Hierarchical Spatial Clustering for Management Zone Delineation

ICPA 2010, Denver, Colorado

Georg Ruß

July 19th, 2010

About me

- German computer scientist
- with interest in (spatial) data mining
- currently using mostly R for spatial data mining
- parts of this talk are going to be my PhD thesis

- last week:"Data Mining in Agriculture" workshop `http://dma2010.de/`
- workshop as a means of bringing together interesting and interested people, not necessarily from agriculture, but rather from the computational, data-driven point of view on precision agriculture

Data Details – Field of Study

Figure: F550 field, depicted on satellite imagery, source: Google Earth

Data Details – Features

- collect a number of geo-coded, high-resolution features such as:
  - N1, N2, N3: nitrogen fertilizer application rates in 2004
  - REIP32, REIP49: vegetation index (red edge inflection point) in 2004
  - Yield: corn yield 2003, winter wheat yield in 2004 and 2007
  - EC25: electrical conductivity of soil in 2004
  - pH, P, K, Mg: soil sampling in 2007
- one field available, 1080 records in $25 \times 25m$-resolution on a hexagonal grid

Data Details – Temporal Aspects

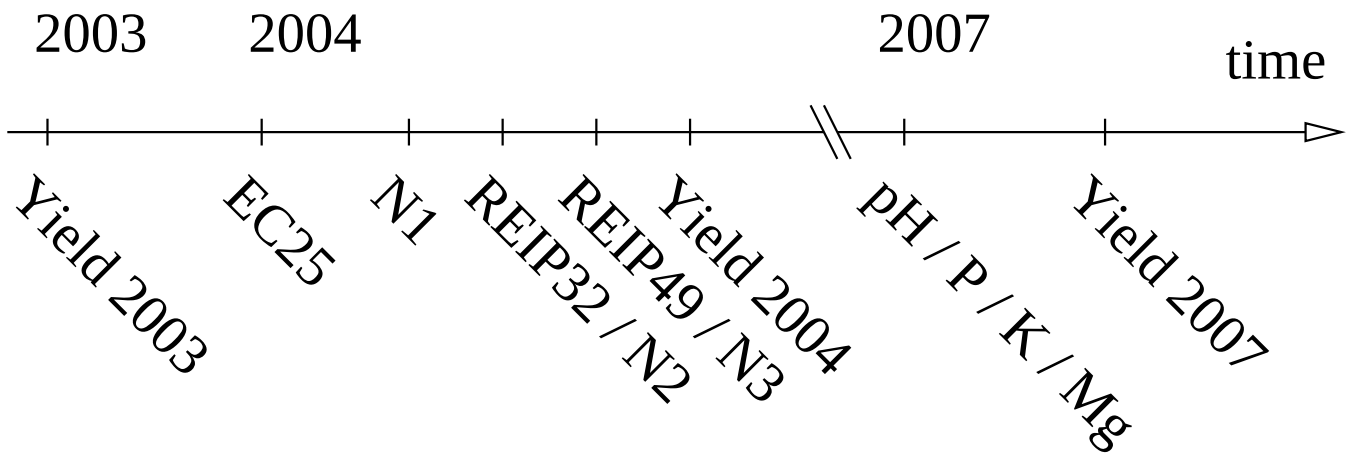
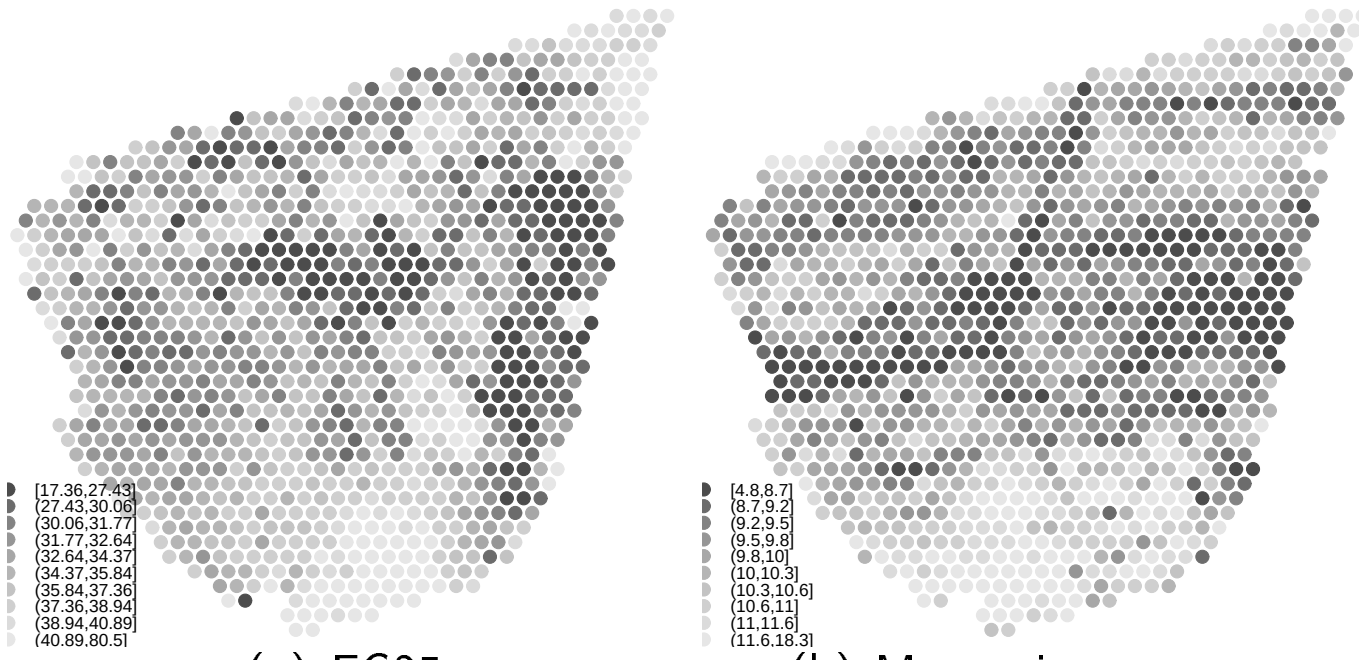


