Trend surfaces Fitting by Ordinary and Generalized Least Squares and Generalized Additive Models

D G Rossiter

罗大维教授

Nanjing Normal University, Geographic Sciences Department
南京师范大学地理学学院
Section of Soil & Crop Sciences, Cornell University
ISRIC-World Soil Information

February 6, 2020

General objective: spatial prediction

- **Objective**: Given a set of **attribute values** at **known points**, **predict** the value of that attribute at other points.
  - Generalize: predict the mean value over some region, e.g., grid cells, polygons.
- **Objective**: **Understand** why the attribute has its spatial distribution.
  - Helps determine the **process** that produced the spatial distribution.
  - Helps select the best modelling approaches.
- This lecture: **trend surfaces** for both objectives.

Universal model of spatial variation

$$Z(\mathbf{s}) = Z^*(\mathbf{s}) + \varepsilon(\mathbf{s}) + \varepsilon'(\mathbf{s}) \tag{1}$$

- $(\mathbf{s})$ a location in space, designated by a **vector** of coördinates
- $Z(\mathbf{s})$ **true** (unknown) value of some property at the location
- $Z^*(\mathbf{s})$ **deterministic** component, due to some known or modelled **non-stochastic** process
- $\varepsilon(\mathbf{s})$ **spatially-autocorrelated stochastic** component
- $\varepsilon'(\mathbf{s})$ pure (“white”) **noise**, no structure

Universal model of spatial variation – trend surface

The **trend surface** presented in this lecture does not separate spatially-correlated residuals from pure noise, so the model is:

$$Z(\mathbf{s}) = Z^*(\mathbf{s}) + \varepsilon'(\mathbf{s}) \tag{2}$$

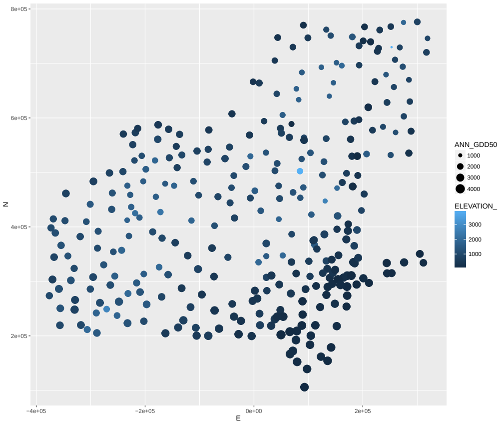
- The deterministic function is of the **coördinates**
- The same mathematics are used if the deterministic function is from a **covariate** which is known at each point **s**.

Example target variable

- Target variable: annual cumulative growing-degree days base 50° F (GDD50)
  - 50° F ≈ 10° C
  - Temperature at which warm-season crop species (e.g., maize, sorghum) can grow
- Predict at every location in region, based on a set of point observations at weather stations with known locations

Example observations

GDD50, Four northeastern US states (NJ, NY, PA, VT)



Q: is there a trend with N and/or E coördinates? With elevation?

Trend surfaces - definition

- One method of **modelling** or **predicting** the values of some **spatially-distributed** variable
- Model and predict using a **continuous mathematical function** of **geographic position**
- **Trend**: varies monotonically (i.e., always increasing or decreasing) with **geographic position**
- **Surface**: **continuous** prediction

Trend surface – physical model

- Target variable varies over space, consistently with coördinates(E, N, H)
- There is a **physical reason** for this
  - temperature: less solar radiation going from S → N, in N hemisphere
  - temperature: less dense atmosphere at higher elevations, holds less heat, so cooler
  - temperature: less seasonal/daily variation near large water bodies, more variation further away

Trend surfaces – conceptual model

- **dependent** variable (to be predicted, to be modelled) is a **function** of the **coördinates**
  - $y = f(x_1, x_2, x_3)$ coördinates
  - e.g., $GDD50 = f(E, N, H)$ (easting, northing, height)
- This function has the same form everywhere in the observation/prediction area
  - a **global** model (vs. **local**)
- So we say the dependent variable has a **geographic trend**
- Example: GDD (dependent variable, to be modelled) are fewer towards the North and at higher elevations (two predictors, independent variables)

Trend surfaces – predictors

- Geographic **coördinates**
  - with respect to some **origin** $(0, 0)$
  - should be **metric** coördinates, with true distances
  - so geographic coördinates (longitude, latitude) must be **transformed**
- For data collected in 3D, include **elevation** above/below some **datum**

Other predictors (not geographic coördinates)

- The same **model forms** can be used with other **global predictors**, not just coördinates
- Examples:
  - Distance from one or more features (urban areas, water bodies . . .)
  - Terrain (slope, aspect, curvature . . .)
  - Land cover / land use
- **The mathematics is the same** as will be presented in this lecture

Models

- A **simplified** representation of reality
- Can **compute** with the model to make **predictions**
- The model will not exactly reproduce reality → lack of fit of observations, these are model **residuals**

Structure vs. noise in reality and the model

- **Reality** – as it exists
  - Reality = f(Structure; Noise)
  - Reality = f(deterministic or stochastic **processes**; **random** variation)
- **Observations** – what we measure
  - Observations = f(Structure; Noise) – as part of reality
  - Observations = f(**model**; **unexplained** variation)
- We want to **match** these

Trend surface example

- **Reality**: Growing Degree Days (GDD) $\approx$ heat available for crop growth
  - GDD = f(coördinates, elevation, “random” variation)
  - “Random variation” = unexplained + observational error
  - **Unexplained**: other factors not known or not measured
    - e.g., aspect, surrounding land cover, nearby water or buildings . . .
- **Trend surface model**:
  - GDD = f(coördinates, elevation) + **noise**

Model forms – 1 – Linear or not

- **Linear**: constant change in independent variable per unit of predictor, does not depend on where in the predictor range
- **Linearizable**: same, with a **transformation** of either independent or predictor variables
- **Non-linear**: change varies with predictor value → smooth function of predictor

Model forms – 2 – Spatial extent

- **Global**: model parameters are the same throughout the range of the predictor
  - e.g., multiple regression
- **Piecewise**: model parameters are different in different parts of the range of the predictor
  - e.g., thin-plate splines
- **Local**: no trend, model from “nearby” observations (e.g., Kriging)

Number of predictors

- **Univariate**: single predictor
- **Bivariate**: two predictors, e.g., geographic coördinates
- **Multivariate**: two or more predictors
  - Must consider **non-independence** of predictors
    - e.g., for linear models, (partial) **co-linearity**: the predictors themselves have a linear relation
  - May consider **interaction** of predictors
    - effect of a combination is more or less than would be predicted considering them separately

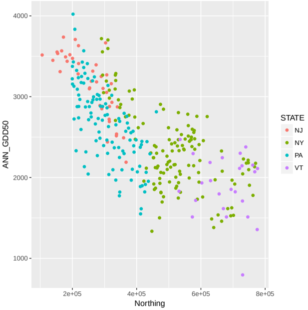
Simple linear regression – concept

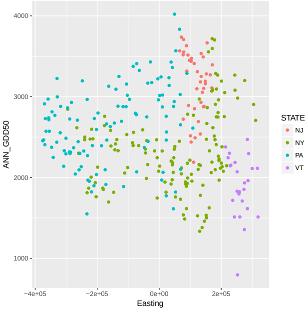
- **Linear** model, **one** predictor
- The dependent variable only depends on one predictor
  - e.g., distance along a transect (1D) or one coördinate(2D)
- The dependence is **linear**
  - constant change in independent variable per unit of predictor
- The model is **global** – it applies throughout the range, all observations are used to calibrate
- **Are these realistic assumptions?**
  - We can check with **model diagnostics**
  - But also think beforehand, based on our **knowledge** of the **process**

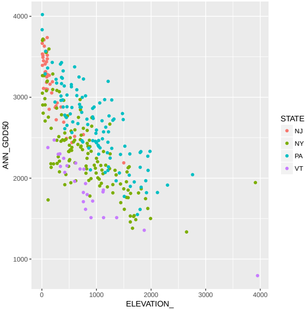
Example: GDD50 physical model

- Why could it depend on *Northing*?
  - Physical principles: sum of solar radiation; longer days in northern hemisphere summer
- Why could it depend on *Easting*?
  - Proxy for distance from ocean with a N/S coastline?
  - Proxy for distance from centre of continent?
- Why could it depend on *elevation*?
  - Physical principles: less air pressure at higher elevations, lower heat capacity
- Which of these would be the most important **single** factor to use in simple regression?
  - Does the study area affect this answer?

