Geostatistical data

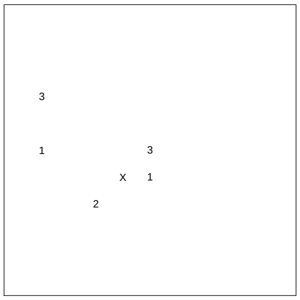
- Quantity measured at a location
  - Assumed characteristic of that location, not a large area
  - Examples:
    elevation, annual rainfall, surface soil pH, O_2 concentration at 3m
- Notation:
  - $\boldsymbol{s}$: location, a vector value.
    - Usually $\boldsymbol{s} = (x, y)$ in some coordinate frame (e.g., longlat or UTM)
    - Written as a vector because details of 1D (beach, line), 2D (earth surface), 3D (ocean, soil, atmosphere) not important
  - $Z(\boldsymbol{s})$: the characteristic at location $\boldsymbol{s}$.
- Geostatistical data
  - $Z(\boldsymbol{s})$ exists everywhere within boundary of study area
  - Generally, no sharp changes (jumps) in $Z(\boldsymbol{s})$
  - $Z(\boldsymbol{s}_1)$ probably different from $Z(\boldsymbol{s}_2)$, but transition is smooth

Many possible goals

- Predict $Z(\boldsymbol{s})$ at unmeasured locations
- Describe spatial pattern in $Z(\boldsymbol{s})$
  - How similar is $Z(\boldsymbol{s})$ to neighboring values?
  - How does that change with distance to neighbor?
- Model relationship between $Z(\boldsymbol{s})$ and covariates

Prediction

- Could be done to fill in a grid so can draw map or use image/contour plot
- Or, done because you need predictions at unmeasured points



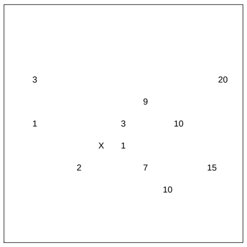
- What is Z at the location marked by X?

Prediction

- One possibility: simple average of Z over entire region
  - Very common in non-spatial contexts
- 1st law of geography (Tobler): everything is related to everything else more closely related to nearby things
- This principle is very important if there is a spatial trend (variation across the region) or some form of spatial pattern.
  - simple average ignores spatial trend and pattern

Prediction

- What if had a bit more data:



- Overall average for region clearly inappropriate
- Consider some form of local average
- We will discuss 3 methods:
  - Inverse distance weighting
  - Spatial trend model
  - Kriging: we'll spend most time / effort on this

Inverse distance weighting

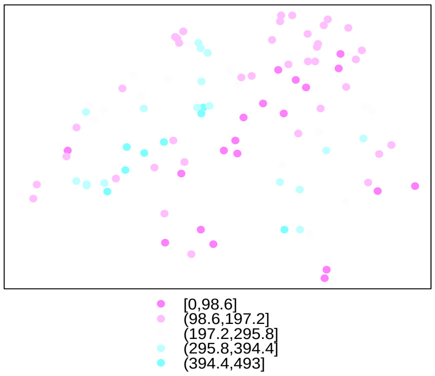
- Concept:
  - Prediction an average that emphasizes nearby values
  - Done by weighting all observations
  - Higher weight to nearby observations
- Notation: $\boldsymbol{s}_i$ is location of i'th observation
  d_{ij} is distance between location i and location j
  $Z(\boldsymbol{s}_i)$ is value of Z at location $\boldsymbol{s}_i$
  $\hat{Z}(\boldsymbol{s}_i)$ is prediction of Z at location $\boldsymbol{s}_i$

$$\begin{aligned}\hat{Z}(\boldsymbol{s}_j) &= \frac{\Sigma_i w_{ij}\, Z(\boldsymbol{s}_i)}{\Sigma_i w_{ij}}, \text{where} \\ w_{ij} &= \frac{1}{d_{ij}^a}\end{aligned}$$

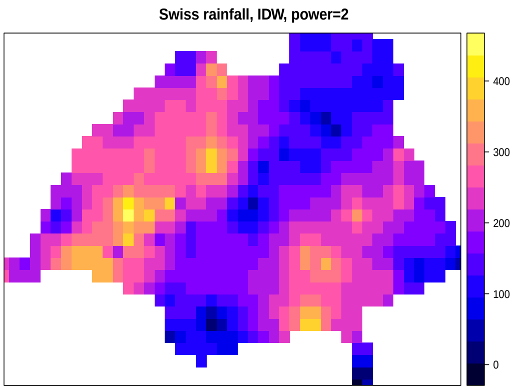
Inverse distance weighting

- a is an arbitrary parameter, commonly 1 or 2
  - $a = 0$ gives you the simple average over entire region (all weights $= 1$)
  - larger values → "more local" estimate, because emphasize shorter distances
  - if $d_{ij} = 0$, i.e. predicting at an observed location, use observed value
- Characteristics:
  - w_{ij} always ≥ 0 and w_{ij}/sum always ≤ 1
  - Sometimes set small values of w_{ij} to 0
  - $\hat{Z}(\boldsymbol{s}_j)$ always within range of observed values
    Some like this; other's don't.
- Problem: have to choose a. Some approaches:
  - Ad hoc (you like the resulting picture),
  - or tradition (your field always uses 2 or 1.5 or ??)
- demonstrate role of a by comparing results for $a = 2$ and $a = 1$

