Assessing Spatial Consistency using Spatio-Temporal Interactions with Generalized Additive Models

June 6, 2024

Payton Miloser

Iowa State University

✉ miloserp@iastate.edu

Are the areas of high yield last year likely to be high yield this year?

Problems:

1. Need to interpret multiple years together, not just pairs of years.
2. Need to locate field areas with consistently high or low yields

Potential Solutions:

1. Expanding beyond the correlation framework
2. View different modeling strategies in combination with spatial statistics

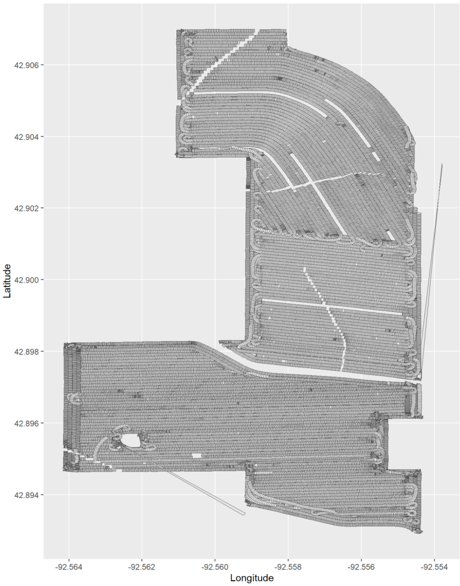
Introduction

Data

Methods

Results

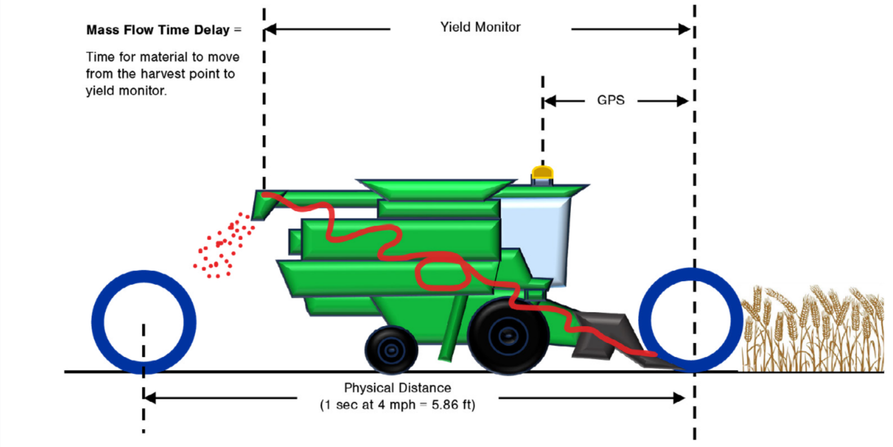
Discussion



- 42719 up to 53375 locations each year
- not in exactly the same location each year

Problems to be Addressed

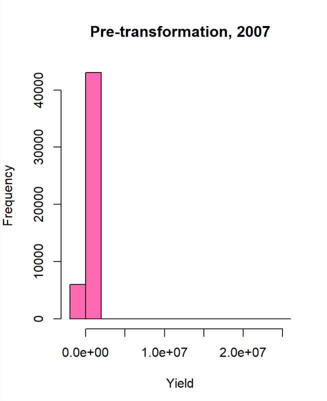
1. The locations will not be in the same place for every year.
2. multiple locations from human error, process error, or relative features of the field.
3. The length of a degree of latitude will not be equal to the length of a degree of longitude!

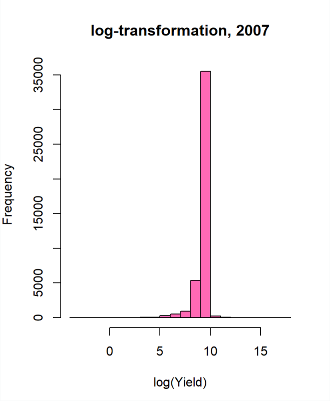

Mass Flow Time Delay =
Time for material to move from the harvest point to yield monitor.
Yield Monitor
GPS
Physical Distance
(1 sec at 4 mph = 5.86 ft)

*Courtesy of Ali Mirzakhani Nafchi and Karishma Kumari

We want to map these yield data under an assumption of multivariate normality → logarithmic transformation

A Gaussian (Normal) process is the joint distribution of all random variables (the yield data), thus it is a distribution over variables with a continuous domain, e.g. time or space.