Relation of GDD with single predictors







Linear? Which is the best single predictor?

Simple linear regression – model

Model form: $y = \beta_0 + \beta_1 x + \varepsilon; \;\; \varepsilon \sim \mathcal{N}(0, \sigma^2)$

- y: *dependent* variable, to be modelled/predicted
- x: *independent* variable, *predictor*
- ε: *error*, lack of fit, noise ...
  - **independently and identically distributed** (IID) from a 0-mean normal distribution with some **error variance** σ^2
- β_1: coefficient for x, "slope" for simple regression
- β_0: centering coefficient, "intercept" for simple regression

Simple linear regression – one observation

Each observation i: $y_i = \beta_0 + \beta_1 x_i + r_i$

- **Same** coefficients β_p at all observations → a **global** model
- Once β_p are known, computed **fitted** values at each point: $\hat{y}_i = \beta_0 + \beta_1 x_i$
- At each point the **residual** lack of fit: $r_i = (y_i - \hat{y}_i)$
- The r_i are *assumed* to be **independently and identically distributed**

Ordinary Least Squares (OLS)

- **Least squares**: parameters β_0, β_1 are selected to **minimize** the **sum of squared residuals**: $\sum_i (y_i - (\beta_0 + \beta_1 x_i))^2$
- This is *not* the only possible optimization criterion!
  - For example, it can be greatly influenced by extreme values, so there are optimization criteria that attempt to fit "most" of the values well, ignoring extremes
  - These are called **robust** regression methods
- **Ordinary**: IID residuals, no weighting of observations, no covariance between residuals

Fitting the simple linear regression by OLS

- Objective: select β_0, β_1 to **optimize** the fit
- Optimization criterion: **minimize** the **sum of squared residuals** $\sum_i (y_i - (\beta_0 + \beta_1 x_i))^2$
  - **squared**, so that $\pm$ residuals are equally influential
  - **ordinary sum**, so all residuals are equally important
- This is not the only possibility! e.g., could **weight** the residuals
  - by their observation precision, spatial correlation ...
- It has strong **model assumptions**

Fitting by OLS

- Minimize $\sum_i \varepsilon_i^2 = \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2$
- Method: take **partial derivatives** with respect to the **two parameters**; solve system of two **simultaneous equations**
- Solution:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

  - $\overline{x}, \overline{y}$ are the **means**
  - $\hat{\beta}_0$ centres the regression on $(\overline{x}, \overline{y})$

Relation to variance/covariance

- Another way to write this:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

- s_{xy} is the sample **covariance**
- s_x^2 is the sample **variance**
- These are unbiased estimates of the population variance/covariance:

$$\hat{\beta}_1 = \frac{\mathrm{Covar}(x, y)}{\mathrm{Var}(x)}$$

- Note that all the error is assumed to be in the dependent variable

OLS linear model fit – 1st order trend on one coördinate

```
> summary(m.ols.n)

Call:
lm(formula = ANN_GDD50 ~ N, data = ne.df)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.320e+03  2.493e+01   93.08   <2e-16
N           -2.554e-03  1.379e-04  -18.52   <2e-16

Residual standard error: 393.7 on 303 degrees of freedom
Multiple R-squared:  0.5311, Adjusted R-squared:  0.5295
```

Trend on N explains 53% of the variability in GDD50 over this area (see next slide)

Evaluating the success of the model fit

Total Sum of Squares TSS: deviation of observations from a null (mean $\overline{z}$) model (no predictors)
$\text{TSS} = \sum_i (z_i - \overline{z})^2$

Residual Sum of Squares RSS: deviation of observations z_i from fitted model predictions $\hat{z}_i$
$\text{RSS} = \sum_i (z_i - \hat{z}_i)^2$

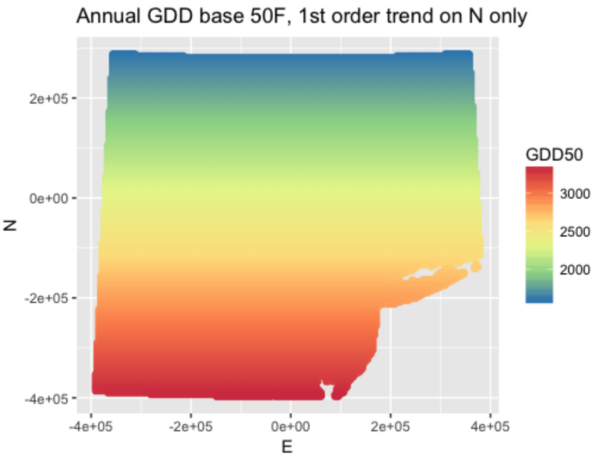
Coefficient of determination (Multiple) $R^2 = 1 - (\text{RSS}/\text{TSS})$

- perfect fit: $R^2 = 1 - 0/1 = 1$
- no fit: $R^2 = 1 - 1/1 = 0$.
- proportion of the variance in the dependent variable explained by the model (i.e., *not* left in the residuals)

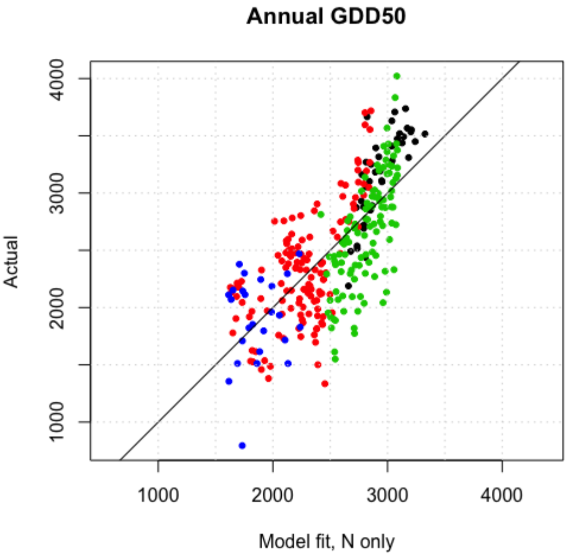
Adjusted evaluation of model fit

- Idea: avoid over-fitting to this dataset (**sample**), so the model is more likely to fit the whole **population** from which the sample is taken
- Idea: avoid over-optimistic estimation of model success
- Adjusted R^2 penalizes R^2 for the number of predictors p in the model (i.e., loss of degrees of freedom), compared to the number of observations n
  - $R^2_{\mathtt{adj}} = 1 - (1 - R^2) \left(\frac{n-1}{n-p-1} \right)$
  - $R^2_{\mathtt{adj}} = 1 - \frac{\mathrm{RSS}/\mathrm{df}_r}{\mathrm{TSS}/\mathrm{df}_t}$
- more $p \rightarrow$ more adjustment
- more $n \rightarrow$ less adjustment
- Somewhat *ad hoc* (empirical), there are more formal ways to evaluate this

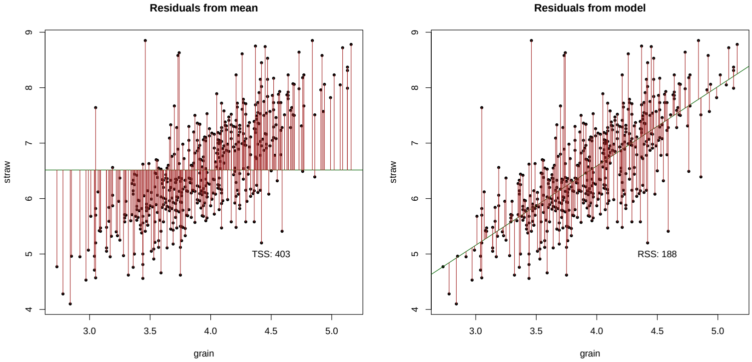
OLS 1st order trend surface, N only



OLS 1st order trend surface, N only



Minimize the squared residuals



If we set $\widehat{\beta}_0 = \overline{y}, \widehat{\beta}_1 = 0$ (left graph) we get a "free" model; the independent variable is not used.
This is the **null model**.

