\mathbf{x}(\mathbf{s} + \mathbf{u})'g(\mathbf{u})d\mathbf{u} - \mathbf{x}_g(\mathbf{s})\mathbf{x}_g(\mathbf{s})' \right\} \boldsymbol{\beta}$$

$cov(Z(s), Z(s + h)), \text{No Trend}$

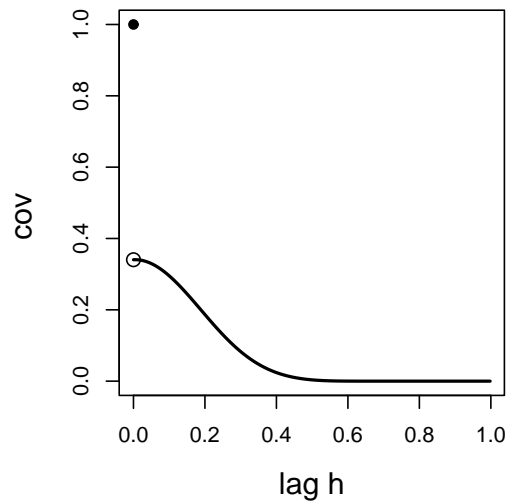
(a) $p(s) = 0$



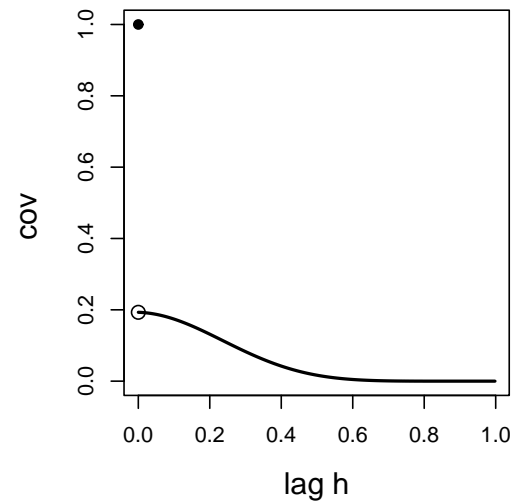
(b) $\psi = 0.05$



(c) $\psi = 0.15$



(d) $\psi = 0.25$



Is it true that we can adjust for location error by carrying out optimal linear prediction using the (non-stationary) covariance

$$K_g(\mathbf{h}) \equiv \begin{cases} C_g(\mathbf{h}); & \mathbf{h} \neq \mathbf{0} \\ v_g(\mathbf{s}); & \mathbf{h} = \mathbf{0} \end{cases}$$

?

Yes, after parameters are estimated and “plugged in”

Inference on β, θ

Assume that $Y(\cdot)$ and $\epsilon(\cdot)$ are Gaussian processes.

What is the *likelihood* of \mathbf{Z}_g ?

When there is no L.E., joint distribution of $\mathbf{Z}_g (= \mathbf{Z})$ is Gaussian.

When there is L.E., joint distribution of \mathbf{Z}_g is a mixture of Gaussians.

Suggestion: Use first two moments and maximize Gaussian likelihood, even though joint distribution is not Gaussian. This gives maximum *quasi-likelihood* estimators (MQLEs) $\hat{\beta}_g, \hat{\theta}_g$ of β, θ .

Kriging Adjusting for Location Error (KALE)

$$p_g(\mathbf{Z}_g; \mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)' \hat{\boldsymbol{\beta}}_g + \mathbf{c}_g(\mathbf{s}_0)' \Sigma_g^{-1} (\mathbf{Z}_g - X_g \hat{\boldsymbol{\beta}}_g), \text{ where}$$

$$\mathbf{c}_g(\mathbf{s}_0) = (c_{g,1}(\mathbf{s}_0), \dots, c_{g,n}(\mathbf{s}_0))'$$

$$c_{g,i}(\mathbf{s}_0) = \int C_Y(\mathbf{s}_i - \mathbf{s}_0 + \mathbf{u}; \hat{\boldsymbol{\theta}}_g) g(\mathbf{u}) d\mathbf{u}$$