Introduction

Data

Methods

Results

Discussion

Spatial smoothing is a tool used to process data by reducing the amount of noise; data points are averaged with their neighbors, thus high frequencies of the signal are removed.

There are two main methodologies we considered for spatial smoothing, both which have different approaches. They are,

1. Geospatial Kriging
2. Thin plate spline interpolation.

Geospatial Kriging

Definition

Geospatial kriging is a method of interpolation and thus spatial smoothing based on Gaussian processes.

Kriging assumes that the distance or the direction of a specific set of sample points with coordinates (x_i, y_i) contain a specific spatial correlation that can be used to assess variability of the surface.

The process can be thought of as the average of the whole aggregate surface of possibilities conditioned on the data.

Geospatial Kriging: the semi-variance

Semi-variance Formula

$$\gamma(h) = \frac{1}{2n} \sum [z(u) - z(u+h)]^2$$

- h (scalar) is the distance or the lag distance
- $z(u)$ is the available data of variable z at location u
- z represents the yield data variable
- $z(u+h)$ represents the available data at location $u+$ some lag distance h

Geospatial Kriging: Variograms

- Empirical variogram: a plot of the calculated semivariance values for pairs of data points against their separation distances (lags).
- Variogram Model: A theoretical model (e.g., spherical, exponential) fitted to the empirical variogram to describe spatial structure.

Used For...

- quantifying spatial dependence
- Parameter estimation for kriging

Geospatial Kriging: kriging estimation

Ordinary Kriging

The ordinary kriging estimator of $z(u_0)$, u_0 being an unobserved location, is defined as the linear unbiased estimator of

$$\hat{z}(u_0) = \sum_{i=1}^{n} \lambda_i z(u_i).$$

This formula is for ordinary kriging with no covariates. It demonstrates the structure of kriging interpolation using weighted predictions for unobserved locations based on known locations.

Geospatial Kriging: kriging estimation

Kriging Weights: Using the variogram to weight observations from surrounding measured values we then predict unmeasured locations.

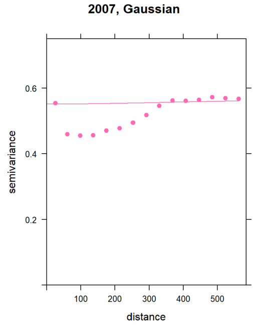
The weights, λ_i come from the variogram,

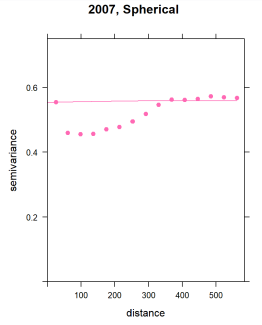
$$\sum_{i=1}^{N} \lambda_i = 1$$

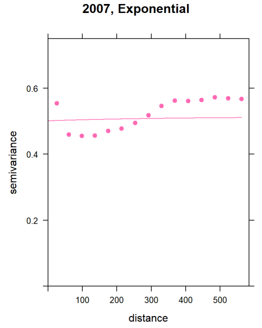
N being the total number of locations. λ_i are higher for nearby locations.

Choosing a Smoothing Method

Problems regarding variogram fit:







Thin Plate Spline Interpolation

Smoothing Spline

A smoothing spline can be defined as a piecewise polynomial function of degree k, that is continuous and has continuous derivatives of orders $1, \dots, k-1$.

