

Fitting spatial + space-time models

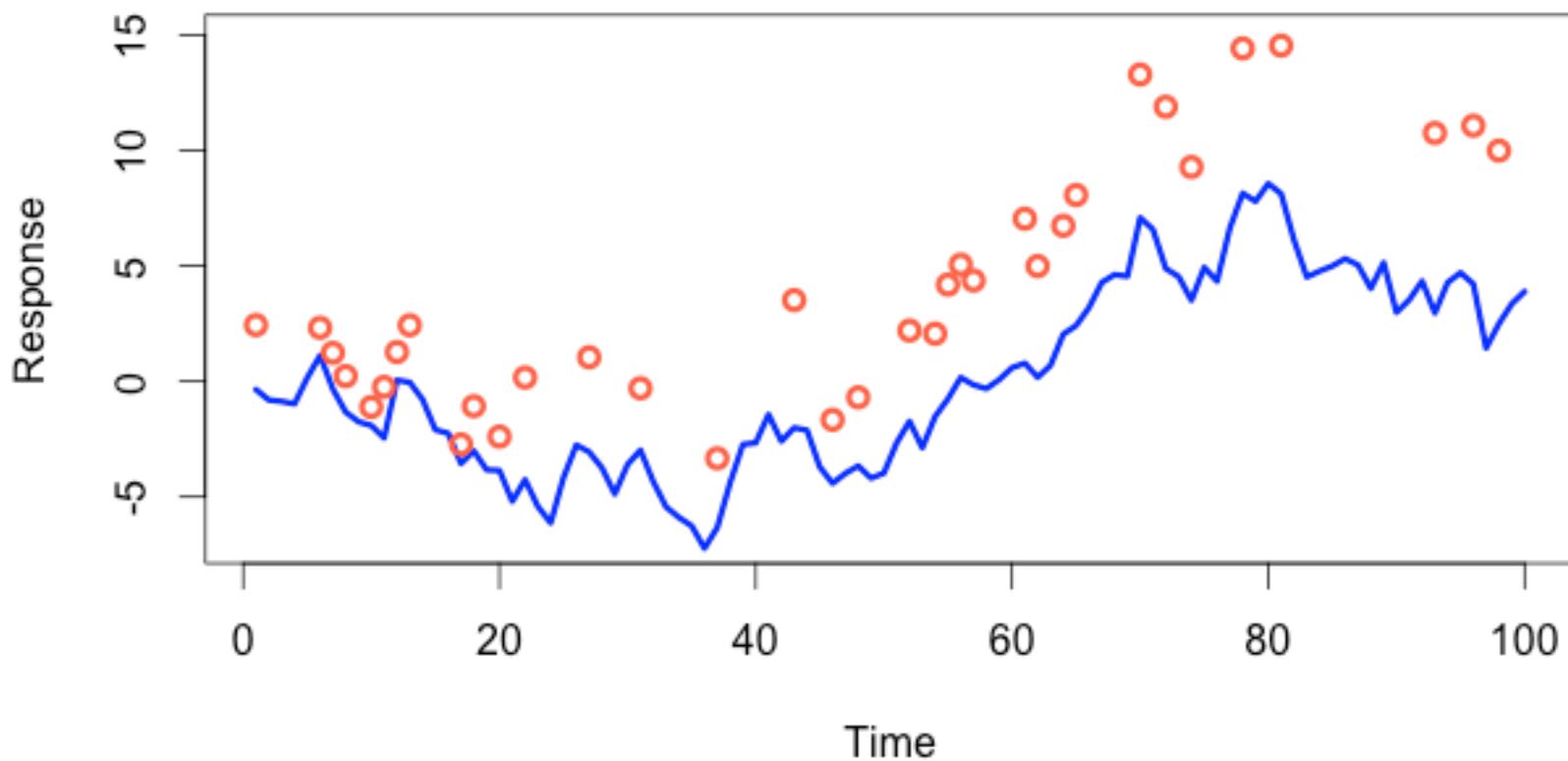
Eric Ward

Feb 21 2017

What we've learned so far

- Time series can be useful for identifying structure, improving precision, and accuracy of forecasts
- Modeling multivariate time series
 - e.g. MARSS() function, with each observed time series mapped to a single discrete state
- Using DFA
 - Structure determined by factor loadings

Response generally the same variable



Making inference about population as a whole involves modeling pink as well as blue time series

Multivariate approach with MARSS

- Time series modeled as independent (allowed to have different states/trends)
- But we estimate the correlation with **Q** matrix
- **Q** = “unconstrained”

$$Q = \begin{bmatrix} q_{1,1} & q_{1,2} & q_{1,3} & q_{1,4} \\ q_{2,1} & q_{2,2} & q_{2,3} & q_{2,4} \\ q_{3,1} & q_{3,2} & q_{3,3} & q_{3,4} \\ q_{4,1} & q_{4,2} & q_{4,3} & q_{4,4} \end{bmatrix}$$

- Obvious problem: $(m * (m+1))/2$ parameters!

Constraints on constraints

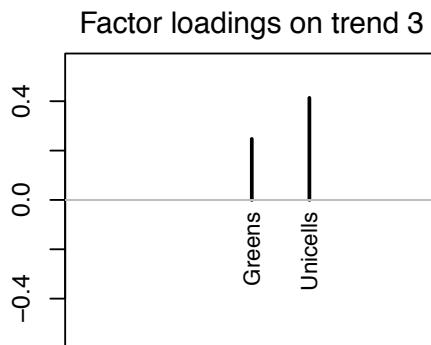
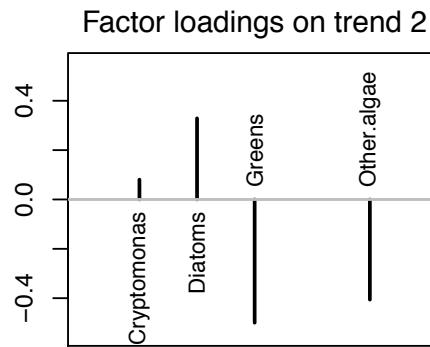
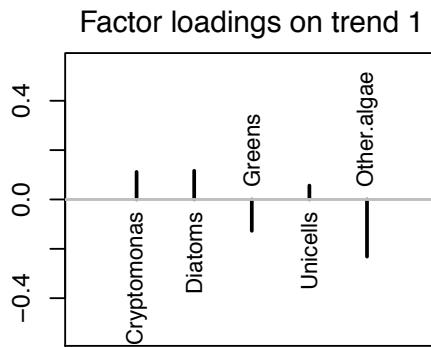
- We'd like to be able to reduce # parameters in Q matrix, and can specify alternate structures, but we have to follow rules
- This can cause problems in MARSS

$$Q = \begin{bmatrix} q_{1,1} & q_{1,2} & 0 & 0 \\ q_{2,1} & q_{2,2} & q_{2,3} & 0 \\ 0 & q_{3,2} & q_{3,3} & q_{3,4} \\ 0 & 0 & q_{4,3} & q_{4,4} \end{bmatrix}$$

- If (1,2) are correlated and (2,3) correlated, (1,3) have to be correlated

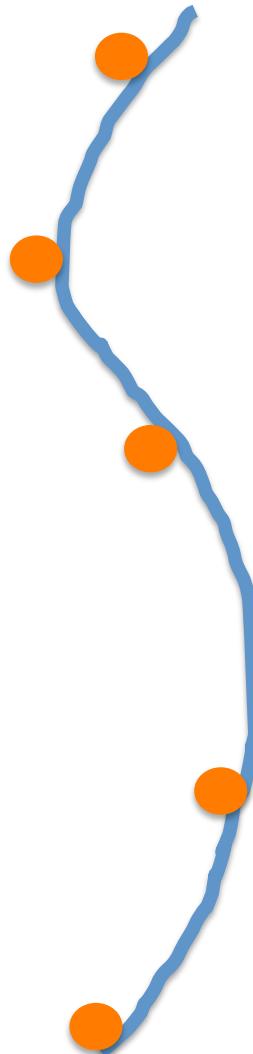
Multivariate approach with DFA

- Estimate how individual observed time series load onto ‘trends’



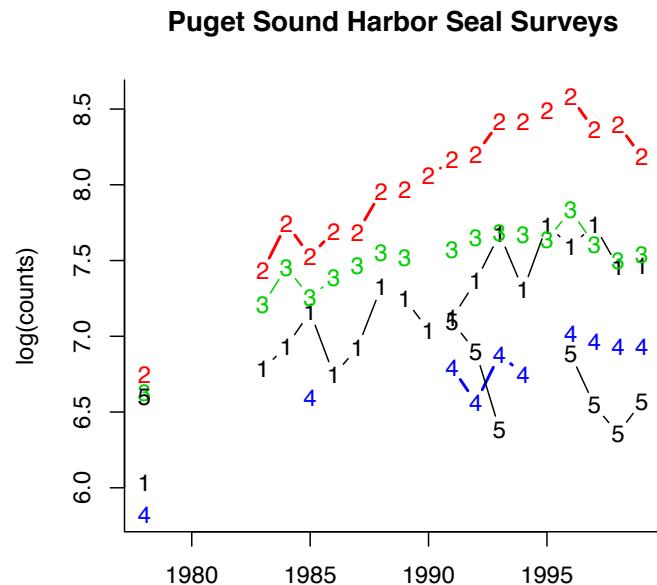
Potential problems with both the MARSS and DFA approach

