Parallel Coordinates for Exploring Properties of Subsets

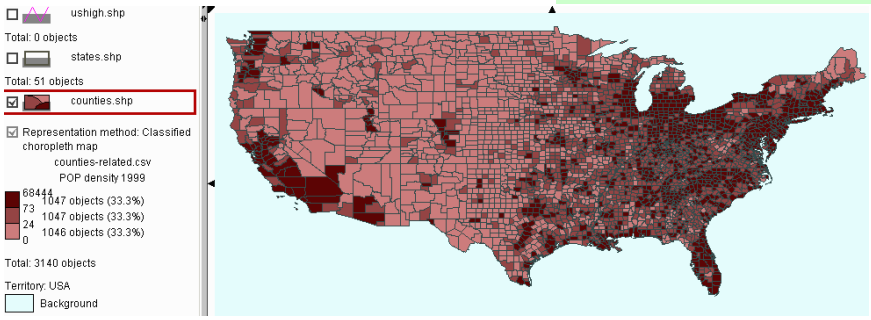
Gennady & Natalia Andrienko
Fraunhofer Institute AIS
Sankt Augustin, Germany

http://www.ais.fhg.de/and

CMVEV conference, London, 13.7.2004

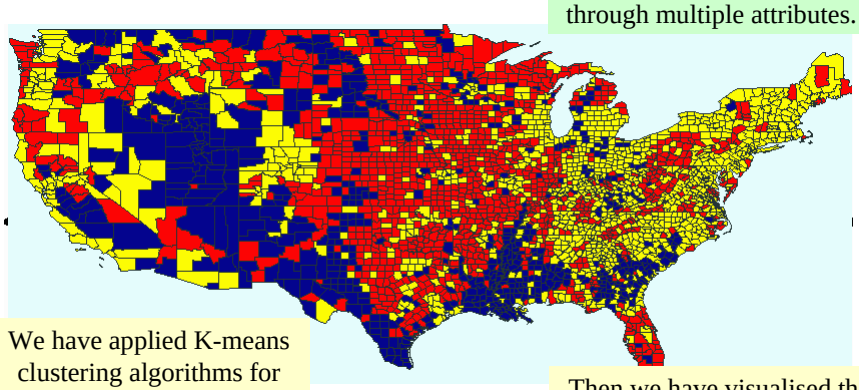
Classification on a map

Goal: group objects by similar values of selected attribute(s).



Cluster Analysis

Goal: group objects by similarity of their characteristics expressed through multiple attributes.



We have applied K-means clustering algorithms for grouping the rectangles according to the proportions of various age groups.

Then we have visualised the results on a map by assigning colours to the groups and painting the areas in the corresponding colours.

Analysis questions

- What are sizes and attribute characteristics of particular subsets
- How subsets are similar / different
- What instances are the “typical” representatives of subsets
- What are the outliers
- What is the spatial distribution of subsets and specific instances, how they relate to other attributes

Requirements

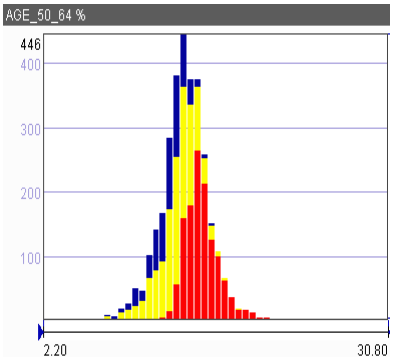
1. Support for large (>>1000 objects) data sets
2. Inform about the distribution of characteristics of the entire set and subsets, and of individual instances: elementary, intermediate, and overall reading levels (Bertin)
3. Flexibility: various classification tools, dynamic re-classification