- we use a cubic spline
- the cubic spline ensures continuous first and second order derivatives which makes the fitted curves very smooth

Tensor Products

Using linear algebra, a tensor product can be defined as $u \otimes v = uv^T$, where u and v are vectors such that

$$u \equiv \begin{bmatrix} a \\ b \\ c \end{bmatrix} \qquad v \equiv \begin{bmatrix} p \\ m \end{bmatrix}$$

and we get the following result.

$$u \otimes v = \begin{bmatrix} a \cdot p \\ a \cdot m \\ b \cdot p \\ b \cdot m \\ c \cdot p \\ c \cdot m \end{bmatrix} = \begin{bmatrix} ap \\ am \\ bp \\ bm \\ cp \\ cm \end{bmatrix}.$$

Thin Plate Spline Interpolation

Bivariate Splines

Using an additive basis framework to predict the Gaussian field as a linear combination of "*a priori* basis functions $\beta_j(u)$", Interpolation using splines extends the formula to two dimensions for $z(u)$ by creating bivariate splines

$$f(x,y) = \sum_{i=1} \sum_{j=1} \beta_j(x)\alpha_i(y)b_{i,j}$$

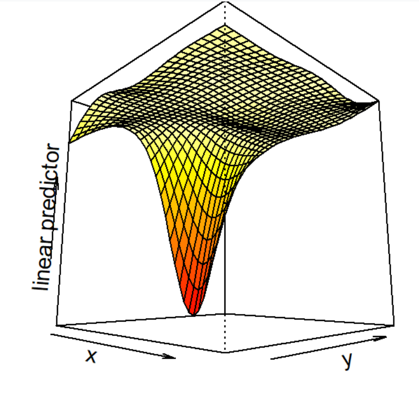
applied to a model,

$$z = f(x,y) + \epsilon$$

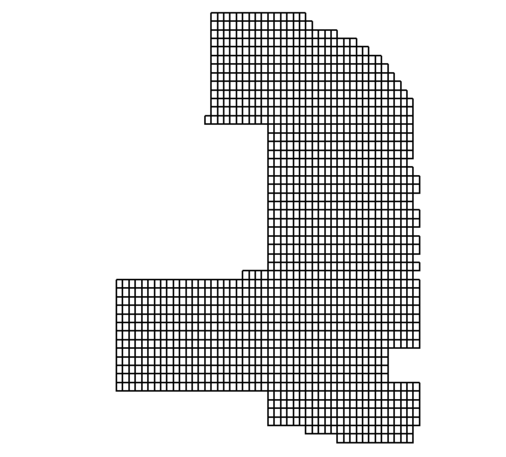
where $f(x,y)$ represents the tensor product of splines for x and y.

Tensor Splines

The tensor product of the splines related to our x and y, say $z = f(x, y)$ where x and y are typically spatial or temporal variables, form the surface when the product is applied.



Generalized Additive Models (GAMs)



- Want to compare across time on a uniform grid.
- The grid was designed to mimic the field with more general locations identified for prediction.

GAMs

Definition

Using the tensor product splines to smooth the data we then are able to simply add up the splines such that for natural cubic splines we get a model of the following form.

$$g(\mu) = \sum_{j=1} \sum_{i=1} \beta_j(u)\alpha_i(u)b_{i,j} + \epsilon$$

GAM Model output

2007 GAM output				
Parametric Coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	9.013	0.004	2545	$< 2e-16$ ***
Smooth Terms	**edf**	-	F	p-value
$te(x,y)$	**23.8**	-	52.12	$< 2e-16$ ***

- **High edf value** → very non-linear relationship between our tensor smooths and the response variable, yield.
- Significant Smoothing term from the ANOVA F-test

H_O : the true function is a flat function (intercept only model)

H_A : the true function is not flat

Benefits of Tensors

- Proved to be computationally efficient for multiple years.
- Solved multiple problems we had previously with the data itself,
    - ex) Tensor product splines do not care about latitude and longitude equality because they treat each dimension independently, constructing the surface by combining univariate splines without assuming any inherent relationship or scale between the latitude and longitude axes.

However...

1. How do we deal with years having different mean yields and different spatial variability?
2. How can we assess whether yields are consistent at each location?

Standardizing & Assessing Consistency

Problem: Variability among space and mean yields.

