Clustering

Sandrien van Ommen

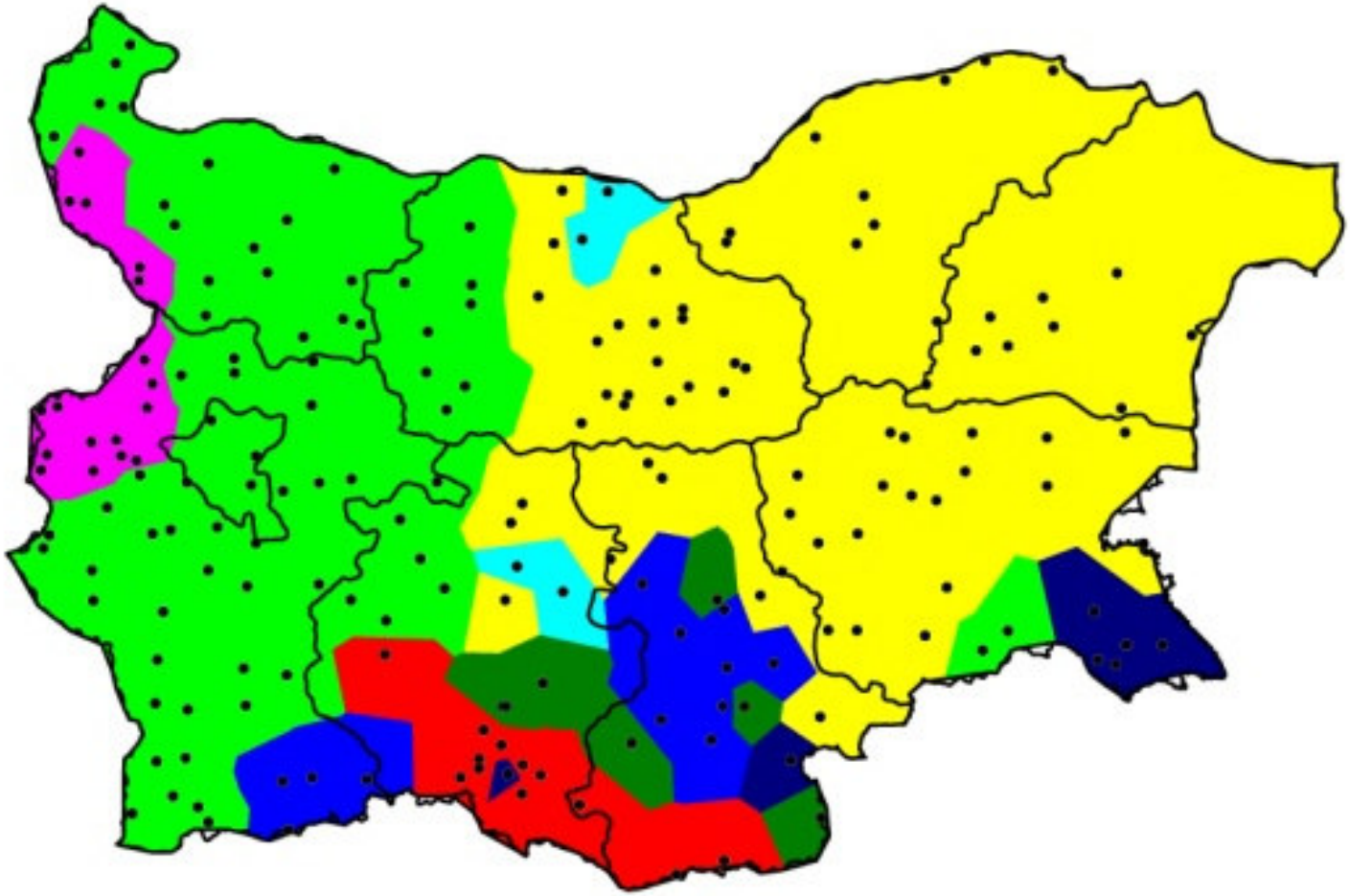
Overview

- Why clustering
- When clustering
- Types of clustering
- Dialects
  - Distances
  - Dutch towns
  - Buldialect
- Conclusion

Why clustering

- To find similarity in your data
  - T-test & Anova = means
  - Correlation & regression = effect size
  - Chi-square = frequencies
  - PCA = summarizing data
  - **Clustering = grouping data on basis of similarity**