- Sites separated by large distances may be grouped together
- Sites close to one another may be found to have very different dynamics



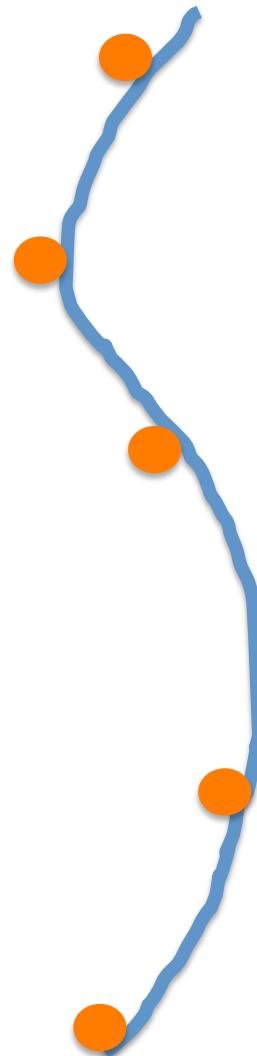
Are there biological mechanisms that may explain this?

- Puget Sound Chinook salmon
 - 21 populations generally cluster into 2-3 groups based on genetics
 - Historically large hatchery programs
- Hood canal harbor seals
 - Visited by killer whales



Motivation of explicitly including spatial structure

- Adjacent sites can be allowed to covary
- Estimated parameters greatly reduced to 2-5



Types of spatial data

- Point referenced data
 - Aka geostatistical data
 - Typically 2-D, but could be 1-D or 3-D (depth, altitude)
 - May be fixed station or random (e.g. trawl surveys)
- Point pattern data
 - Spatially referenced based on outcomes
 - Presence of a species
 - Inference focused on describing clustering (or not)
- Areal data
 - Locations occur in blocks
 - counties, management zones, etc.

Computationally convenient approaches

- CAR (conditionally autoregressive models)
 - Better suited for Bayesian methods
 - Goal of both is to write the distribution of a single prediction analytically in terms of the joint (y_1, y_2)
- SAR (simultaneous autoregressive models)
 - Better suited for ML methods
 - Simultaneously model distribution of predicted values
- ‘Autoregressive’ in the sense of spatial dependency / correlation between locations

CAR models (Besag 1991)

$$Y_i = BX_i + \phi_i i + \varepsilon_i$$

BX Regression coefficients

$\phi_i i$ Spatial component (aka Markov Random Field)

ε_i Residual error term

- Create spatial adjacency matrix W , based on neighbors, e.g.
- $W(i,j) = 1$ if neighbors, 0 otherwise
- W often row-normalized (rows sum to 1)

	1	2	3	4	5
1	0	1	0	0	0
2	1	0	1	0	0
3	0	1	0	1	0
4	0	0	1	0	1
5	0	0	0	1	0



- Besag et al. 1991

$$\phi_k | \phi_{-k} \sim N\left(\frac{\sum_{i=1}^n w_{ki} \phi_i}{\sum_{i=1}^n w_{ki}}, \frac{\sigma^2}{\sum_{i=1}^n w_{ki}}\right)$$

- Leroux et al. 1999

$$\phi_k | \phi_{-k} \sim N\left(\frac{\rho \sum_{i=1}^n w_{ki} \phi_i}{\rho \sum_{i=1}^n w_{ki} + 1 - \rho}, \frac{\sigma^2}{\rho \sum_{i=1}^n w_{ki} + 1 - \rho}\right)$$

- In matrix form $y \sim N(0, (I - \rho W)^{-1} \tilde{D})$

$$\tilde{D}_{ii} = \sigma_i^2$$

- Implemented in R: **spdep, CARBayes**

SAR models

- Simultaneous autoregressive model

$$y \sim N\left(0, (I - \rho W)^{-1} \tilde{D} (I - \rho W)^{!-1}\right)$$

$$\tilde{D}_{ii} = \sigma_i^2$$

- Remember that the CAR was

$$y \sim N\left(0, (I - \rho W)^{-1} \tilde{D}\right)$$

$$\tilde{D}_{ii} = \sigma_i^2$$

Commonalities of both approaches

- Adjacency matrix W can also instead be modified to include distance
- Models spatial dependency as a function of single parameter ρ
- Models don't include time dimension in spatial field
 - One field estimated for all time steps

Problems with these approaches

- Wall (2004) “*A close look at the spatial structure implied by the CAR and SAR models*”

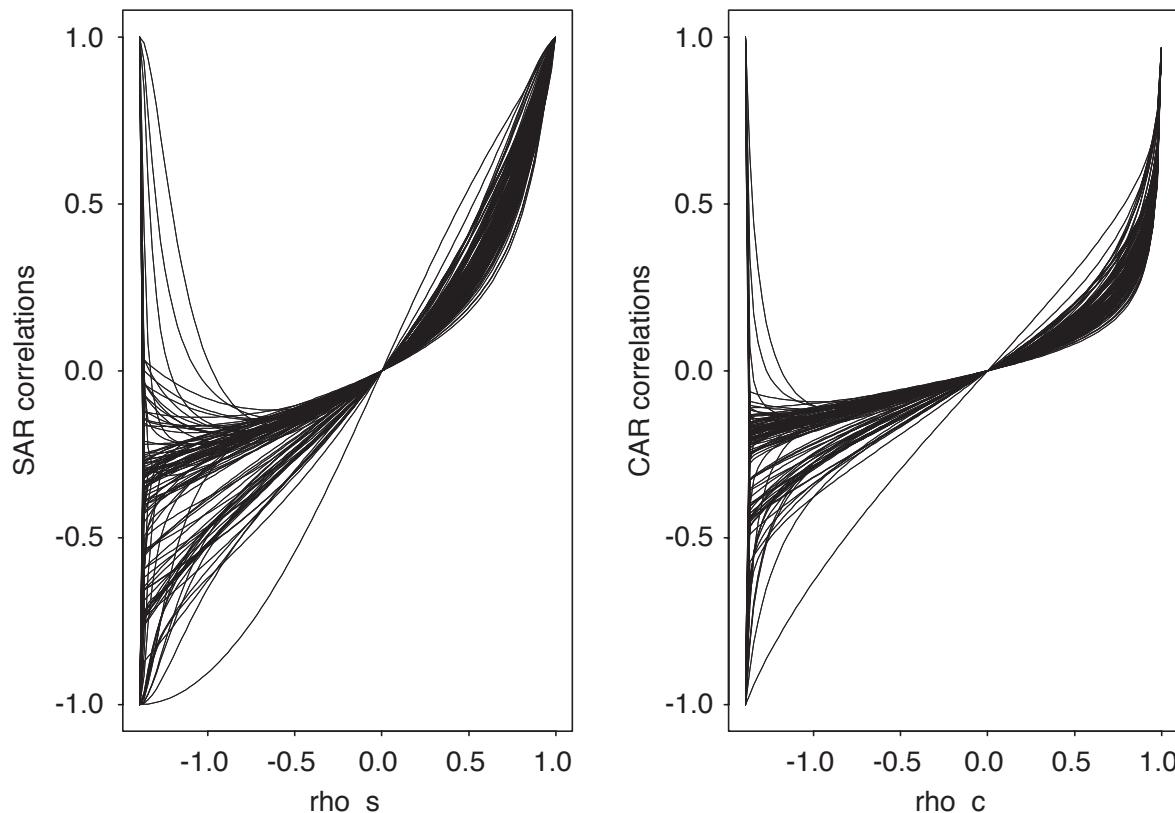


Fig. 5. Lines in both plots represent the implied model correlations between first order neighbors in the US lattice based on the SAR model (left) and CAR model (right) as functions of the respective parameters ρ_s and ρ_c .

