Data Mining of Manufacturing Data

Application of GSOM

Georg Ruß

The University of Melbourne, Australia

Otto-von-Guericke-Universität Magdeburg, Germany

Outline

- Given task
- Data mining process
- Preprocessing
- Quality measuring
- Achievements

Given task

- Given manufacturing data:
  1. Get preprocessed data / preprocess data
  2. Apply GSOM to it
  3. Check clustering on maps
  4. Find distinguishing attributes
- Two-class clustering problem
- Threshold between classes is known
- No further knowledge about the data itself; blind

Manufacturing data

	REF	C1	C2	. . .	C92	X93	. . .	X131
1	9628	1.00E-09	2.21E-02	. . .	15	SMO	. . .	Dec-18-95
2	9496	1.20E+00	2.21E-02	. . .	33	SMO	. . .	Dec-19-95
3	9336	1.05E-09	2.34E-02	. . .	33	SMO	. . .	Dec-19-95
4	8300	1.36E-09	2.26E-02	. . .	14	SMO	. . .	Dec-19-95
5	9480	-1.00E-08	2.50E-02	. . .	23	SMO	. . .	Dec-19-95
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
16381	9728	-2.69E-08	2.24E-02	. . .	17	SMO	. . .	Feb-26-96

Preprocessing steps

- *old:*
  - Pruning
  - Normalisation
  - Expansion of categorical data
- *new [additional]:*
  - Elimination of outliers
  - Sampling

Simulation schedule

Dataset	Part	SF	GP, LL	SP1, LL	SP2, LL	dimensions
OLD	sampled (5 sets)	0.7	1-10, 100	0	0	808x1638
			5	10	[0,10]	
			10	10	[0,10]	
	complete	0.7	1-10, 100*	0	0	808x16380
			5	10	[0,10]	
			10	10	[0,10]	
		0.8	1-10	0	0	
		0.9	1-10	0	0	
NEW	complete	0.8	1-10	0	0	808x15980
		0.9	1-10	0	0	
		0.1-0.9	1	0	0	

Clustering quality measure [1]

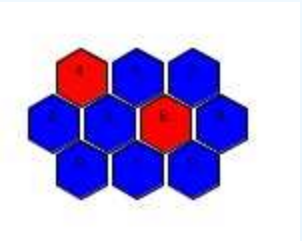
- Problem: subjectiveness of clustering quality
- Solution: CQ measure distills the quality into a single number
- Represents the deviation of the data distribution on the map from the original data distribution
- Range: from 0 to 1
  - 0: no clustering / no deviation
  - 1: perfect clustering / maximum deviation

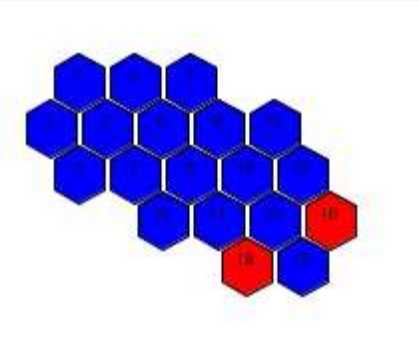
Clustering quality measure [2]

- Calculation:

