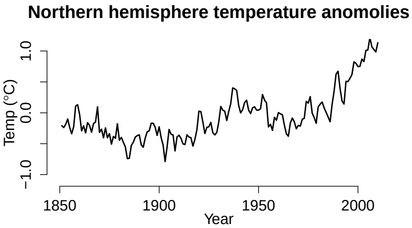
Some Statistical Problems in Climate Reconstruction

Dan Cervone

April 15, 2014

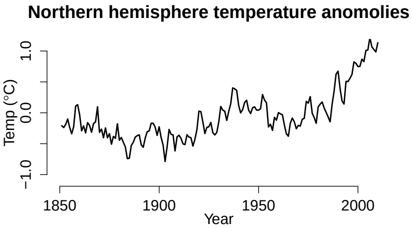
Historical Global Temperature Reconstruction

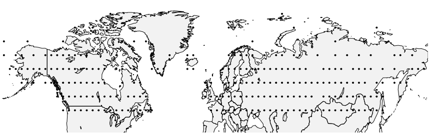
Data: CRUTEMv3



Historical Global Temperature Reconstruction

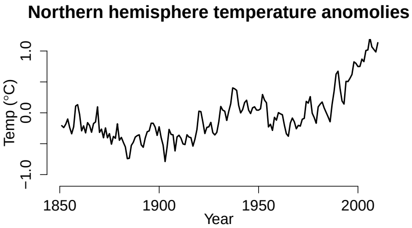
Data: CRUTEMv3

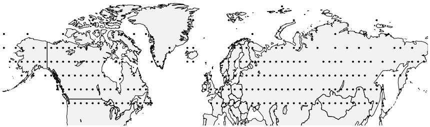


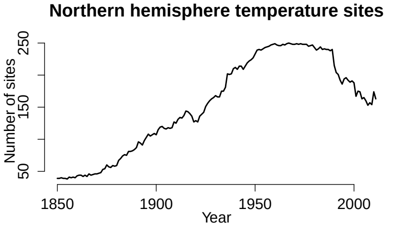


Historical Global Temperature Reconstruction

Data: CRUTEMv3

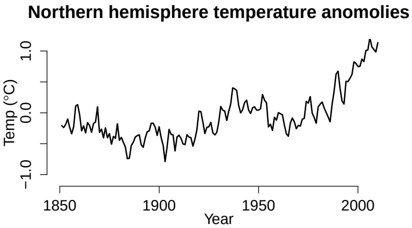


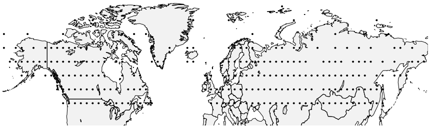




Historical Global Temperature Reconstruction

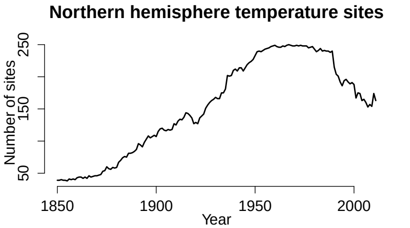
Data: CRUTEMv3





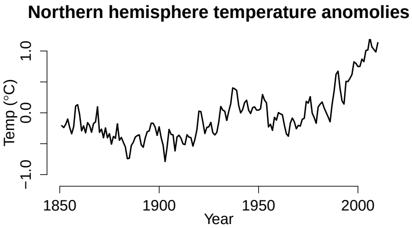
What is the estimand?

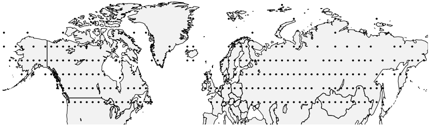
- Interpolate gaps in observational record
- Extrapolate before 1850