Prediction variance - I

- The fit of the line to the points is not exact, i.e., the estimated parameters $\widehat{\beta}_p$ are uncertain
- So any **predictions** made with the equation are also uncertain.
- The **prediction variance** depends on
  1. the variance of the **regression** $s^2_{Y.x}$; and
  2. the **distance** $(x_0 - \overline{x})$ of the predictand at value x_0 from the **centroid** of the regression, $\overline{x}$
- The first term is the uncertainty of the regression parameters.
- The second term shows that the further from the centroid of the regression, the more any error in estimating the slope of the line will affect the prediction.

Prediction variance - II

Then the estimation variance is:

$$s^2_{Y_0} = s^2_{Y.x}\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]$$

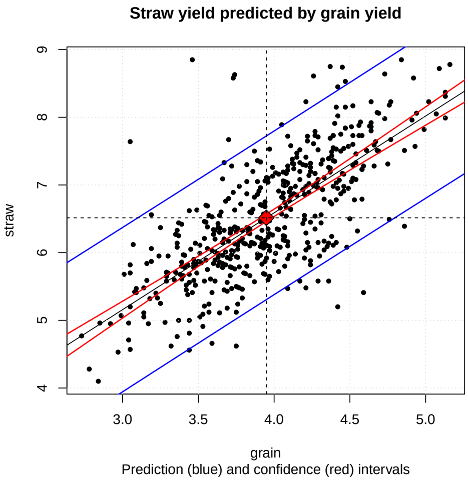
This shows that if we try to predict "too far" $(x_0 - \overline{x})^2$ from the centroid $\overline{x}$, the uncertainty will be so large that any prediction is meaningless.

The variance of the regression $s^2_{Y.x}$ is computed from the residuals:

$$s^2_{Y.x} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

The better the fit, the smaller the uncertainty in the regression parameters.

Visualizing OLS uncertainty



Note more error away from centroid.

Multiple linear regression - I

- Extend to p predictors:
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$
  - e.g., two coördinates, maybe with their interaction or powers
- More easily written in matrix notation
  - $\mathbf{y} = \mathbf{X}\beta + \varepsilon$
  - $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$
  - $\mathbf{X}$ is the **design matrix**
  - β is the **coefficient vector**
  - $\mathbf{I}$ is the **identity** matrix: diagonals all 1, off-diagonals all 0
    - Notice that this means there is **no correlation** among the errors!
    - This is the assumption we will relax in **generalized** least squares (GLS)

Multiple linear regression - II

- The matrix notation for **simple** linear regression can be expanded as:

$$y = [1 \ x] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \varepsilon$$

- The matrix notation for **multiple** linear regression can be expanded as:

$$y = \begin{bmatrix} 1 \ x_1 \ x_2 \ \ldots \ x_p \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \ldots \\ \beta_p \end{bmatrix} + \varepsilon$$

- In the expanded design matrix $\mathbf{X}$, the **1** and $\mathbf{x_i}$ are column vectors of the predictors.

Solution

- Solve for β by minimizing the **sum of squares of the residuals**: $S = \varepsilon^T \varepsilon = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$
- This expands to

$$\begin{aligned} S &= \mathbf{y}^T\mathbf{y} - \beta^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta \\ S &= \mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta \end{aligned}$$

- Minimize by finding the **partial derivative** with respect the the unknown coefficients β, setting this equal to $\mathbf{0}$, and solving:

$$\begin{aligned} \frac{\partial}{\partial \beta^T} S &= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta \\ \mathbf{0} &= -\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\beta \\ (\mathbf{X}^T\mathbf{X})\beta &= \mathbf{X}^T\mathbf{y} \\ (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\beta &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ \hat{\beta}_{\mathrm{OLS}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

Analogy with simple OLS

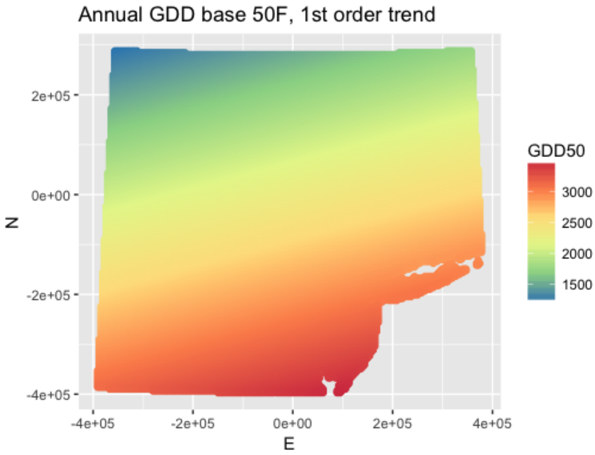
- $(\mathbf{X}^T\mathbf{X})$ is the matrix equivalent of s_x^2, the variance of the predictor $\mathbf{x}$
  - Dimensions: $[p, n] \cdot [n, p] = [p, p]$, i.e., the product-crossproduct matrix of the predictors
  - Products are positive, crossproducts may be positive or negative
- taking the matrix inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ is the matrix equivalent of division: $1/s_x^2$
- $\mathbf{X}^T\mathbf{y}$ is the matrix equivalent of s_{xy}, i.e., the covariance between predictor and predictand.
  - Dimensions: $[1, n] \cdot [n, 1] = [1, 1]$, i.e., a scalar

OLS linear model fit – 1st order trend on two coördinates

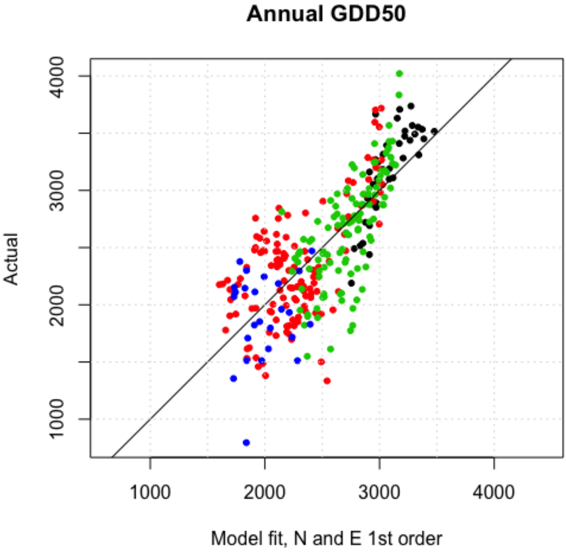
```
> summary(m.ols.ne)

Call: lm(formula = ANN_GDD50 ~ N + E, data = ne.df)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.706e+03  6.154e+01   60.21  < 2e-16
N           -2.818e-03  1.370e-04  -20.58  < 2e-16
E            7.480e-04  1.210e-04    6.18 2.07e-09

Residual standard error: 371.5 on 302 degrees of freedom
Multiple R-squared:  0.5837, Adjusted R-squared:  0.5809
```