Figure: timeline of data acquisition

Spatial Autocorrelation

Are (spatial) data records independent of each other?
(Do we have spatial autocorrelation?)



(a) EC25 (b) Magnesium content

Figure: F550, EC25 and Magnesium readings

Management Zone Delineation

- A common task in agriculture:
  - subdivide the field into smaller zones
  - zones are rather homogeneous
  - zones are spatially mostly contiguous
  - similarity between zones is low
- from a data mining perspective: *spatial clustering*

Literature Approaches

- mostly non-spatial algorithms are used
  - no spatial contiguity
  - small islands, outliers, etc.
  - black-box models
  - fuzzy c-Means, k-Means, etc.
- spatial contiguity is not always required, but desirable
- spatial autocorrelation is usually neglected rather than exploited

(good summary in “Geostatistical Applications for PA”, chapter 8, see exhibitions, my approach falls into the VIIIth category there, called “modeling”)

Spatial Contiguity Constraint

- spatial clustering = clustering with a spatial contiguity constraint
- $\longrightarrow$ constrained clustering
- Keep it simple and understandable:
  - hierarchical clustering
  - agglomerative clustering
- Idea:
  1. split field into small zones which are homogeneous
  2. iteratively merge these zones obeying similarity and spatial constraint

Spatial Tessellation

- k-Means clustering on the data points' coordinates
  - due to spatial autocorrelation, adjacent points are likely to be similar
  - this ensures homogeneity of these small zones
  - k is user-controllable and easy to understand
    - homogeneous field: smaller k
    - heterogeneous field: higher k
- much more flexible than grid-based approaches