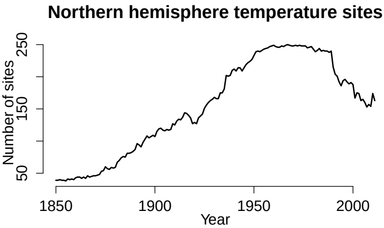


Historical Global Temperature Reconstruction

Data: CRUTEMv3

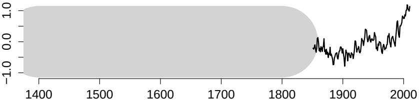






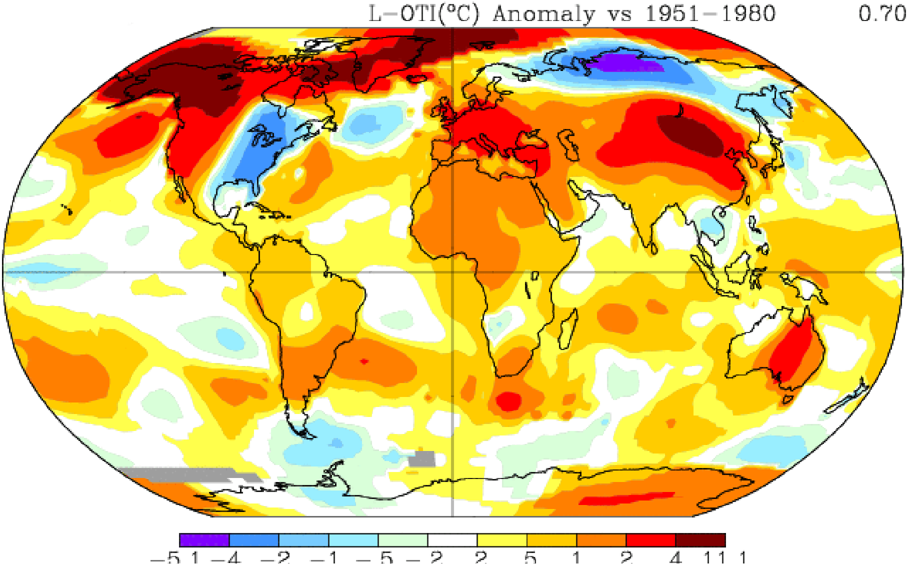
What is the estimand?

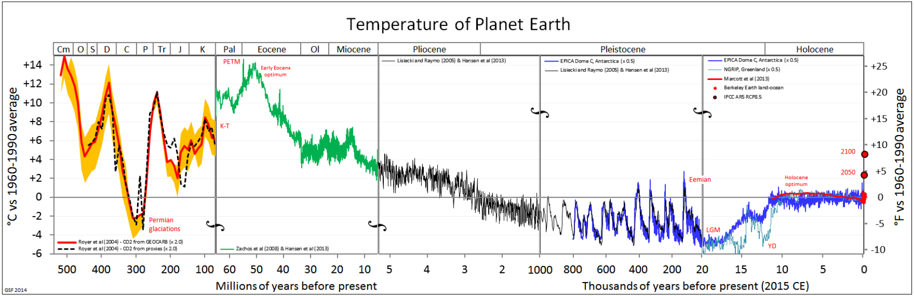
- Interpolate gaps in observational record
- Extrapolate before 1850



All the moments each moment

Image: NASA/GISS





[image source: wikimedia commons]

Proxies:

- $^{18}O/^{16}O$, ocean sediment
- Ice cores
- Varves (rock sediment)
- Tree rings

Example: BARCAST

Tingley & Huybers 2010, 2013

Spatiotemporal temperature reconstruction using temperature record and proxies:

$$\mathbf{T}_t = \begin{pmatrix} \mathbf{T}_{o,t} \\ \mathbf{T}_{p,t} \end{pmatrix} \text{ at locations } \mathbf{S} = \begin{pmatrix} \mathbf{S}_o \\ \mathbf{S}_p \end{pmatrix}$$

Example: BARCAST

Tingley & Huybers 2010, 2013

Spatiotemporal temperature reconstruction using temperature record and proxies:

$$\mathbf{T}_t = \begin{pmatrix} \mathbf{T}_{o,t} \\ \mathbf{T}_{p,t} \end{pmatrix} \text{ at locations } \mathbf{S} = \begin{pmatrix} \mathbf{S}_o \\ \mathbf{S}_p \end{pmatrix}$$

- $\mathbf{T}_o$ are temperatures at locations of temperature records $\mathbf{S}_o$.
- $\mathbf{T}_p$ are temperatures at locations of proxy records $\mathbf{S}_p$.

Example: BARCAST

Tingley & Huybers 2010, 2013

Spatiotemporal temperature reconstruction using temperature record and proxies:

$$\mathbf{T}_t = \begin{pmatrix} \mathbf{T}_{o,t} \\ \mathbf{T}_{p,t} \end{pmatrix} \text{ at locations } \mathbf{S} = \begin{pmatrix} \mathbf{S}_o \\ \mathbf{S}_p \end{pmatrix}$$

- $\mathbf{T}_o$ are temperatures at locations of temperature records $\mathbf{S}_o$.
- $\mathbf{T}_p$ are temperatures at locations of proxy records $\mathbf{S}_p$.