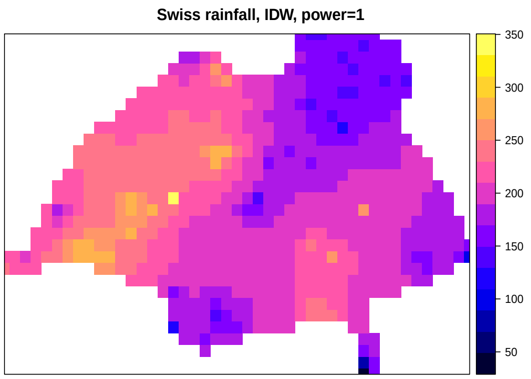
Swiss rainfall data



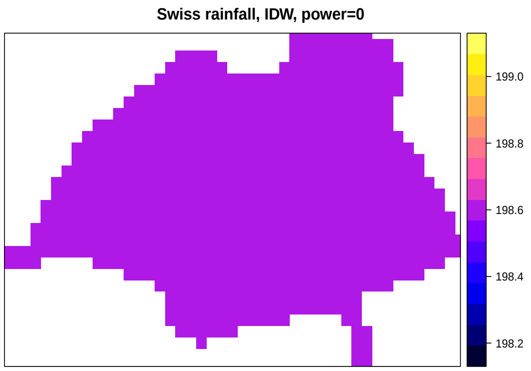
Swiss rain, IDW, power=2



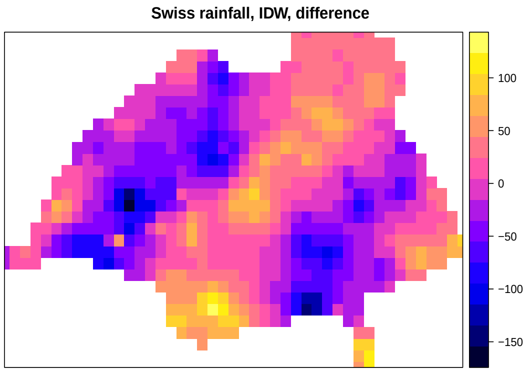
Swiss rain, IDW, power=1



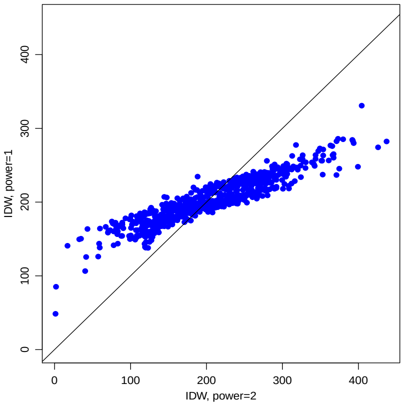
Swiss rain, IDW, power=0



Swiss rain, IDW, difference



Swiss rain, IDW, comparison



Spatial trend surface

- Assume some function of X and Y coordinates fits the data
- often a low order polynomial (linear or quadratic)

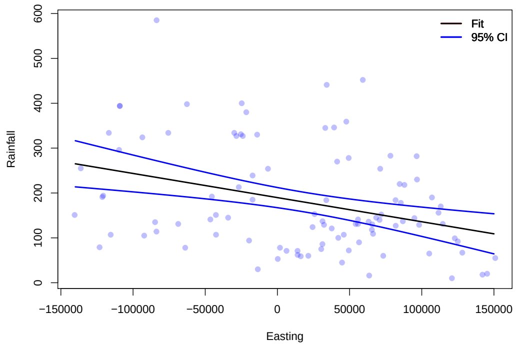
$$\begin{aligned} Z(\boldsymbol{s}_i) &= \beta_0 + \beta_1 X_i + \beta_2 Y_i + \varepsilon_i, \text{ or} \\ Z(\boldsymbol{s}_i) &= \beta_0 + \beta_1 X_i + \beta_2 Y_i + \beta_3 X_i^2 + \beta_4 Y_i^2 + \beta_5 X_i Y_i + \varepsilon_i, \end{aligned}$$