Example



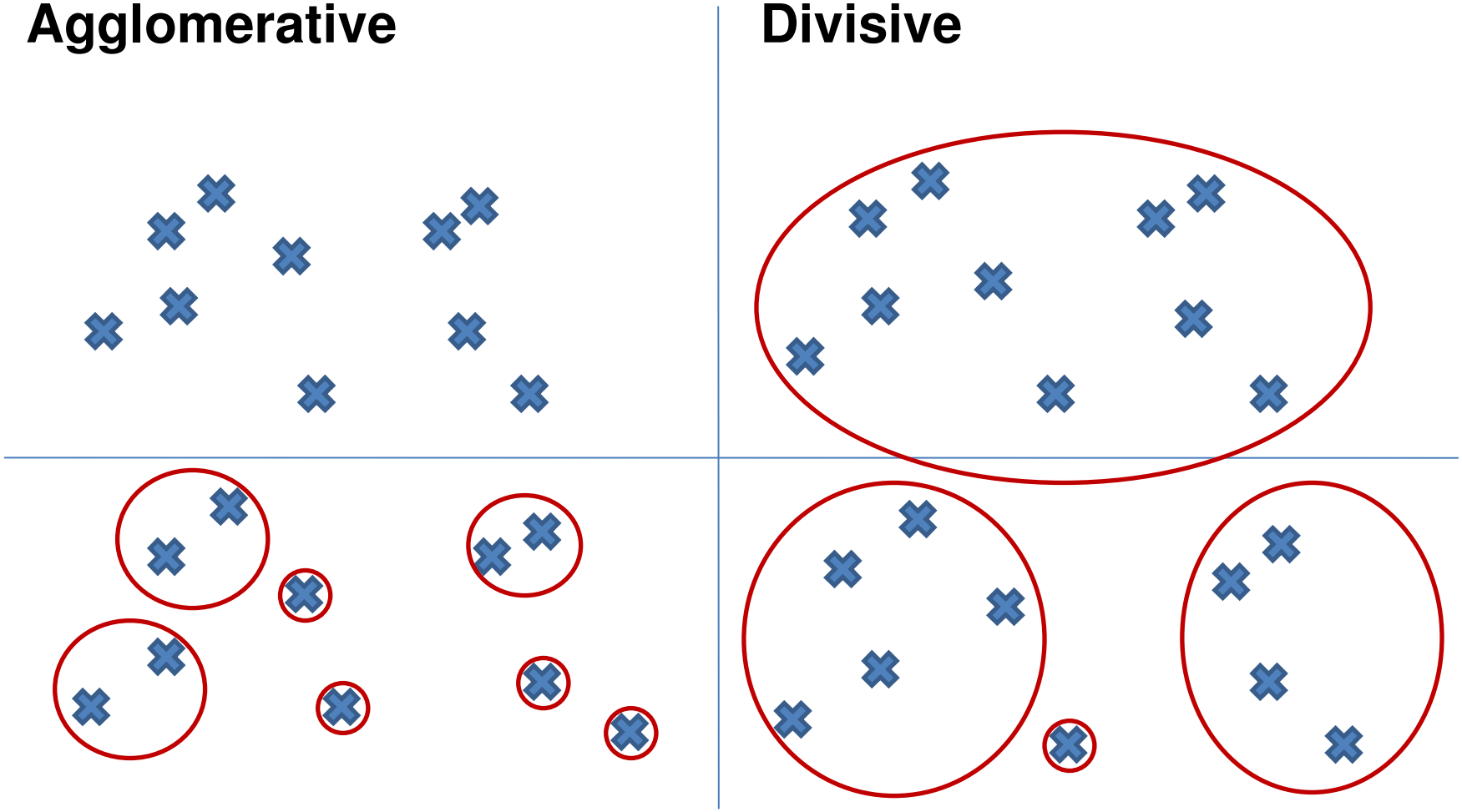
When clustering

- Market research: determining populations
- Biology: group gene families
- Social Network Analysis
- **Linguistics:**
  - Dialect differences/-areas
  - ‘neighborhoods’ in semantics/syntax/phonology
  - Language models
- …

Types of clustering

- Agglomerative- / Divisive clustering
- Soft (disjunctive)- / Hard clustering
- Hierarchical- /Non- hierarchical (flat) clustering
- Single-link vs. Complete-link clustering