With t indexing years,

$$\mathbf{T}_t - \mu\mathbf{1} = \alpha(\mathbf{T}_{t-1} - \mu\mathbf{1}) + \boldsymbol{\epsilon}_t$$

$$\boldsymbol{\epsilon}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, K(\mathbf{S}, \mathbf{S}))$$

$$K(\mathbf{s}, \mathbf{s}^*) = \tau^2 \exp(-\gamma||\mathbf{s} - \mathbf{s}^*||^2)$$

Example: BARCAST

Tingley & Huybers 2010, 2013

"Errors in variables":

- True temperatures $\mathbf{T}$ are not observed.
- Measurement error for temperature sites $\mathbf{W}_{o,t} \sim \mathcal{N}(\mathbf{T}_{o,t}, \sigma_o^2 \mathbf{I})$.
- Linear model for proxies $\mathbf{W}_{p,t} \sim \mathcal{N}(\mu_p \mathbf{1} + \mathbf{T}_{p,t}\beta_p, \sigma_p^2 \mathbf{I})$.

Example: BARCAST

Tingley & Huybers 2010, 2013

"Errors in variables":

- True temperatures $\mathbf{T}$ are not observed.
- Measurement error for temperature sites $\mathbf{W}_{o,t} \sim \mathcal{N}(\mathbf{T}_{o,t}, \sigma_o^2 \mathbf{I})$.
- Linear model for proxies $\mathbf{W}_{p,t} \sim \mathcal{N}(\mu_p \mathbf{1} + \mathbf{T}_{p,t}\beta_p, \sigma_p^2 \mathbf{I})$.
- $(\mathbf{W} \;\; \mathbf{T})'$ is just a huge multivariate normal!

Example: BARCAST

Tingley & Huybers 2010, 2013

"Errors in variables":

- True temperatures $\mathbf{T}$ are not observed.
- Measurement error for temperature sites $\mathbf{W}_{o,t} \sim \mathcal{N}(\mathbf{T}_{o,t}, \sigma_o^2 \mathbf{I})$.
- Linear model for proxies $\mathbf{W}_{p,t} \sim \mathcal{N}(\mu_p \mathbf{1} + \mathbf{T}_{p,t}\beta_p, \sigma_p^2 \mathbf{I})$.
- $(\mathbf{W} \;\; \mathbf{T})'$ is just a huge multivariate normal!

Inference with Gibbs sampling or EM:

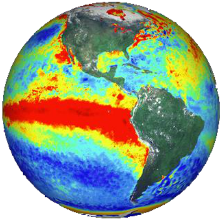
- Update latent $\mathbf{T}$.
- Update parameters $\tau^2, \gamma, \mu, \alpha, \mu_p, \beta_p, \sigma_o^2, \sigma_p^2$.

Example: BARCAST

Tingley & Huybers 2010, 2013

Difficulties:

- Spatiotemporal nonstationarity and anisotropy.

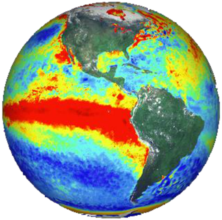


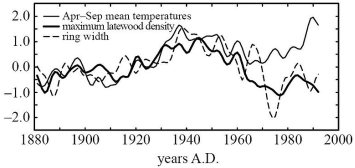
Example: BARCAST

Tingley & Huybers 2010, 2013

Difficulties:

- Spatiotemporal nonstationarity and anisotropy.
- Model inhomogeneity.



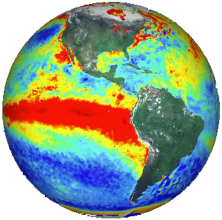


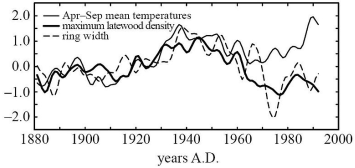
Example: BARCAST

Tingley & Huybers 2010, 2013

Difficulties:

- Spatiotemporal nonstationarity and anisotropy.
- Model inhomogeneity.
- Uncertainty in spatial referencing.



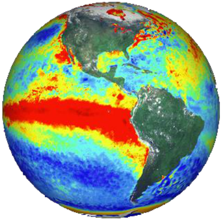


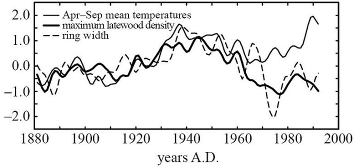
Example: BARCAST

Tingley & Huybers 2010, 2013

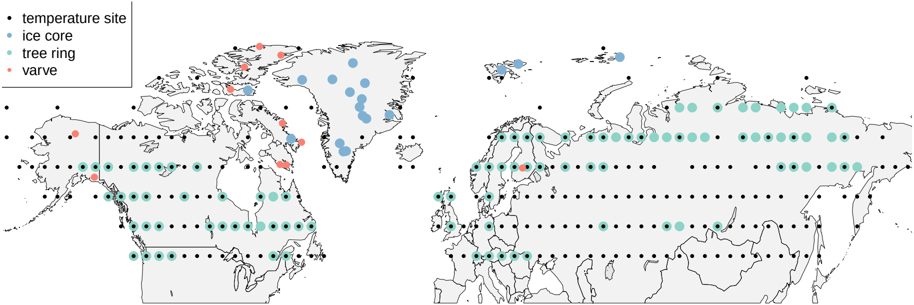
Difficulties:

- Spatiotemporal nonstationarity and anisotropy.
- Model inhomogeneity.
- Uncertainty in spatial referencing.
- ...