Alternative to CAR & SAR: model elements of Q as functions

- Create matrix D, as pairwise distances
- This can be 1-D, or any dimension
 - We'll use Euclidian distances for 2-D

$$D = \begin{bmatrix} 0 & |x_1 - x_2| & |x_1 - x_3| & |x_1 - x_4| \\ |x_1 - x_2| & 0 & |x_2 - x_3| & |x_2 - x_4| \\ |x_1 - x_3| & |x_1 - x_3| & 0 & |x_3 - x_4| \\ |x_1 - x_4| & |x_1 - x_4| & |x_3 - x_4| & 0 \end{bmatrix}$$



Next specify spatial covariance function (aka “kernel”)

- Exponential

$$Cov(x_1, x_2) = \sigma^2 \exp(-d_{1,2} / \theta) \quad Cov(x_1, x_2) = \sigma^2 \exp(-\theta \cdot d_{1,2})$$

- Squared exponential*

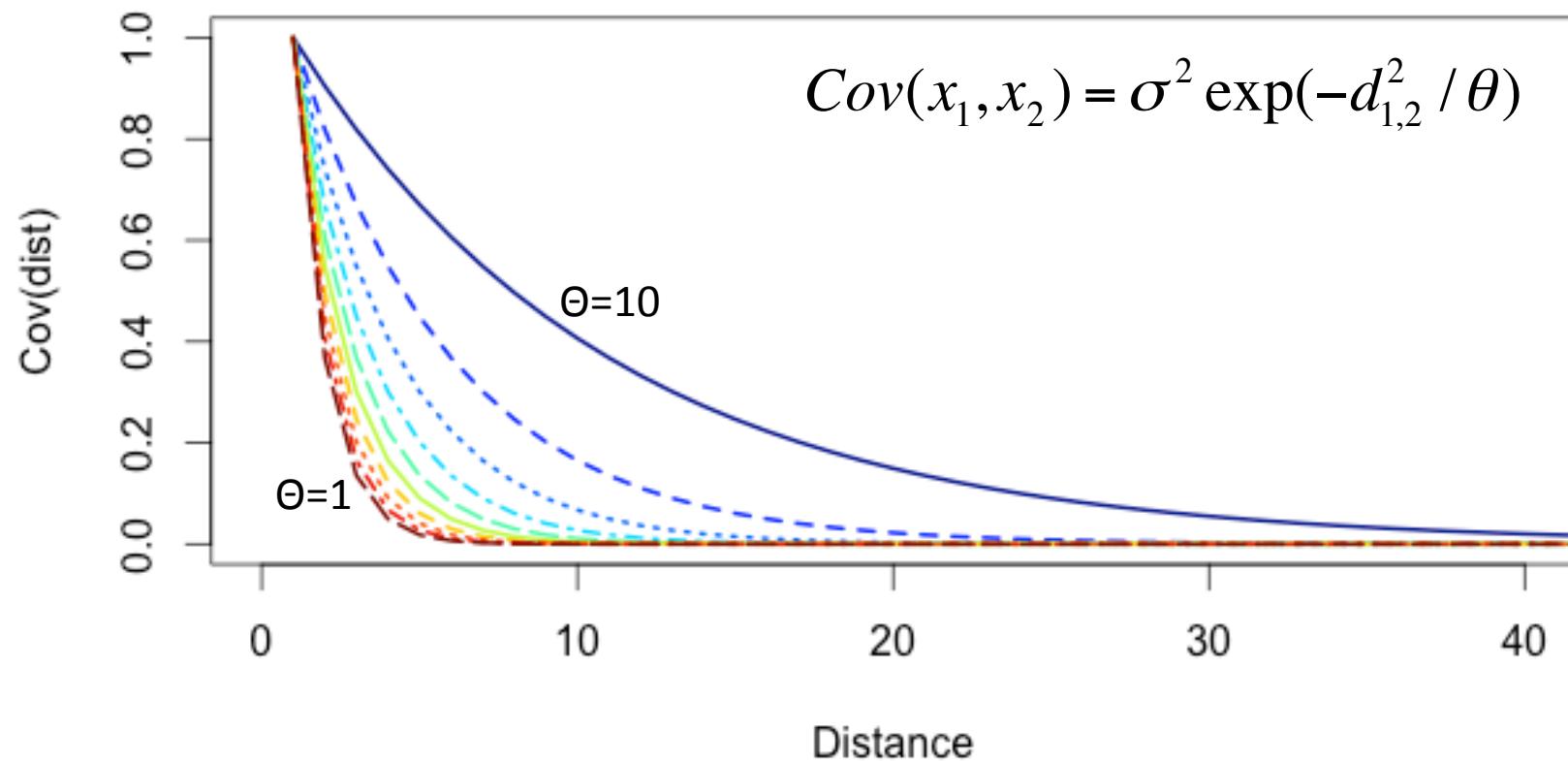
$$Cov(x_1, x_2) = \sigma^2 \exp(-d_{1,2}^2 / \theta)$$

- Matern family

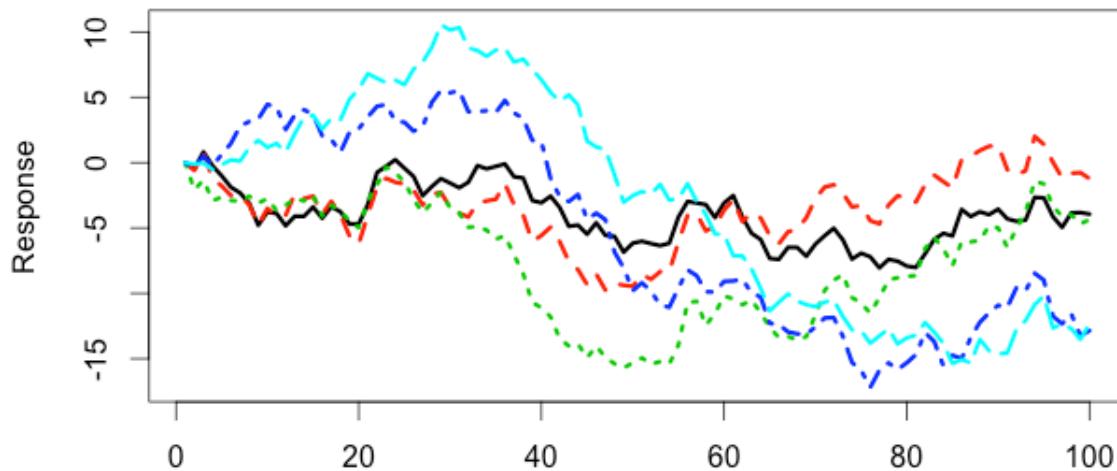
$$Cov(x_1, x_2) = \sigma^2 \frac{1}{\Gamma(v) 2^{v-1}} \left(\sqrt{2v} \frac{d_{1,2}}{\rho} \right) K_v \left(\sqrt{2v} \frac{d_{1,2}}{\rho} \right)$$