Solution: centered and standardized the log(yield) within each year. We subtracted each in-year mean yield from every yield observation and then divided by the standard deviation,

$$Scaled(z) = \frac{z_i - \bar{z}_i}{s_{z_i}}$$

for $i = 1, \ldots n$ for each year 2007 through 2011

Standardizing & Assessing Consistency

Example from 2007 Data:

obs (z_i)	**mean yield ($\bar{z}_i$)**	**sd (s_{z_i})**	**scaled(z)**
4.91	9.01	0.75	-5.49
9.13	9.01	0.75	0.167
17.05	9.01	0.75	10.78

- Locations can be consistently low in yield or consistently high in yield!
  - Calculate the standard deviation at every single location across all five years
  - For high and low measures of this consistency we can use the mean yield at each location across the five years

Introduction

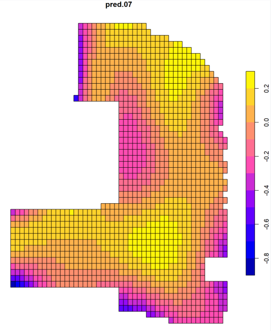
Data

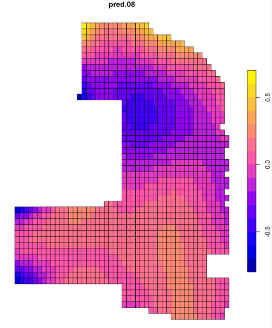
Methods

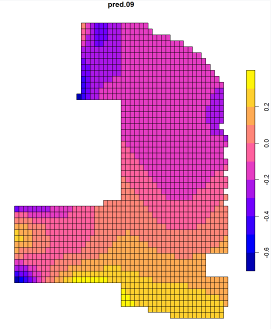
Results

Discussion

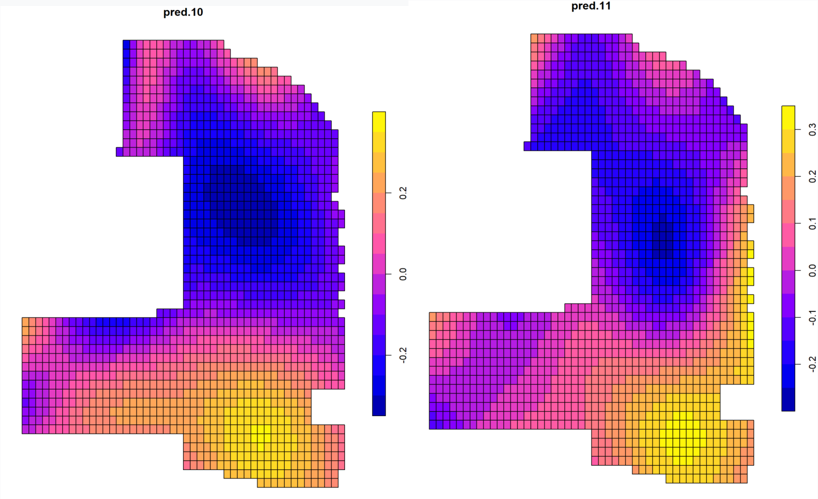
GAM Yield Predictions

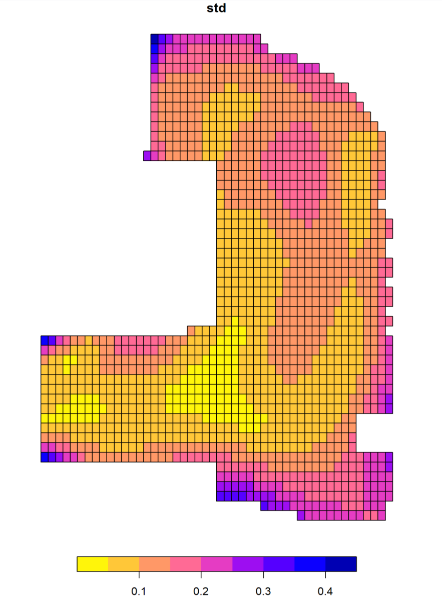




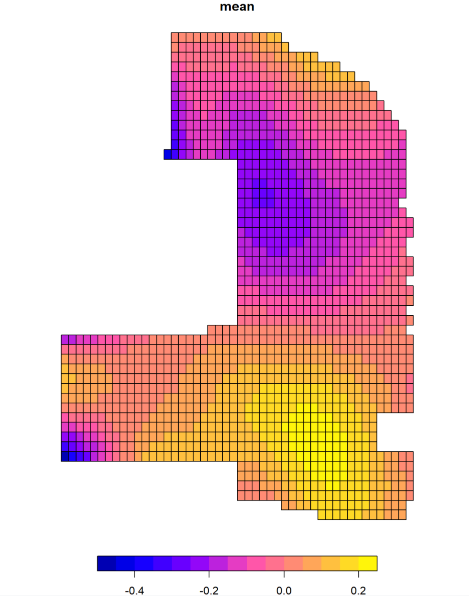


GAM Yield Predictions



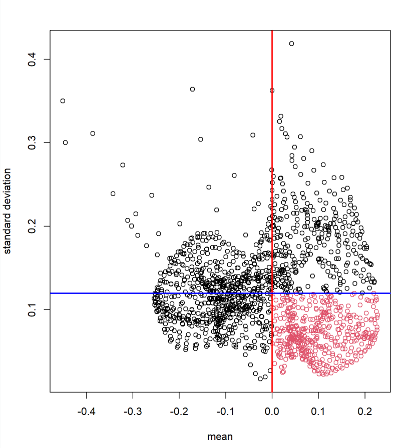


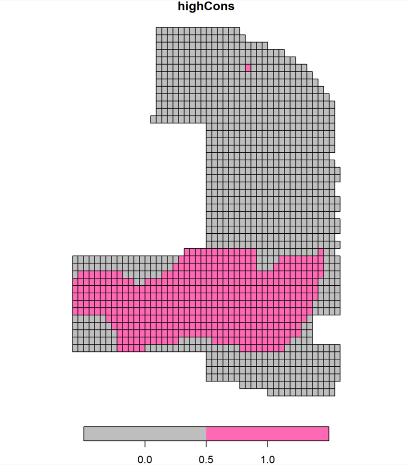
- Lower standard deviation = more consistent
- If each year were exactly the same as the other we would expect a standard deviation of $\sigma = 0$ at the respective locations.