Trend on N and E explains 58% of the variability in GDD50 over this area

OLS 1st order trend surface, N and E



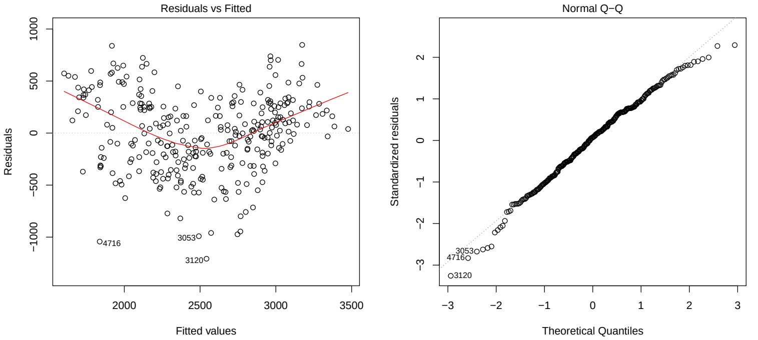
OLS 1[st] order trend surface, N and E



Regression diagnostics

- We can always solve the OLS equation! but recall that the OLS solution depends on **assumptions.**
- So, must check that the model assumptions are satisfied; including **non-spatial**:
  - residuals are approximately **normally distributed**
  - no relation between residuals and fitted **values** (i.e., mean residual should be 0 no matter what the fitted value)
  - no difference in **spread** of residuals at different fitted values
- . . . and **spatial**:
  - for OLS, independent residuals (spatial, temporal, observation sequence . . .)
  - for trend surfaces this implies **no spatial dependence**

Checking non-spatial diagnostics – graph



residuals vs. fits

theoretical vs. actual quantile

estimating normal σ^2 from residuals

Detail: standardized residuals

- The Quantile-Quantile ('QQ') plot compares **standardized** residuals with the same number of points drawn from a Normal distribution
- Standardization adjusts the residuals to distribute as $\mathcal{N}(0, 1)$ with equal variance.
- They are computed as:

$$r'_i = \frac{r_i}{s \cdot \sqrt{1 - h_{ii}}}$$

r_i: unstandardized residuals; s: sample standard deviation of the residuals; h_{ii}: diagonal entries of the "hat" matrix $V = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

Detail: residual standard deviation

The sample standard deviation of the residuals is computed as:

$$s = \sqrt{\frac{1}{(n-p)} \cdot \sum r_i^2}$$

n: number of observations; p number of predictors

This is an overall measure of the variability of the residuals, and so can be used to standardize the residuals to $\mathcal{N}(0, 1)$.

Detail: "Hat" matrix

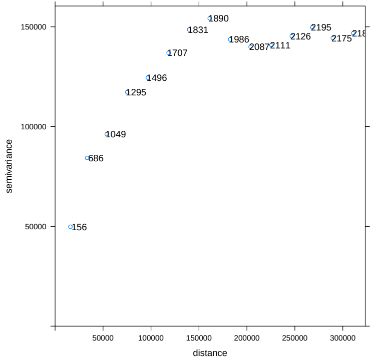
- The "hat" matrix $V = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is another way to look at linear regression.
- When this multiplies the observed vector y it produces the fitted values $\hat{y}$; it "puts the hat symbol on" – the "hat" symbol signifies "estimated" or "predicted"
- The hat value for an observation is the diagonal element $V[i,i] = h_{ii}$; it gives the overall leverage of that observation
- $\sqrt{1 - h_{ii}}$ in the denominator: high influence (large h_{ii}) the denominator is small and so the standardized residual is increased.
- Thus the standardized residuals are higher for points with high influence on the regression coefficients.

Checking non-spatial diagnostics – interpretation

- There is a relation between residuals and fitted values: residuals at both extremes are *positive* (under-predictions); in the mid-range most residuals are *negative* (over-predictions)
  - Mean residual is not 0 through the range of fitted values
- Extreme residuals are *not* from a normal distribution.
- This linear model is *not* justified – it is not reliable for predictions, especially at the extremes
  - add a quadratic term?
  - or are E, N coördinates not sufficient predictors?
  - add **elevation**?
  - fit **piecewise** or with **smooth function** of the predictor?
  - add local deviations by Regression Kriging (**RK**)?

Checking for spatial independence of residuals

Empirical variogram of residuals, `ANN_GDD50 ~ N + E`:



There is definitely **spatial dependence**! I.e., closer separation in **geographic** space → closer separation in **feature** (attribute) space. Range about 150 km.

Higher-order polynomial trend surfaces

- Multiple regression can also use higher-order terms of predictors in a **polynomial** of the predictors
- E.g., 2nd order:

$$y = \beta_0 + \beta_1 E + \beta_2 N + \beta_3 E^2 + \beta_4 N^2 + \beta_5 (E * N) + \varepsilon$$

- Higher-order terms allow closer fit – but will only be justified if the form of the surface matches the form of the phenomenon being modelled
- Should **not** be extrapolated – higher-order terms lead to extreme predictions outside the range of calibration
- Solve by OLS as with any multiple regression

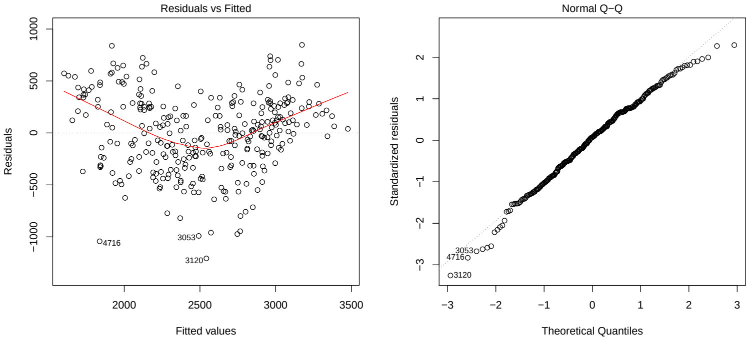
Example with GDD50

Four orders, p-values from the nested ANOVA – is the additional complexity statistically-significant?

- 1^st^order (N only), adjusted $R^2 = 0.530$, p-value ≈ 0
- 1^st^ order (N, E); adjusted $R^2 = 0.584$, p-value ≈ 0
- 2^nd^ order (N, E); adjusted $R^2 = 0.687$, p-value ≈ 0
- 3^rd^ order (N, E); adjusted $R^2 = 0.709$, p-value 0.0002
- 4^th^ order (N, E); adjusted $R^2 = 0.718$, p-value 0.0825

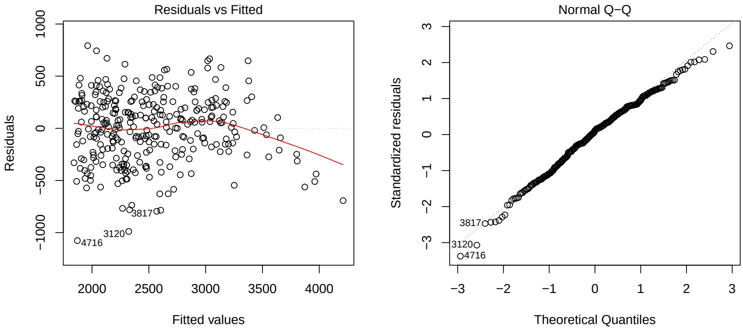
Question: What physical reason could there be for a higher-order trend surface for GDD50 over this region?

Regression diagnostics – 1st order trend



Relation of fits vs. residuals: positive residuals at highest/lowest fits

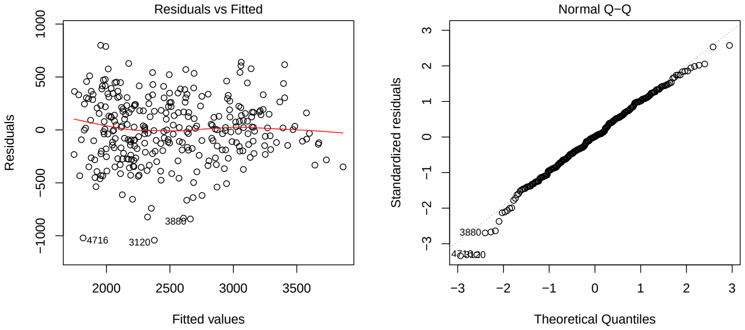
Regression diagnostics – 2nd order trend



Relation of fits vs. residuals seen in 1st order trend has been removed

But systematic over-prediction of highest values

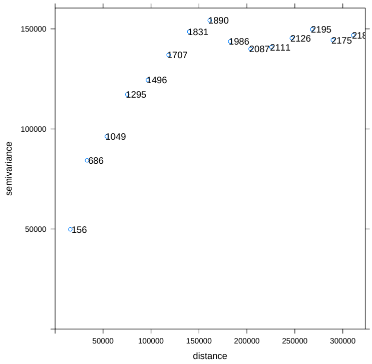
Regression diagnostics – 3rd order trend



No relation of fits vs. residuals
Just a few very poor fits

Checking for spatial independence of residuals

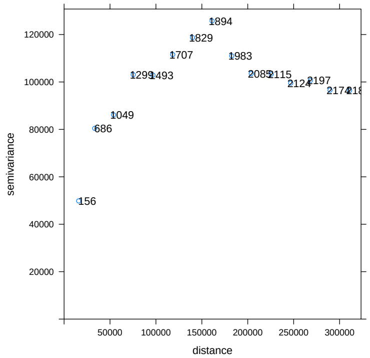
Empirical variogram of residuals, 1st order trend surface



Clear **spatial dependence**! I.e., closer separation in **geographic** space → closer separation in **feature** (attribute) space. Range about 150 km.

