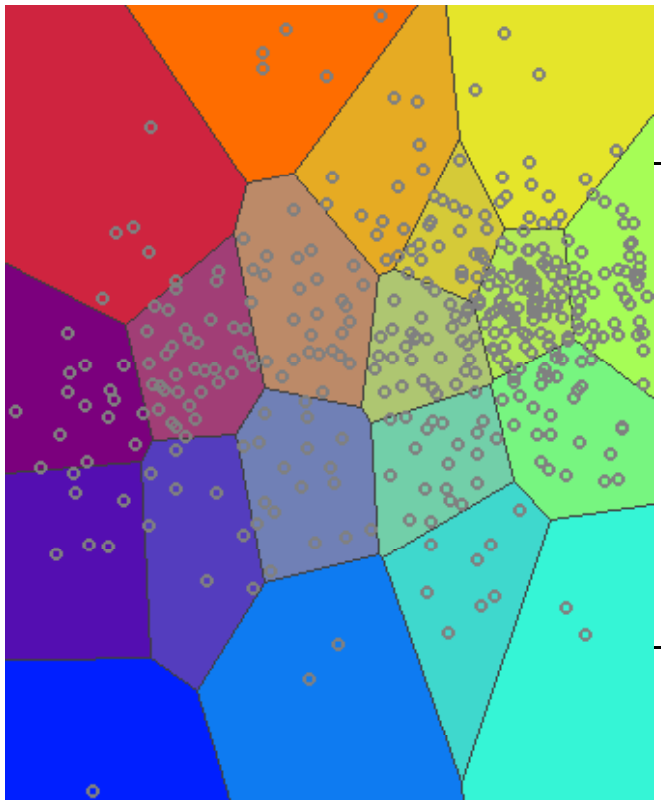
Sammon's Projection for Clustering Complex Geographical Objects



Fraunhofer IAIS

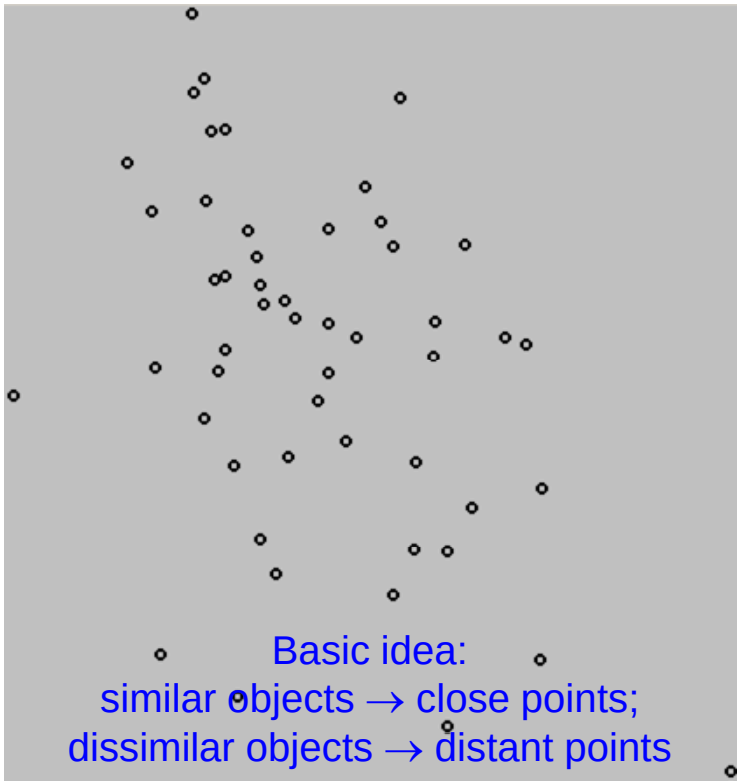
Gennady Andrienko & Natalia Andrienko

http://geoanalytics.net

Projection

Set of objects

Projection space



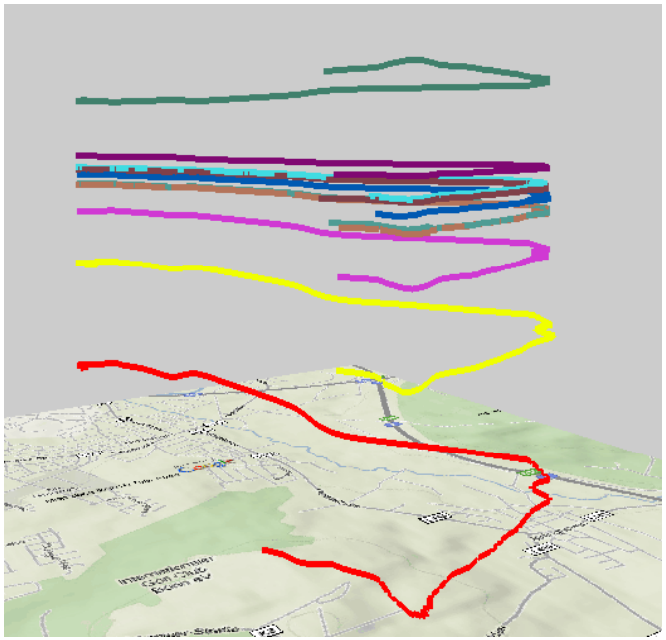
This requires a numeric measure of (dis)similarity

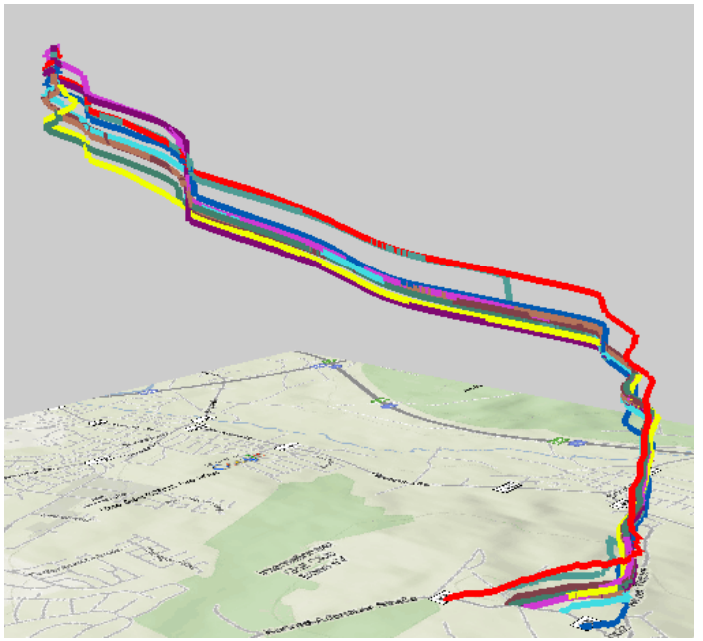
Measuring (dis)similarity

- Purpose
- Material
- Colour
- Size
- Shape
- Weight
- Producer
- …

- Approach 1: use feature vectors
  - Describe objects by values of N numeric attributes (features) chosen according to the analysis goals
  - Feature vector == list of N attribute values
  - Dissimilarity == distance between the vectors in the N-dimensional abstract space of the possible combinations of the attribute values (e.g. Euclidean distance)
- However, objects may have complex properties that cannot be adequately represented by numeric features
  - Approach 2: devise an ad-hoc *distance function*
    - i.e. algorithm to measure dissimilarity