$$\Sigma_g = (\sigma_{g,ij})$$

$$\sigma_{g,ij} = \int C_Y(\mathbf{s}_j - \mathbf{s}_i + \mathbf{v} - \mathbf{u}; \hat{\boldsymbol{\theta}}_g) g(\mathbf{u}) g(\mathbf{v}) d\mathbf{u} d\mathbf{v}; i \neq j$$

$$\sigma_{g,ii} = v_g(\mathbf{s}_i; \hat{\boldsymbol{\beta}}_g, \hat{\boldsymbol{\theta}}_g)$$

Adjusting versus Ignoring

We shall compare kriging **adjusting** for location error (**KALE**) to kriging **ignoring** location error (**KILE**).

Adjusting for L.E., involves computing location-adjusted quantities $\mathbf{c}_g(\mathbf{s}_0)$ and Σ_g . These are then used to find the optimal linear predictor for $Y(\mathbf{s}_0)$, namely the **KALE** predictor $p_g(\mathbf{Z}_g; \mathbf{s}_0)$.

When location error is (inappropriately) **ignored**, we obtain the **KILE** predictor $p(\mathbf{Z}_g; \mathbf{s}_0)$.

Simulation study to compare Adjusting versus Ignoring

We performed a large scale simulation study comparing classical kriging **adjusting** for location error (**KALE**) with classical kriging **ignoring** location error (**KILE**). Full details: Cressie and Kornak (2003).

Simulation-study Conclusions

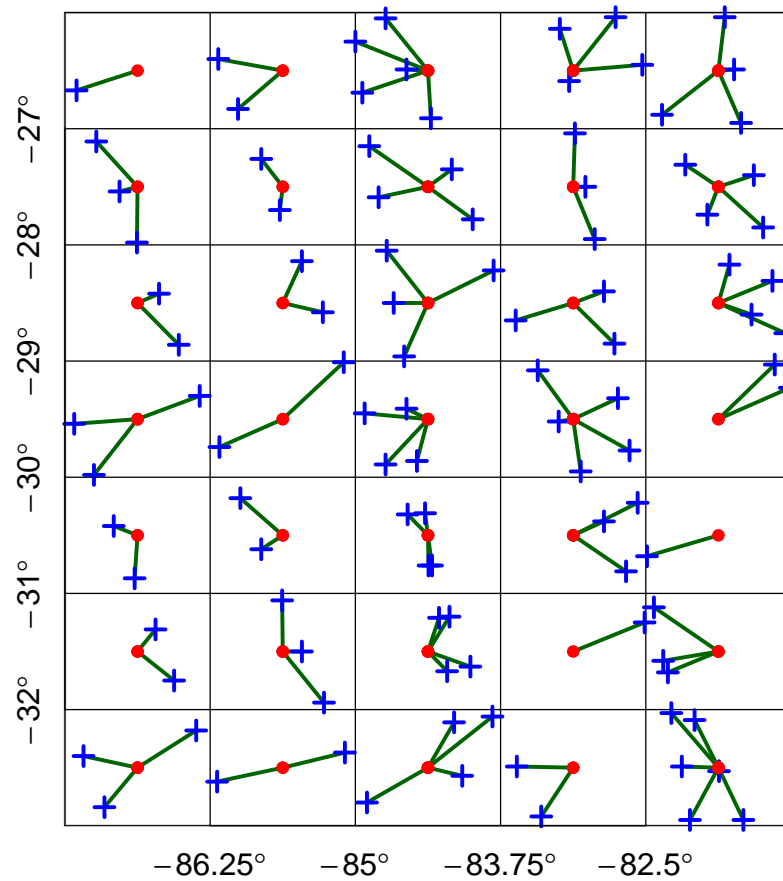
- RMSPE increases with increasing L.E.
- KALE gives reduced RMSPE over KILE
- RMSPE improvement of KALE over KILE increases with increasing L.E.
- KALE gives largest RMSPE improvement for prediction locations close to, or at intended sites
- RMSPE is not affected by the addition of a linear trend term