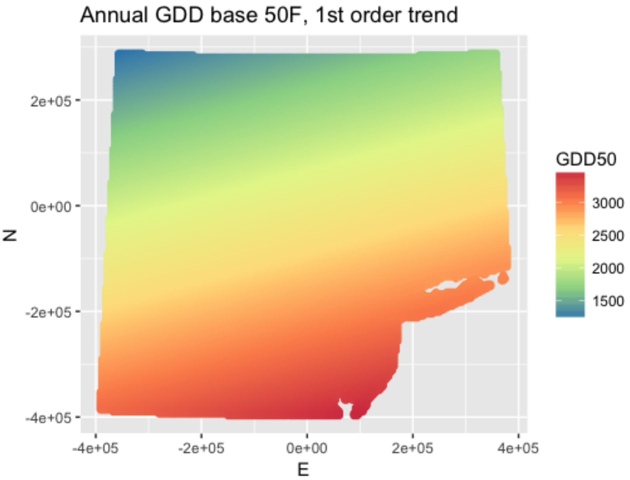
Checking for spatial independence of residuals

Empirical variogram of residuals, 2nd order trend surface

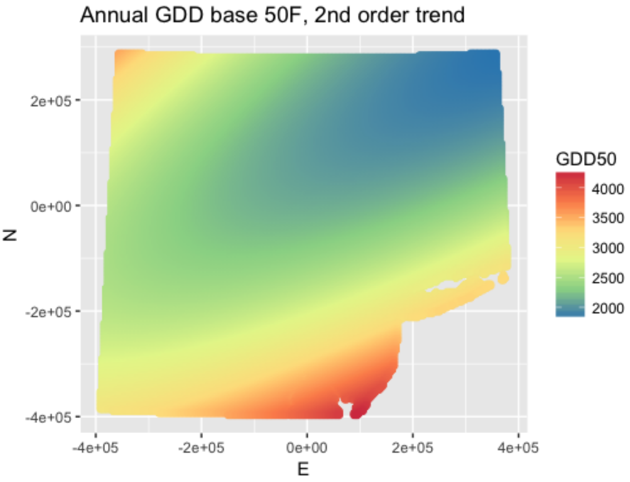


Same as 1st order, spatial dependence to about 150 km.
Total sill reduced from 150 000 to 120 000 GDD^2

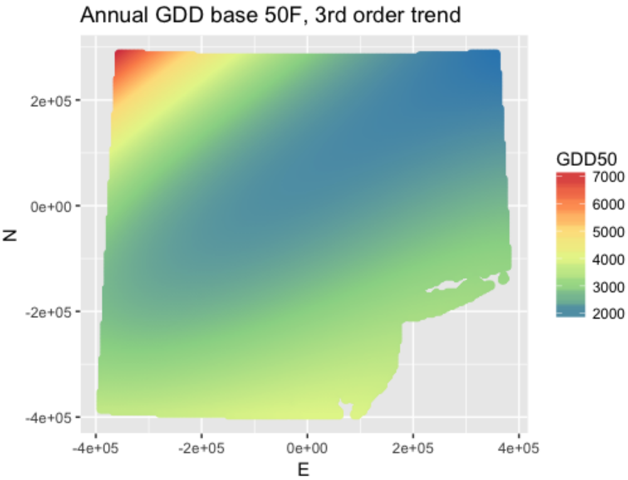
1st order trend



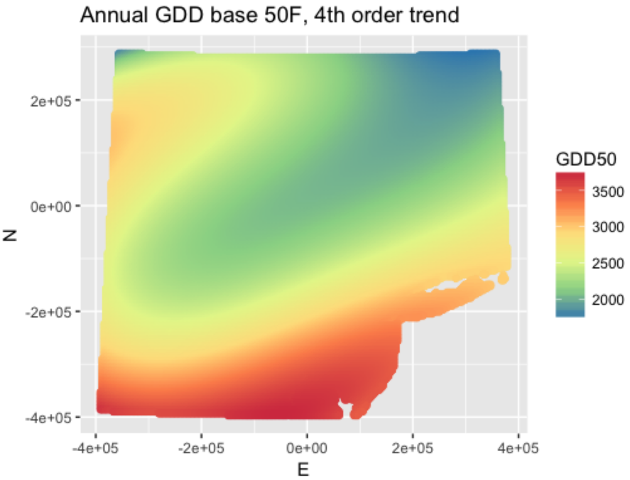
2nd order trend



3rd order trend



4th order trend

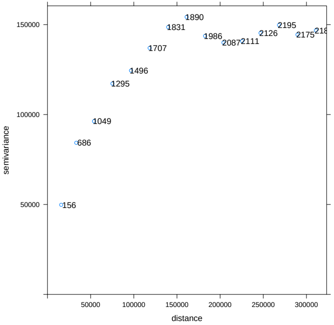


Generalized least squares (GLS)

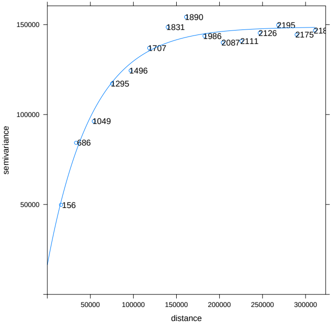
- The OLS fit to a linear model is only optimum if the **residuals** (what the model does not explain) are **independent**.
- In most trend surfaces this is not realistic: **Nearby** residuals tend to be **similar**.
- Physical reason: the **"unexplained"** part of the residual is due to some **spatially-correlated** factor that is not in the model.
  - GDD example: model uses coördinates , but GDD also is affected by elevation, slope and aspect (solar radiation), and maybe nearby land cover (urban area, forest . . .).
  - These are not in our model.
  - But these effects are themselves **spatially correlated** at some scales.

Evidence for spatial correlation of residuals from the OLS fit

Empirical variogram



Variogram model



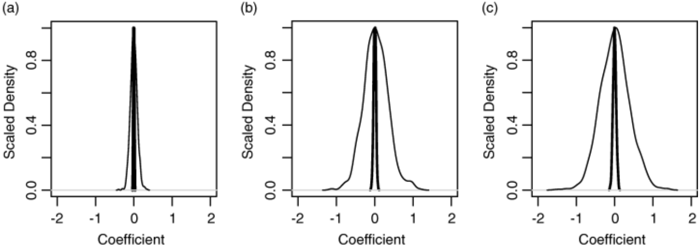
The residuals are not independent.

Effective range 155 km: exponential model fit

a = 51 600 m; total sill 148 800 GDD^2, nugget 16 470 GDD^2

OLS is imprecise under spatial correlation – 1

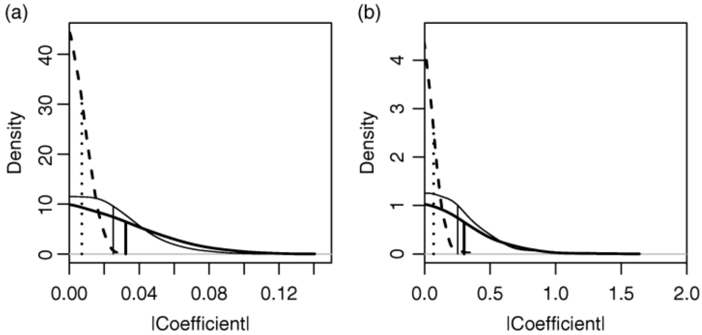
Unbiased but *imprecise*, shown in a simulation study (known regression parameters all 0).