- only used for predicting $\hat{Z}(\boldsymbol{s}_i)$
- not doing any test or inference, so don't worry about correlation in ε's
- accounting for correlation, i.e. using GLS, will give better estimates of $\hat{\beta}$'s.
- Potential advantages over inverse distance weighting:
  - Can estimate Var ε
  - Can construct prediction intervals for $\hat{Z}(\boldsymbol{s}_i)$:
  - $\hat{Z}(\boldsymbol{s}_i) \pm T_{1-\alpha/2} \sqrt{s^2 + \text{Var } \hat{Z}(\boldsymbol{s}_i)}$

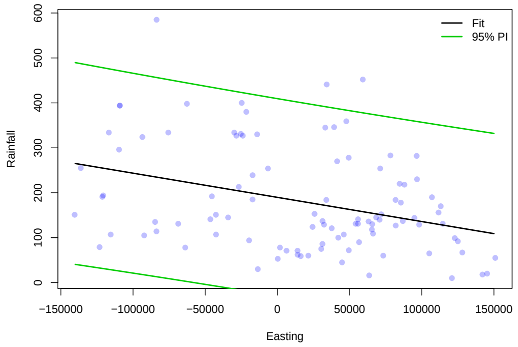
Spatial trend surface

- Confidence and Prediction intervals
- Reminder: can interpret a fitted regression line two ways
  - Predict average Y at some new X
    - Uncertainty only in regr. coefficients ("the line")
    - If I collected a second set of n obs, how similar is $\hat{Y}$?
    - Decreases as n increases
    - se of a mean, confidence interval for $\hat{Y}$
  - Predict a new observation at some new X
    - Uncertainty in both the line and obs around the line
    - Before sampling a new obs, how accurate is prediction?
    - Usually, very similar to s (rMSE), never smaller
    - sd of a predicted observation, prediction interval for $\hat{Y}$

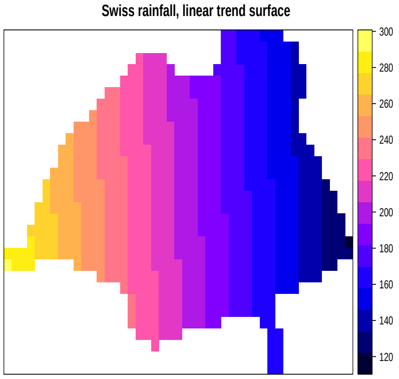
Confidence interval: Swiss rain



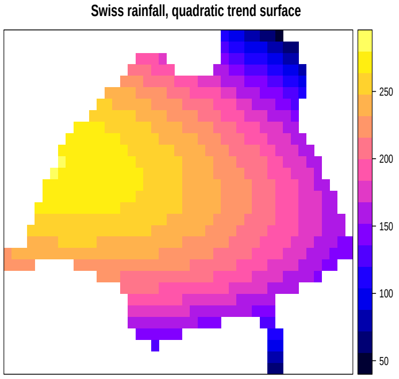
Prediction interval: Swiss rain



Swiss rain, linear TS

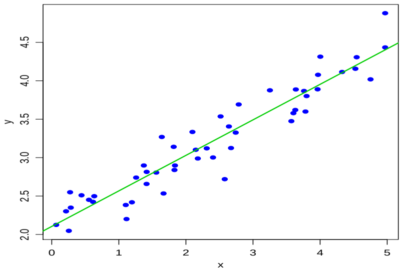


Swiss rain, quadratic TS



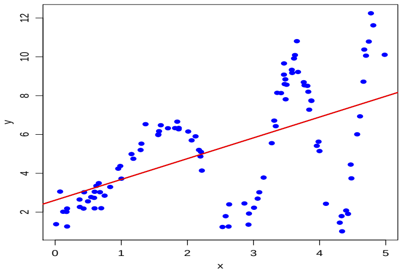
Splines: more flexible regression functions

- Concepts only. Details require a lot of intricate math and stat theory
- Consider response Y and one predictor X
- Sometimes a simple model is great



Splines: more flexible regression functions

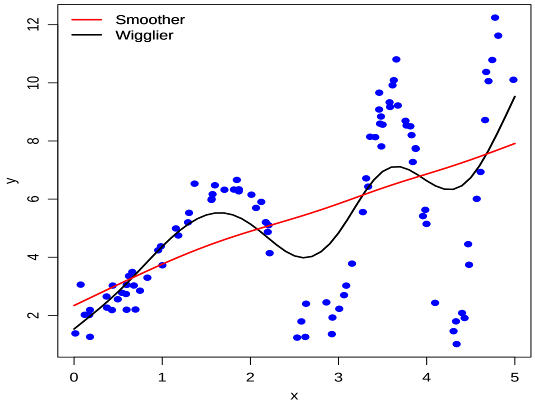
- And sometimes not



Splines: more flexible regression functions

- How to model relationship between Y and X, especially to predict?
  - If subject-matter-based model, use it!
  - Fit a higher order polynomial (quadratic, cubic)
  - Non-parametric regression: smooth the data
- Various NP regression methods. Focus on smoothing splines
- Concept: put together many models for small pieces of the data
- Need to choose number of small pieces
  - fewer pieces: smoother curve, closer to linear
  - more pieces: wigglier curve, closer to data