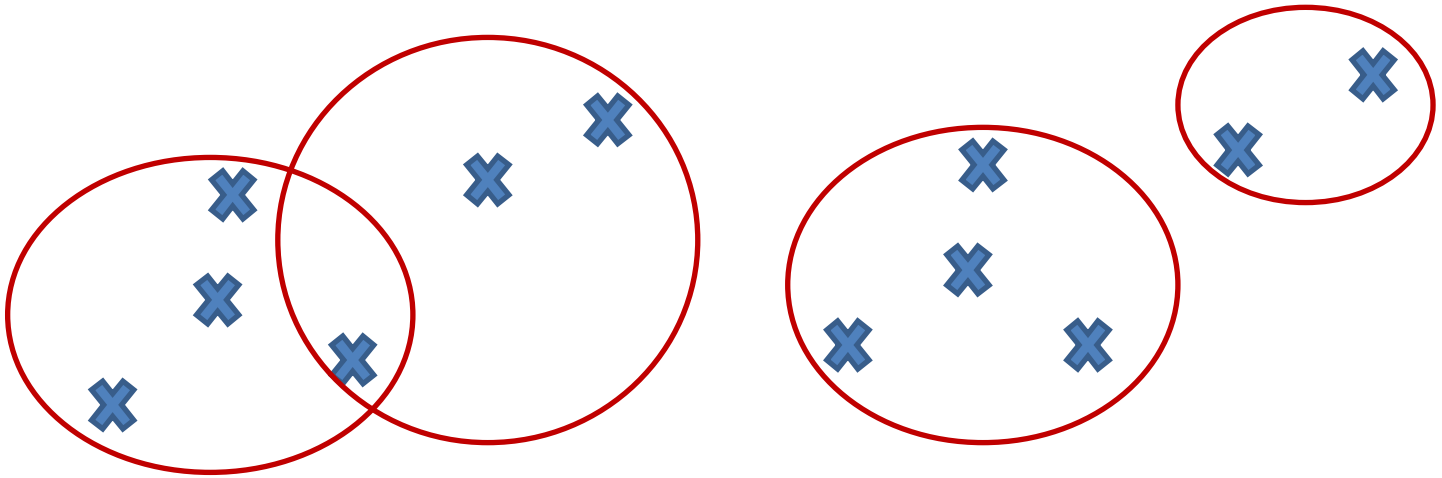
Types of Clustering



Types of clustering

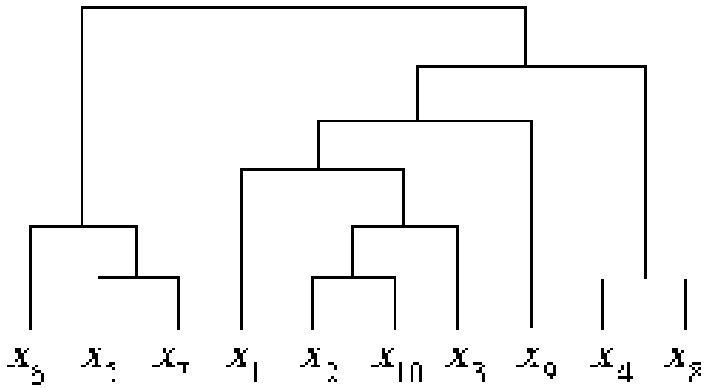
Soft

Hard

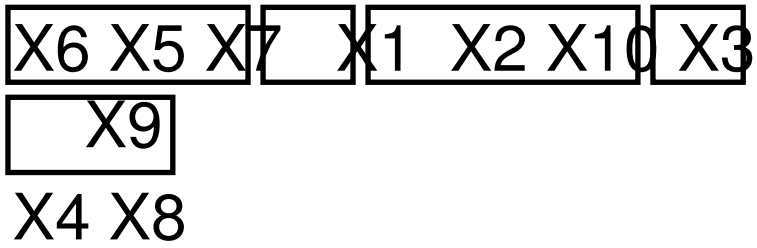


Types of Clustering

Hierarchical

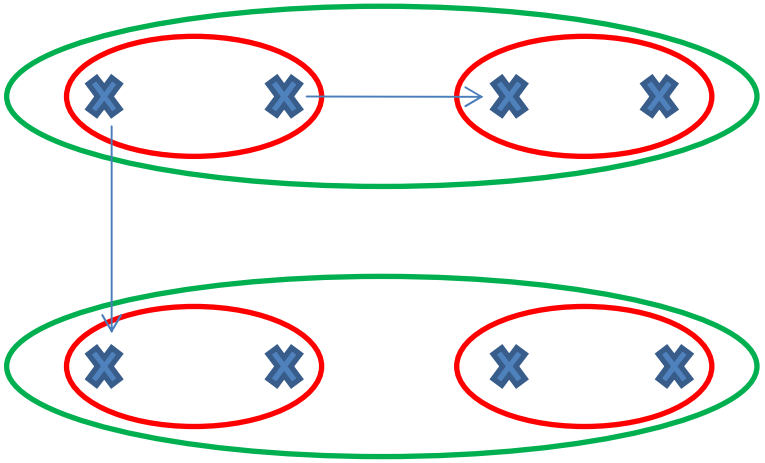


Flat

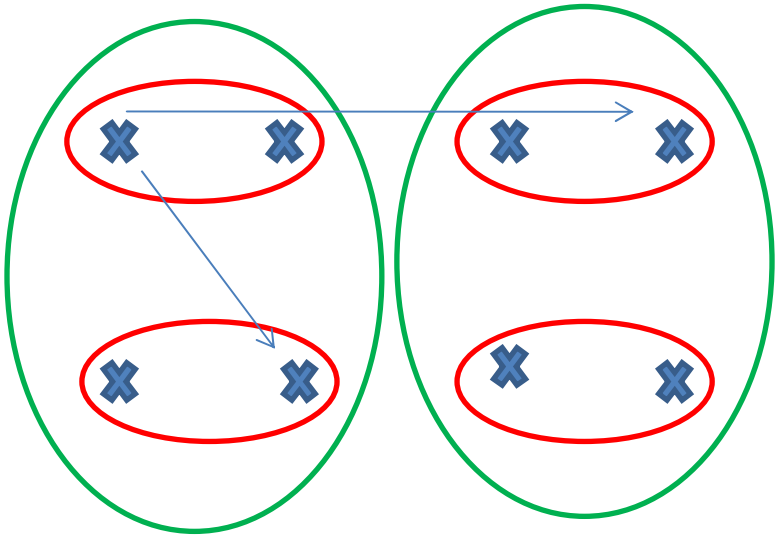


Types of Clustering

Single-link



Complete-link

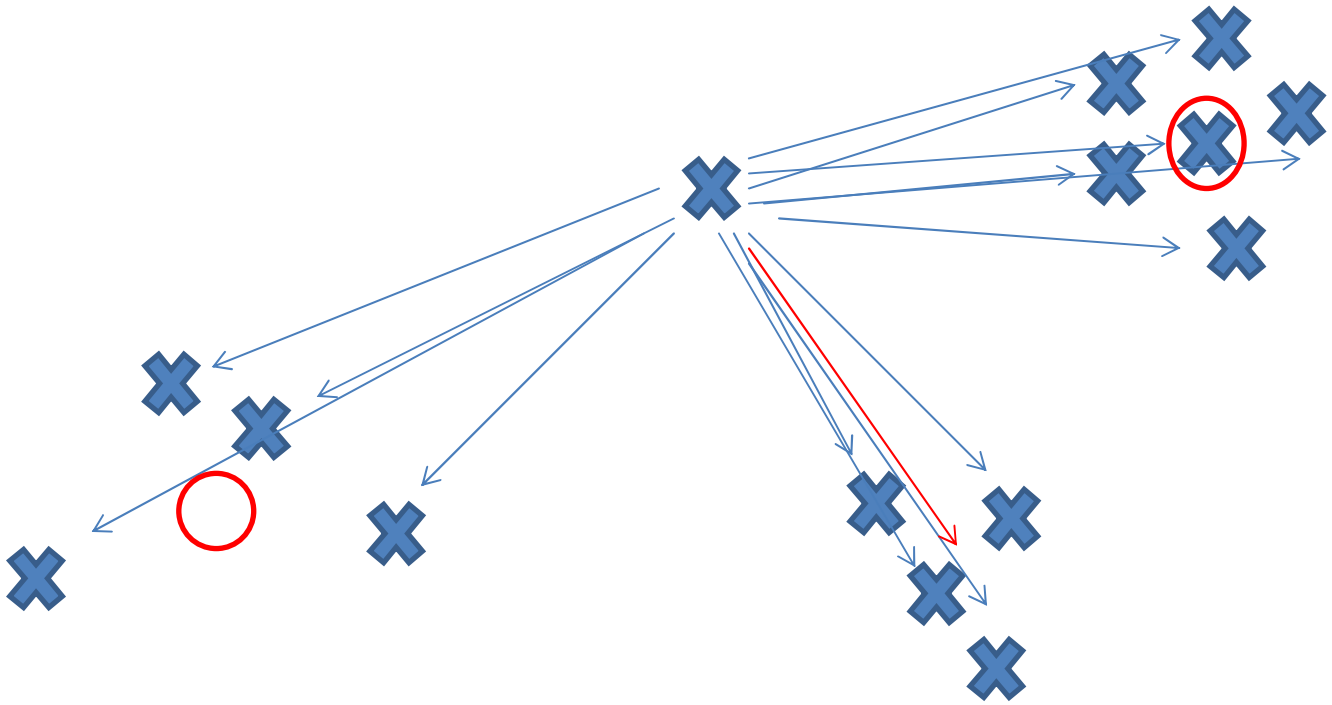


Types of Clustering

- Single link: $d_{k[ij]} = \min(d_{ki}; d_{kj})$
- Complete link: $d_{k[ij]} = \max(d_{ki}; d_{kj})$
- UPGMA (unweighted pair group method using arithmetic averages):
  - $d_{k[ij]} = ((n_i / (n_i + n_j)) \times d_{ki}) + ((n_j / (n_i + n_j)) \times d_{kj})$
- WPGMA (weighted):
  - $d_{k[ij]} = (1/2 \times d_{ki}) + (1/2 \times d_{kj})$

Types of Clustering

- Group average, centroids, medoids



Types of Clustering

- Centroids:
  - UPGMC
    - $d_{k[ij]} = ((n_i / (n_i + n_j)) \times d_{ki}) + ((n_j / (n_i + n_j)) \times d_{kj}) - ((n_i \times n_j) / (n_i + n_j)^2 \times d_{ij})$
  - WPGMC
    - $d_{k[ij]} = (1/2 \times d_{ki}) + (1/2 \times d_{kj}) – (1/4 \times d_{ij})$

Types of Clustering

Flat:

- K-means
- L1
- The EM algorithm (soft, iterative)

  – Make use of centroids (or medoids)

Dialects

- Clustering = grouping data based on similarity
- Similarity is calculated based on distance
  - $s(x,y) = 1 / (1 + d(x,y))$
- for grouping dialects: calculate distances
  - Sequence comparison (pronunciation distance)
  - Average pronunciation distance between villages

Dialects

- Sequence comparison:
  - Levenshtein Distance
    - Based on ‘string-changing operations’
      - Deletions, insertions, substitions
    - Weight assigned to each operation
    - Calculation of the minimum cost