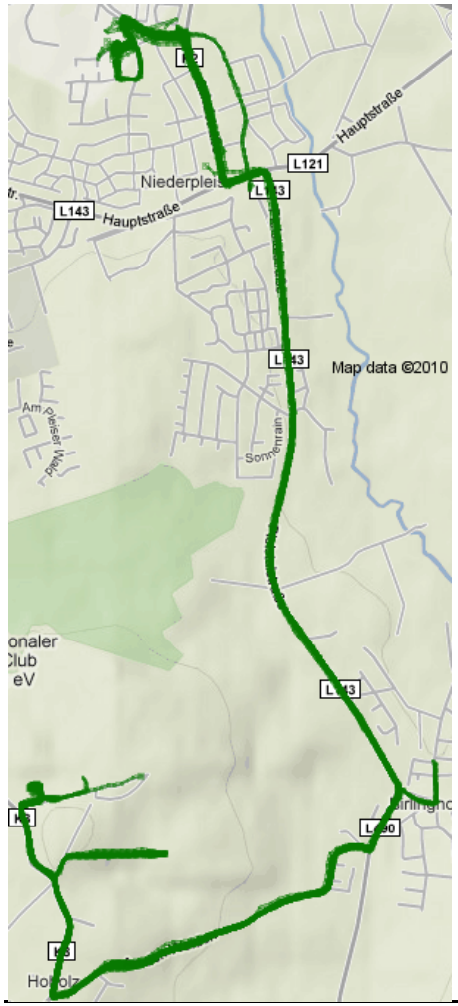
Example of complex geographical objects: trajectories





- Trajectory: sequence (t_1,s_1), (t_2,s_2), … (t_i – time moment, s_i – position in space)
- Similarity: routes (ignoring time), dynamics (relative time), coincidence (absolute time) – no adequate representation by feature vectors
- Complexities: different sequence lengths, different time spacing, measurement errors and gaps, …

Example of specific distance function: route similarity



- Finds *corresponding* points in two trajectories
- Computes the average distance between the corresponding points
- Accumulates the length of the corresponding parts
- Accumulates the deviations of non-corresponding points
- Penalty factor = (cumulative deviation) / (corresponding length)
- Penalty distance = (cumulative deviation) * (penalty factor)
- Final distance = average distance + penalty distance

Projection techniques

- Principal components analysis (PCA)
- Self-organizing map (SOM) – projection onto a discrete space (regular grid)

➢ Require input in form of feature vectors

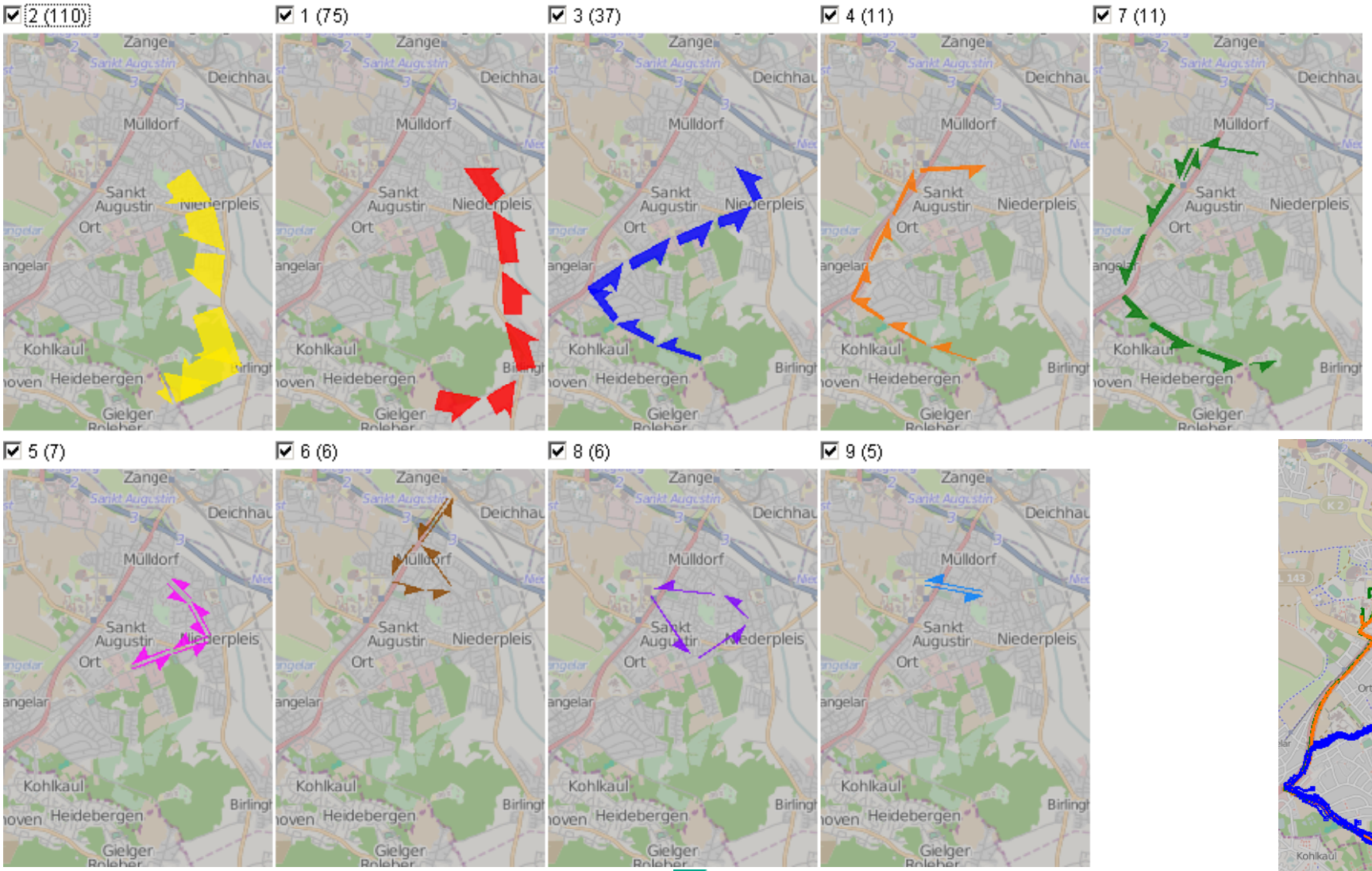
- Multi-dimensional scaling (MDS)
- Sammon's projection (a.k.a. Sammon's mapping)