Location uncertainty



- Tree locations uncertain for many older specimens
- Ice cores subject to glacial flow

Gaussian Processes

For $\mathbf{s} \in \mathcal{S}$, $X(\mathbf{s}) \sim \mathcal{GP}(0, K(\mathbf{s}, \mathbf{s}))$ means for any $\mathbf{s}_1, \dots \mathbf{s}_p \in \mathcal{S}$,

$$\begin{pmatrix} X(\mathbf{s}_1) \\ \vdots \\ X(\mathbf{s}_p) \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K(\mathbf{s}_1, \mathbf{s}_1) & \dots & K(\mathbf{s}_1, \mathbf{s}_p) \\ \vdots & \ddots & \\ K(\mathbf{s}_p, \mathbf{s}_1) & & K(\mathbf{s}_p, \mathbf{s}_p) \end{pmatrix}\right),$$

$K(\, , \,)$ is a covariance function, e.g. $K(\mathbf{s}, \mathbf{s}^*) = \tau^2 \exp(-\gamma ||\mathbf{s} - \mathbf{s}^*||^2)$.

Gaussian Processes

For $\mathbf{s} \in \mathcal{S}$, $X(\mathbf{s}) \sim \mathcal{GP}(0, K(\mathbf{s}, \mathbf{s}))$ means for any $\mathbf{s}_1, \dots \mathbf{s}_p \in \mathcal{S}$,

$$\begin{pmatrix} X(\mathbf{s}_1) \\ \vdots \\ X(\mathbf{s}_p) \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K(\mathbf{s}_1, \mathbf{s}_1) & \dots & K(\mathbf{s}_1, \mathbf{s}_p) \\ \vdots & \ddots & \\ K(\mathbf{s}_p, \mathbf{s}_1) & & K(\mathbf{s}_p, \mathbf{s}_p) \end{pmatrix}\right),$$

$K(\, , \,)$ is a covariance function, e.g. $K(\mathbf{s}, \mathbf{s}^*) = \tau^2 \exp(-\gamma ||\mathbf{s} - \mathbf{s}^*||^2)$.

- Interpolation of X at unobserved location $\mathbf{s}^*$

Gaussian Processes

For $\mathbf{s} \in \mathcal{S}$, $X(\mathbf{s}) \sim \mathcal{GP}(0, K(\mathbf{s}, \mathbf{s}))$ means for any $\mathbf{s}_1, \dots \mathbf{s}_p \in \mathcal{S}$,

$$\begin{pmatrix} X(\mathbf{s}_1) \\ \vdots \\ X(\mathbf{s}_p) \end{pmatrix} \sim \mathcal{N} \left(\mathbf{0}, \begin{pmatrix} K(\mathbf{s}_1, \mathbf{s}_1) & \dots & K(\mathbf{s}_1, \mathbf{s}_p) \\ \vdots & \ddots & \\ K(\mathbf{s}_p, \mathbf{s}_1) & & K(\mathbf{s}_p, \mathbf{s}_p) \end{pmatrix} \right),$$

$K(\ ,\ )$ is a covariance function, e.g. $K(\mathbf{s}, \mathbf{s}^*) = \tau^2 \exp(-\gamma||\mathbf{s} - \mathbf{s}^*||^2)$.

- Interpolation of X at unobserved location $\mathbf{s}^*$
- $X(\mathbf{s}^*)|X(\mathbf{s}) \sim$ $\mathcal{N}(K(\mathbf{s}^*, \mathbf{s})K(\mathbf{s}, \mathbf{s})^{-1}X(\mathbf{s}), K(\mathbf{s}^*, \mathbf{s}^*) - K(\mathbf{s}^*, \mathbf{s})K(\mathbf{s}, \mathbf{s})^{-1}K(\mathbf{s}, \mathbf{s}^*))$

