Data Mining of Agricultural Yield Data: A Comparison of Regression Models

Georg Ruß

July 20th, 2009

Outline

Motivation

Available Data
- Data Details
- Data Overview

Points of interest

Advanced Regression Techniques
- Detailed Results

Motivation: Precision Agriculture

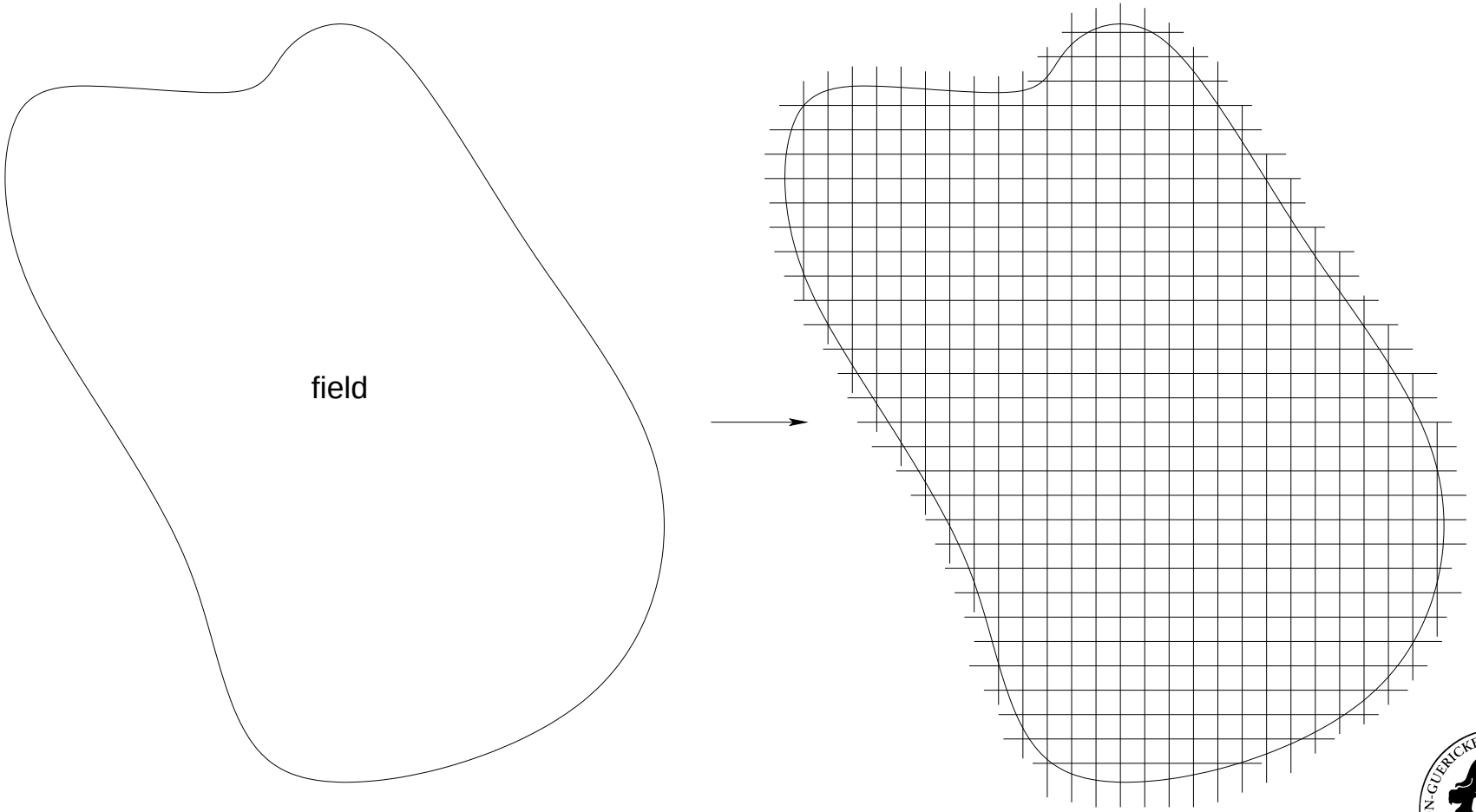


Figure: Precision Farming: from uniform field to small scale area

Motivation: Precision Agriculture

- precision agriculture
  - divide field into small-scale parts
  - treat small parts independently instead of uniformly
  - cheap data collection
  - GPS-based technology
- lots of data (sensors, imagery, GPS-tagged)
- use data mining to
  - improve efficiency
  - improve yield

Data Flow Model

Figure: Data Mining Context

Nitrogen Fertilizer

- easy to measure when manuring
- three points into the growing season where nitrogen fertilizer is applied
- three attributes: N1, N2, N3

Vegetation Measuring

- Red Edge Inflection Point
- first derivative value along the red edge region
- aerial photography or tractor-mounted sensor
- larger value means more vegetation
- measured before N2 and N3
- two attributes: REIP32, REIP49

Electric Conductivity

- measure apparent conductivity of soil down to 1.5m
- uses commercial sensors
- one attribute: EM38