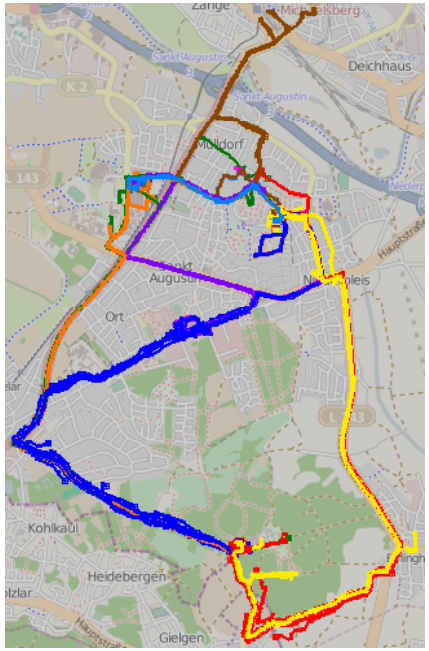
➢ Can be applied to a pre-defined *distance matrix*, which can be computed using an arbitrary distance function

⇒ Suitable for complex objects requiring specific distance functions

Use of projection to explore clustering results

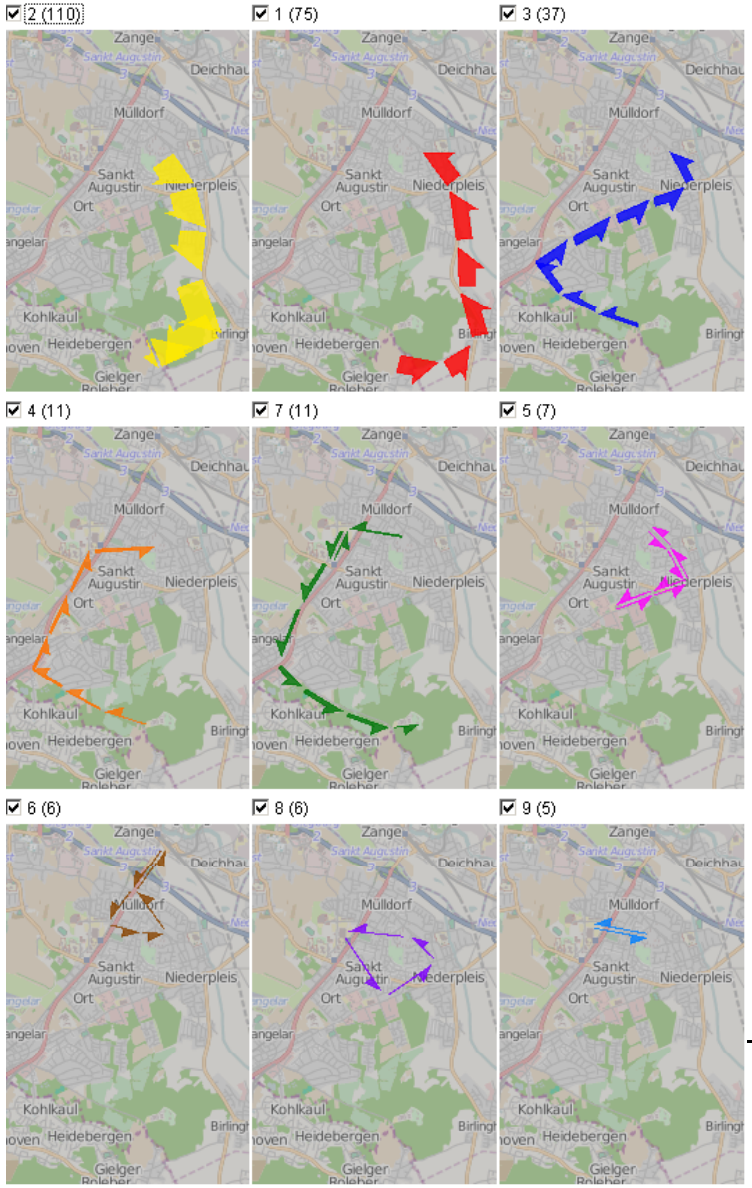
Density-based clusters of trajectories by route similarity





Fraunhofer IAIS

Exploring clustering results



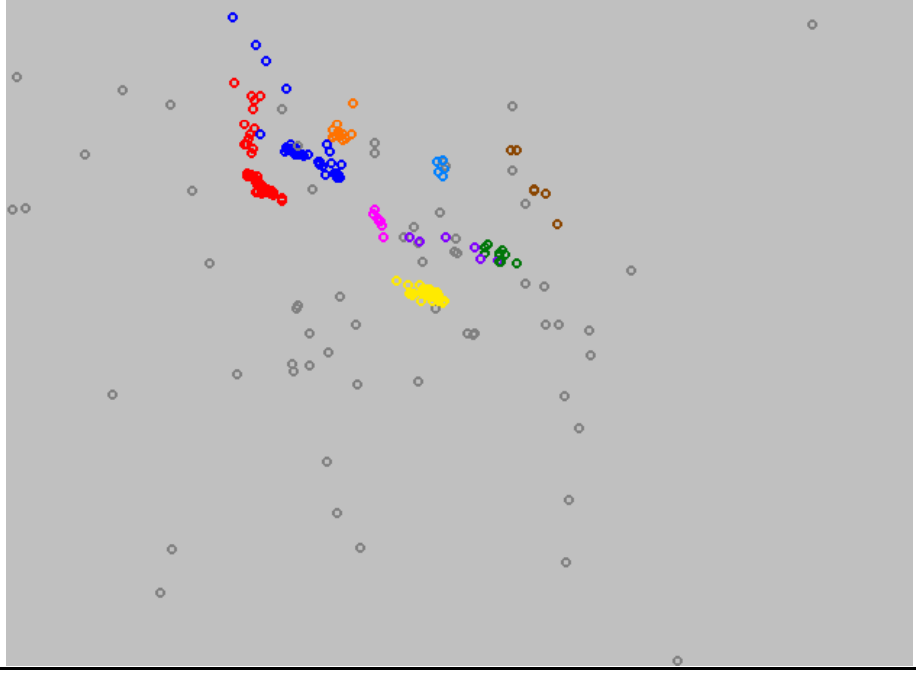
Are the clusters well-separated?

Are they compact?

How different are the clusters?