Dialects

Hamming Distance:

æ	ə	f	t	ə	n	ʉ	n
æ	f	t	ə	r	n	u	n
	1	1	1	1		1	

Levenshtein Distance:

æəftənʉn	delete ə	1
æftənʉn	insert r	1
æftərnʉn	subst. ʉ/u	1
æftərnun		
		3

Distance Matrix

		∅		æ		ə		f		t		ə		n		ɪ		n	
		0		1		2		3		4		5		6		7		8	
∅	0	**0**		1	**1**	2	**2**	3	**3**	4	**4**	5	**5**	6	**6**	7	**7**	8	**8**
æ	1	1		0	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9
		1		2	**0**	1	**1**	2	**2**	3	**3**	4	**4**	5	**5**	6	**6**	7	**7**
f	2	2		2	1	1	2	1	3	3	4	4	5	5	6	6	7	7	8
		2		3	**1**	2	**1**	2	**1**	2	**2**	3	**3**	4	**4**	5	**5**	6	**6**
t	3	3		3	2	2	2	2	2	1	3	3	4	4	5	5	6	6	7
		3		4	**2**	3	**2**	3	**2**	3	**1**	2	**2**	3	**3**	4	**4**	5	**5**
ə	4	4		4	3	2	3	3	3	3	2	1	3	3	4	4	5	5	6
		4		5	**3**	4	**2**	3	**3**	4	**2**	3	**1**	2	**2**	3	**3**	4	**4**
r	5	5		5	4	4	3	3	4	4	3	3	2	2	3	3	4	4	5
		5		6	**4**	5	**3**	4	**3**	4	**3**	4	**2**	3	**2**	3	**3**	4	**4**
n	6	6		6	5	5	4	4	4	4	4	4	3	2	3	3	4	3	5
		6		7	**5**	6	**4**	5	**4**	5	**4**	5	**3**	4	**2**	3	**3**	4	**3**
u	7	7		7	6	6	5	5	5	5	5	5	4	4	3	3	4	4	4
		7		8	**6**	7	**5**	6	**5**	6	**5**	6	**4**	5	**3**	4	**3**	4	**4**
n	8	8		8	7	7	6	6	6	6	6	6	5	5	4	4	4	3	5
		8		9	**7**	8	**6**	7	**6**	7	**6**	7	**5**	6	**4**	5	**4**	5	**3**

Alignment

		∅	æ	ə	f	t	ə	n	ʉ	n
		0	1	2	3	4	5	6	7	8
∅	0	0	←	←	←	←	←	←	←	←
æ	1	↑	↖ 0	← 1	←	←	←	←	←	←
f	2	↑	↑	↖	↖ 1	←	←	←	←	←
t	3	↑	↑	↖ ↑	↖ ↑	↖ 1	←	←	←	←
ə	4	↑	↑	↖	↖ ↑ ←	↑	↖ 1	←	←	←
r	5	↑	↑	↑	↖	↑	↑ 2	↖	↖ ←	↖ ←
n	6	↑	↑	↑	↖ ↑	↖ ↑	↑	↖ 2	↖ ←	↖
u	7	↑	↑	↑	↖ ↑	↖ ↑	↑	↑	↖ 3	↖ ↑ ←
n	8	↑	↑	↑	↖ ↑	↖ ↑	↑	↑	↖ ↑	↖ 3

Example

- 5 Dutch towns
- Data collected, transcribed, distances computed

	Grouw	Haarlem	Delft	Hattem	Lochem
Grouw		42	44	46	47
Haarlem			16	36	38
Delft				38	40
Hattem					21
Lochem					

Example

Group-average agglomerative Clustering

$$d_{k[ij]} = \frac{d_{ki} + d_{kj}}{2}$$

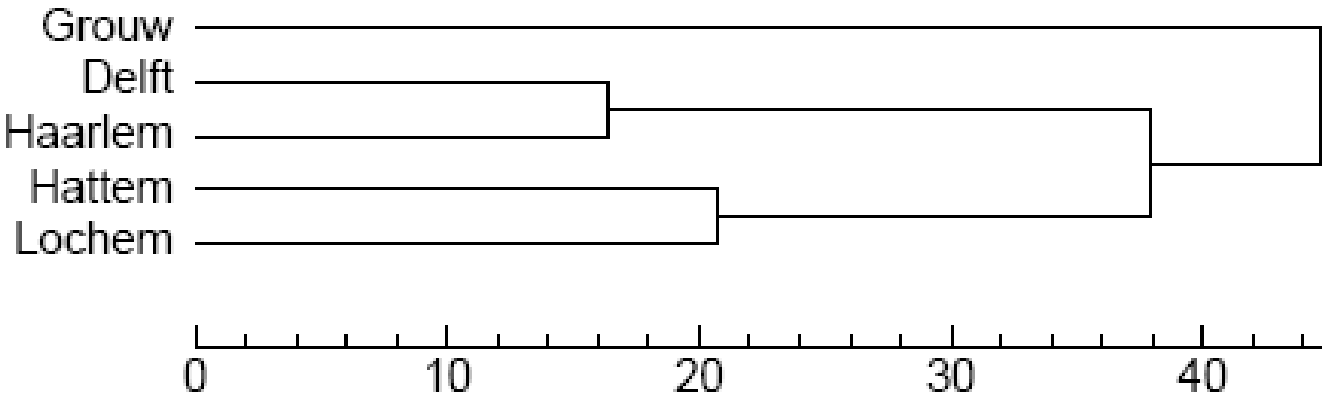
$$d_{Grouw,\ [Haarlem\ \&\ Delft]} = \frac{d_{Grouw,\ Haarlem} + d_{Grouw,\ Delft}}{2}$$