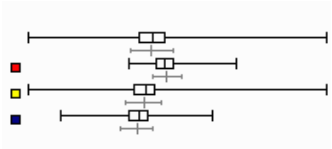
Usual approach

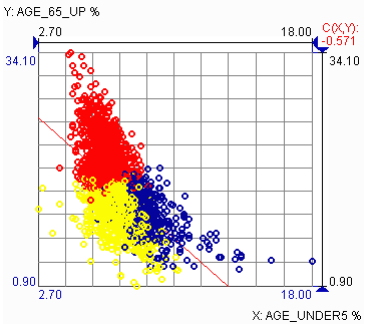
Coordination of multiple displays via

- Data update coordination
- Brushing
- Multiplication of displays

Visual techniques for studying subsets

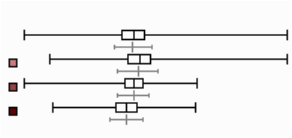


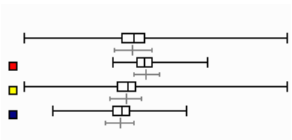




General statistics Vs. individual instances

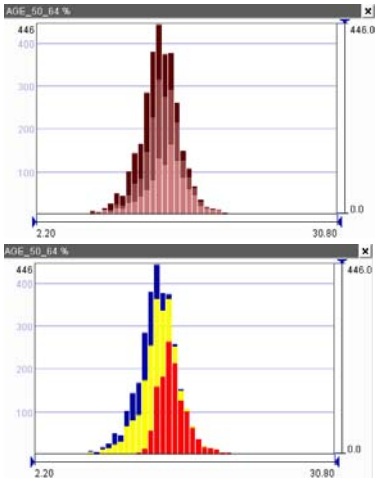
Box plots





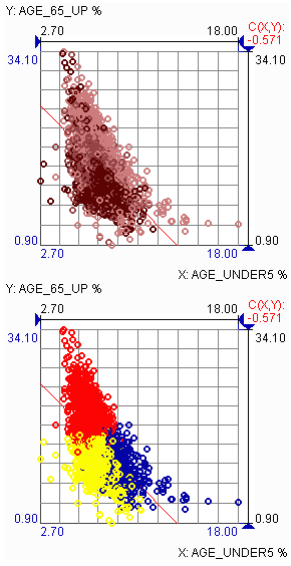
- Support only overall and intermediate levels (no access to individual instances)
- Too coarse aggregation
- Multiple plots are needed for several attributes

Histograms



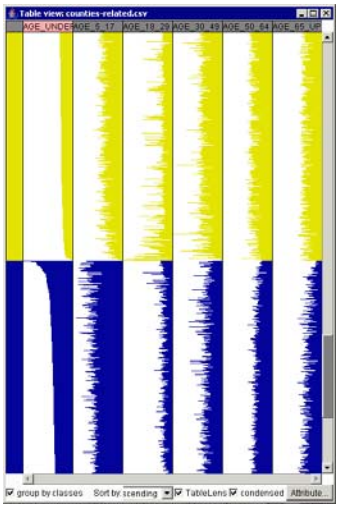
- Shapes depend on granularity
- Support comparison of a class to the entire set, but not comparison of classes
- No access to individual instances
- Need of zooming for seeing details

Scatter-plots



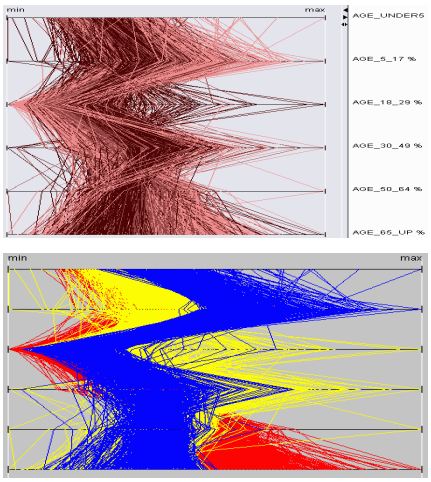
- Overplotting!
- Data binning could help, but it excludes the possibility to represent several classes within a single display

Table lens



- Only small data sets (<1000 instances) are supported

Parallel Coordinates Plot

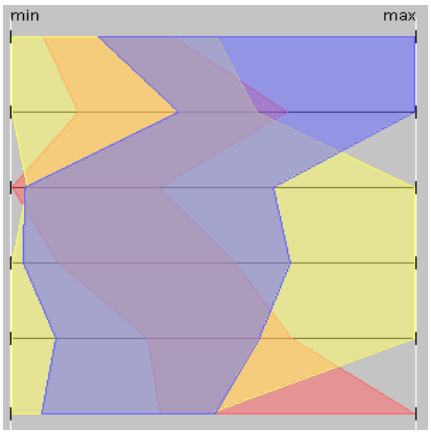


- (N-1) 2D projections
- Overplotting
- Difficult identification

<u>Approaches:</u>

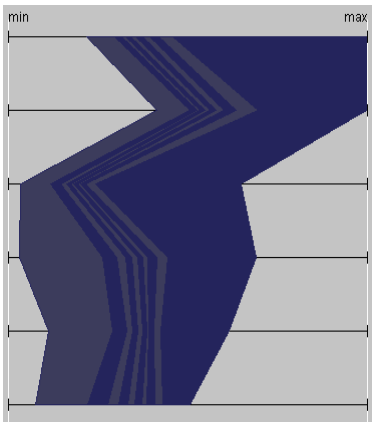
1. Average/median lines
2. Envelopes
3. Line densities
4. Transparency