*Also referred to as radial basis function, or Gaussian

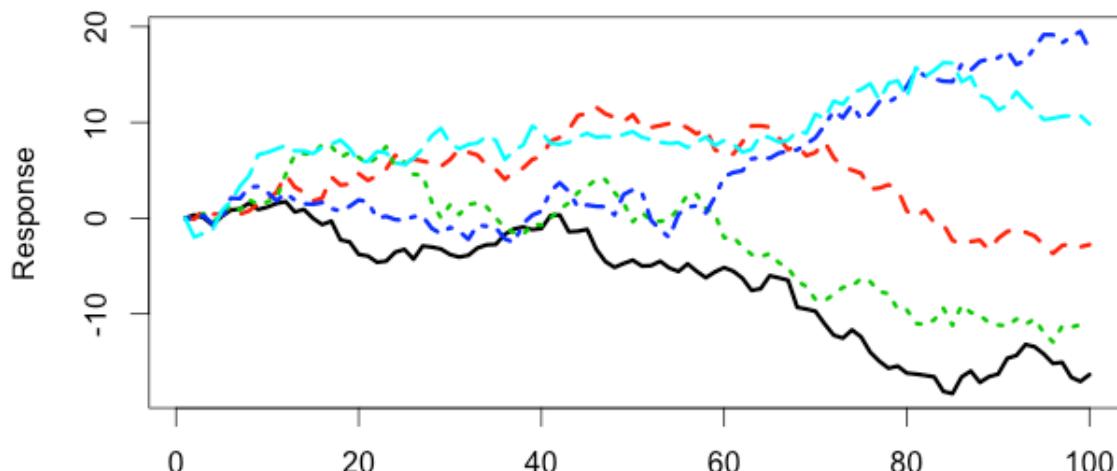
Squared exponential



Theta = 10

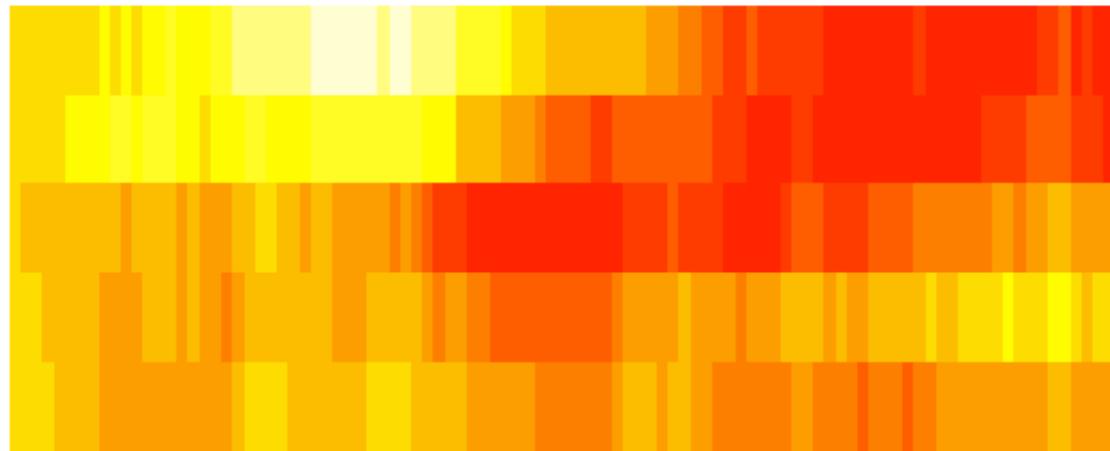


Theta = 1



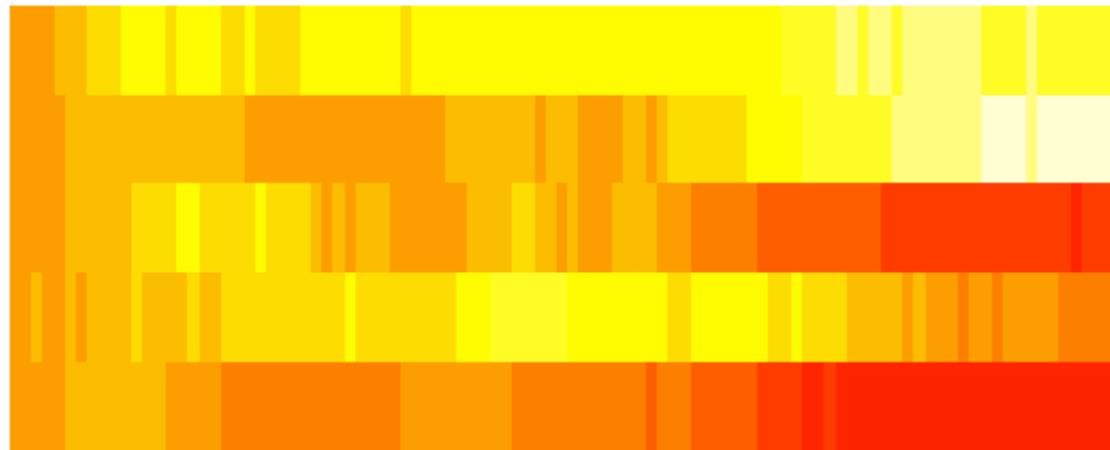
Theta = 10

Response



Theta = 1

Response



Considerations for time series models

- Should spatial dependency be included
- How to model it
 - Constant
 - Time varying
 - Autoregressive
 - Random walk
 - Independent variation?

Model-based geostatistical approaches

- 1. Generalized least squares
- 2. Bayesian methods in spBayes
- 3. INLA models
- 4. Spatial GAMs

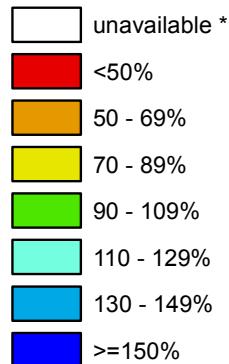
Method 1: using gls()

- Generalized least squares function
 - Similar syntax to lm, glm, etc.
- Flexible correlation structures
 - corExp()
 - corGaus()
 - corLin()
 - corSpher()
- Allows irregularly spaced data / NAs
 - unlike Arima(), auto.arima(), etc.

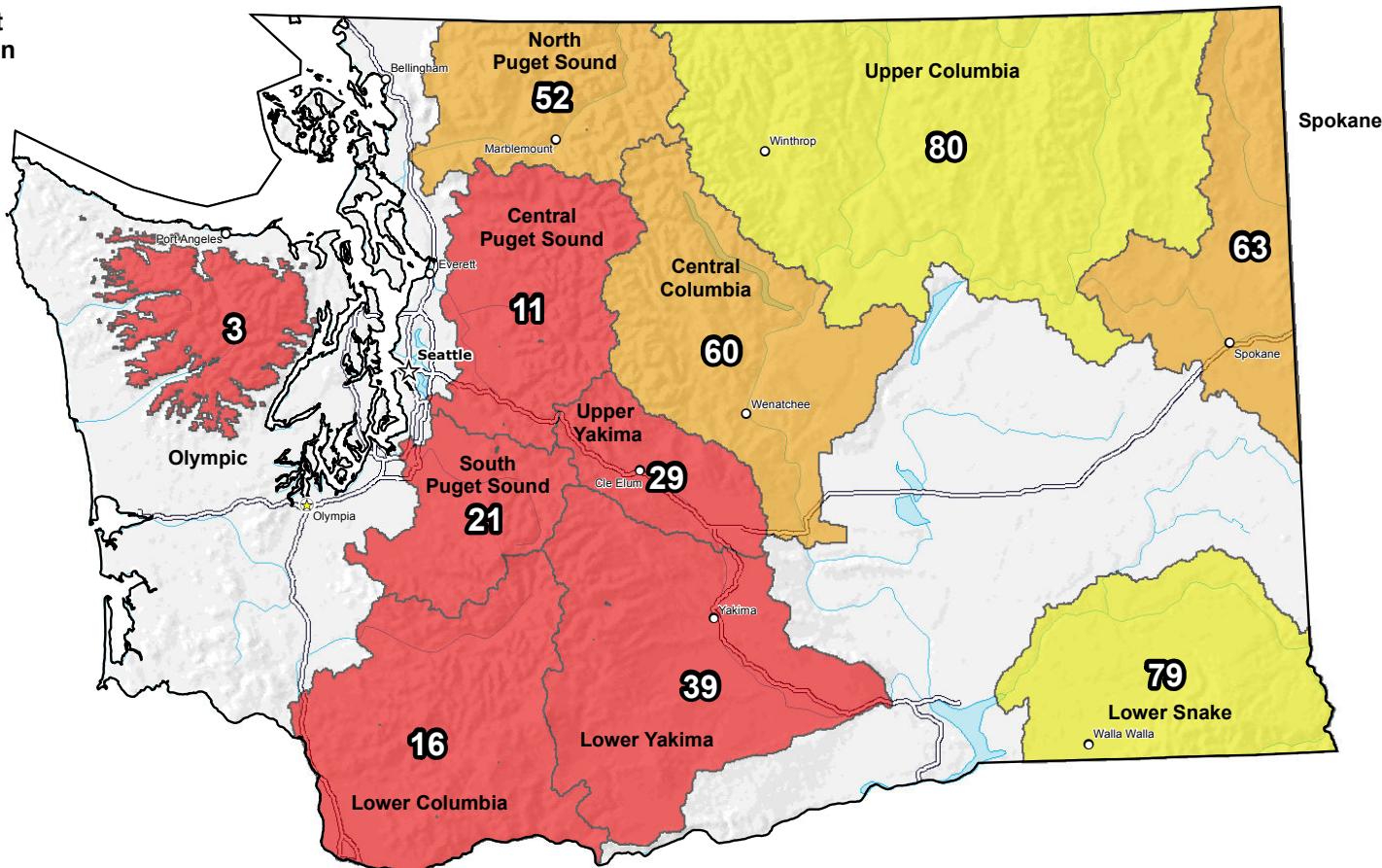
Washington SNOTEL Current Snow Water Equivalent (SWE) % of Normal

Feb 16, 2015

Current Snow Water
Equivalent (SWE)
Basin-wide Percent
of 1981-2010 Median



* Data unavailable at time
of posting or measurement
is not representative at this
time of year



0 10 20 40 60 80 100 Miles

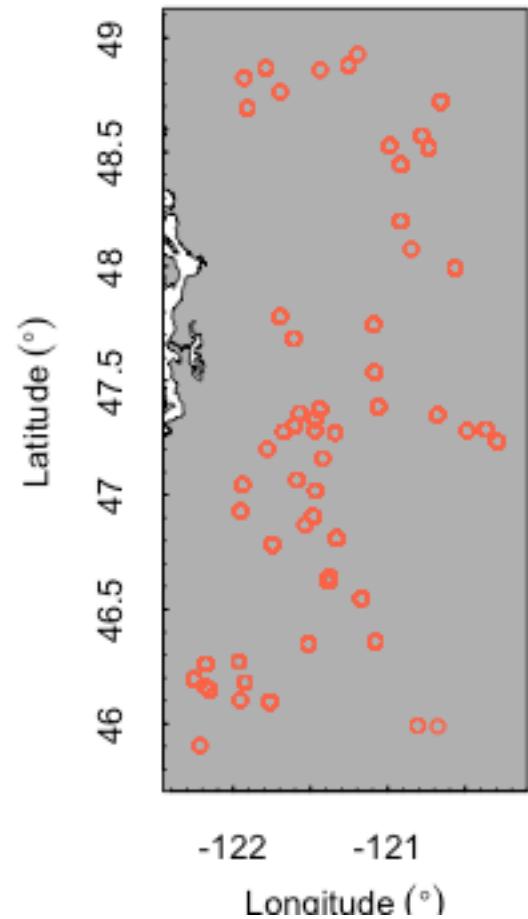
Provisional Data
Subject to Revision



The snow water equivalent percent of normal represents the current snow water equivalent found at selected SNOTEL sites in or near the basin compared to the average value for those sites on this day. Data based on the first reading of the day (typically 00:00).