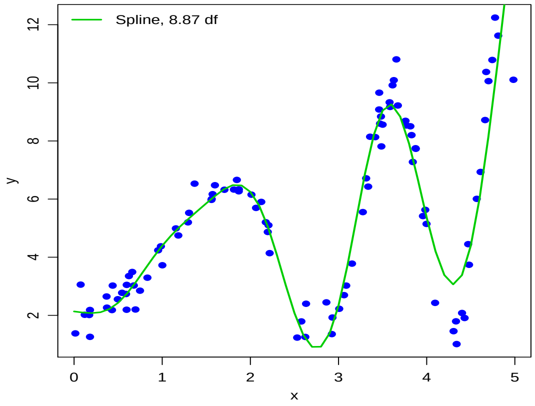
Spline fits



Spline fits

- How to choose how smooth/wiggly?
- not easy
- Too smooth is obviously bad
- Extemely wiggly effectively connects the dots
  - also bad - predictions of new obs. are inaccurate
  - overfitting the observed data
  - treating "noise" as signal.
- One common solution: cross validation
  - leave out an obs, fit a model, predict left obs.
  - put back, leave out next obs, $\cdots$
  - right choice is the value that makes good preds. of all the left-out obs
- splines require more data than when you know the model

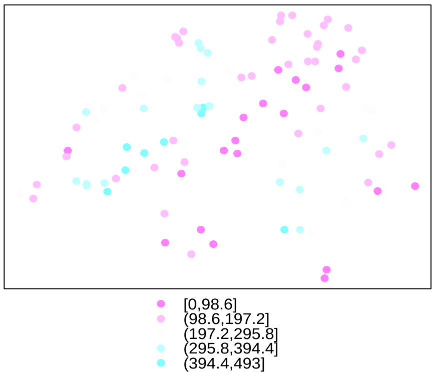
Spline fits



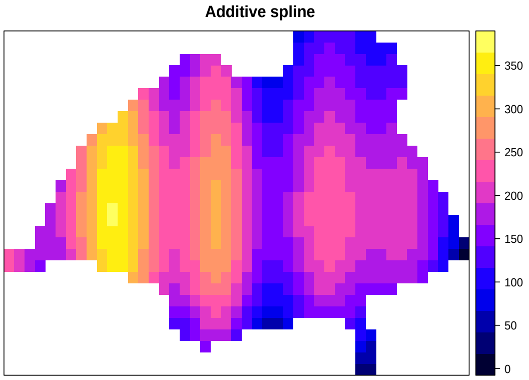
Spline fits

- Two ways to extend to spatial data
- Additive splines
  - spline fn of X coordinate: describes pattern in X
  - spline fn of Y coordinate: describes pattern in Y
  - Add them together
- depends on axis directions. Assumes pattern along the axes

Swiss rainfall data



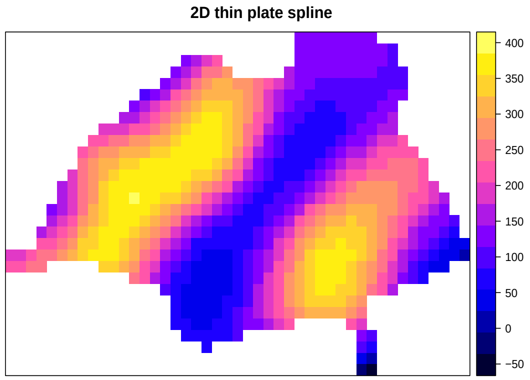
Swiss rain, Spline fit: s(x) + s(y)



Spline fits

- Two ways to extend to spatial data
- thin plate spline
  - think of a sheet of paper or thin sheet of metal
  - drape over the data, allow to wiggle
  - models pattern in all directions simultaneously
  - not dependent on axis directions
  - requires much more data than additive splines

Swiss rain, Spline fit: s(x,y)