Gaussian Processes

For $\mathbf{s} \in \mathcal{S}$, $X(\mathbf{s}) \sim \mathcal{GP}(0, K(\mathbf{s}, \mathbf{s}))$ means for any $\mathbf{s}_1, \dots \mathbf{s}_p \in \mathcal{S}$,

$$\begin{pmatrix} X(\mathbf{s}_1) \\ \vdots \\ X(\mathbf{s}_p) \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K(\mathbf{s}_1, \mathbf{s}_1) & \dots & K(\mathbf{s}_1, \mathbf{s}_p) \\ \vdots & \ddots & \\ K(\mathbf{s}_p, \mathbf{s}_1) & & K(\mathbf{s}_p, \mathbf{s}_p) \end{pmatrix}\right),$$

$K(\, , \,)$ is a covariance function, e.g. $K(\mathbf{s}, \mathbf{s}^*) = \tau^2 \exp(-\gamma ||\mathbf{s} - \mathbf{s}^*||^2)$.

- Interpolation of X at unobserved location $\mathbf{s}^*$
- $X(\mathbf{s}^*)|X(\mathbf{s}) \sim \mathcal{N}(K(\mathbf{s}^*, \mathbf{s})K(\mathbf{s}, \mathbf{s})^{-1}X(\mathbf{s}), K(\mathbf{s}^*, \mathbf{s}^*) - K(\mathbf{s}^*, \mathbf{s})K(\mathbf{s}, \mathbf{s})^{-1}K(\mathbf{s}, \mathbf{s}^*))$
- Kriging: BLUP without normality assumption

GP interpolation with location error

Errors in variables: $X(\mathbf{s}^*)|X(\mathbf{s}) \sim \mathcal{N}(\mathbf{b}'X(\mathbf{s}), v^2)$

- Observe $\tilde{X}(\mathbf{s}) = X(\mathbf{s}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \perp X(\mathbf{s})$.

GP interpolation with location error

Errors in variables: $X(\mathbf{s}^*)|X(\mathbf{s}) \sim \mathcal{N}(\mathbf{b}'X(\mathbf{s}), v^2)$

- Observe $\tilde{X}(\mathbf{s}) = X(\mathbf{s}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \perp X(\mathbf{s})$.
- Still a regression problem: $X(\mathbf{s}^*)|\tilde{X}(\mathbf{s}) \sim \mathcal{N}(\tilde{\mathbf{b}}'\tilde{X}(\mathbf{s}), \tilde{v}^2)$

GP interpolation with location error

Errors in variables: $X(\mathbf{s}^*)|X(\mathbf{s}) \sim \mathcal{N}(\mathbf{b}'X(\mathbf{s}), v^2)$

- Observe $\tilde{X}(\mathbf{s}) = X(\mathbf{s}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \perp X(\mathbf{s})$.
- Still a regression problem: $X(\mathbf{s}^*)|\tilde{X}(\mathbf{s}) \sim \mathcal{N}(\tilde{\mathbf{b}}'\tilde{X}(\mathbf{s}), \tilde{v}^2)$
- Berkson errors: $\boldsymbol{\epsilon} \perp \tilde{X}(\mathbf{s})$ (not satisfied)

GP interpolation with location error

Errors in variables: $X(\mathbf{s}^*)|X(\mathbf{s}) \sim \mathcal{N}(\mathbf{b}' X(\mathbf{s}), v^2)$

- Observe $\tilde{X}(\mathbf{s}) = X(\mathbf{s}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \perp X(\mathbf{s})$.
- Still a regression problem: $X(\mathbf{s}^*)|\tilde{X}(\mathbf{s}) \sim \mathcal{N}(\tilde{\mathbf{b}}' \tilde{X}(\mathbf{s}), \tilde{v}^2)$
- Berkson errors: $\boldsymbol{\epsilon} \perp \tilde{X}(\mathbf{s})$ (not satisfied)

Is i.i.d. error in $\mathbf{s}$ just i.i.d. error in $X(\mathbf{s})$?

- $X(\mathbf{s} + \mathbf{u}) = X(\mathbf{s}) + \boldsymbol{\epsilon}$

GP interpolation with location error

Errors in variables: $X(\mathbf{s}^*)|X(\mathbf{s}) \sim \mathcal{N}(\mathbf{b}'X(\mathbf{s}), v^2)$

- Observe $\tilde{X}(\mathbf{s}) = X(\mathbf{s}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \perp X(\mathbf{s})$.
- Still a regression problem: $X(\mathbf{s}^*)|\tilde{X}(\mathbf{s}) \sim \mathcal{N}(\tilde{\mathbf{b}}'\tilde{X}(\mathbf{s}), \tilde{v}^2)$
- Berkson errors: $\boldsymbol{\epsilon} \perp \tilde{X}(\mathbf{s})$ (not satisfied)