Estimated regression coefficients for 1000 simulations, with **increasing spatial autocorrelation from (a) to (c)**. GLS estimates are illustrated by the thick line and the thin line gives the OLS results. **Ecography**, 30(6): 845

`https://doi.org/10.1111/j.2007.0906-7590.05338.x`

OLS is imprecise under spatial correlation – 2



Density plot of absolute values of regression coefficients estimated by (a) GLS and (b) OLS for 1000 simulations. True parameter values underlying the simulations are 0 in all cases. The dashed, thin and thick lines represent estimates of parameters for covariates with **low, intermediate and high autocorrelation**, respectively.

Other uses for GLS

- Residuals are correlated in **time**, e.g., hydrologic or climate **time series**
- Residuals depend on the sequence of observation (e.g., an instrument drifts out of calibration)
- Residuals depend on the observer

GLS conceptual model - I

- **OLS** model: **independent** residuals:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$$

- **GLS** model: the residuals are a **random variable** η that has a **covariance structure**:

$$\mathbf{y} = \mathbf{X}\beta + \eta,\ \eta \sim \mathcal{N}(0, \mathbf{V})$$

- $\mathbf{V}$ is a positive-definite **variance-covariance matrix** of the model residuals.

GLS conceptual model - II

This is called a **mixed** model:

- The coefficients β are **fixed** effects, because their effect on the dependent variable is fixed once the parameters are known.
- The covariance parameters η are called **random** effects, because their effect on the dependent variable is stochastic, depending on a **random variable** with these parameters.
- In the OLS conceptual model the random effects ε are the **same** for all observations, in GLS they have a **covariance** between each pair.

Variance-covariance matrix

The variance-covariance matrix of the residuals in GLS:

$$V = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1}^2 & \sigma_2^2 & \cdots & \sigma_{2,n} \\ & & \cdots & \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_n^2 \end{bmatrix}$$

In the OLS case this is just:

$$V = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

Estimating V

- *How to estimate all these variances and covariances?* We only have one sample, not the whole population.
- **Assumption 1**, *homoscedascity* of the variances: $\sigma_i^2 = \sigma^2, \ \forall i$
  - i.e., each observation's variance is from the same distribution
  - so $\mathbf{V} = \sigma^2 C$, where σ^2 is the variance of the residuals and **C** is the correlation matrix.
- **Assumption 2**, between-observation covariances follow some **function**
  - so once we have **one** function we can compute the covariances between **all** the residuals
  - geostatistics: covariances in C depend only on the **separation distance** d between them:
    - $\sigma_{i,j}^2 = C(x_i, x_j) = f(d(x_i, x_j))$
  - we get this information from the variogram or correlogram

Optimization criterion - I

- As in OLS we want to **minimize** the sum-of-squares of the residuals $S = \varepsilon^T \varepsilon$.
- However, the error vectors can now *not* be assumed to be **spherically** distributed around the 0 expected value
- So the distance measure, previously estimated by the sum-of-squares, must be **generalized**
- Generalize by taking into account the **covariance V** between error vectors.

Optimization criterion - I

- Generalized estimate of S:

$$S = (\mathbf{y} - \mathbf{X}\beta)^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta)$$

- Dimensions: $[1, n] \cdot [n, n] \cdot [n, 1] = [1, 1]$, i.e., a scalar
- This reduces to the OLS formulation of S when $\mathbf{V} = \mathbf{I}$

GLS Solution

- Expanding the equation for S, taking the partial derivative with respect to the parameters, setting equal to zero and solving we obtain:

$$\begin{aligned}
\frac{\partial}{\partial \beta} S &= -2\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} + 2\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\beta \\
0 &= -\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} + \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\beta \\
\widehat{\beta}_{\mathrm{GLS}} &= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}
\end{aligned}$$

- This reduces to the OLS estimate $\widehat{\beta}_{\mathrm{OLS}}$ of Equation 3 if there is no covariance, i.e., $\mathbf{V} = \mathbf{I}$.

Key point GLS vs. OLS

- Regression coefficients $\widehat{\beta}$ now depend on the **observations** and also the **covariance of the model residuals.**
- For geographic trend surfaces the covariance is the **spatial correlation.**
- So if there is spatial dependence of the residuals, the GLS regression coefficients $\widehat{\beta}_{\text{GLS}}$ will differ from the OLS coefficients $\widehat{\beta}_{\text{OLS}}$.
- **Clustered** observations have less influence on the regression coefficients
  - especially at the extreme values of independent variable (high-leverage)

Computing the GLS coefficients

- Problem: we need to know **V before** we can solve the GLS equation for the the regression coefficients $\widehat{\beta}_{\text{GLS}}$.
- But if **V** is estimated from the spatial correlation structure of the regression **residuals** $(\mathbf{y} - \mathbf{X}\beta)$ we need to know the regression coefficients β **before** we can compute a variogram to model the spatial correlation of the residuals.
  - *"Which came first, the chicken or the egg?"*
- Solution 1: iteration
- Solution 2: REML

GLS solution by iteration

1. Compute $\widehat{\beta}_{\text{OLS}}$ by OLS
2. Compute and model the empirical variogram from the OLS residuals
3. Compute $\widehat{\beta}_{\text{GLS}}$ by GLS, using the variogram model to build the correlation structure **V**
4. Repeat step (2) using the empirical variogram from the **GLS** residuals
5. Repeat step (3) to get a new estimate of $\widehat{\beta}_{\text{GLS}}$
6. Repeat steps (4) and (5) until there is no significant change in $\widehat{\beta}_{\text{GLS}}$.
    - In practice this almost always converges after only a few iterations.
    - But it has no theoretical basis.

GLS solution by REML - I

- A method to compute $\widehat{\beta}_{\text{GLS}}$ and the covariance structure in one pass.
- REML = "Residual maximum likelihood"
- Method:
  1. express V in terms of the parameters $\theta = [\sigma^2, s, a]$ of its covariance function.
     - σ^2 = total sill, s = nugget proportion, a = range.
  2. Maximum likelihood (MLE): find the values of θ that are **most likely** (in a defined probabilistic sense) to have **produced the observed values**, given the model.
  3. Once these are known, compute $\widehat{\beta}_{\text{GLS}}$ by GLS.

GLS solution by REML - II

- The trick is to reduce the unknown β to a **sufficient statistic** that allows the MLE of just the random effects θ.
- Lark, R. M., & Cullis, B. R. (2004).
  Model based analysis using REML for inference from systematically sampled data on soil.

  European Journal of Soil Science, 55(4), 799–813.

  `https://doi.org/10.1111/j.1365-2389.2004.00637.x`

GLS solution by REML - III

- The log-likelihood of the regression and covariance parameters is:

$$\ell(\beta, \theta|\mathbf{y}) = c - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)$$

  where c is a constant and **V** is built from the variance parameters θ and the distances between the observations.

- Integrate out the *nuisance parameters* β and express the likelihood as:

$$\ell(\theta|\mathbf{y}) = \int \ell(\beta, \theta|\mathbf{y})\, d\beta$$

- This can be solved for θ by maximum likelihood.

Difference between OLS and GLS coefficients

- This depends on the **strength of spatial correlation.** If none, OLS = GLS. As strength increases, possible change in coefficients increases
- Also depends on the **configuration** of the observations: If evenly-spaced grid, OLS = GLS. More clustering, more possible change in coefficients
- Also depends on the **data values** of the response variable at clusters – if these are extreme values the cluster has more influence on the OLS coefficients

Specifying the GLS mode in R

```
library(nlme)  ## this includes the gls method
m.gls.ne <- gls(model=ANN_GDD50 ~ N + E,
                data=ne.df,
                correlation=corExp(value=c(50000, 0.1),
                    nugget=TRUE,
                    form=~E + N))
```

- Correlation structure is typically initialized from a variogram model fit to the OLS residuals, but can be directly specified.
- If there is consistent spatial structure the solution is not so sensitive to the starting values.
- The nugget, if present, is specified as a proportion of the total sill.