Spatial Tessellation

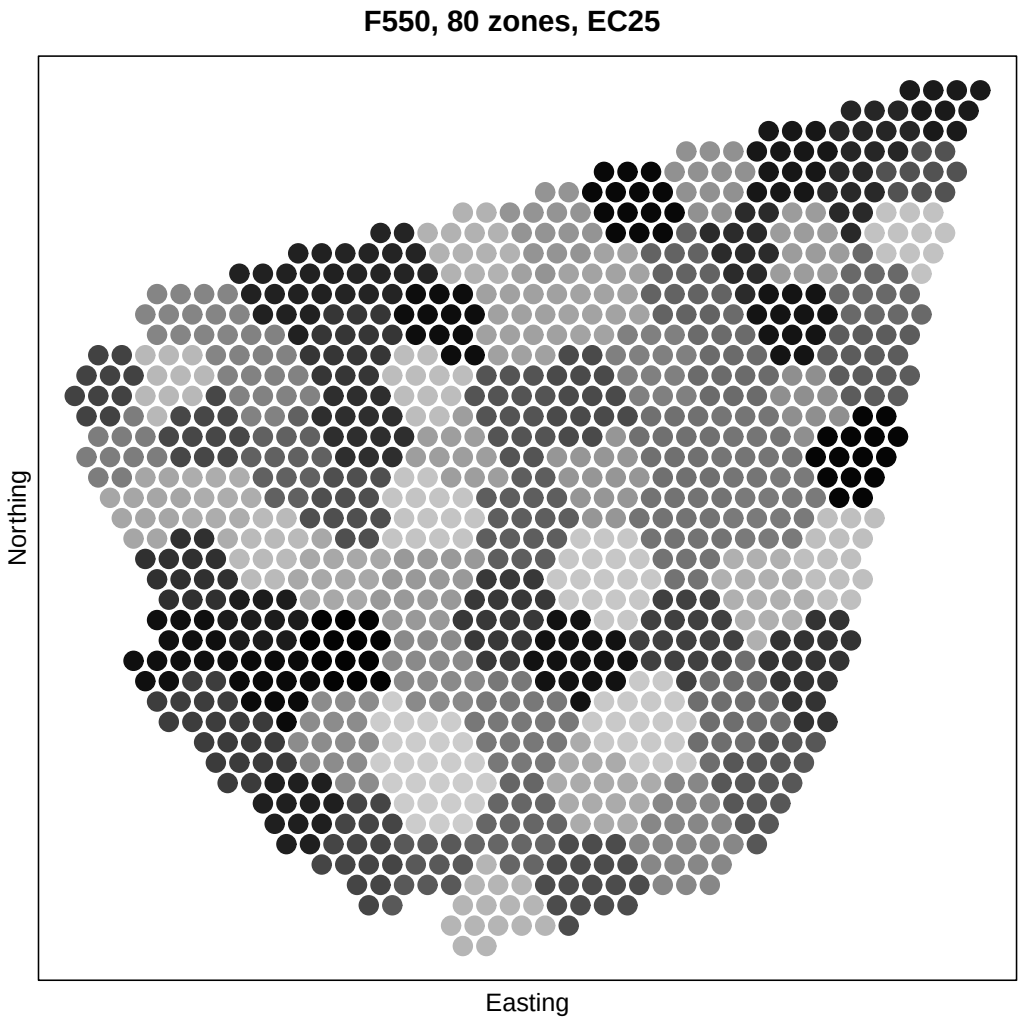
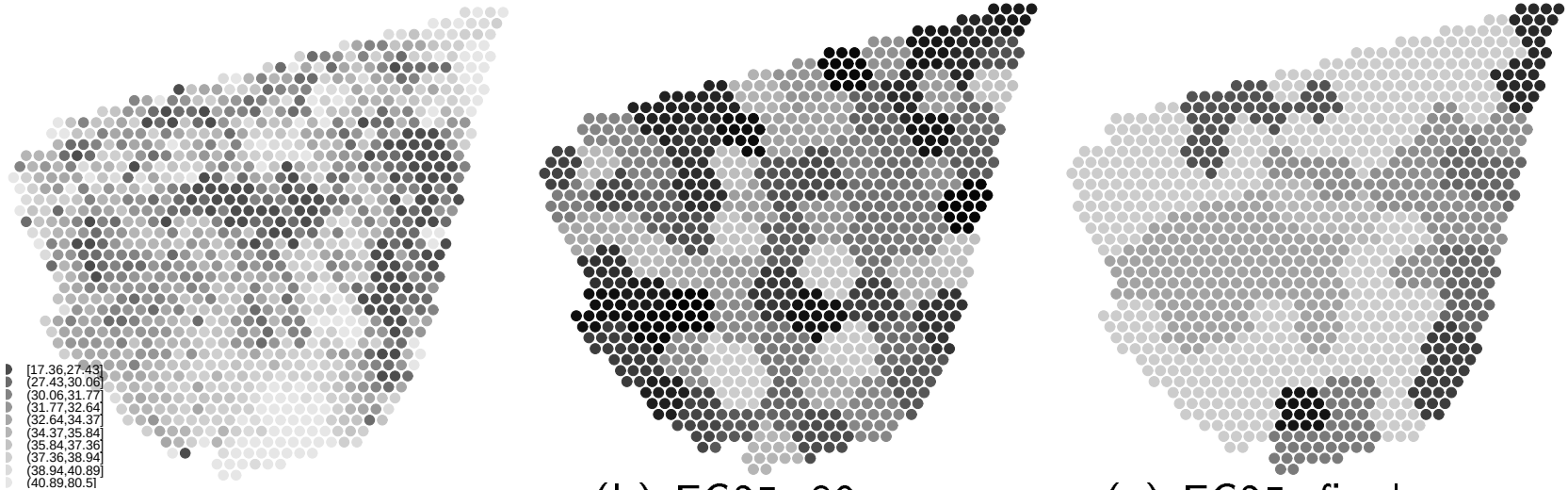


Figure: Tessellation of F550 using k-means, $k = 80$ (grey shades are for illustration only, no further meaning here)

Hierarchical Agglomerative Constrained Clustering

- principle: merge only adjacent zones, if they are similar enough
  - this ensures spatial contiguity
  - $\rightarrow$ spatial constraint, non-adjacent zones *cannot link*
- once non-adjacent zones become much more similar than adjacent ones, they may be merged
  - introduce a user-controllable *contiguity factor cf*
  - $cf \geq 2$: high contiguity
  - $cf \in [1, 2]$: low contiguity
  - $cf \leq 1$: no contiguity