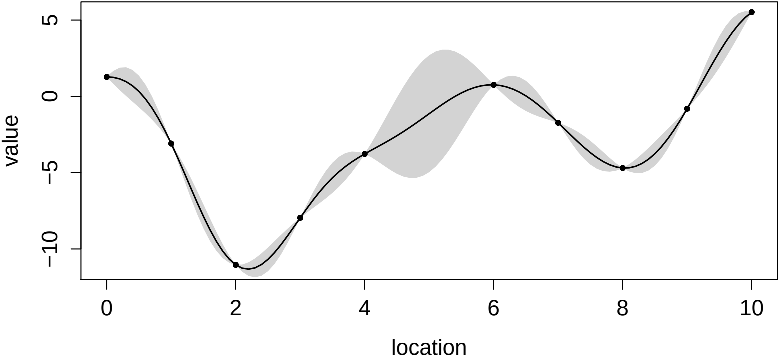
Is i.i.d. error in $\mathbf{s}$ just i.i.d. error in $X(\mathbf{s})$?

- $X(\mathbf{s} + \mathbf{u}) = X(\mathbf{s}) + \boldsymbol{\epsilon}$
- $\boldsymbol{\epsilon} \sim \mathcal{N}((K_{u,s}K_{s,s}^{-1} - \mathbf{I})X(\mathbf{s}), K_{u,u} - K_{u,s}K_{s,s}^{-1}K_{s,u})$ where $K_{u,s} = K(\mathbf{s} + \mathbf{u}, \mathbf{s})$, etc.
- $\boldsymbol{\epsilon} \not\perp X(\mathbf{s})$ and $\boldsymbol{\epsilon} \not\perp X(\mathbf{s} + \mathbf{u})$

GP interpolation with location error

Illustration

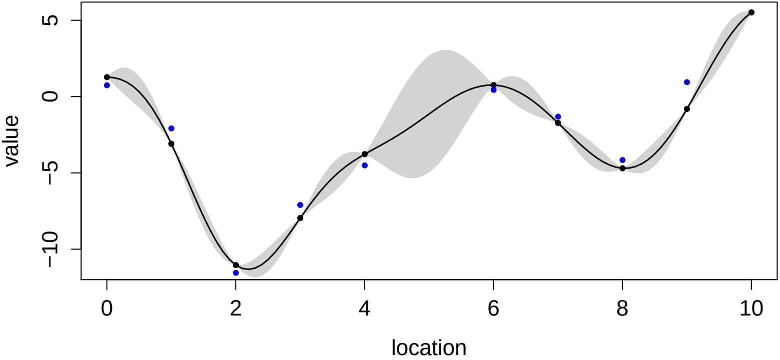
i.i.d. additive error for **measurements** X



GP interpolation with location error

Illustration

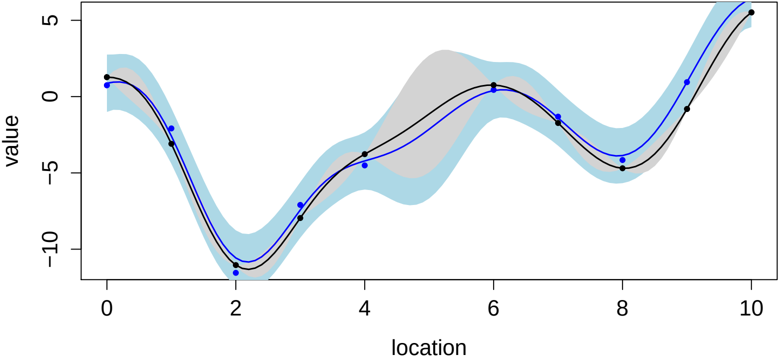
i.i.d. additive error for **measurements** X



GP interpolation with location error

Illustration

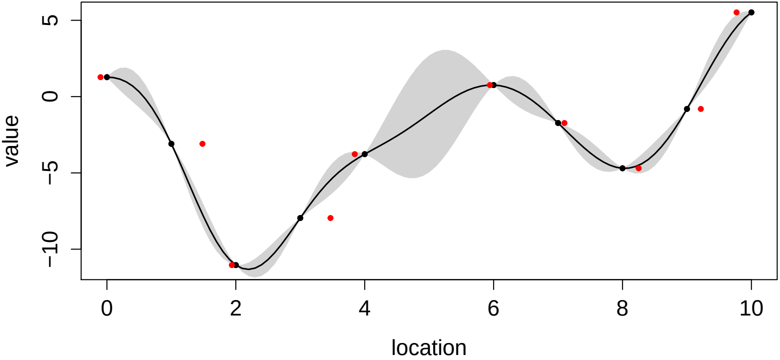
i.i.d. additive error for **measurements** X