Total Column Ozone

- TOMS instrument on Nimbus 7 satellite – usually a complete but spatially *irregular* coverage of globe in 1 day
- Data put on *regular* grid (1.25° lon \times 1° lat)
- Because data are massive, grid centers are used as data locations (intended locations) rather than actual locations (which are available!)
- Consider 7.25° lon \times 7° lat region off the coast of Chile on Oct 1, 1988:

Satellite Data:

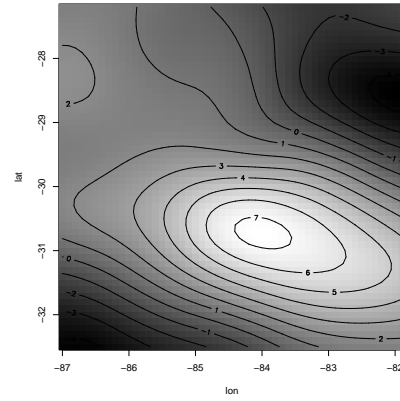
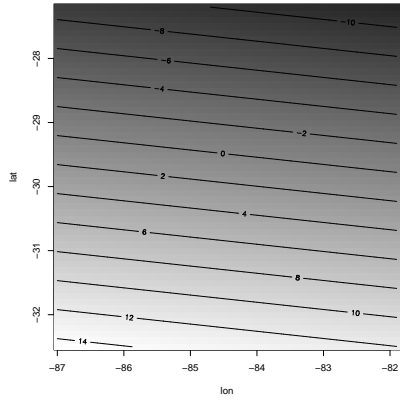
Intended Locations (**dots**) and Observed Locations (**pluses**)



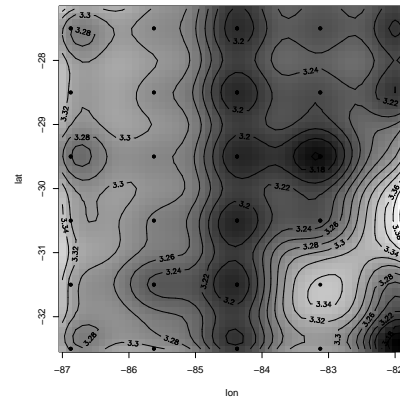
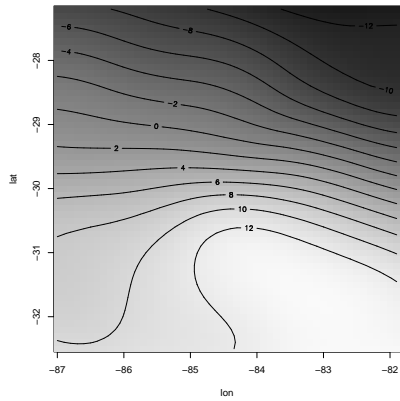
Total Column Ozone - ctd

- L.E. appears to be (marginally) uniform on a $1.25^{\circ} \times 1^{\circ}$ lon/lat rectangle (this L.E. distribution is subsequently assumed to be g in the analysis)
- The process is modeled as having linear trend in the lon and lat directions
- The spherical covariance function is used to describe the spatial dependence and is defined in terms of great-arc distances

$$\mathbf{x}_g(\cdot)' \hat{\boldsymbol{\beta}}_g - 307.3 \quad + \quad \hat{\nu}(\cdot)$$



$$= p_g(\mathbf{Z}_g; \cdot) - 307.3 \quad \pm \quad \{\text{MSPE}_g(\cdot)\}^{1/2}$$



Summary

- Location error should be accounted for when there is *significant* location error in the data (reduces RMSPE).
- For further details: “Spatial statistics in the presence of location error with an application to remote sensing of the environment”, Cressie and Kornak (2003), forthcoming in *Statistical Science*.