$$= \frac{42 + 44}{2}$$

$$= 43$$

Example

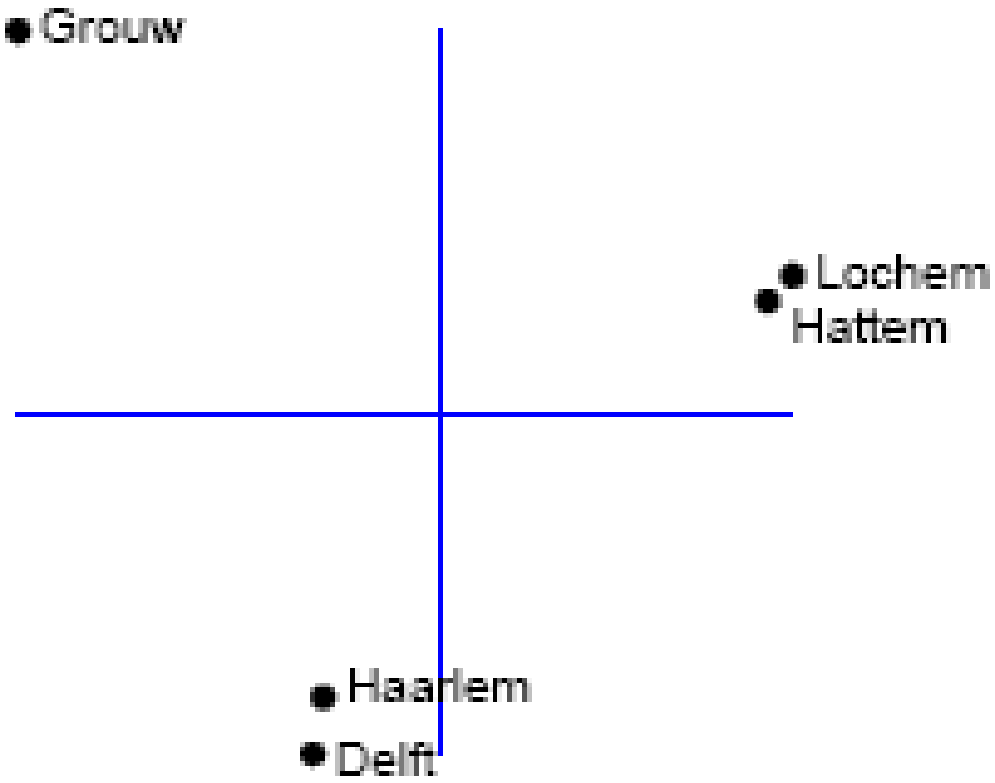
	Grouw	Haarlem & Delft	Hattem	Lochem
Grouw		**43**	46	47
Haarlem & Delft			**37**	**39**
Hattem				21
Lochem				



(Dendrogram)

Example

MultiDimensional Scaling (MDS)



Dialects

- The data: from project: “Buldialect: Measuring linguistic unity and diversity in Europe”
  - Levenshtein Distance: What is the minimum cost?
    - Deletions/insertions
    - Substitutions
  - Vowel-vowel and consonant-consonant alignment
  - 156 word pronunciations
  - 197 villages in Bulgaria

Buldialect

- Dialect distances between villages = Average of all distances in pronunciation
- Many objects!
  - Distances in pronunciation
  - -> distances between villages

  (n x (n-1)) / 2 = values

  -> (197 x (197-1))/2 = 19306 distance values

Buldialect

- Let's try some of the different methods
  - Single link
  - Complete link
  - UPGMA
  - WPGMA
  - (MDS)