GP interpolation with location error

Illustration

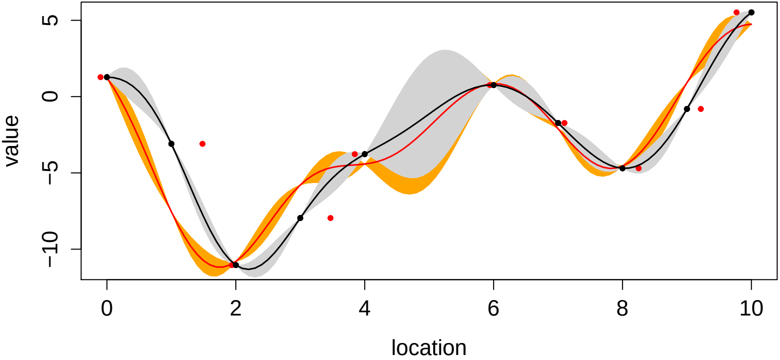
i.i.d. additive error for **locations** s



GP interpolation with location error

Illustration

i.i.d. additive error for **locations** s



Kriging

Cressie & Kornak 2003; Fanshawe & Diggle 2010

Location errors induce (non-Gaussian) process by convolution:

- $X_g(\mathbf{s}) = X(\mathbf{s} + \mathbf{u}), \mathbf{u} \sim g(\mathbf{u})$

Kriging

Cressie & Kornak 2003; Fanshawe & Diggle 2010

Location errors induce (non-Gaussian) process by convolution:

- $X_g(\mathbf{s}) = X(\mathbf{s} + \mathbf{u}), \mathbf{u} \sim g(\mathbf{u})$
- $X_g(\mathbf{s})$ is mean 0 with covariance function:
  $K_g(\mathbf{s}, \mathbf{s}) = \int K(\mathbf{s}+\mathbf{u}, \mathbf{s}+\mathbf{u}) g(\mathbf{u})$, or $K_g(\mathbf{s}, \mathbf{s}^*) = \int K(\mathbf{s}+\mathbf{u}, \mathbf{s}^*) g(\mathbf{u})$

Kriging

Cressie & Kornak 2003; Fanshawe & Diggle 2010

Location errors induce (non-Gaussian) process by convolution:

- $X_g(\mathbf{s}) = X(\mathbf{s} + \mathbf{u}), \mathbf{u} \sim g(\mathbf{u})$
- $X_g(\mathbf{s})$ is mean 0 with covariance function: $K_g(\mathbf{s}, \mathbf{s}) = \int K(\mathbf{s} + \mathbf{u}, \mathbf{s} + \mathbf{u}) g(\mathbf{u})$, or $K_g(\mathbf{s}, \mathbf{s}^*) = \int K(\mathbf{s} + \mathbf{u}, \mathbf{s}^*) g(\mathbf{u})$
- Kriging gives BLUP for $X(\mathbf{s}^*)$ given $X_g(\mathbf{s})$.

Typically $K_g(\mathbf{s}, \mathbf{s})$ evaluated by Monte Carlo.

Kriging

With measurement error in locations, the BLUP:

- Inadmissible under squared error loss
- First two moments give invalid interval coverage
- Requires Monte Carlo, generally $\mathcal{O}(n^3)$

Kriging

With measurement error in locations, the BLUP:

- Inadmissible under squared error loss
- First two moments give invalid interval coverage
- Requires Monte Carlo, generally $\mathcal{O}(n^3)$

We should use the BnLUP $E[X(\mathbf{s}^*)|X_g(\mathbf{s})]$!

- Dominates BLUP
- First two moments give valid interval coverage
- Easily implemented with HMC, generally $\mathcal{O}(n^3)$
- Easily extends to inference for parameters

Simulation

Compare interpolation using BLUP, BnLUP, and no adjustment for location measurement error.

- $\mathbf{s} = \{0, 1, \ldots, 4, 6, \ldots, 10\}$ and $\mathbf{s}^* = \{5, 11\}$
- $\mathbf{u} \sim \text{Unif}(-\theta_u, \theta_u)$.
- Other combinations of all other parameters.

Simulation

Compare interpolation using BLUP, BnLUP, and no adjustment for location measurement error.

- $\mathbf{s} = \{0, 1, \ldots, 4, 6, \ldots, 10\}$ and $\mathbf{s}^* = \{5, 11\}$
- $\mathbf{u} \sim \text{Unif}(-\theta_u, \theta_u)$.
- Other combinations of all other parameters.