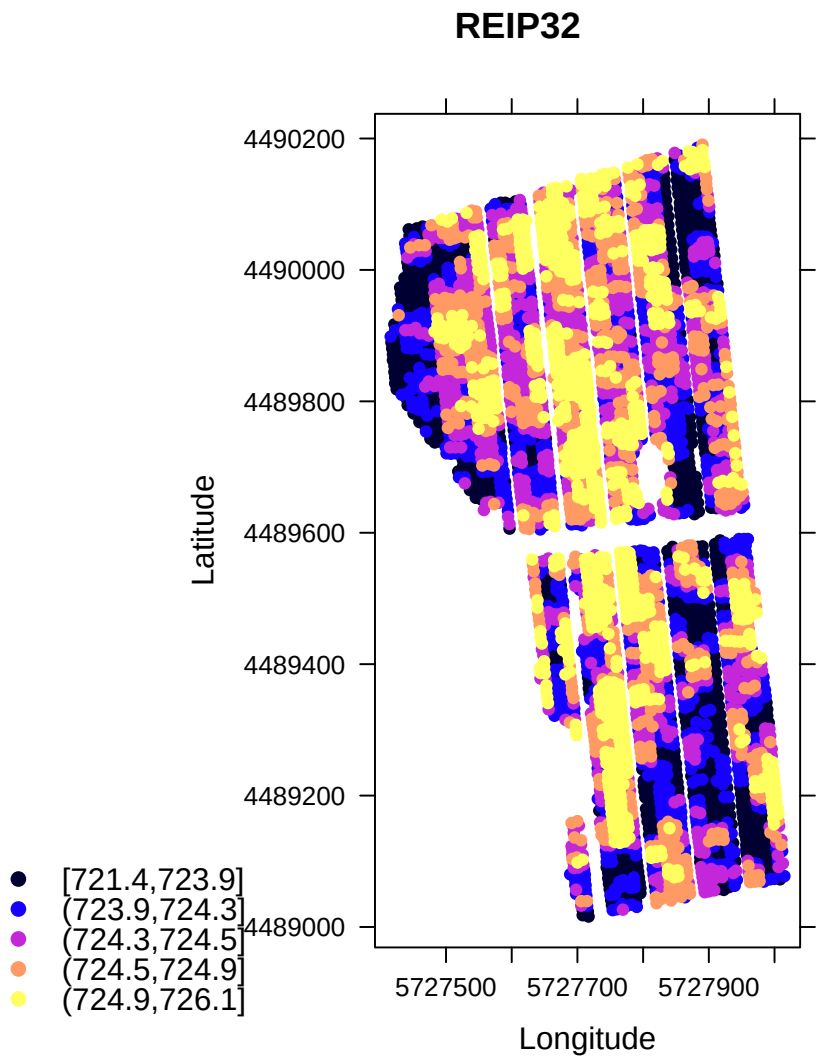
Yield

- measure yield when harvesting
- data from 2003 (previous year) and 2004 (current year)
- two attributes: Yield03, Yield04

Table: Attributes overview

Attr.	**min**	**max**	**mean**	**std**
N1	0	100	57.7	13.5
N2	0	100	39.9	16.4
N3	0	100	38.5	15.3
REIP32	721.1	727.2	725.7	0.64
REIP49	722.4	729.6	728.1	0.65
EM38	17.97	86.45	33.82	5.27
Yield03	1.19	12.38	6.27	1.48
Yield04	6.42	11.37	9.14	0.73

Illustration



Research Questions

- How much does *fertilization* influence current-year yield?
- Is there a correlation between data attributes that influences yield?
- How well can modeling techniques predict current year's yield?
- Which predictor model would be best suited to the agriculture data?

$\longrightarrow$ find suitable *regression* models.

Advanced Regression Techniques

- yield prediction: task of multi-dimensional regression
- establish suitable regression techniques (current work)
- evaluate those techniques (current work)
- use those techniques (future work)

Regression Task

$$T = \{\{x_1, \ldots, x_n\}, y_i\}_{i=1}^{N} \tag{1}$$

- approximate underlying function via cross-validation learning
  - train models on input/output-combinations (supervised learning)
  - estimate residuals on test set
  - error measure: rmse, mae

Regression Techniques

- Multi-Layer Perceptron Networks
- Radial Basis Function Networks
- Regression Tree
- Support Vector Regression

$\longrightarrow$ estimate approximation quality for models and compare models

Error Measures

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=j}^{n}(y_i - y_{a,i})^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_{a,i}| \quad (3)$$

Model Parameter Estimation

- evaluate and fine-tune models on one data set (F04)
- run models with the same parameter settings on the remaining data sets (F330, F131)
- check whether the model ranking is the same
- (model parameters in the paper)

Modeling Results / Error Values

Error Measure / Model	F04		F131	F131net	F330
MAE MLP:	0.3706		0.2468	0.2300	0.3576
RMSE MLP:	0.4784		0.3278	0.3073	0.5020
MAE RBF:	0.3838		0.2466	0.2404	0.3356
RMSE RBF:	0.5031		0.3318	0.3205	**0.4657**
MAE REGTREE:	0.4380		0.2823	0.2530	0.4151
RMSE REGTREE:	0.5724		0.3886	0.3530	0.6014
MAE SVR:	**0.3446**		**0.2237**	**0.2082**	**0.3260**
RMSE SVR:	**0.4508**		**0.3009**	**0.2743**	0.4746

Table: Results of running different models on different data sets. The best predictive model for each data set is marked in **bold** font.

Results

- SVR best model in terms of error values
- SVR best in terms of computational cost
- so-far reference model MLP can be replaced by SVR
- RegTree performs worst, but may be considered as an alternative with explanatory power
- model parameters can be fine-tuned manually on one data set and carried over to other (comparable) data sets

Future Work

- improve model understandability (support vectors?)
- combine model output (ensemble methods?)
- use models in a feature selection task
- check whether spatial autocorrelation in the data is an issue