Ways to view results

- Dendrograms
- Maps

Single-link

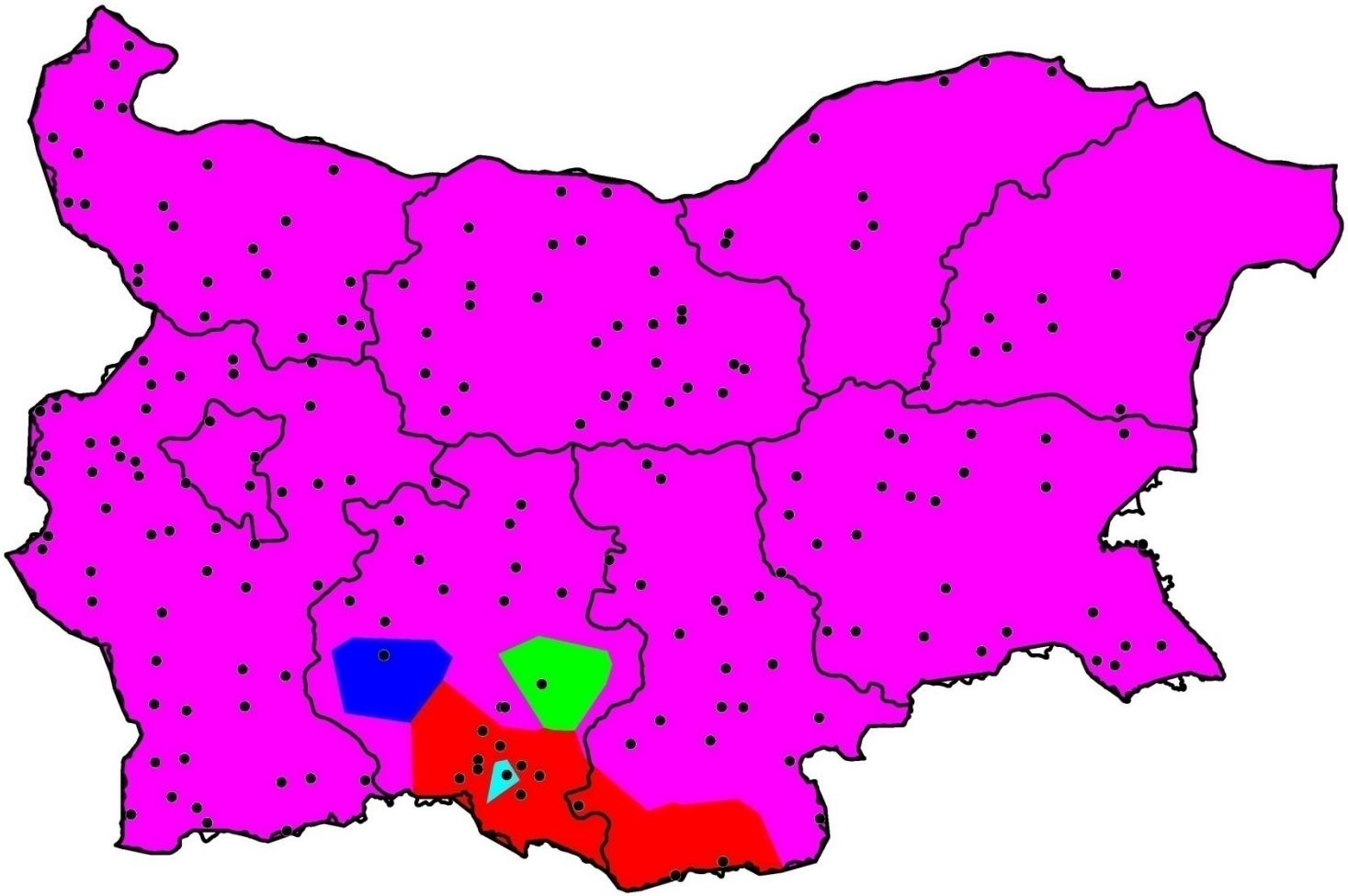
Dendrogram

Space contracting

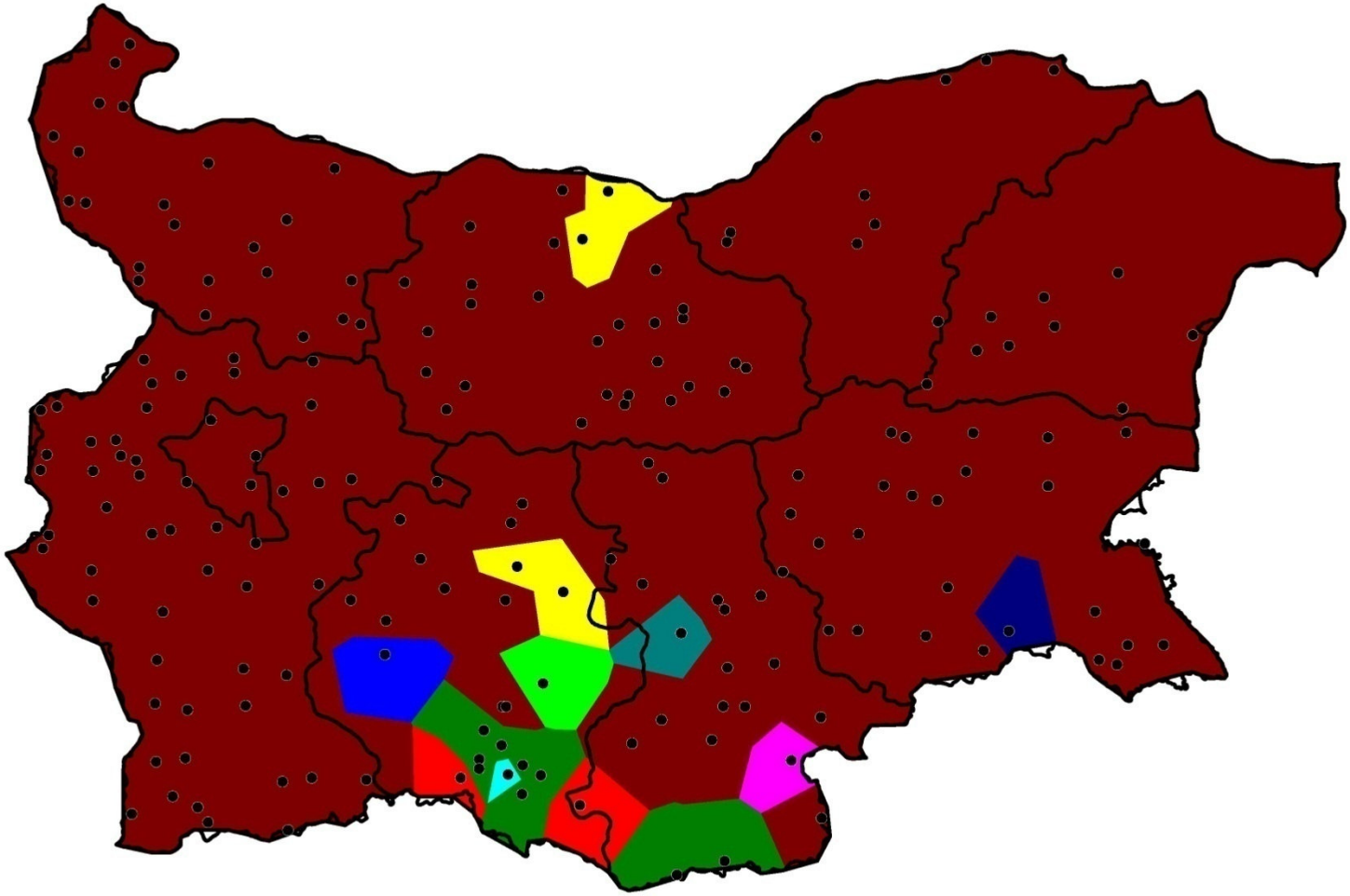
– Unequally sized clusters

– Outliers visible

Single-link



Single-link



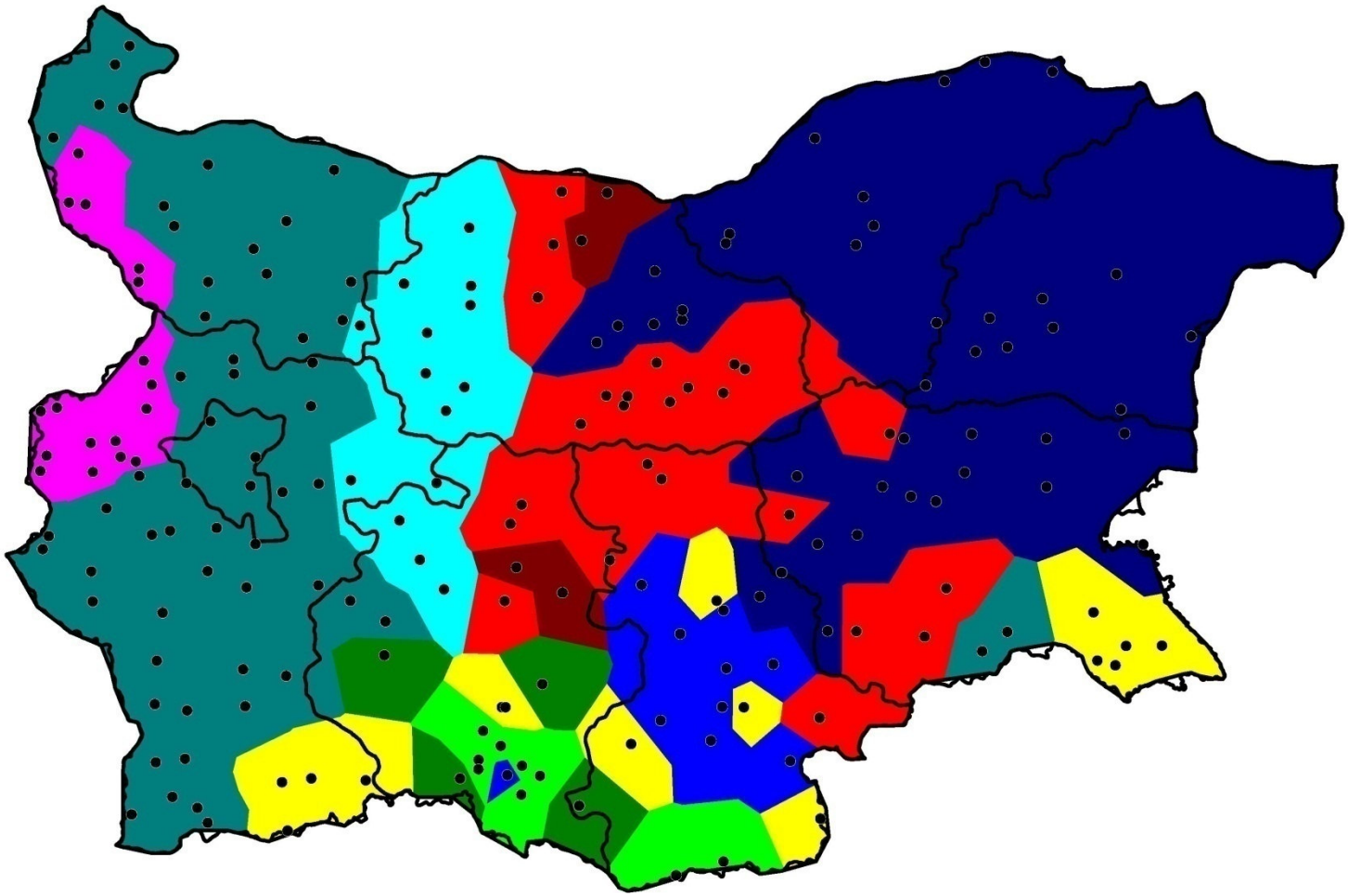