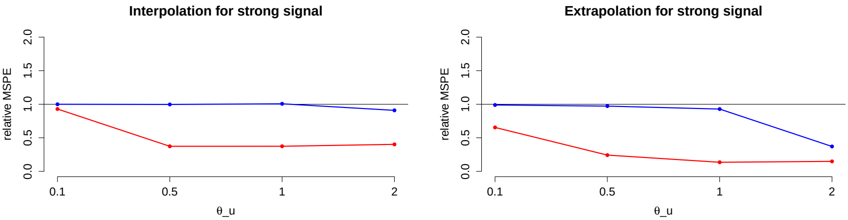
Two cases of particular interest:

- Strong signal: high autocorrelation (γ small) and small "nugget" variance σ^2
- Weak signal: low autocorrelation (γ large) or large "nugget" variance σ^2

All parameters fixed and known in simulations.

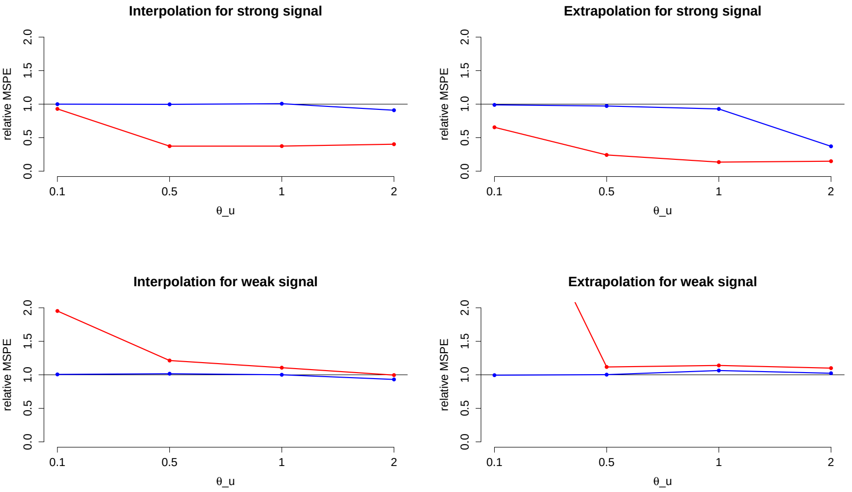
Simulation results

Mean squared prediction error vs. oracle predictor



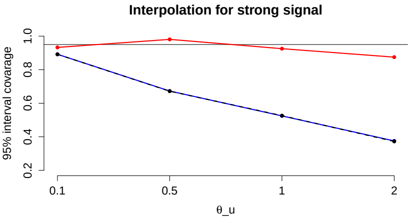
Simulation results

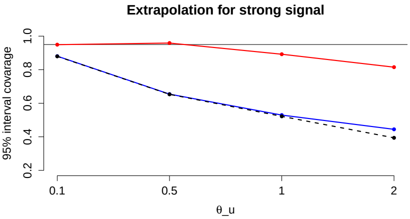
Mean squared prediction error vs. oracle predictor



Simulation results

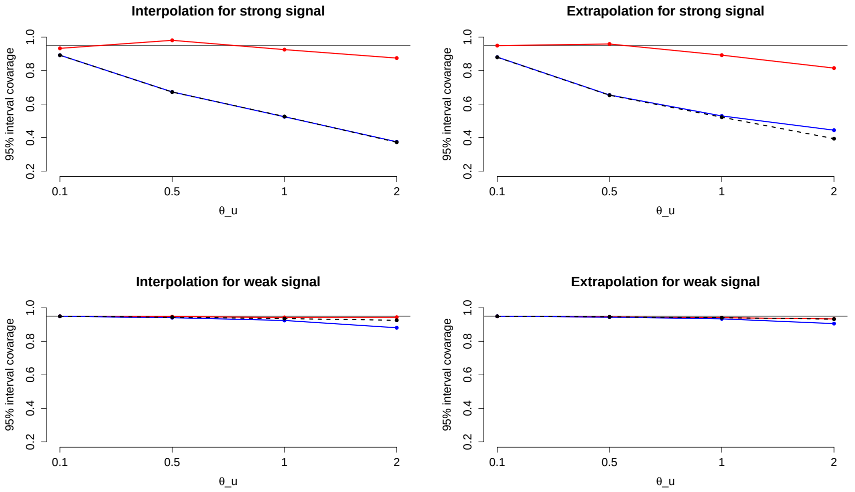
95% Interval coverage





Simulation results

95% Interval coverage



Next steps

Application to climate data:

- Location errors in covariates referencing
- Large-scale model
- Influence of extreme value estimation

Next steps

Application to climate data:

- Location errors in covariates referencing
- Large-scale model
- Influence of extreme value estimation

Also:

- (Stochastic) EM implementation
- BnLUP without normality assumption?