Models for data

- IDW: no model
- trend surface: $\boldsymbol{Z(s)} = \beta_0 + f(\boldsymbol{s}) + \varepsilon$
  - all the spatial "action" is in $f(\boldsymbol{s})$
  - Have to choose form of $f(\boldsymbol{s})$
    - $\beta_1 X + \beta_2 Y$
    - $\beta_1 X + \beta_2 Y + \beta_3 X^2 + \beta_4 Y^2 + \beta_5 XY$
    - $s(X) + s(Y)$
    - $s(X, Y)$
  - given form of model, can easily estimate unknown parameters, e.g., β_1, β_2, or the parameters in $s()$.
- kriging: simple, ordinary $\boldsymbol{Z(s)} = \beta_0 + \varepsilon$
  - ε are correlated. nearby observations more so.
  - all the spatial "action" is in the correlations
- universal kriging: $\boldsymbol{Z(s)} = \beta_0 + f(\boldsymbol{s}) + \varepsilon$, ε correlated

Kriging

- Original motivation:
  - underground gold mining. gold content varies along a rock face
  - want to predict where highest gold content
  - and / or average gold content in an area
  - sample a small fraction of the rock face
  - prediction problem: predict $\hat{\boldsymbol{Z(s)}}$ at new locations given data
- Danie Krige: treat $\boldsymbol{Z(s)}$ as spatially correlated collection of r.v.'s
- derive optimal predictor
- original paper: 1951, So. African mining journal
- procedure now known as kriging

Kriging

- Simplest setup of the problem:
  $\boldsymbol{Z(s)} = \beta_0 + \varepsilon$, $\varepsilon \sim N(\underline{0}, \boldsymbol{\Sigma})$, β_0, $\boldsymbol{\Sigma}$ known
- In words:
  - observations are spatially correlated r.v.'s
  - mean β_0 known
  - covariances (or correlations) between all pairs of obs. $\boldsymbol{\Sigma}$, known
- Can derive: Kriging is the optimal linear predictor
  - No other linear combination of the observations has a smaller variance
- predictions are weighted average of the obs.
- weights are functions of the spatial pattern
  - When little spatial pattern, → regional average
  - When strong spatial pattern, → local average
- weights can be > 1 or < 0
  - predictions can exceed range of observations

Kriging notation

- Use vectors and matrices to describe the data $\boldsymbol{Z(s)}$, their means $\boldsymbol{\mu}$, and their variance-covariance matrix, $\boldsymbol{\Sigma}$.
- $\boldsymbol{Z(s)} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix}$
- $P(\boldsymbol{s}_0)$ is a function that predicts $Z(\boldsymbol{s}_0)$

Kriging as making a good prediction

- What should we choose for $P(\boldsymbol{s}_0)$?
- Want a "good" prediction. Need to measure how good or how bad.
- In general, define a loss function,$L()$, that tells us how to measure good/bad.
- Kriging: use squared error loss

$$L\left(Z(\boldsymbol{s}_0), P(\boldsymbol{s}_0)\right) = \left(Z(\boldsymbol{s}_0) - P(\boldsymbol{s}_0)\right)^2$$

- $P(\boldsymbol{s}_0)$ depends on the data, so $L()$ is a random variable
- so define a good predictor as one that minimizes E $L()$
- That predictor is E $Z(\boldsymbol{s}_0) \mid \boldsymbol{Z(s)}$

Simple Kriging

- Kriging model: $\boldsymbol{Z(s)} = \boldsymbol{\mu}(s) + \varepsilon(s)$
  where $\varepsilon(s)$ are correlated (spatial pattern)
- $\boldsymbol{\mu}(s)$ is known, initially assume $\boldsymbol{\Sigma}$ is known
- can derive:

$$P(\boldsymbol{s}_0) = \mu(\boldsymbol{s}_0) + \boldsymbol{\sigma}' \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{Z(s)} - \boldsymbol{\mu}(s)\right)$$

,

  - Derivation done in S&G, p. 223
  - $\boldsymbol{\sigma}$ is vector of covariances: Cov $(Z(\boldsymbol{s}_0), \boldsymbol{Z(s)})$
  - $\boldsymbol{\Sigma}$ is the Var-Cov matrix of the observations
- Least Squares regression: same loss function
  - Used to writing $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - Algebra: same as $\hat{Y}_i = \bar{Y} + \sigma_{xy}(\sigma_X^2)^{-1}(X_i - \bar{X})$

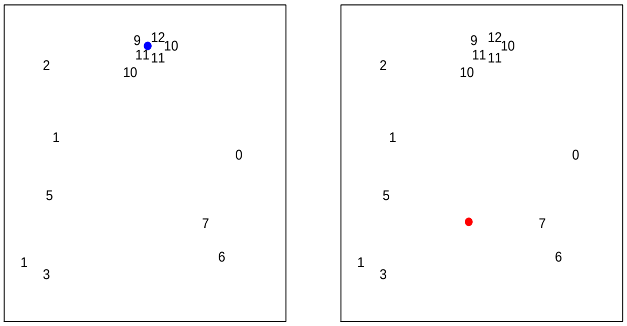
Simple Kriging

- This is the best predictor if $\boldsymbol{Z(s)}$ is Gaussian
- best linear predictor if $\boldsymbol{Z(s)}$ is not Gaussian
- Can also estimate prediction variance

$$\sigma^2(\boldsymbol{s}_0) = \sigma^2 - \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}$$