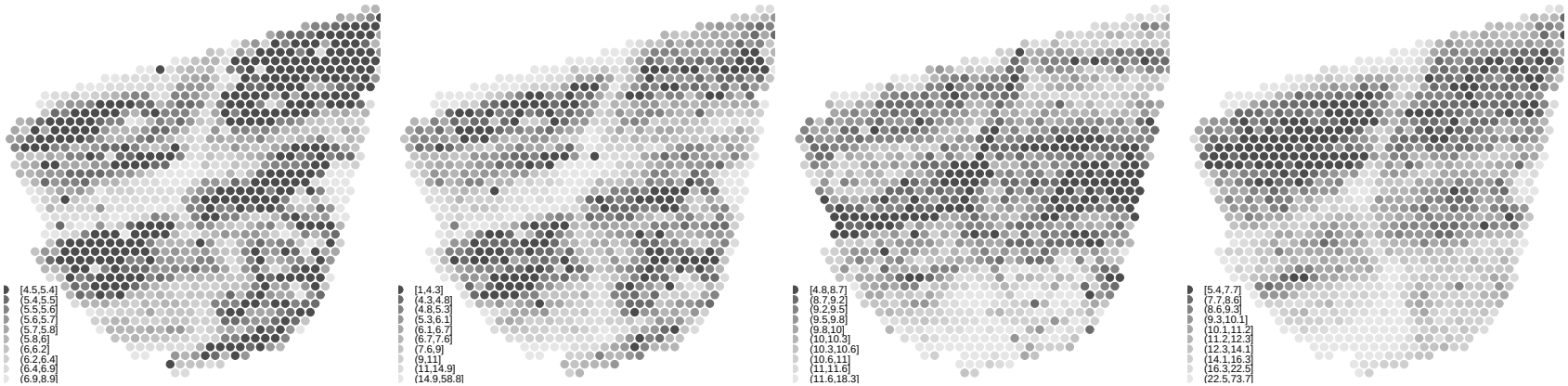
HACC – 1D example



(a) EC25 plot (b) EC25, 80 zones (c) EC25, final zones

Figure: F550, EC25 clustering

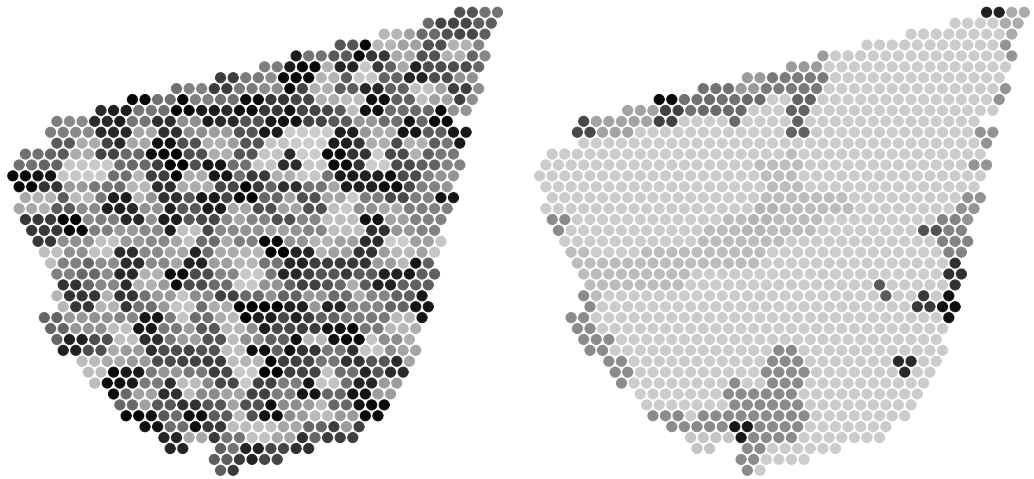
HACC – 4D example



(a) pH plot (b) P plot (c) Mg plot (d) K plot

Figure: F550, four attributes

HACC – 4D example (cont.)



(a) F550-4D, beginning (b) F550-4D, ten zones

Figure: F550, management zones

- actually, 3 zones (when comparing attribute values)
  - low pH, low P, low Mg, low K (largest zone)
  - high pH, high P, high Mg, high K (border zones)
  - high pH, high P, low Mg, high K (middle, from left)

Summary

- precision agriculture as a data-driven approach
- spatial, geo-referenced data records in large amounts
- management zone delineation solved as a spatial clustering approach
- from a computer scientist's point of view: important difference between spatial and non-spatial data treatment $\Rightarrow$ use models which are fit for spatial tasks

Time for . . .

Questions?

Next Workshop *Data Mining in Agriculture* likely in 2011 (NYC)

- contact: `georg.russ@ieee.org`
- slides, R scripts and further info at `http://research.georgruss.de`