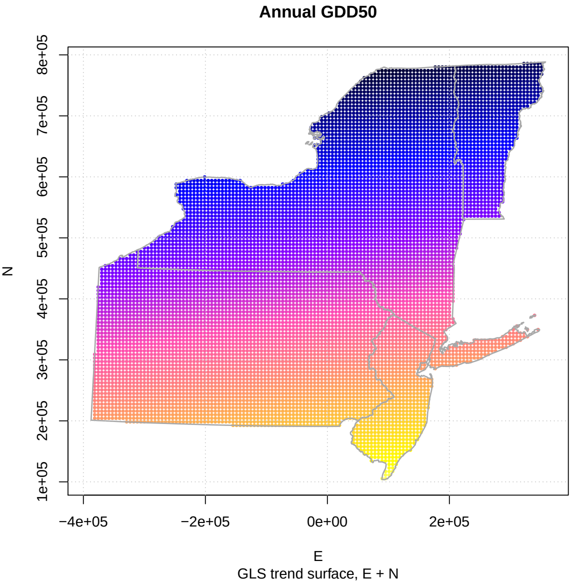
Example GLS R model fit

```
  Model: ANN_GDD50 ~ E + N
       AIC      BIC    logLik
  4380.513 4399.065 -2185.256

Correlation Structure: Exponential spatial correlation
 Formula: ~E + N
 Parameter estimate: range 36007.4

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 3516.002 155.08352 22.671668  0.0000
N             -0.002   0.00033 -7.234058  0.0000
E              0.000   0.00029  1.255212  0.2104

Residual standard error: 381.3984
Degrees of freedom: 305 total; 301 residual
```

GLS model fit: spatial structure

- The REML fit found a range **parameter** 36 km
- Recall, the **exponential** model range parameter is 1/3 of the **effective range**, where the semivariance reaches 95% of the sill
    - The exponential model is asymptotic to the sill parameter and never reaches it
- The variogram model estimate of the range was fit to 155 km; $36 * 3 = 108$ km
- So in this case the REML fit a somewhat shorter range of spatial correlation of the residuals than the estimate from the OLS residuals.
    - Note that the estimate from the OLS variogram is based on a sub-optimal model, so this correction is to be expected.

GLS trend surface



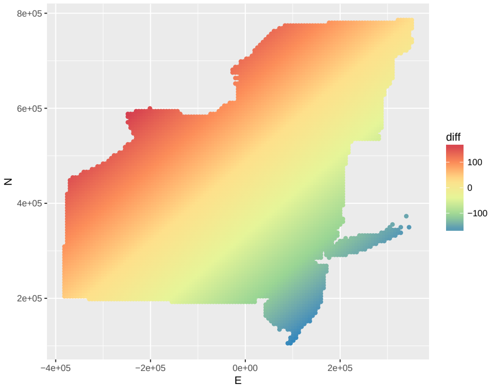
Difference between GLS and OLS fits

```
> round(coef(m.gls.ne) - coef(m.ols.ne),6)
(Intercept)           N           E
-189.859956    0.000449   -0.000380

> 100*((coef(m.gls.ne) - coef(m.ols.ne))/coef(m.ols.ne))
(Intercept)           N           E
  -5.123233  -15.942841  -50.802335

> AIC(m.ols.ne); AIC(m.gls.ne)
[1] 4480.302
[1] 4380.513
```

Coefficients change by about -16% (N) and -51% (E), so GLS surface is **less steep** in both dimensions.

AIC (**A**kaike's **I**nformation **C**riterion) is lower (better) for GLS

GLS – OLS trend surfaces



GLS surface is higher in the NW, lower in SE

Conclusion

- The assumptions of OLS require that the residuals from the model fit be **independently** and **identically** distributed, usually following a **normal** distribution.
- In this case, OLS gives one kind of optimum fit.
- In many geographic applications such as **trend surfaces** the residuals have **spatial correlation** – check for this with a variogram of the residuals.
- In that case GLS computes correct regression coefficients.
- The advantage of the REML method vs. iteration to compute the GLS fit is that REML computes both the regression parameters and the spatial correlation parameters.

Generalized Additive Models (GAM)

- Problem: what if a relation is:
  - not linear over the whole range of predictor/predictand ...
  - not linearizable by a transformation of the predictor over its whole range?
- One solution: GAM as an extension of linear models

GAM as extension of linear models

Each term in the linear sum of predictors need not be the predictor variable itself, but can be an **empirical smooth function** of it.

So instead of the **linear additive** model of *k* predictors:

$$y_i = \beta_0 + \sum_k \beta_k x_{k,i} + \varepsilon_i \tag{3}$$

we allow additive *functions* f_k of the predictors:

$$y_i = \beta_0 + \sum_k f_k(x_{k,i}) + \varepsilon_i \tag{4}$$

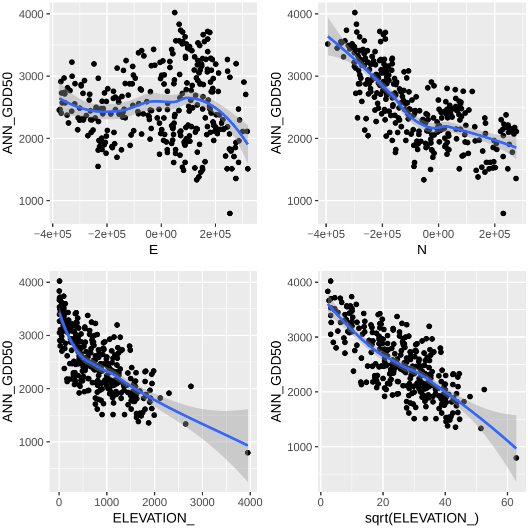
Advantages

- **Non-linear** relations in nature can be fit, without any need to try transformations or to fit piecewise regressions.
  If this is a better model fit, it should result in better predictions.
- The model is **additive**, so the marginal contribution of each predictor to the model fit can be determined.
- **Interactions** can be included via 2D (etc.) **surfaces**

Disadvantages

- An **empirical** fit, *no theory*
  - but shape of marginal fits can suggest causes
- *Can not* be extrapolated beyond the range of calibration.
- The choice of smooth function, and the degree of smoothness, is **arbitrary**
  - the degree of smoothness determined by cross-validation.

Empirical smooth relations predictand/predictor



Empirical smoothers

- loess Local Polynomial Regression Fitting
- Fit at each point using some subset of the points
  - fitting method: default **weighted least squares**
  - proportion of points to use controlled by span parameter (default 0.75)
  - tricubic weighting, proportional to $(1 - (\frac{d}{d_{\max}})^3)^3$
  - degree of polynomial, default 2 (quadratic)
- With all these choices, fit is **empirical**
- Analyst must subjectively match smoothness of fit to smoothness of real-world relation

GAM model formulation for the 2D trend surface

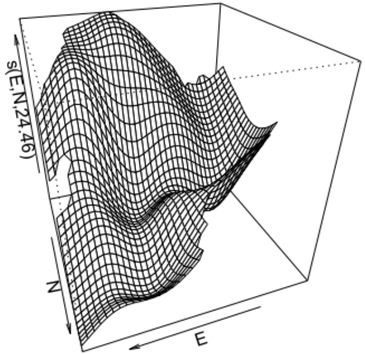
- `gam` function of the `mgcv` package
- call:
  `gam(ANN_GDD50 ~ s(E, N), data=ne.df)`
- Predictor: 2D thin-plate spline of the coördinates `s(E,N)`