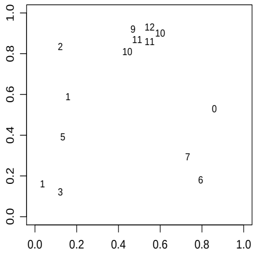
  - This is the variance in the prediction conditional on (i.e., given) observed values
- Looks a bit unusual: usually add variances
- $\boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}$ large when prediction loc. is highly corr. with nearby locs
- Reduces uncertainty in the prediction

Simple Kriging



Simple Kriging: example

- Understanding the prediction in a simple situation
  - Our population: constant mean.
  - Our data: not same value because of random variation



Simple Kriging: example

- The prediction is:

$$P(\boldsymbol{s}_0) = \mu + \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z(s)} - \boldsymbol{\mu})$$

- This is a weighted average of the deviations from the mean, $\boldsymbol{\mu}$

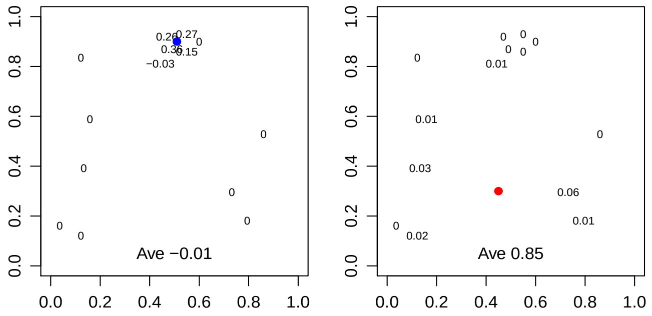
$$P(\boldsymbol{s}_0) = \mu + \boldsymbol{w}\,(\boldsymbol{Z(s)} - \boldsymbol{\mu})$$

  where the weights, $\boldsymbol{w}$, depend on the correlations: $\boldsymbol{w} = \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}$
- Can rewrite as a weighted average of n $Z(s)$ values and the mean, μ

$$P(\boldsymbol{s}_0) = \boldsymbol{w}\boldsymbol{Z(s)} + (1 - \Sigma\boldsymbol{w})\boldsymbol{\mu}$$

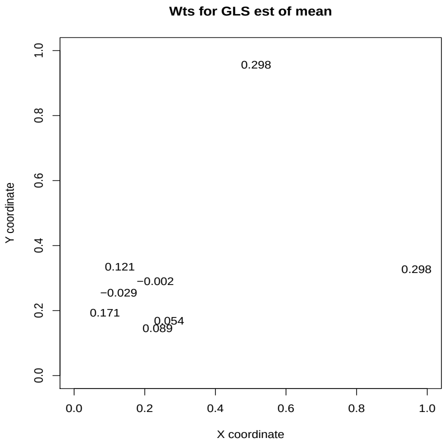
- look at those weights for predictions at two observations:
  - blue location: close to measured locations
  - red location: distant from all measured locations