Prepared by:
USDA/NRCS National Water and Climate Center
Portland, Oregon
<http://www.wcc.nrcs.usda.gov>

- We'll use Snow Water Equivalent (SWE) data in Washington state
- 70 SNOTEL sites
 - We'll focus only on Cascades
- 1981-2013
- Initially start using just the February SWE data
- 1518 data points (only 29 missing values!)



We can use AIC to evaluate different correlation structures

```
mod.exp = gls(Feb ~ elev, correlation =  
corExp(form=~lat+lon,nugget=T), data = y[which(is.na(y  
$Feb)==F & y$Water.Year==2013),])  
AIC(mod.exp) = 431.097
```

```
mod.gaus = gls(Feb ~ elev, correlation =  
corGaus(form=~lat+lon,nugget=T), data =  
y[which(is.na(y$Feb)==F & y$Water.Year==2013),])  
AIC(mod.gaus) = 433.485
```

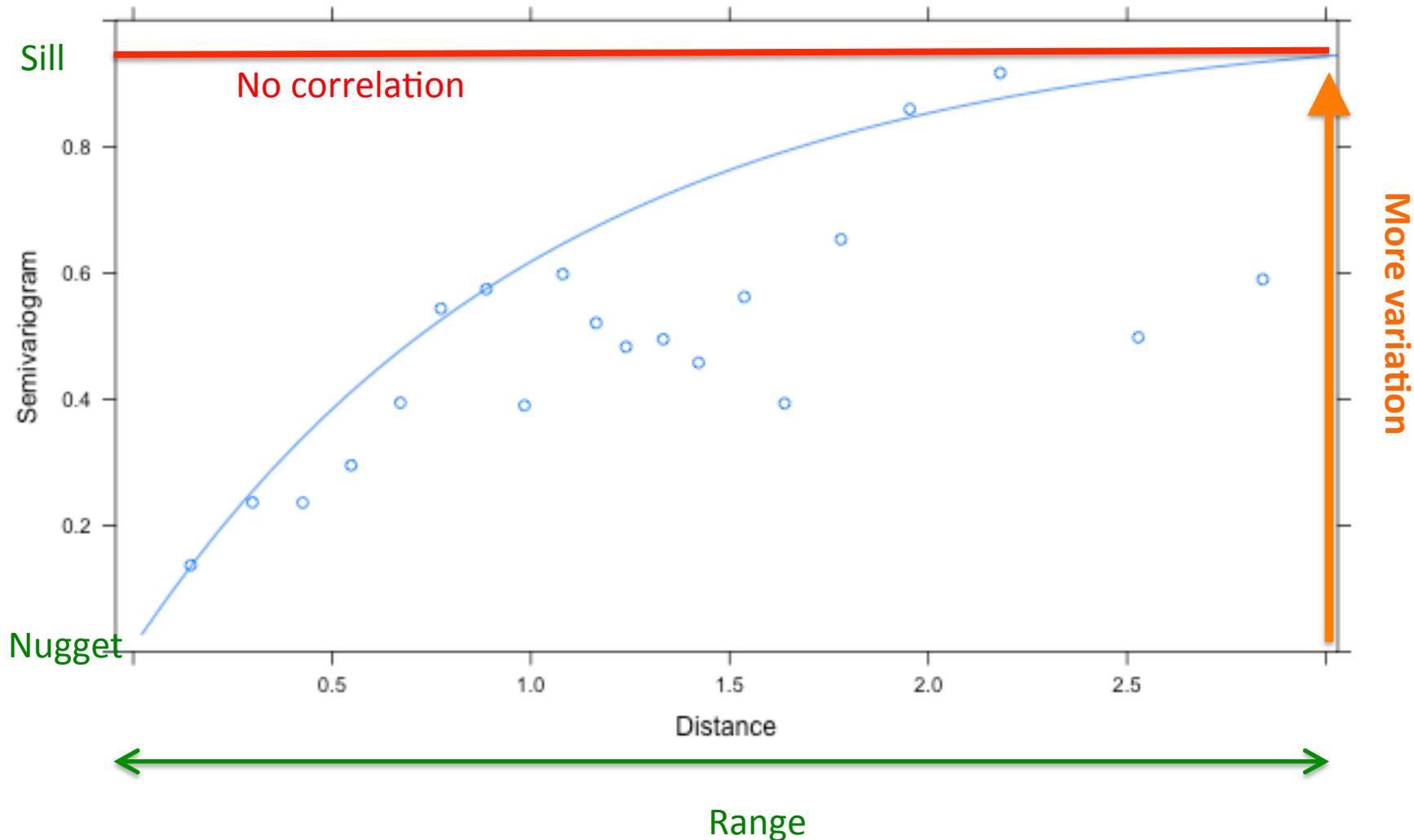
Diagnostics: fitting variograms

```
var.exp <- Variogram(mod.exp, form =~ lat+lon)  
plot(var.exp,main="Exponential",ylim=c(0,1))
```

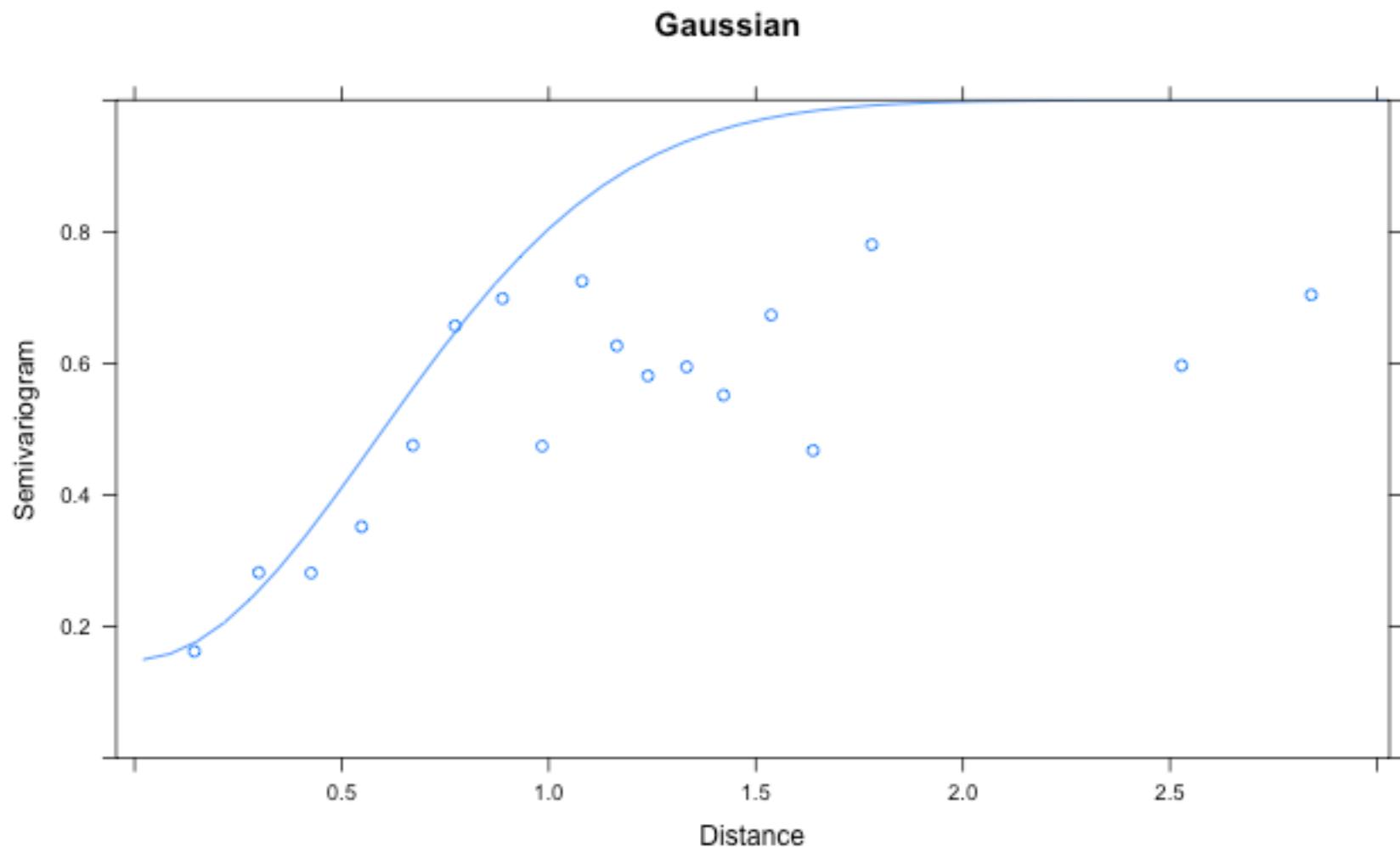
```
var.gaus <- Variogram(mod.gaus, form =~ lat+lon)  
plot(var.gaus,main="Gaussian",ylim=c(0,1))
```

Exponential variogram

Exponential



Gaussian variogram



Extensions of these models

- corExp and corGaus spatial structure useful for wide variety of models / R packages

Linear/non-linear mixed effect models

- lme() / nlme() in nlme package

Generalized linear mixed models

- glmmPQL() in MASS package

Generalized additive mixed models

- gamm() in mgcv package

Method 2: spBayes

- Slightly more complicated syntax:
- Specify:
 - a. Priors on parameters
 - b. Tuning parameters for Metropolis sampling (jumping variance)
 - c. Starting / initial values
 - d. Covariance structure (“exponential”, “gaussian”, “matern”, etc)
 - e. Number of MCMC iterations / burn-in, etc.