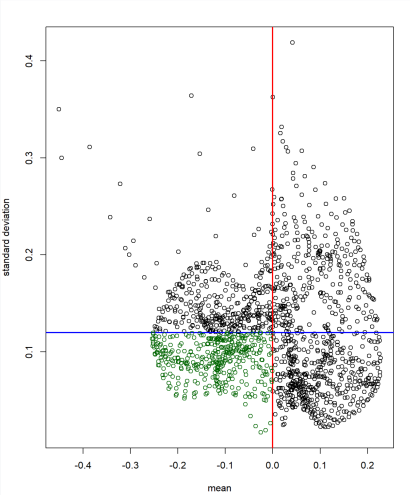
Taking the mean of each individual location across all five years results in the mean yield across all five years…

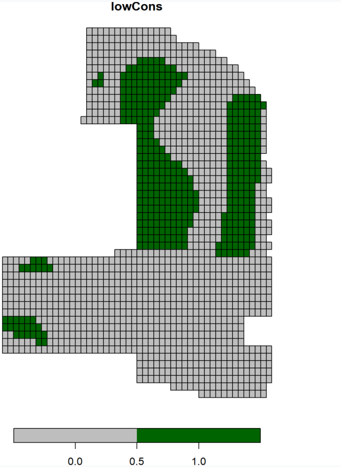
High, Consistent Locations





Low, Consistent Locations





Introduction

Data

Methods

Results

Discussion

In Summary...

Instead of pairwise correlation based comparisons we improved a technique to be used in precision agriculture based data.

Tensor product splines and GAMs give agricultural data a more flexible application and allow for multiple years to be compared simultaneously.

We can locate high/low consistent areas using spatial ag. data!

Future Extensions?

1. The edge effect is certainly an area of interest; this may be supported in the future by adding a buffer around the edges for interpolation or cleaning up the edges
2. Spline predictions can also give us the standard error associated with each location. It may be enough to filter by some high percentile of standard errors, e.g. remove locations with prediction standard errors in the top 0.1% of standard errors
3. Variance Component Analysis for mixed-effects models (as we have with the spatial-temporal effects), estimates the contribution of each random effect to the variance of yield.

$$Y_i = \mu_i + \tau_i + \epsilon_i$$

Thank you!

Questions?