Overview of subsets - 1



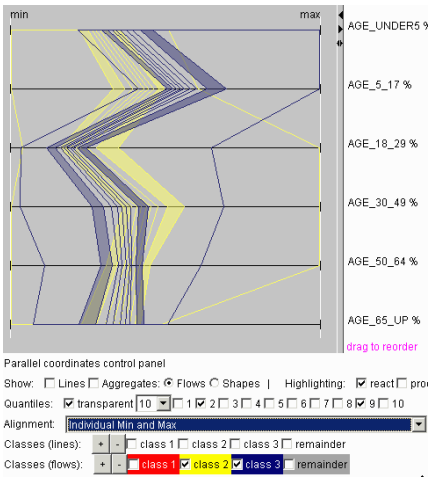
Outlines of clusters

Overview of subsets - 2

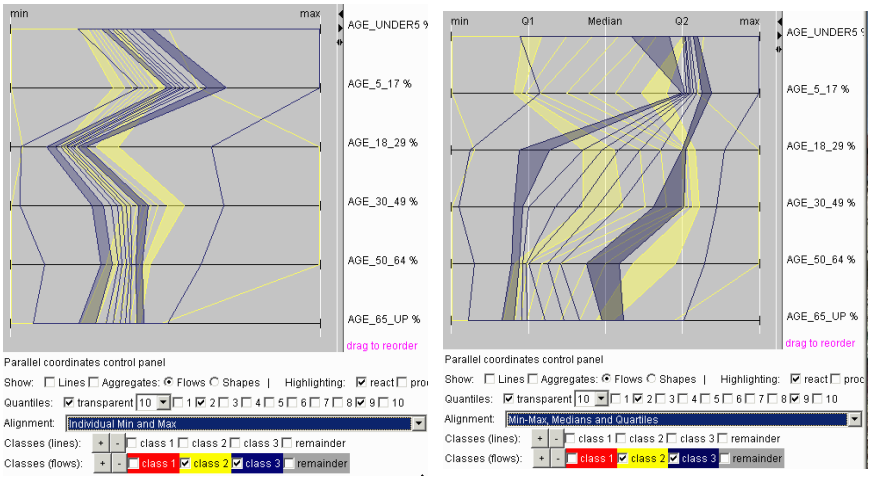


More detail for a selected subset

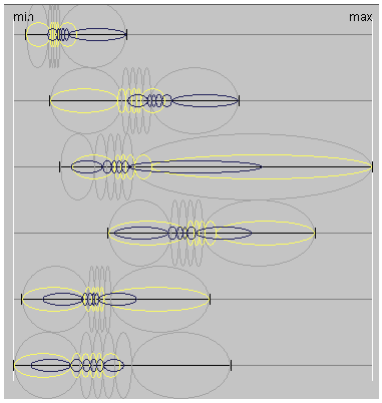
Compare two subsets - 1

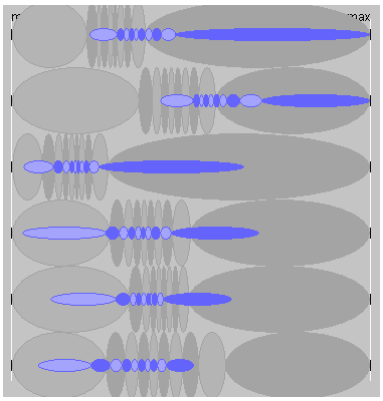


Compare two subsets - 2



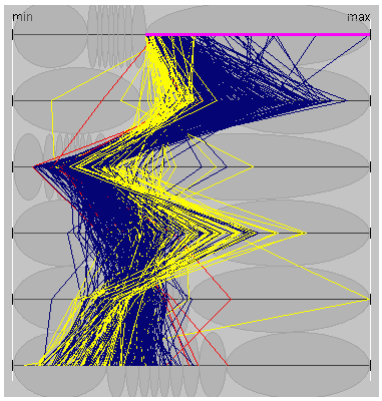
Alternative representation

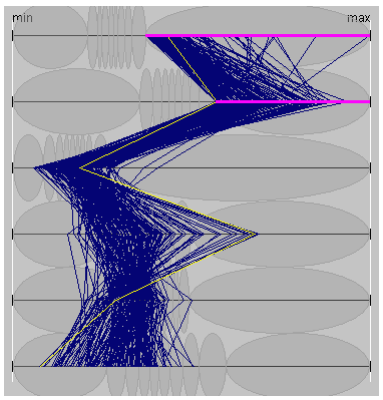




Cluster statistics combined with cluster sizes

Selection of lines with specific properties





Questions and tools

- What are sizes and attribute characteristics of particular subsets
- How subsets are similar / different
- What instances are the “typical” representatives of subsets
- What are the outliers
- What is the spatial distribution of subsets and specific instances

- Subset outlines
- Details for selected subsets, visual overlapping
- Two drawing modes: stripes and ellipses
- Statistics-based scaling of PCP
- Interactive procedure for selecting instances with specific characteristics
- Dynamic linking to other displays and maps

Conclusions - PCP

- Aggregated representation combined with easy access to individual instances; dynamic linking to other displays & maps
- Scalable approach: complexity is M*N*log(N)
- It is possible to implement the method in a client-server mode so that complete data loading is avoided

Conclusions - General

- Straightforward approaches to coordination (brushing etc.) often fail when dealing with large and very large data sets
- Further work on adapting various displays is needed
- Special attention to time-series data & time plots
- Combining visualisation with computations (statistics, data mining)

Further information:

demos, papers, projects, other methods

http://www.ais.fhg.de/and

and

gennady.andrienko@ais.fraunhofer.de

Thank you!