Example with WA SWE data

```
# This syntax is dependent on model parameters. See vignette
```

```
priors <- list("beta.Norm"=list(rep(0,p), diag(1000,p)), "phi.Unif"=c(3/1,  
3/0.1), "sigma.sq.IG"=c(2, 2), "tau.sq.IG"=c(2, 0.1))
```

```
# Phi is spatial scale parameter, sigma.sq is spatial variance, tau.sq = residual  
starting <- list("phi"=3/0.5, "sigma.sq"=50, "tau.sq"=1)
```

```
# variance of normal proposals for Metropolis algorithm
```

```
tuning <- list("phi"=0.1, "sigma.sq"=0.1, "tau.sq"=0.1)
```

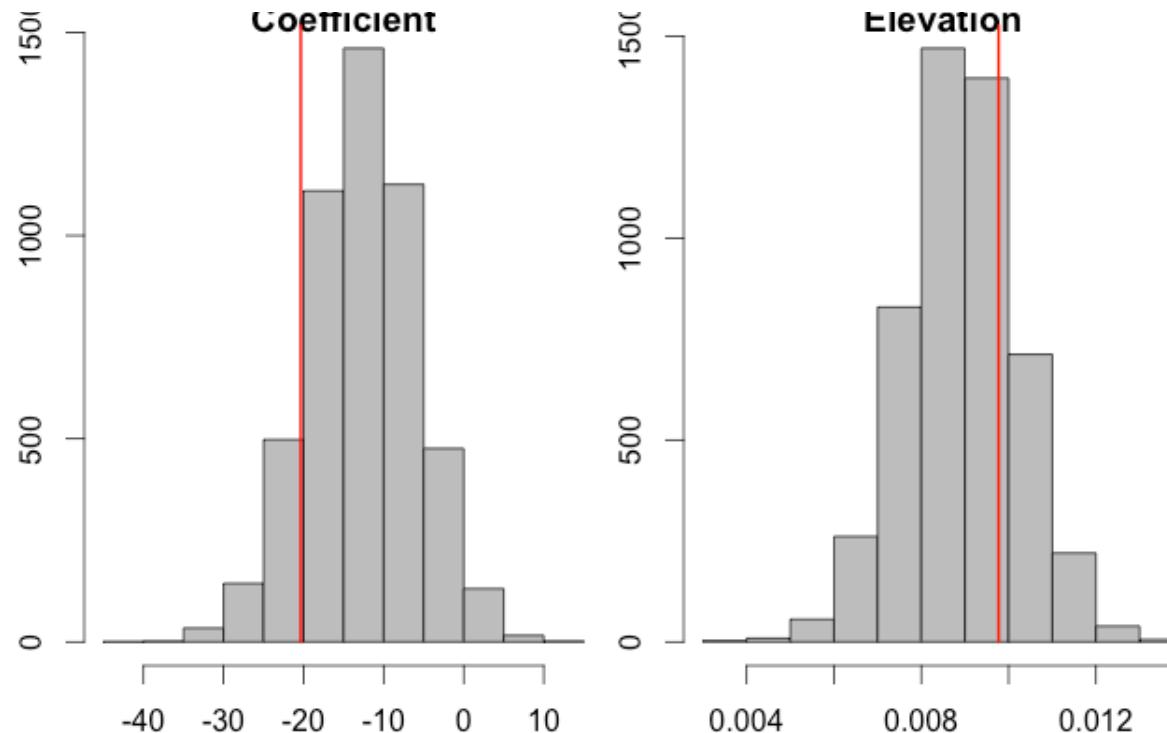
```
m.1 <- spLM(y~X-1, coords=cords,n.samples=10000, cov.model =  
"exponential", priors=priors, tuning=tuning, starting=starting)
```

Coefficients need to be extracted

```
##recover beta and spatial random effects
```

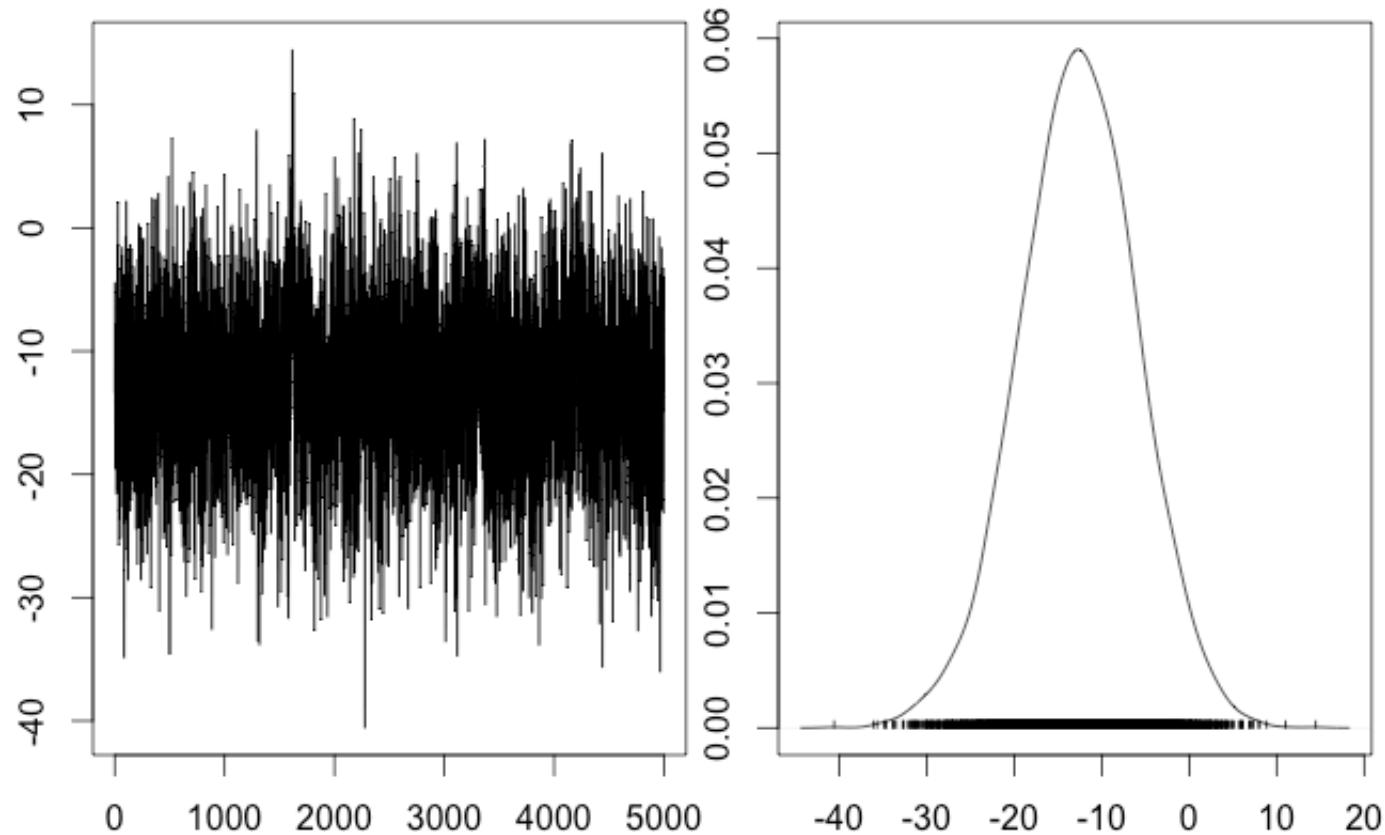
```
burn.in <- 5000
```

```
m.1 <- spRecover(m.1, start=burn.in, verbose=FALSE)
```



Standard MCMC diagnostics

- Output is of class ‘mcmc’



Other spatial models in spBayes

Non-Gaussian errors

spGLM

Multivariate

spMvLM()

spMvGLM()