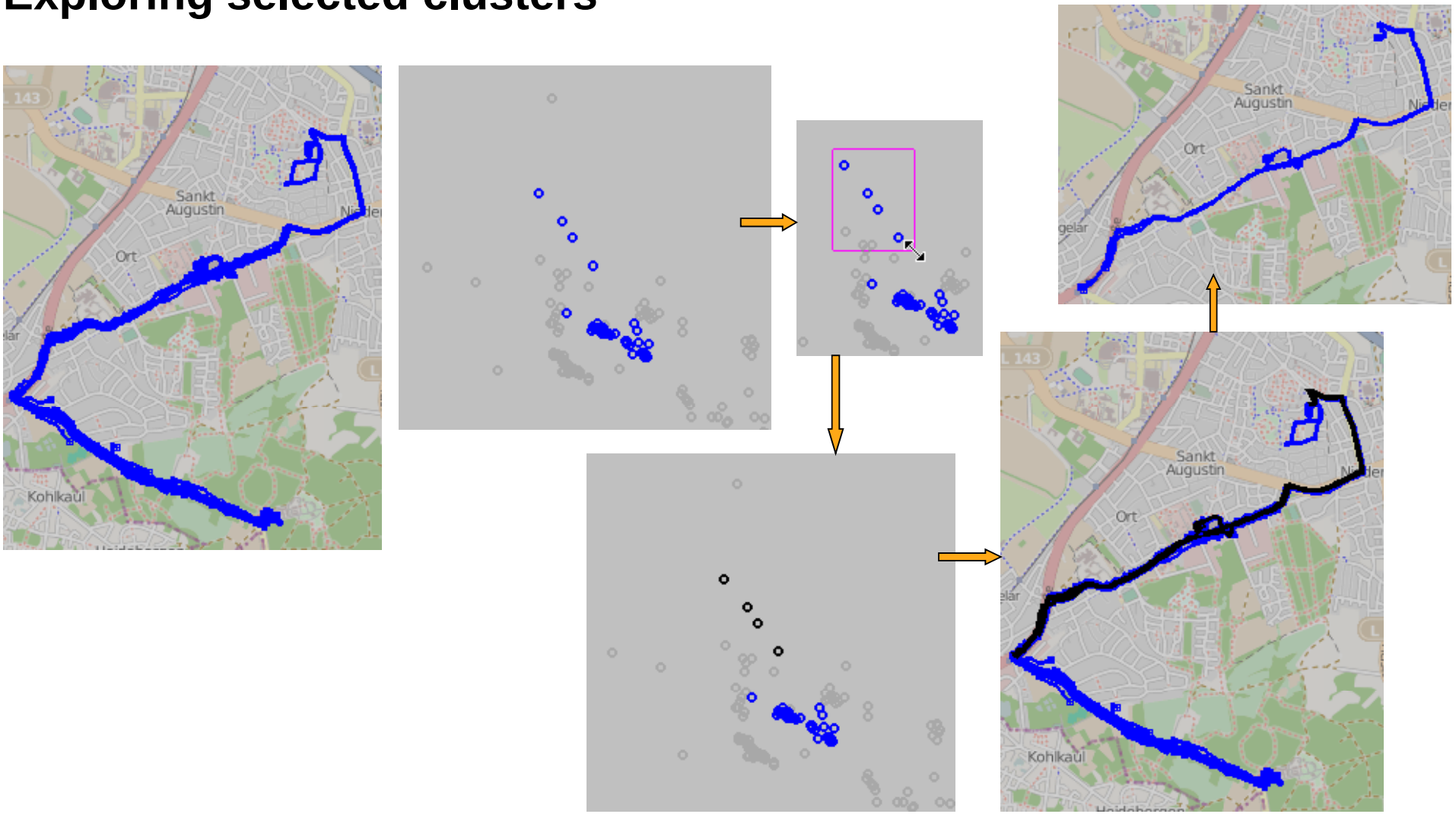
How far are the clusters from the noise?

How sensitive are the results to the parameters?