Complete-link

Dendrogram

Space dilating

- Balanced clustering
- Clusters often not easy to interpret

Complete-link



UPGMA

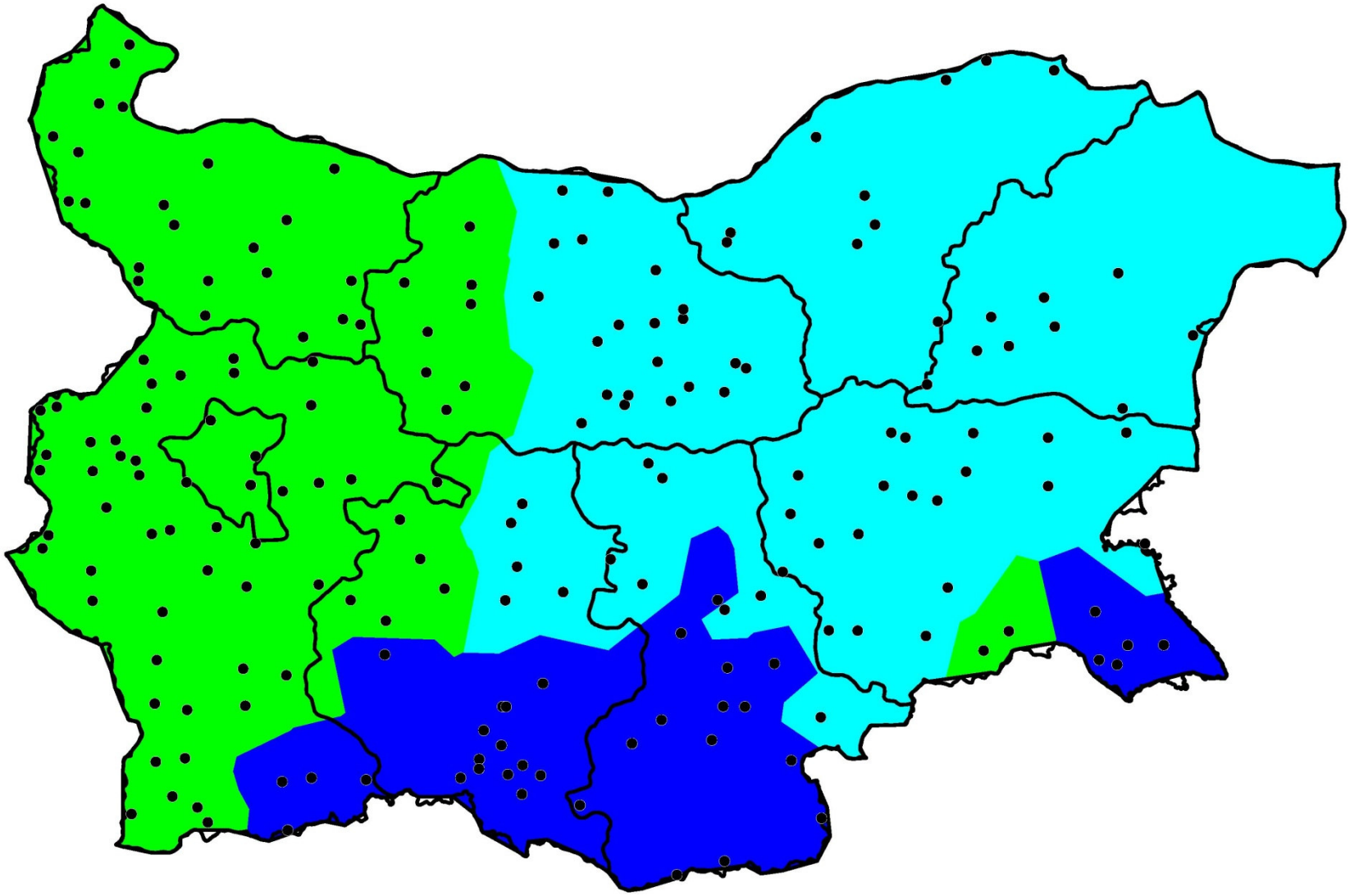
Dendrogram

Space conserving

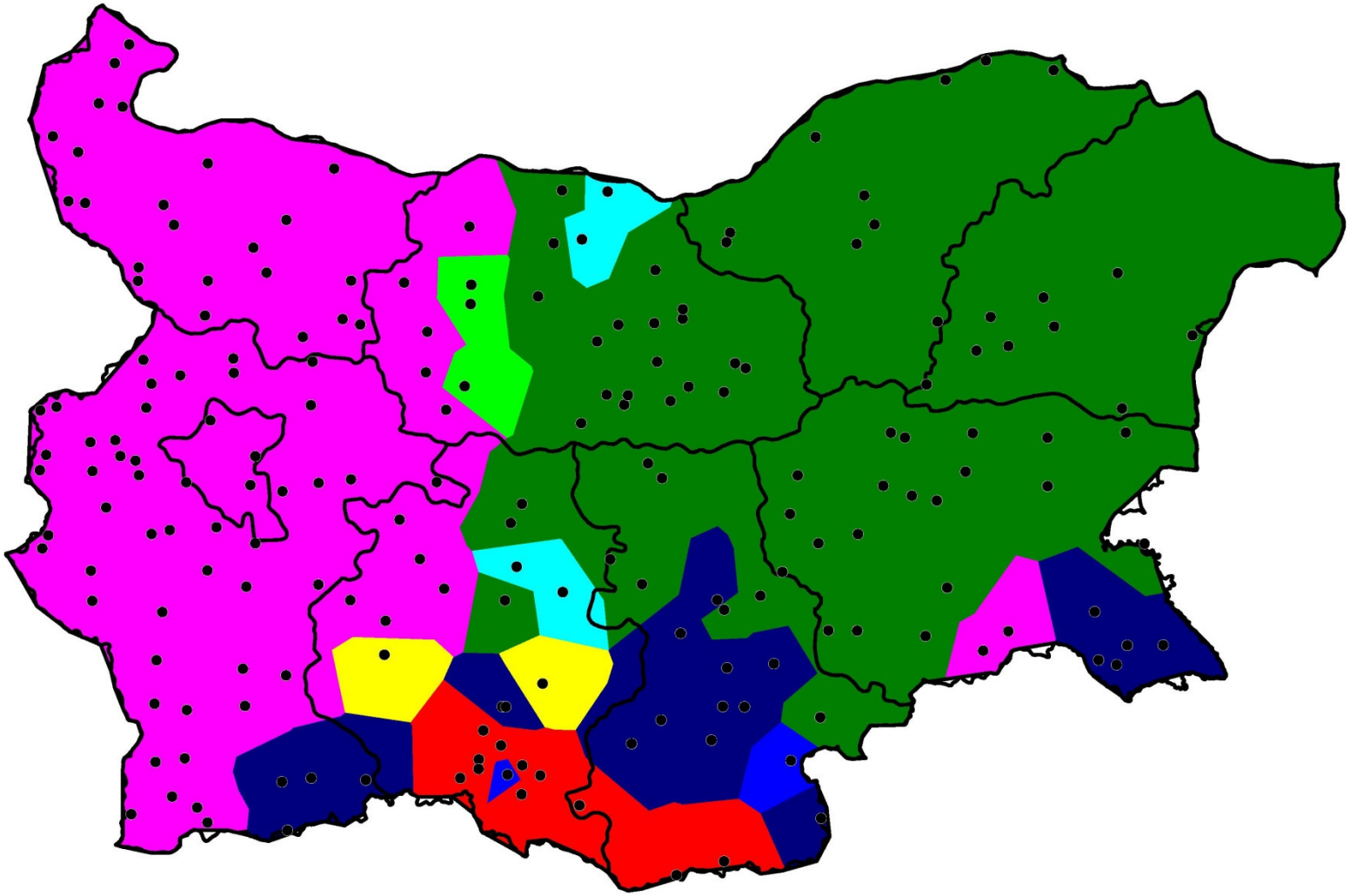