Dynamic linear model (DLM)

spDynLM

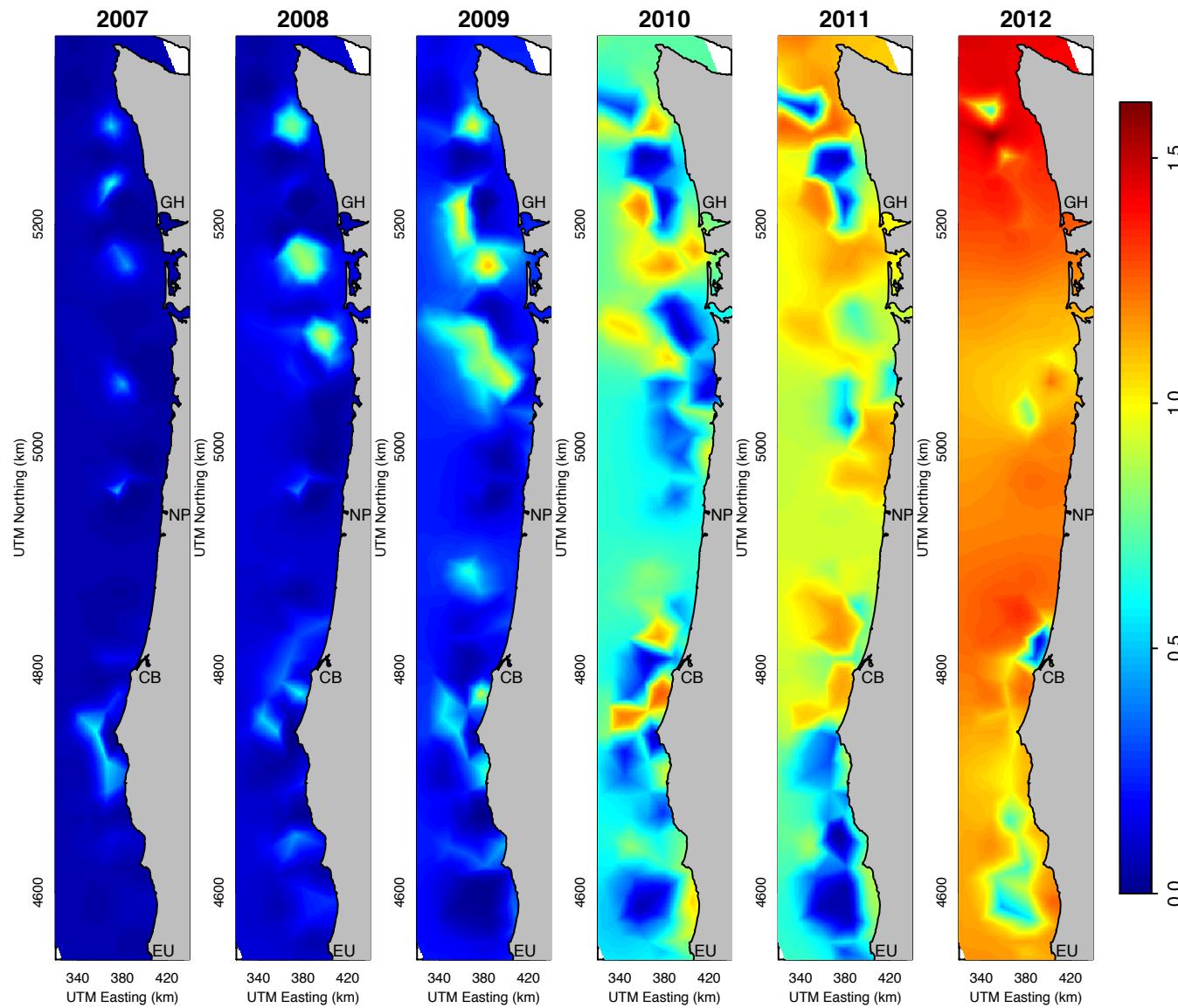
Method 3: models in INLA

- INLA = Integrated Nested Laplace Approximation
- R-INLA software = interface between R and the GMRFLib C library, etc.
- Get it here:
<http://www.r-inla.org/>

Motivation of INLA

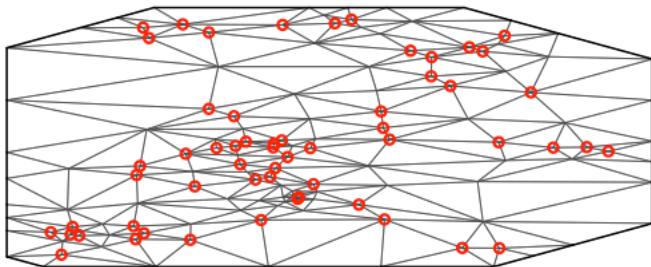
- Some problems contain simple spatial structure
 - e.g. the harbor seal data in MARSS() with only 5 – 7 time series
- Others are much more complex
 - WA SWE data
 - Fisheries survey data (1000s of points)
- Including time-varying spatial fields becomes very computationally difficult
- Doing all of the above in a Bayesian setting can be prohibitive, but we can use Laplace approximation

Varying spatial fields example: eulachon-shrimp bycatch

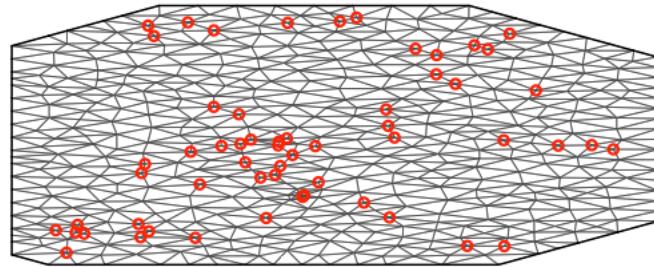


INLA's approximation: SNOTEL data

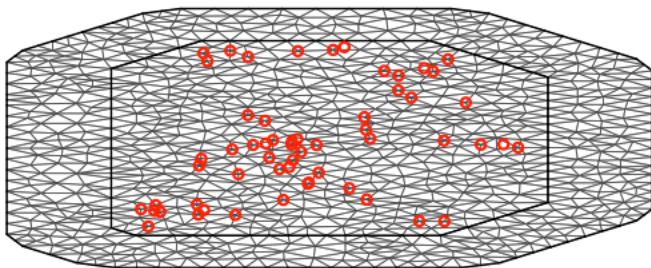
Constrained refined Delaunay triangulation



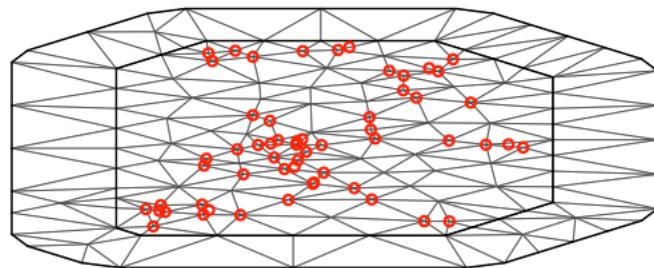
Constrained refined Delaunay triangulation



Constrained refined Delaunay triangulation



Constrained refined Delaunay triangulation



How many points fall on vertices? Is the boundary area large enough? Choosing this must be done very carefully!

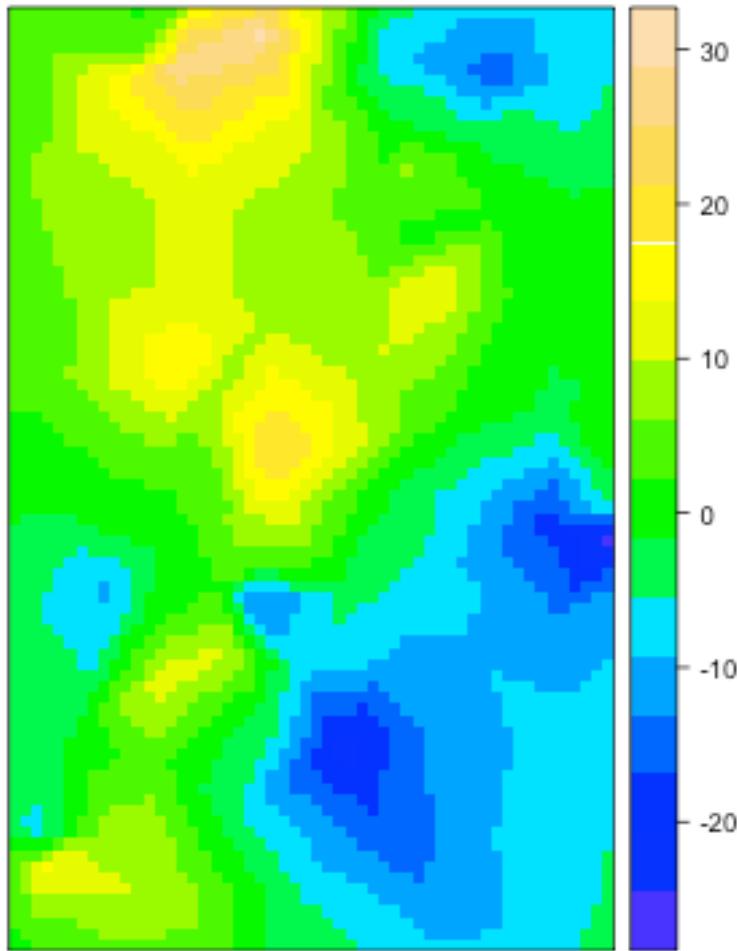
Estimation done via maximum likelihood

- Estimates seems similar to those from gls() and spBayes()
- Year included as numeric here (not significant)