Fraunhofer IAIS

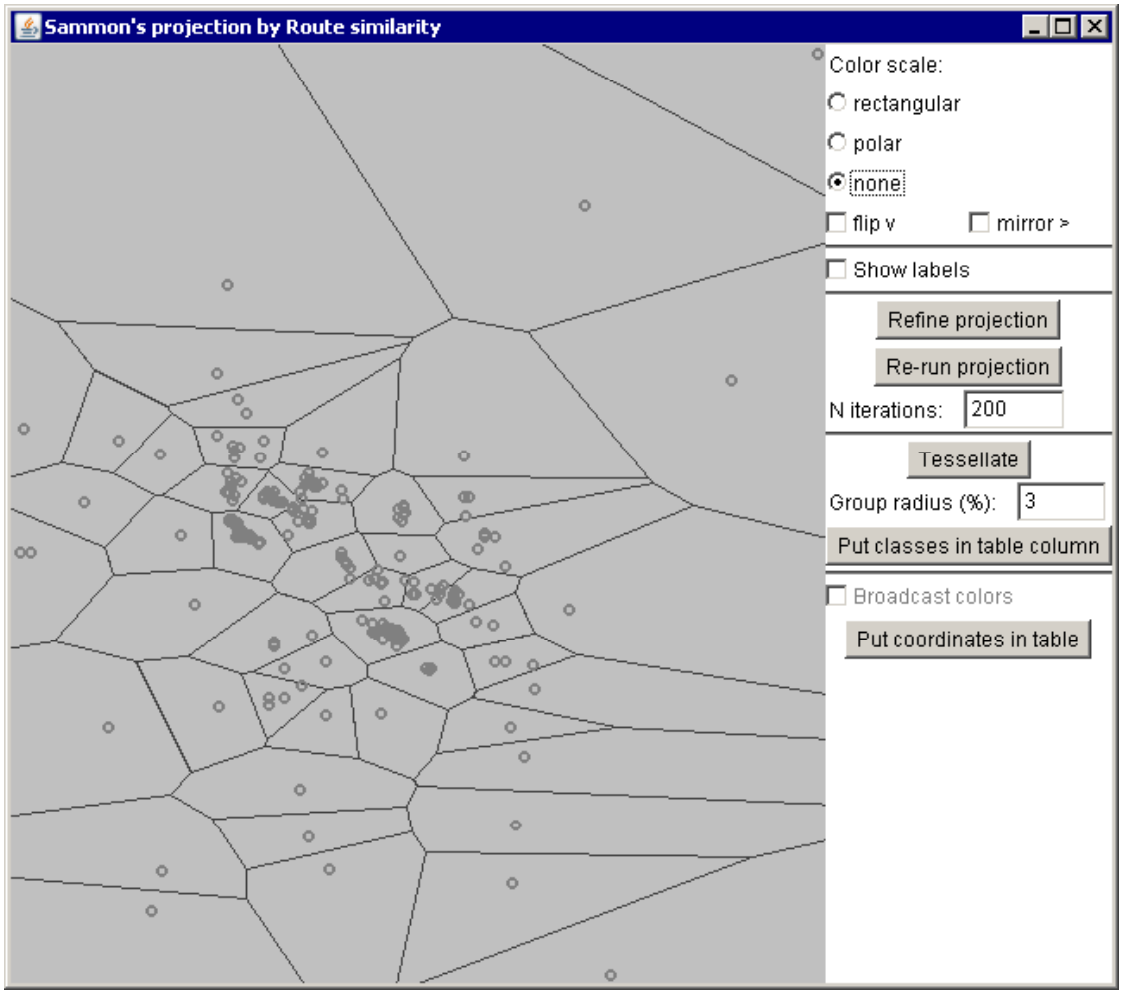
Exploring selected clusters



Fraunhofer IAIS

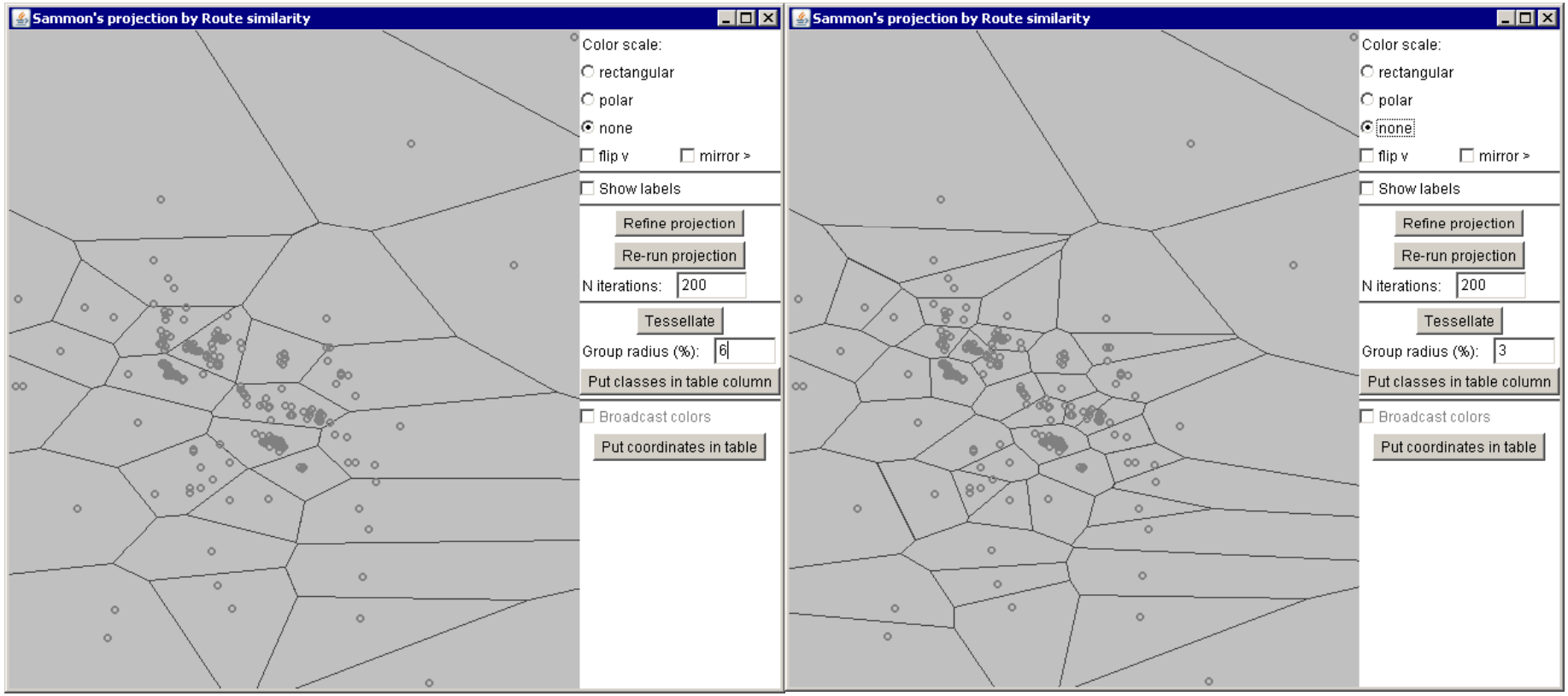
Use of projection to define clusters

Tessellation of the projection area



- Each polygon defines a cluster

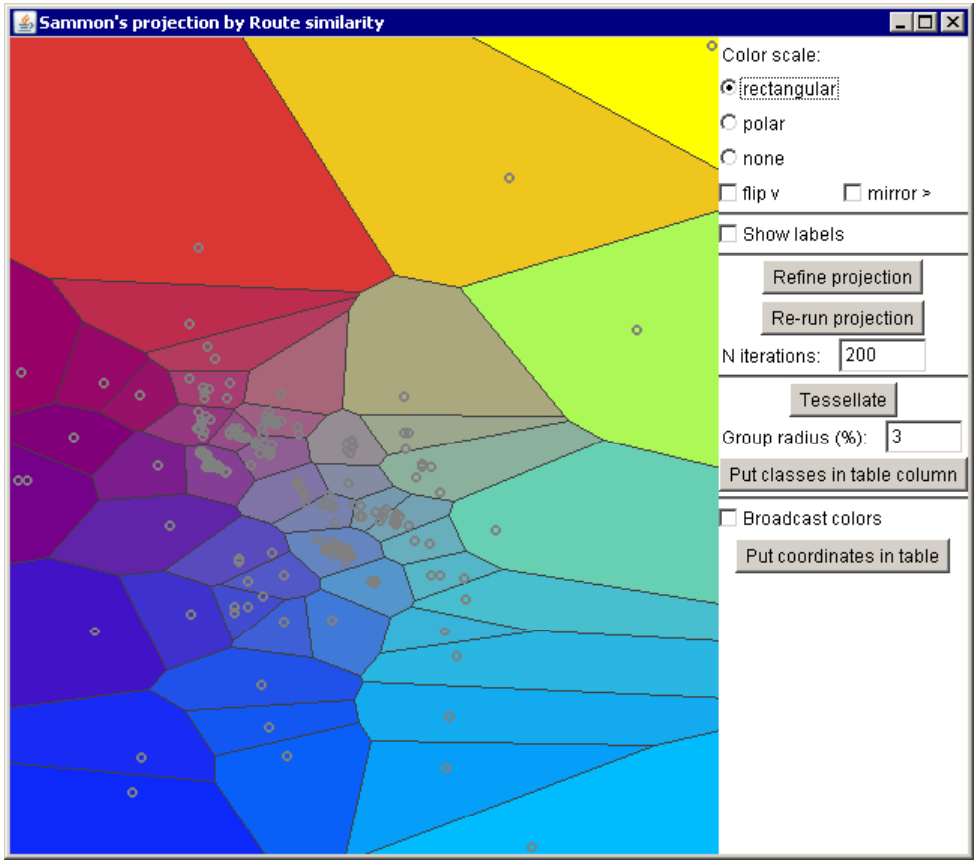