GAM model summary – 2D trend

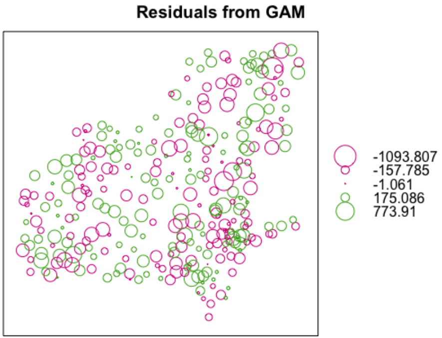
```
Parametric coefficients:
            Estimate Std. Error
(Intercept) 2517.518      9.986
---
Approximate significance of smooth terms:
                 edf Ref.df     F
s(E,N)         24.46   27.8 36.98
---
R-sq.(adj) =  0.771
```

Compare: $R^2{}_{\mathrm{GAM}} = 0.771, R^2{}_{\mathrm{OLS}} = 0.584$; adjusts “locally”

Fitted 2D geographic trend

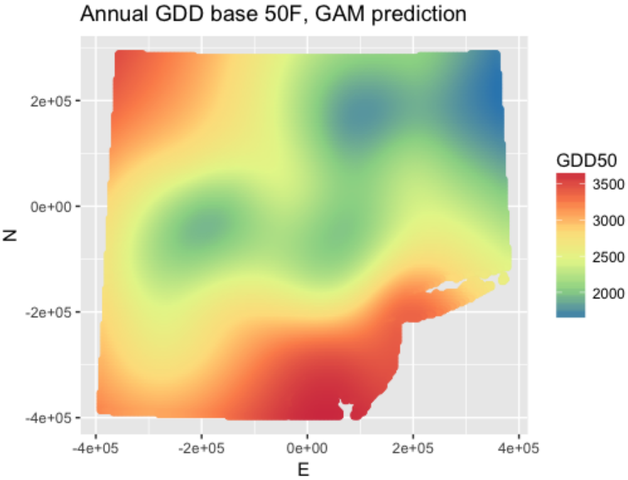


Spatial correlation of GAM residuals

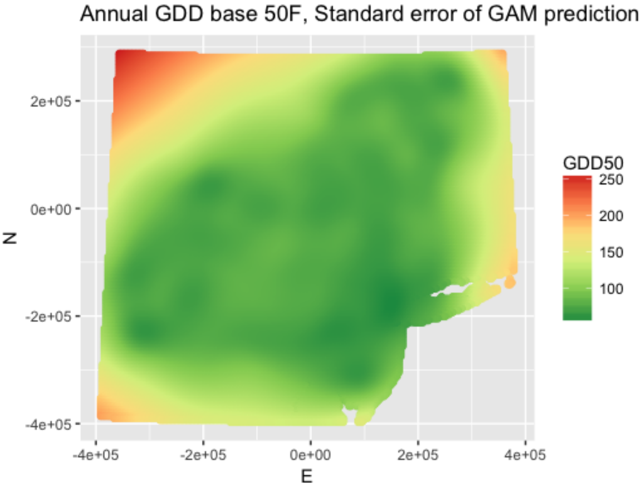


Some spatial correlation at finer scale than GAM smoother

GAM predictions – 2D trend



Standard errors of GAM 2D trend predictions



GAM model formulation for the trend surface – 2D trend + 1D elevation

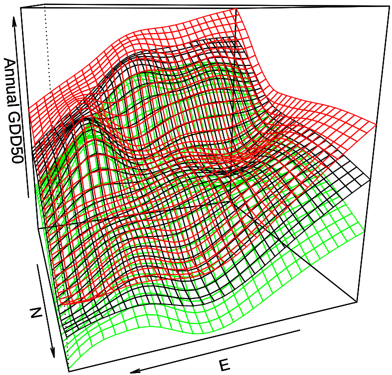
- call:
  `gam(ANN_GDD50 ~s(E, N)+s(ELEVATION_),data=ne.df)`
- Term 1: 2D thin-plate spline of the coördinates `s(E,N)`
- Term 2: 1D spline of the elevation `s(ELEVATION_)`

GAM model summary – 2D trend + 1D elevation

```
Parametric coefficients:
            Estimate Std. Error
(Intercept) 2517.518      9.986
---
Approximate significance of smooth terms:
                 edf Ref.df    F
s(E,N)        23.529 27.300 37.8  <2e-16
s(ELEVATION_)  8.521  8.922 51.6
---
R-sq.(adj) =  0.908
```

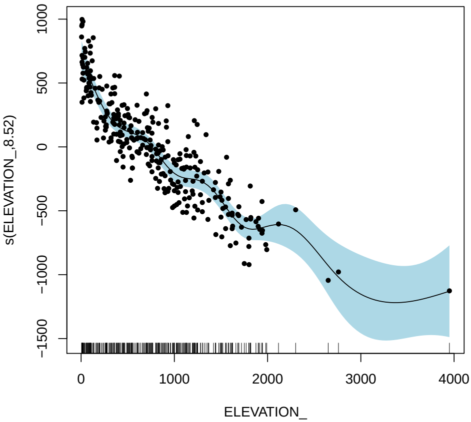
Adding elevation greatly improves the model; it also modifies the fit for the 2D trend term

Fitted 2D geographic trend – with s.e.



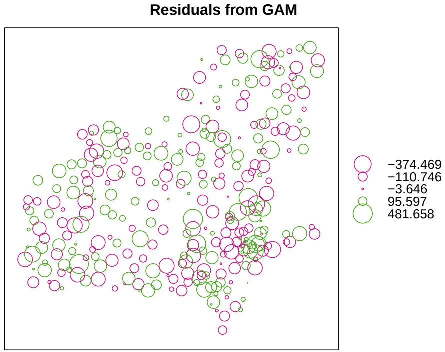
red/green are +/− 1.96 s.e.

Fitted 1D relation with elevation



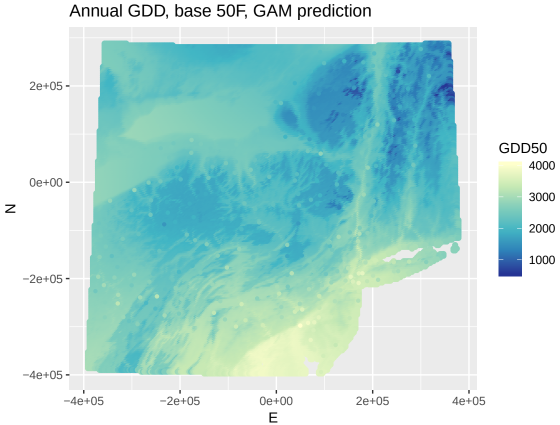
Wide confidence interval at the high elevations – few points → large uncertainty

Spatial correlation of GAM residuals

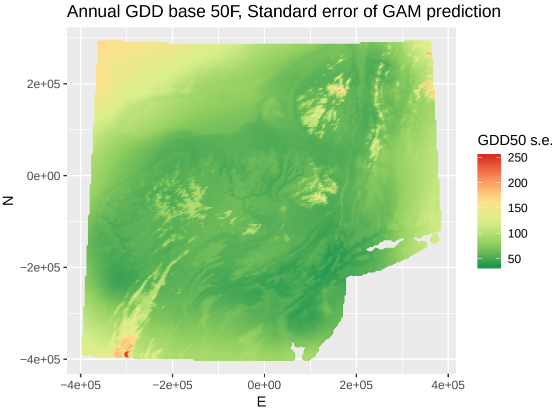


No residual spatial correlation, elevation term has removed it (finer-scale smooth)

GAM predictions – 2D trend + elevation



Standard errors of GAM 2D trend + elevation predictions



Conclusion: GAM for trend surfaces

- Good fit, adjusts within the region
- No theory, smoothers are empirical
- Independent marginal effect of predictors: 2D trend, 1D elevation
- Removes spatial dependence of OLS residuals at the range of the empirical smoother, but not finer
    - So, could refine map by OK of the residuals

GAM References

- Hastie T *et al.* **The elements of statistical learning data mining, inference, and prediction**. Springer, 2nd ed edition, 2009. ISBN 9780387848587; §9.1
- James G *et al.* **An introduction to statistical learning: with applications in R**. Springer, 2013. ISBN 9781461471370; §7.7
- Venables, W & B Ripley. **Modern Applied Statistics with S**. Fourth Edition. Springer, 2002. ISBN 0-387-95457-0; §8.8

End