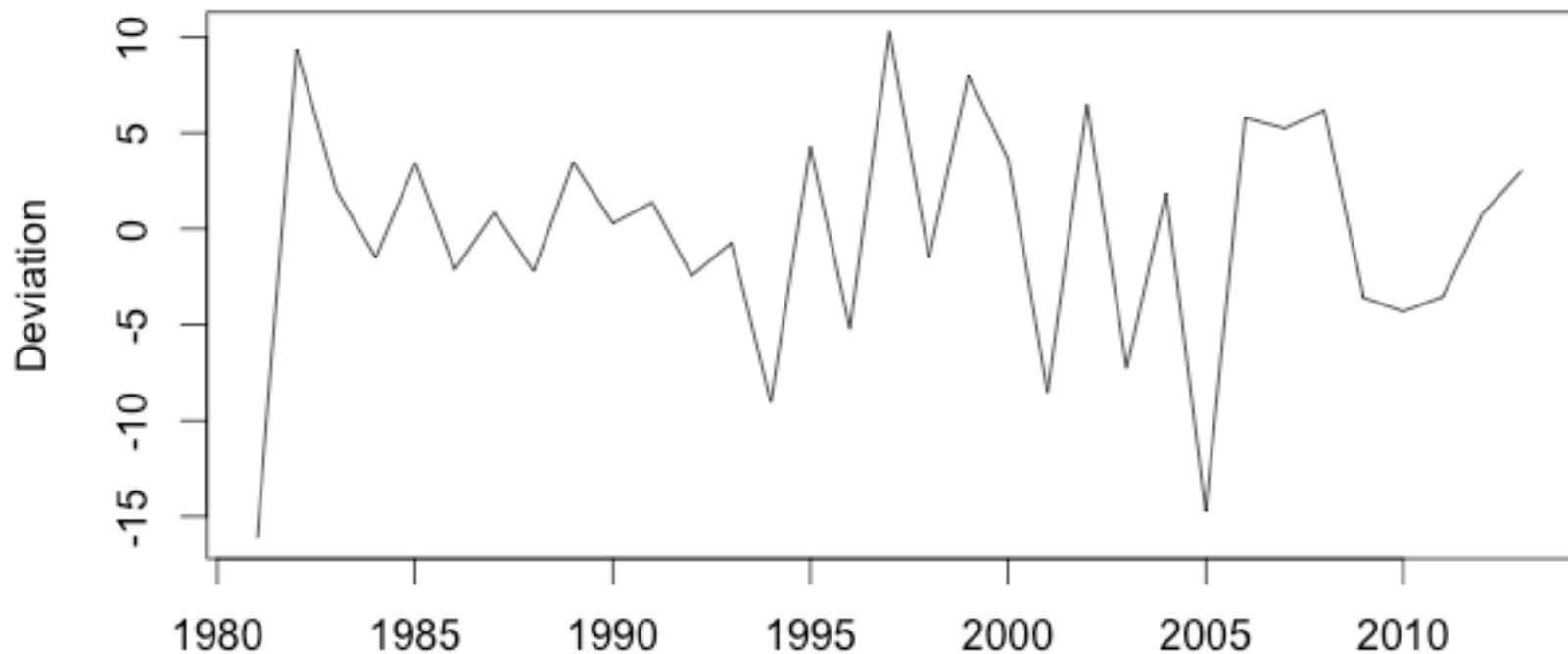
```
> res2pos$summary.fixed
      mean        sd   0.025quant   0.5quant   0.975quant       mode       kld
Intercept -14.576549539 2.788559e+01 -69.319288831 -14.579598755 40.13298938 -14.583354523 2.924022e-16
elev        0.010464086 8.931255e-04   0.008740241   0.010453876  0.01224473   0.010432282 6.838504e-13
year       -0.004199109 1.409759e-02  -0.031884876  -0.004197175  0.02345057  -0.004192103 5.540561e-16
> |
```

- Alternatively, we can include year in spatial field
- Year can also be included as factor

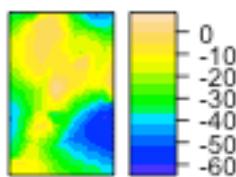
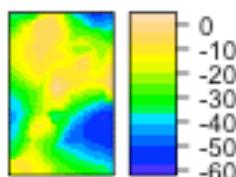
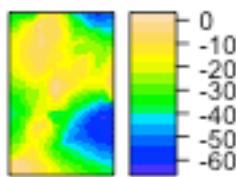
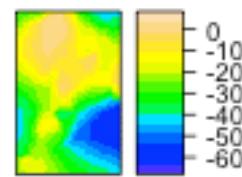
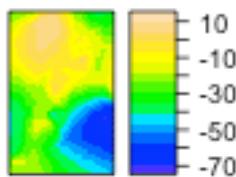
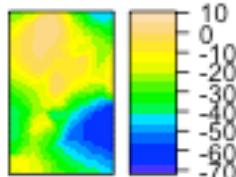
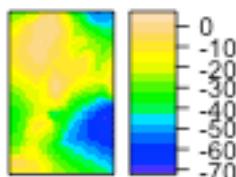
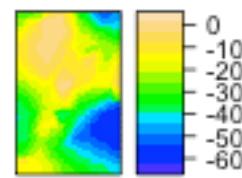
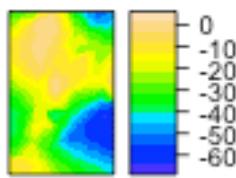
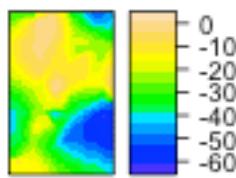
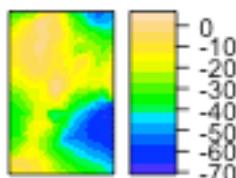
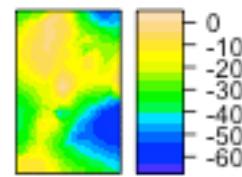
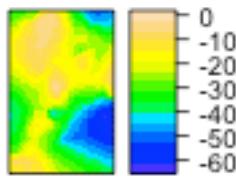
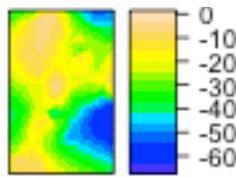
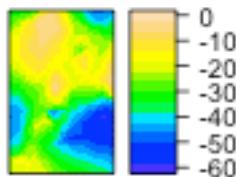
Projecting INLA estimates to surface



Fixed effect (factor) estimates



Spatial SWE fields by year



Additional output / diagnostics

- Plots of observed v fitted values
- Prediction to new locations
- Samples from posterior distribution via Laplace approximation

Approach 4: Spatial GAMs

- Previous approaches modeled errors as correlated
- Spatial GAMs generally model mean as spatially correlated
- Somewhat analogous to AR vs MA models

Examples of 2D GAMs

```
g = gam(y ~ s(lat, lon)) # basis spline
```

```
g = gam(y ~ te(lat, lon)) # tensor product spline
```

```
g = gam(y ~ te(lat, lon, by = as.factor(year)) #  
smooths by year
```

Examples of recent NOAA work

Ward, E.J., G.R. Pess, K. Anlauf-Dunn, and C.E. Jordan. 2012. Applying time series models with spatial correlation to identify the scale of variation in habitat metrics related to threatened coho salmon (*Oncorhynchus kisutch*) in the Pacific Northwest. Canadian Journal of Fisheries and Aquatic Sciences, 69(11): 1773-1782

Shelton, A.O., J.T. Thorson, E.J. Ward, and B.E. Feist. 2014. Spatial, semi-parametric models improve estimates of species abundance and distribution. Canadian Journal of Fisheries and Aquatic Sciences, 71: 1655-1666

Thorson, J.T., H. Skaug, K. Kristensen, A.O. Shelton, E.J. Ward, J. Harms, and J. Benante. 2015. The importance of spatial models for estimating the type and strength of density dependence. Ecology, In press

Ward, E.J., K.N. Marshall, A. Sedgley, T. Ross, T. Hass, S.F. Pearson, G. Joyce, N.J. Hamel, P.J. Hodum, R. Faucett. 2015. Using citizen-science data to identify local hotspots of seabird occurrence. PeerJ 2:e704

Ward, E.J., J.H. Anderson, T.J. Beechie, G.R. Pess, and M.J. Ford. 2015. Increasing hydrologic variability threatens depleted anadromous fish populations. *In press, Global Change Biology*