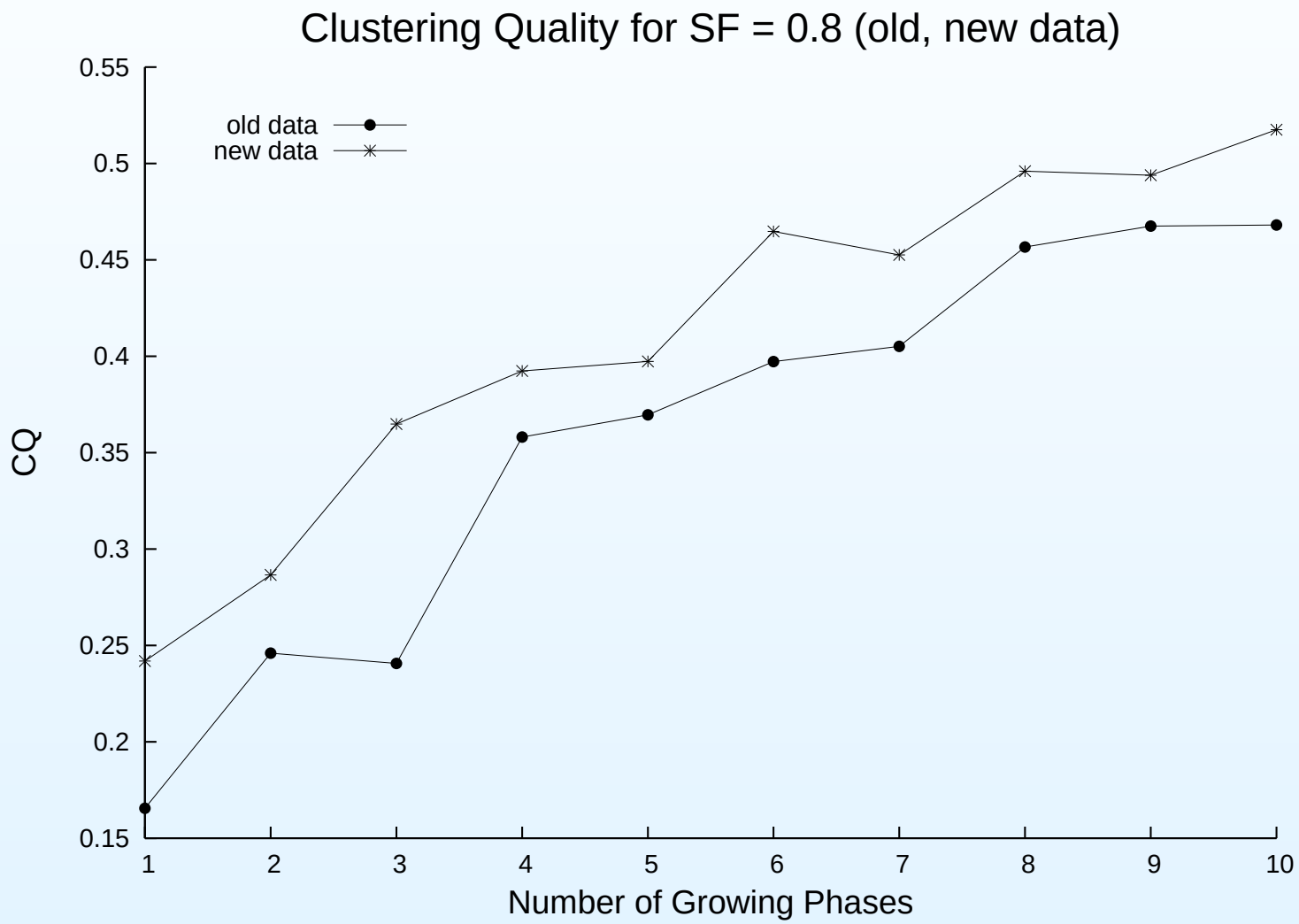
$$CQ = \sum_i \max\left\{ \frac{\frac{g_i}{n_i} - \frac{G}{N}}{1 - \frac{G}{N}} * \frac{n_i}{N}, 0 \right\} + \sum_i \max\left\{ \frac{\frac{b_i}{n_i} - \frac{B}{N}}{1 - \frac{B}{N}} * \frac{n_i}{N}, 0 \right\}$$

- Tested with 59x100- and 59x1000-dimensional datasets
- Hits maps with CQ = 1:



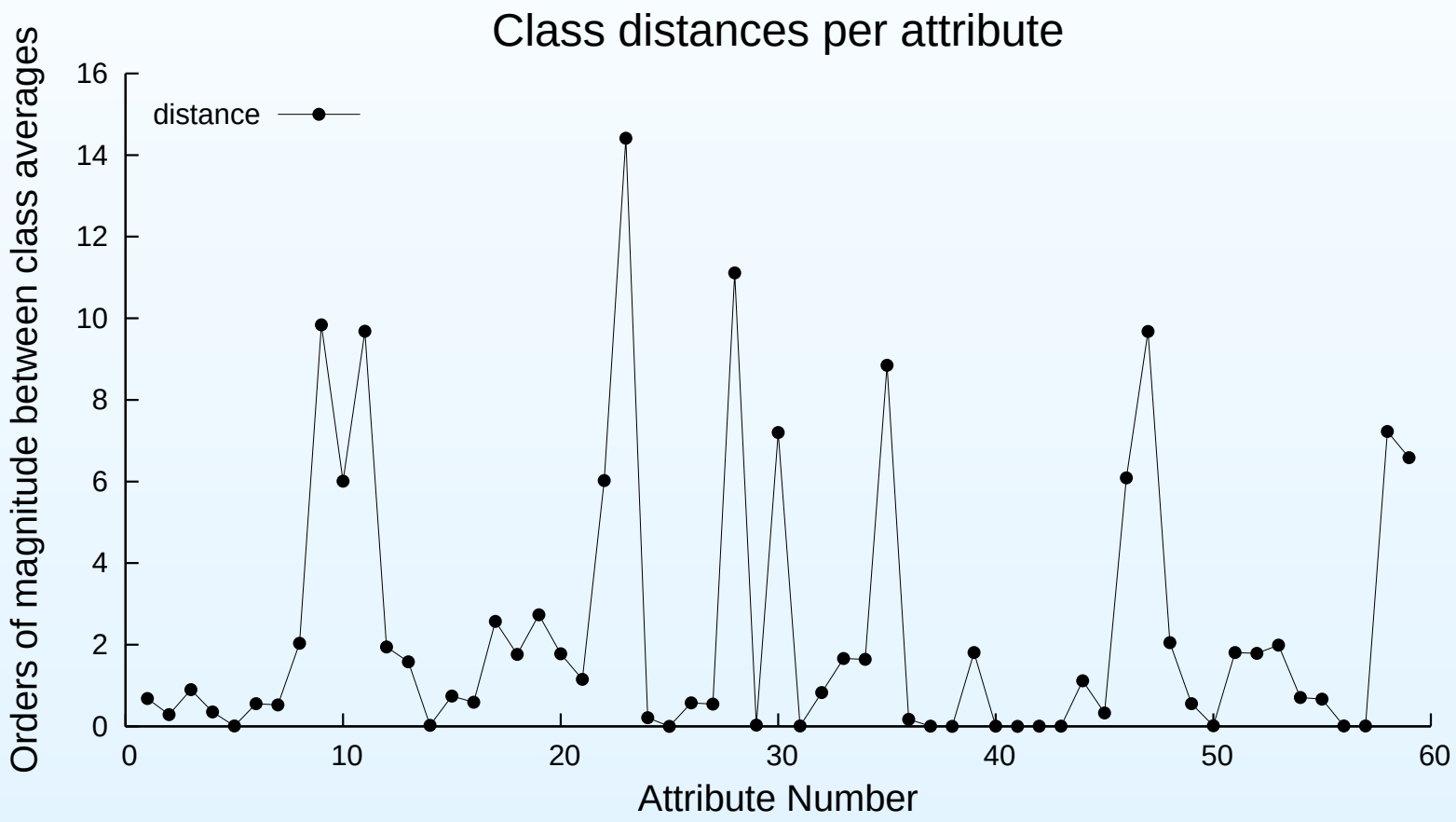


Improvements after additional preprocessing



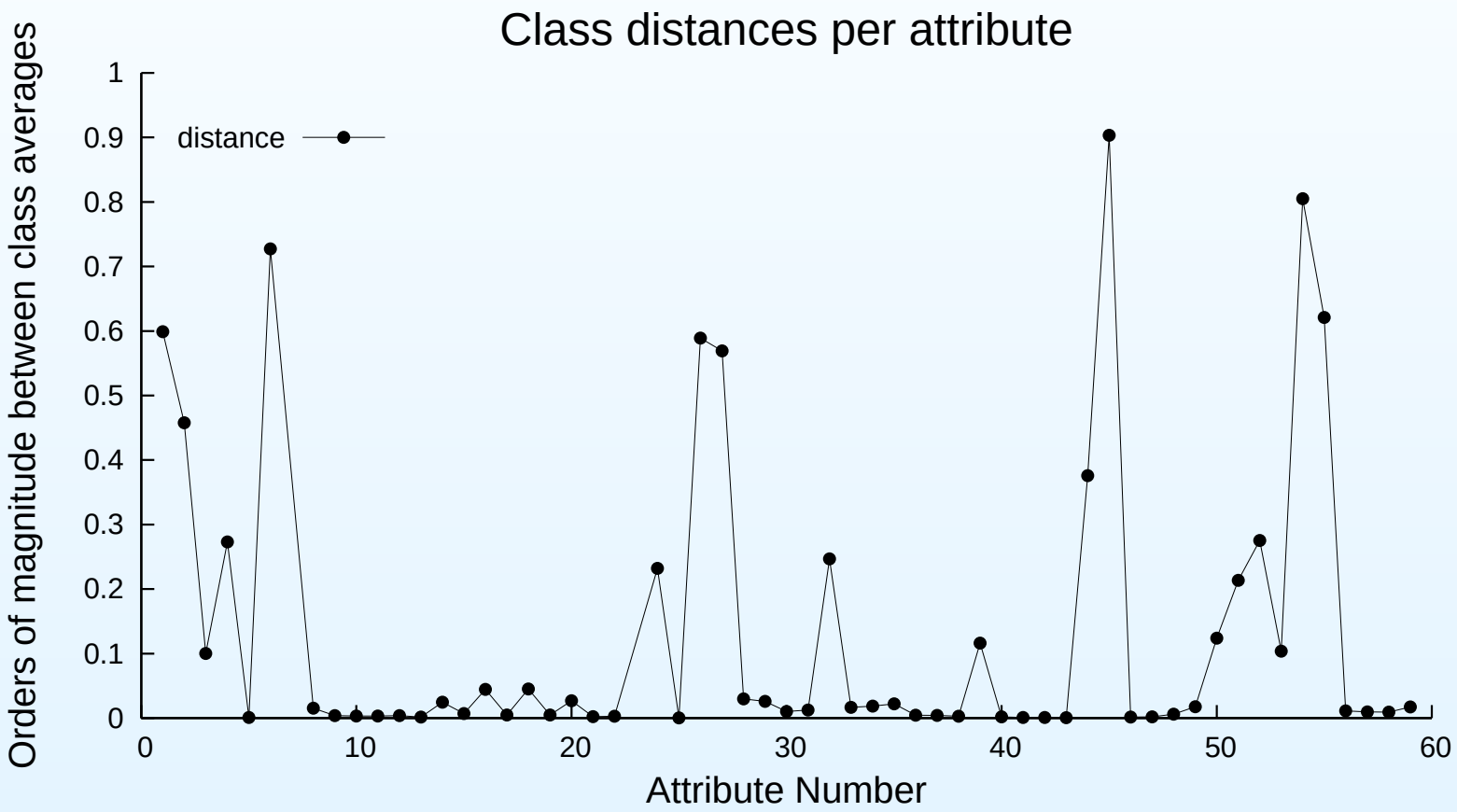
Finding distinctive attributes [1]

- Distance between attribute averages [with outliers]:



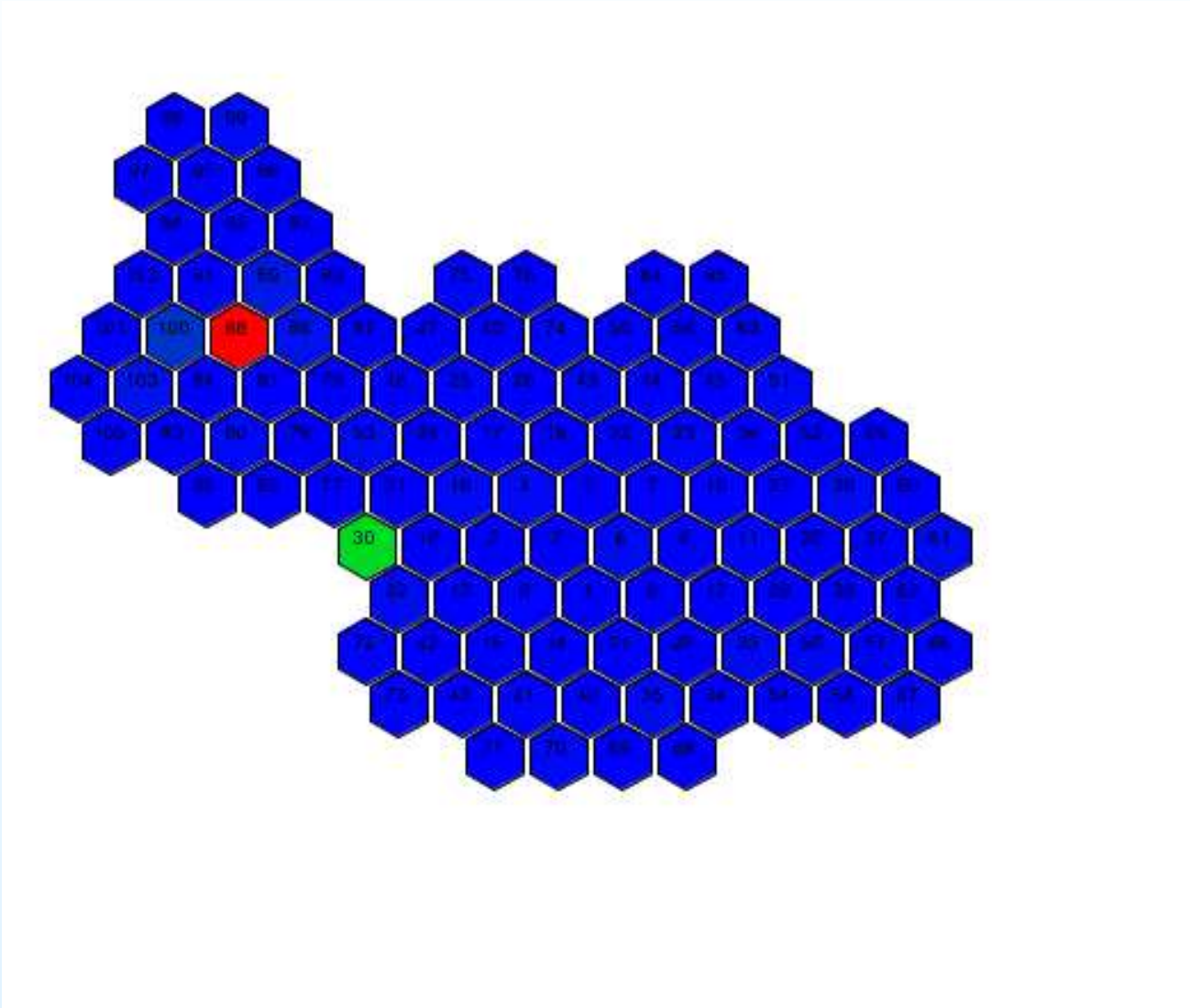
Finding distinctive attributes [2]

- Distance between attribute averages [without outliers]:



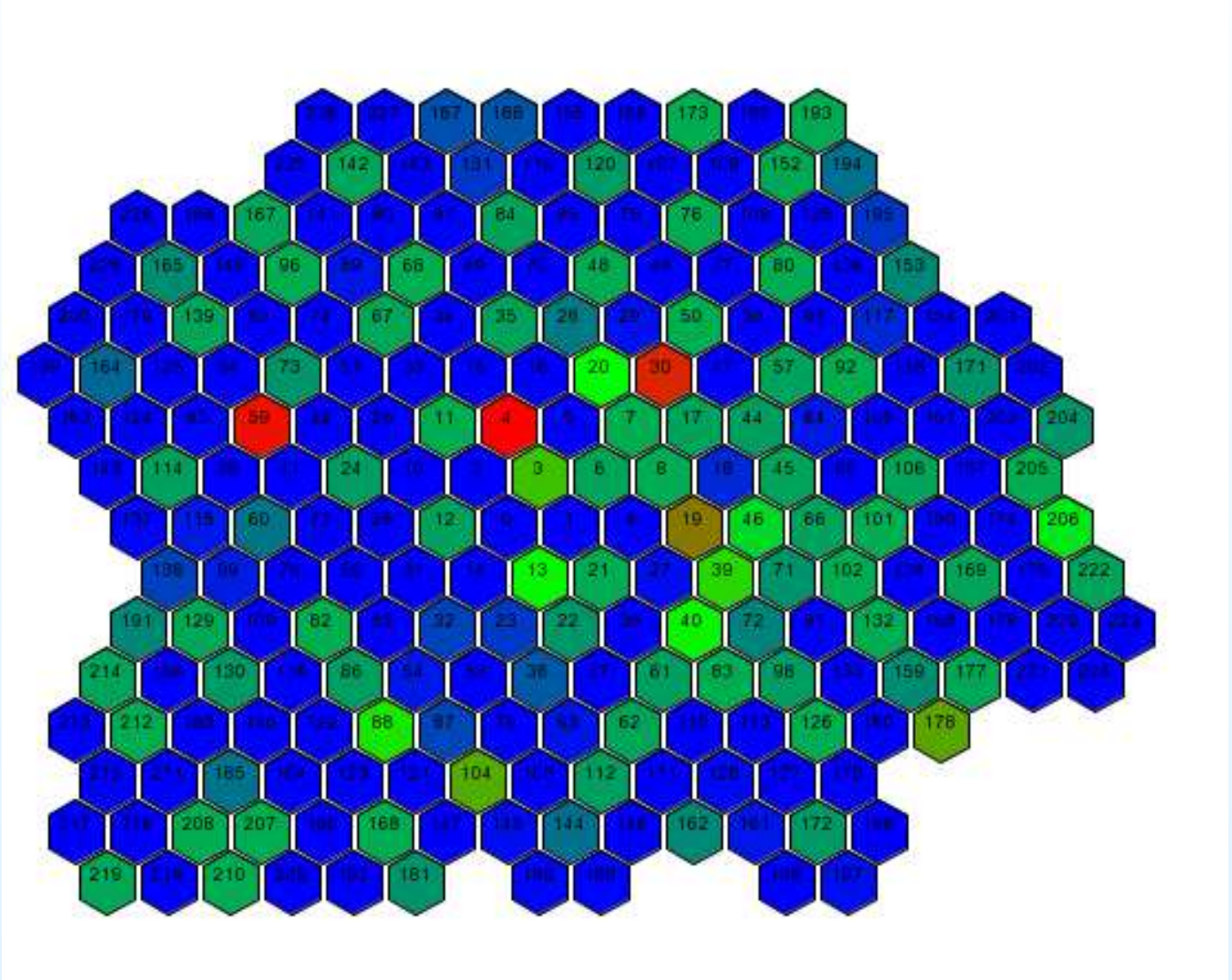
Resulting maps [1]

- Prototypical desired map:



Resulting maps [2]

- Prototypical resulting map:



Summary: data mining steps