Simple Kriging: example

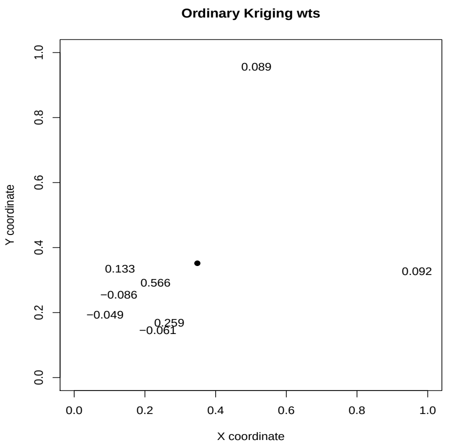


Ordinary Kriging

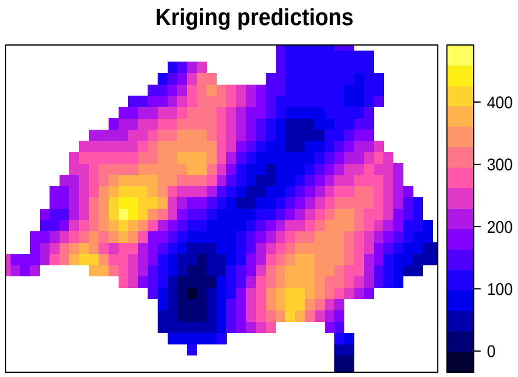
- The problem with simple kriging is that $\mu(\boldsymbol{s}_0)$ usually not known
- Ordinary Kriging: estimate $\hat{\mu}(\boldsymbol{s}_0)$
- slightly different statistical properties
  - no best linear predictor
  - but O.K. is best linear unbiased predictor
$$P(\boldsymbol{s}_0) = \hat{\mu}(\boldsymbol{s}_0) + \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z(s)} - \hat{\mu})$$
, where $\hat{\mu}(\boldsymbol{s}_0)$ is estimated by generalized least squares, GLS
  - For $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$, OLS: $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$
  - In general, GLS: $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$
  - To estimate $\hat{\mu}$, $\hat{\mu} = (\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z(s)}$
- Consequence of GLS is less weight on obs. correl. with others
- Picture on next slide



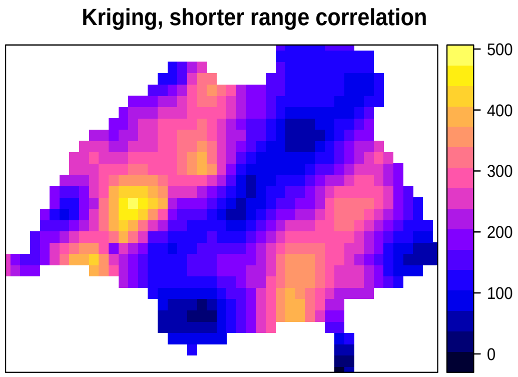
- Useful insight: $\boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}$ is a row vector, so
$$P(\boldsymbol{s}_0) = \hat{\mu}(\boldsymbol{s}_0) + \boldsymbol{\lambda}(\boldsymbol{Z(s)} - \hat{\mu})$$
- values in $\boldsymbol{\lambda}$ depend on covariance btwn obs. values and covariance between prediction location and obs. values
- high for obs. close to prediction location
- values in $\boldsymbol{\lambda}$ may be negative, when obs. are "shadowed"
- Picture on next slide.
- Prediction variance:
$$\sigma^2(\boldsymbol{s}_0) = \sigma^2 - \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} + \frac{(1 - \mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma})^2}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}$$
- S.K. prediction variance + addn. variance because est. μ.