Choosing cluster sizes

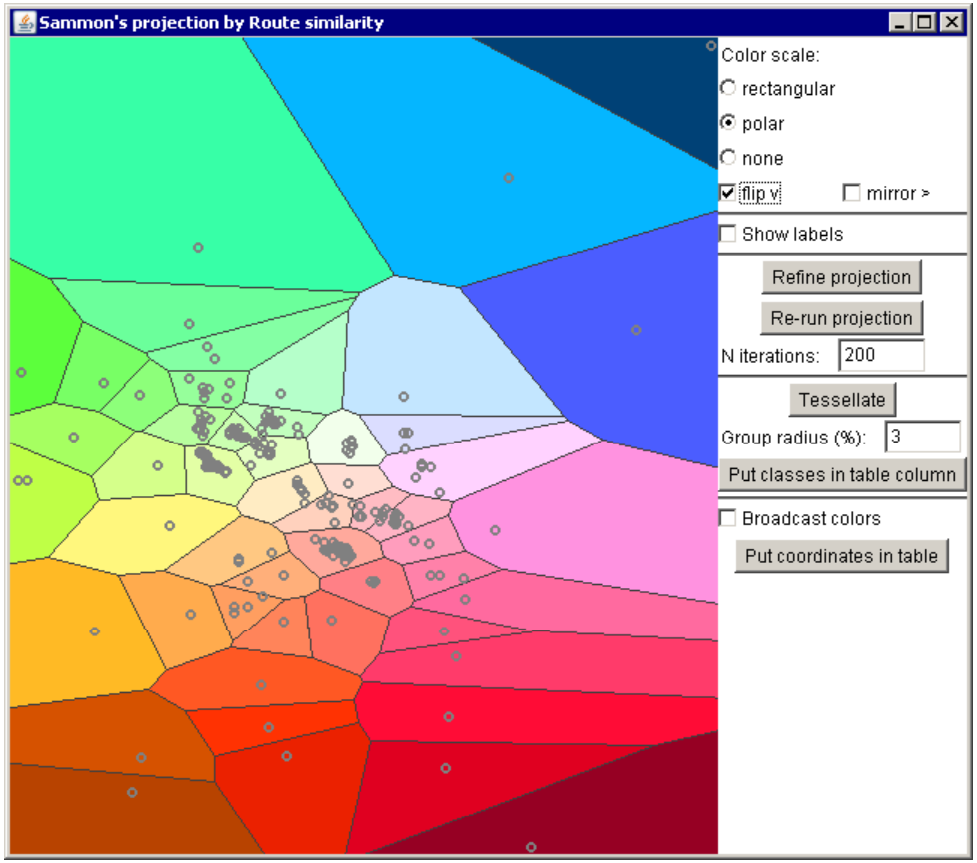


The user may interactively change the sizes (radii) of the clusters.

Further possible extension: interactive refinement or joining of selected clusters by direct manipulation in the projection display.

