HACC-spatial: Hierarchical Agglomerative Spatially Constrained Clustering

Georg Ruß

Otto-von-Guericke-Universität Magdeburg, Germany
russ@dma-workshop.de

DMA Workshop, Sep 3rd, 2011
Precision Agriculture

- GPS technology used in site-specific, sensor-based crop management
- combination of agriculture and information technology
- data-driven approach to agriculture
- lots of data analysis tasks
Data Details – Example Field

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 $25 \times 25 \text{m}$-resolution on a hexagonal grid
Data Details – Temporal Aspects

![timeline of data acquisition](image)

**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

- → 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
Spatial Contiguity Constraint

- spatial clustering = clustering with a spatial contiguity constraint
- → constrained clustering
- Keep it simple and understandable:
  - hierarchical clustering
  - agglomerative clustering
- Idea:
  1. (optionally) split field into small zones which are homogeneous
  2. iteratively merge clusters obeying similarity and spatial constraint
Optional 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 $k$
Optional 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 objects/clusters, if they are similar enough
  - this ensures spatial contiguity
  - → spatial constraint, non-adjacent clusters cannot link

- once non-adjacent clusters 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
Plots for different predictor variables

(a) F631: EC25

(b) F610: EC25

(c) F440: REIP32
F631, EC25 clustering (low/high spat. contig.)

(a) F631, EC25, 50 clusters

(b) F631, EC25, 50 clusters

(c) F631, EC25, 30 clusters

(d) F631, EC25, 30 clusters
F610, EC25, tolerance against missing data

Figure: HACC-spatial on F610 using EC25
F440, different contiguity settings (low to high)
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
- important difference between spatial and non-spatial data treatment ⇒ use models which are fit for spatial tasks
Time for . . .

Questions?

Next Workshop *Data Mining in Agriculture* in 2012:
http://dma-workshop.de

▶ contact: russ@dma-workshop.de
▶ slides, R scripts and further info at
http://research.georgruss.de
Survey on “Data Mining in Agriculture”

- Third paper in this workshop
- by Antonio Mucherino, author of the “Data Mining in Agriculture” book (Springer, 2009)
Survey

- mainly about Antonio’s ...
  - biclustering on wine fermentation data
- ... and my work:
  - yield prediction
Wine fermentation

- measure metabolites:
  - glucose
  - fructose
  - organic acids
  - glycerol
  - ethanol . . .

- Try to predict problematic fermentations from the above variables
  - cluster known fermentations (normal, slow, stuck), assign score
  - for new fermentations: find best cluster and predict outcome
  - obtain a score for the probability of a fermentation to become problematic
Yield prediction

► collect spatial, high-resolution data
  ► vegetation indices
  ► fertilizer data
  ► previous yields
  ► sensor data
  ► digital elevation model

► (try to) predict yield
  ► regression task
  ► use different regression models
  ► develop spatial regression
  ► as a basis for: assessing a variable’s importance for yield prediction
  ► → spatial variable importance
  ► (ongoing part of my PhD thesis)
Other topics

- automatic recognition and grading of fruit (data mining and image processing)
- detection and analysis of animal sounds (data mining and audio signal processing)
- classification of flower species
- estimation of soil properties and soil types
- disease outbreaks, water consumption
- ...