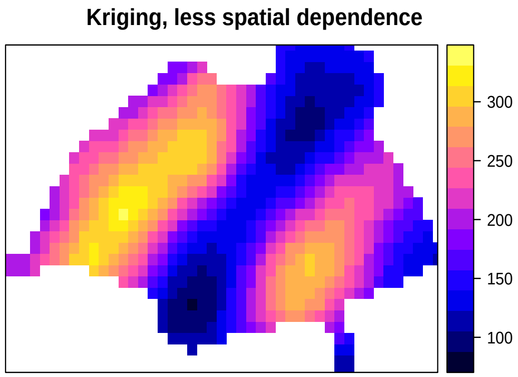
Swiss rainfall



Swiss rainfall

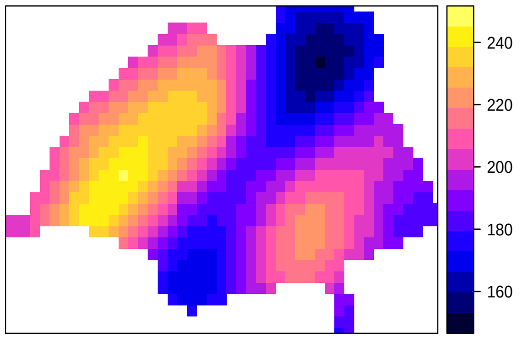


Swiss rainfall



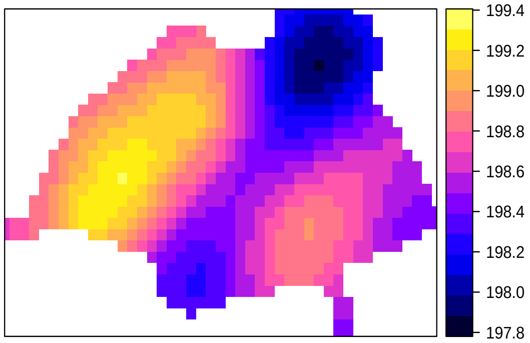
Swiss rainfall

Kriging, small spatial dependence



Swiss rainfall

Kriging, almost no spatial dependence

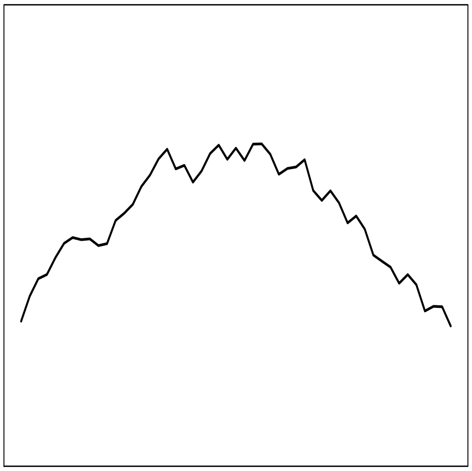


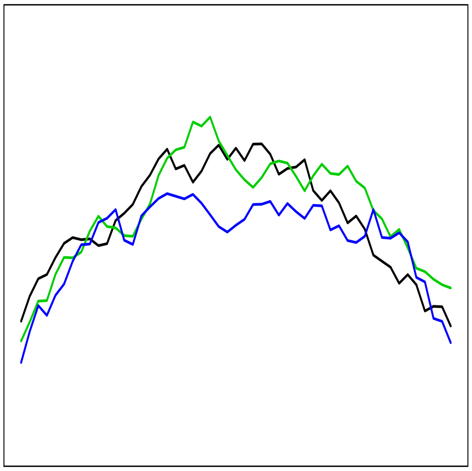
Universal Kriging

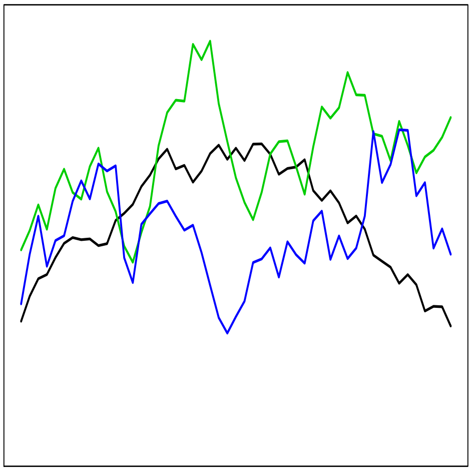
- generalize O.K. to any regression model for local mean
- model: $\boldsymbol{Z(s)} = \boldsymbol{X}(s)\beta + \varepsilon(s)$
- i.e. trend + random variation
  - No unique decomposition
  - Generally consider trend as fixed, repeatable, pattern
  - and random variation to be non-repeatable pattern
- Measure Z at 50 spatial locations. What is the sample size?

Universal Kriging

- generalize O.K. to any regression model
- model: $\boldsymbol{Z(s)} = \boldsymbol{X}(s)\beta + \varepsilon(s)$
- i.e. trend + random variation
  - No unique decomposition
  - Generally consider trend as fixed, repeatable, pattern
  - and random variation to be non-repeatable pattern
- Measure Z at 50 spatial locations. Q: What is the sample size?
- A: ONE. You have one realization of that spatial pattern
- Makes it very difficult to distinguish fixed and random components
- Operationally:
  - trend is the variability that can be predicted by $\boldsymbol{X(s)}$
  - random variation is that which can not
- Choice of $\boldsymbol{X(s)}$ is really important!







Universal Kriging

- notice that spatial variation accounts for lack of fit to trend model
  - Two competing explanations
  - Defer discussion until we talk about spatial linear models
- Should be able to anticipate the predictor:

$$P(\boldsymbol{s}_0) = \boldsymbol{X}(s)\hat{\beta}_{GLS} + \lambda \left(\boldsymbol{Z(s)} - \boldsymbol{X}(s)\hat{\beta}_{GLS} \right)$$

- and the prediction variance:

$$\sigma^2(\boldsymbol{s}_0) = \sigma^2 - \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} + \text{term for Var } \boldsymbol{X}(s)\hat{\beta}$$

- the term for Var $\boldsymbol{X}(s)\hat{\beta}$ is complicated, not too informative

$$\hat{\beta}_{GLS} = (\boldsymbol{X}(s)'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(s))^{-1}\boldsymbol{X}(s)'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z(s)}$$

Comparison of spatial prediction methods

- Inverse distance weighting
  - more weight to nearby locs
  - wts relative to other nearby locs
  - if no other nearby locs, will still average the more distant locs
  - no easy way to estimate uncertainty in prediction
- Trend surfaces
  - depend on specified model form
  - model is a global model
  - splines based on global est of smoothing param.
    - although there are local extensions
  - estimate doesn't depend on number of nearby locs

Comparison of spatial prediction methods

- Kriging:
  - based on correlations among observations
  - estimated from global properties
  - big advantage: estimate depends on number of nearby locs
    - nearby points: prediction more like the local ave.
    - no nearby points: prediction more like the global ave.
  - and data determines how smooth
  - theory: best predictor
    - my experience: not compelling because assumptions never met