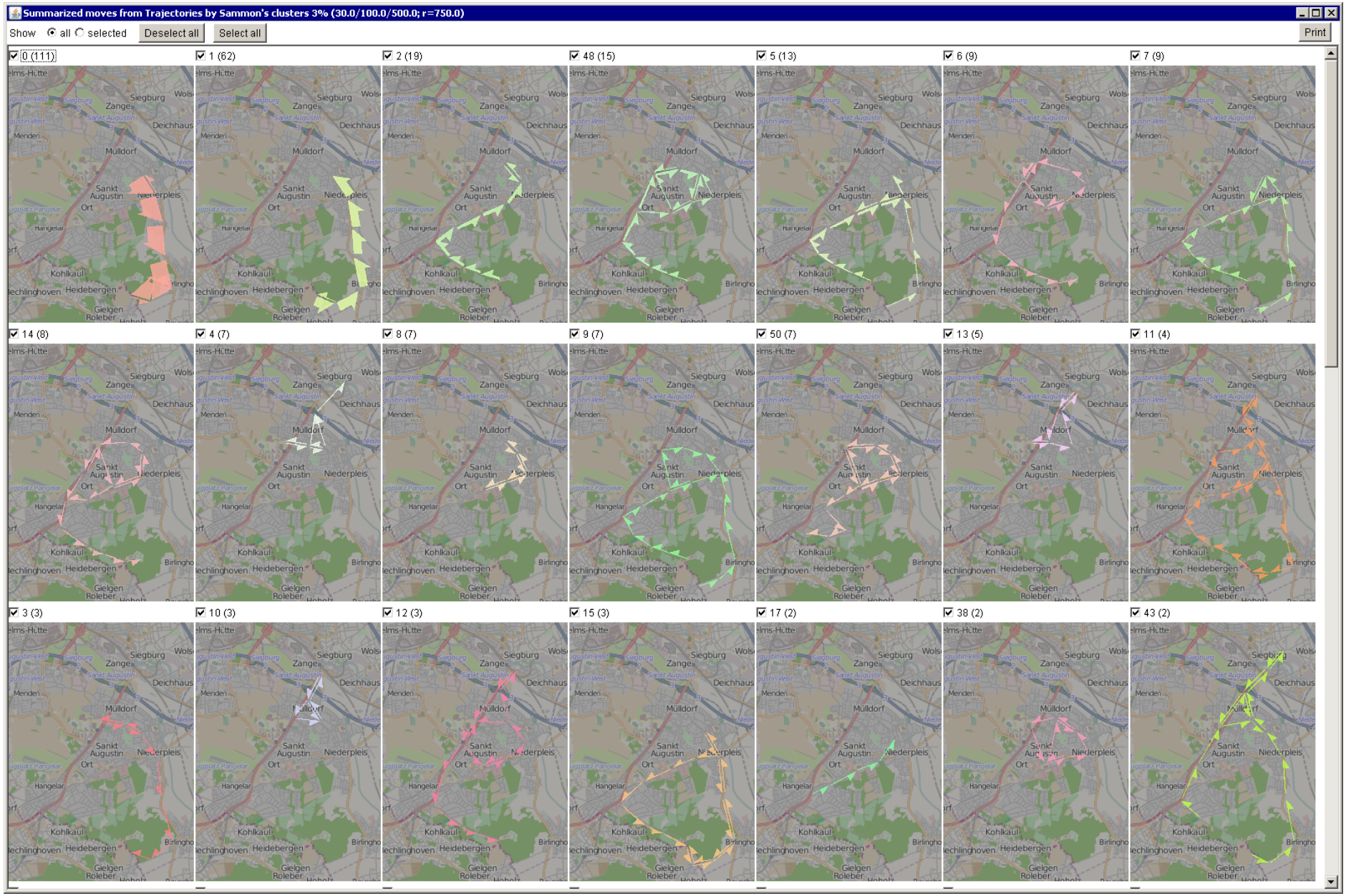
Assigning colours to clusters



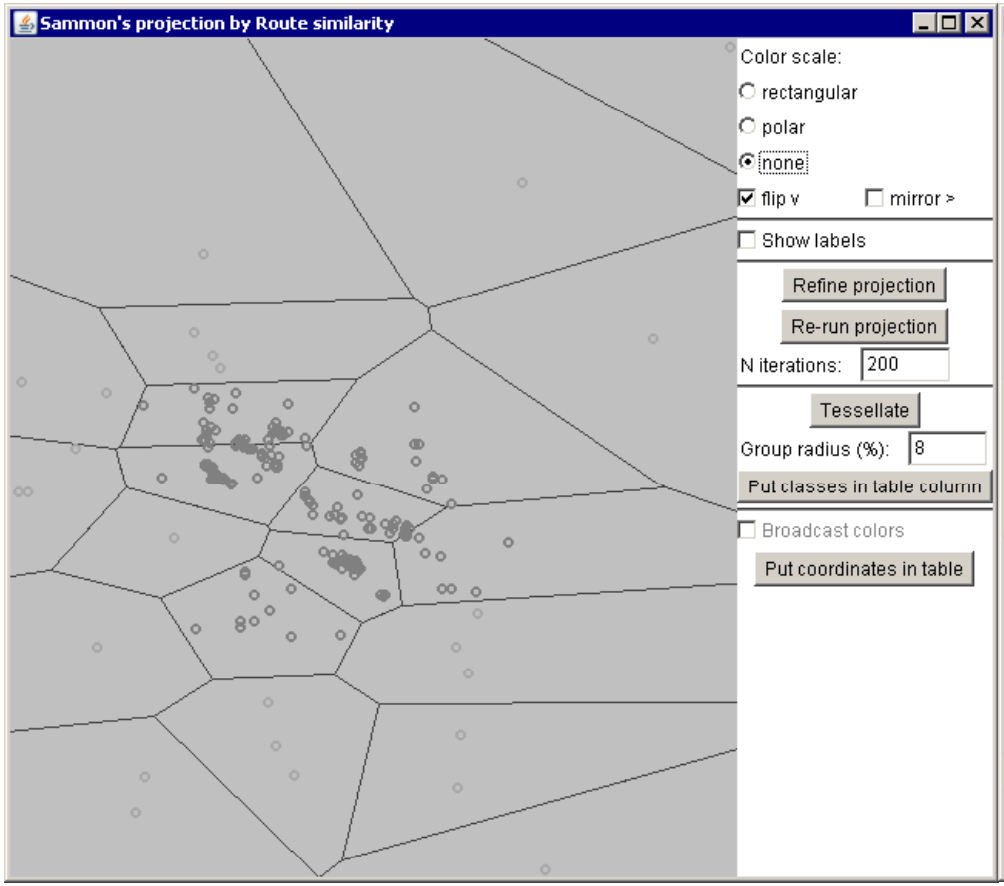


Fraunhofer IAIS

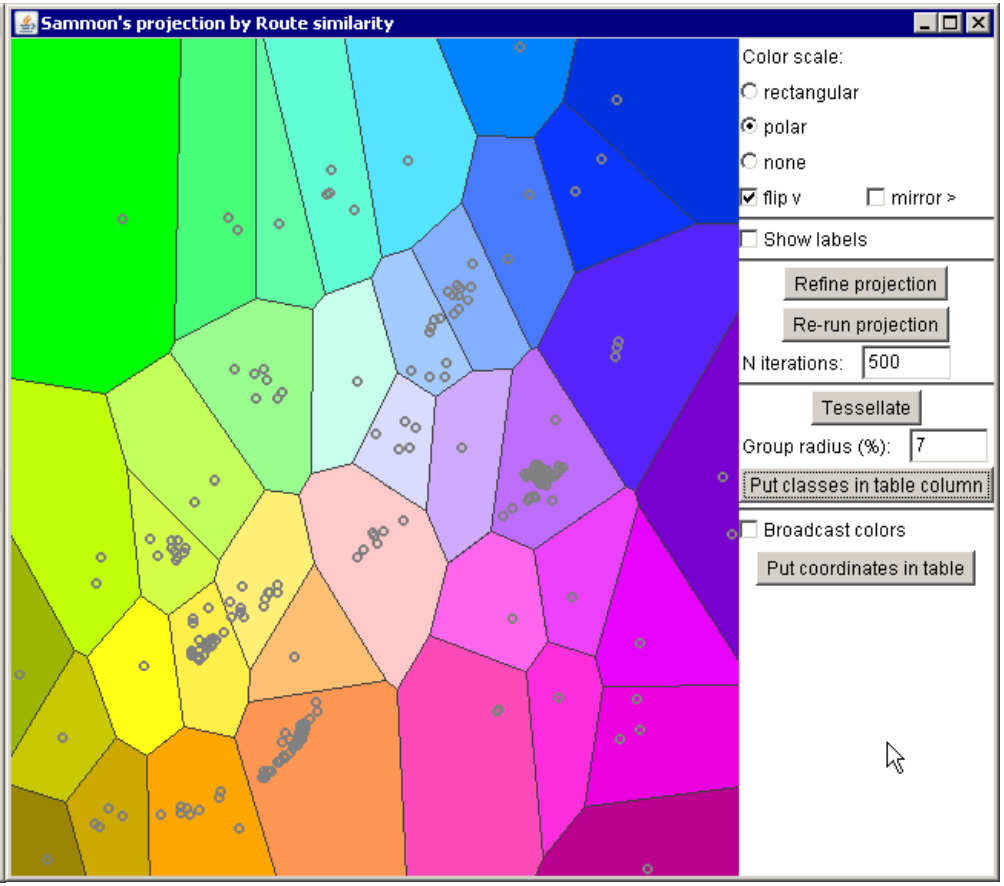
Clusters of trajectories defined by means of projection



Progressive clustering with the use of projection



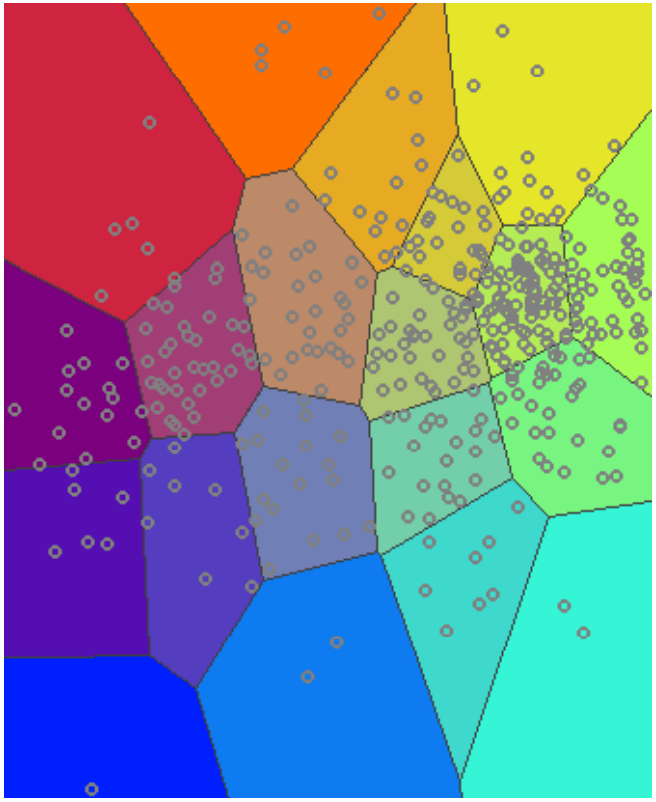
Step 1

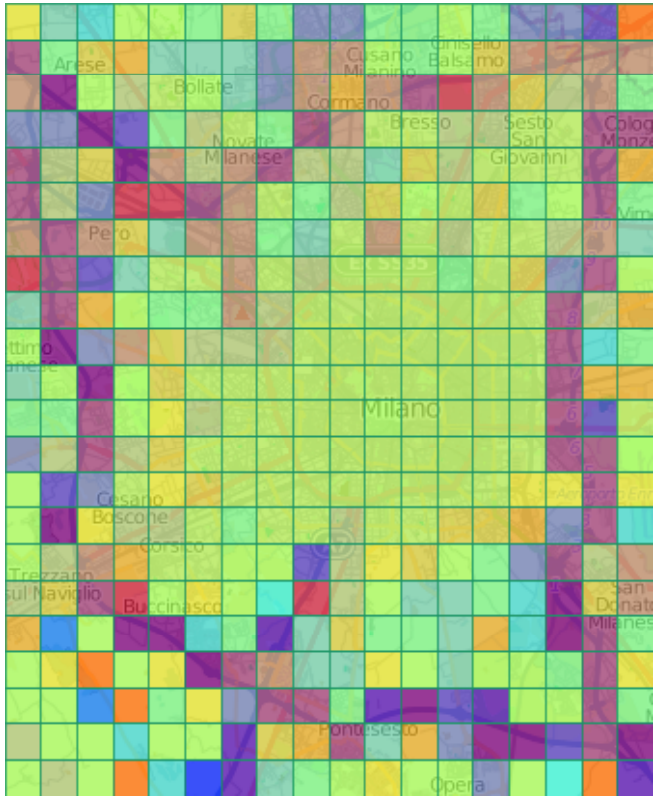


Step 2

Results of step 2

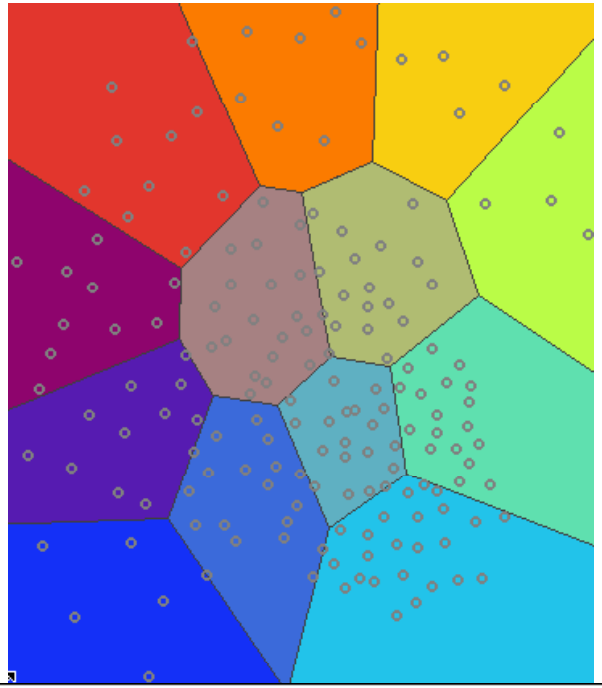
Applicability to other types of objects (example)



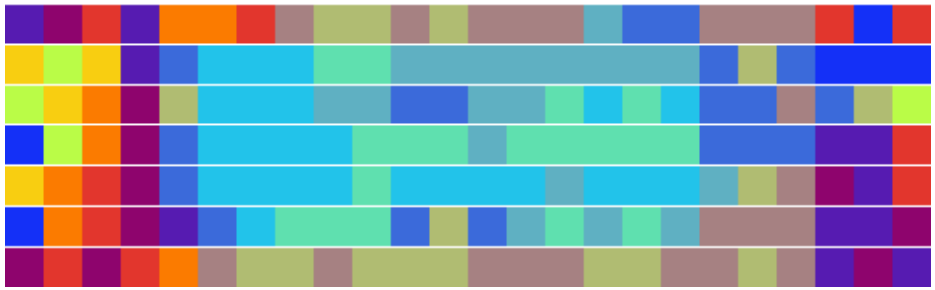


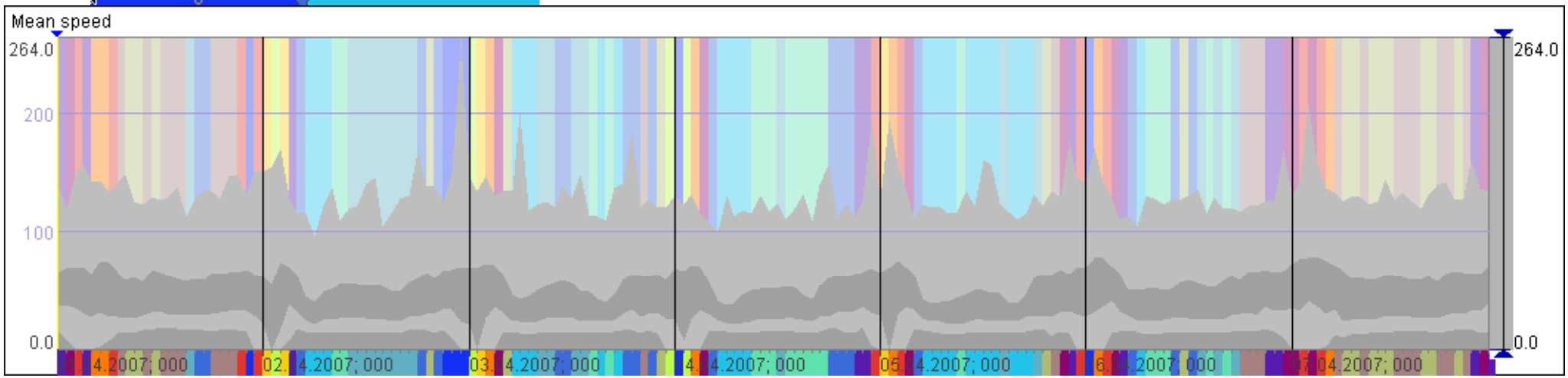
Clustering of spatial compartments (grid cells) according to temporal variations of aggregate characteristics of movement (e.g. average speeds)

Applicability to other types of objects (example)



Clustering of time intervals according to spatial distributions of aggregate movement characteristics





Conclusion

- Similarity of complex objects (in particular, spatio-temporal objects) often needs to be measured in specific ways *(not as distances in N-dimensional abstract space)*
  - ⇒ Specific distance functions are created (e.g. for trajectories, time series, …)
  - ⇒ Difficulty for cluster analysis: only a few of the existing clustering methods can work with arbitrary distance functions
- Projection techniques (Sammon's projection, MDS) can arrange objects in an abstract space according to distances among them in terms of any distance function
  - ⇒ Each object gets a position in the abstract space, defined by coordinates; similar objects have close positions
  - ⇒ Clustering can be now applied to the positions (coordinates) of the objects
    - + Most of the existing clustering algorithms (e.g. k-means, hierarchical, SOM) become applicable to this type of data
    - + Clusters of similar objects can also be built by direct grouping of points in the projection space or by tessellation of the projection space