- $d_{k[ij]} = ((n_i / (n_i + n_j)) \times d_{ki}) + ((n_j / (n_i + n_j)) \times d_{kj})$

- Linkage between groups
- Averages instead of extreme values
- Number of elements in cluster taken into account

UPGMA



UPGMA



WPGMA

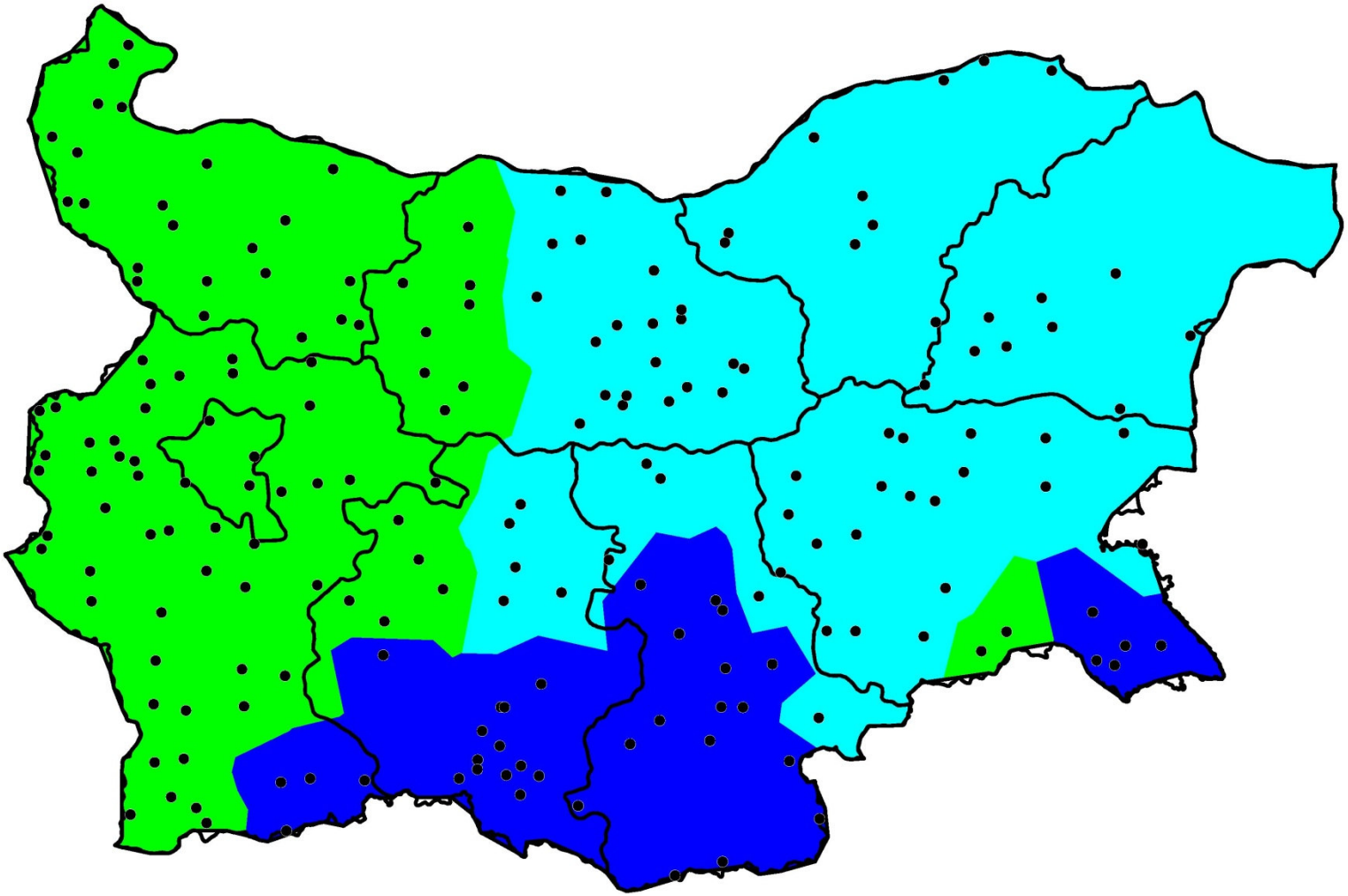
Dendrogram

Space conserving

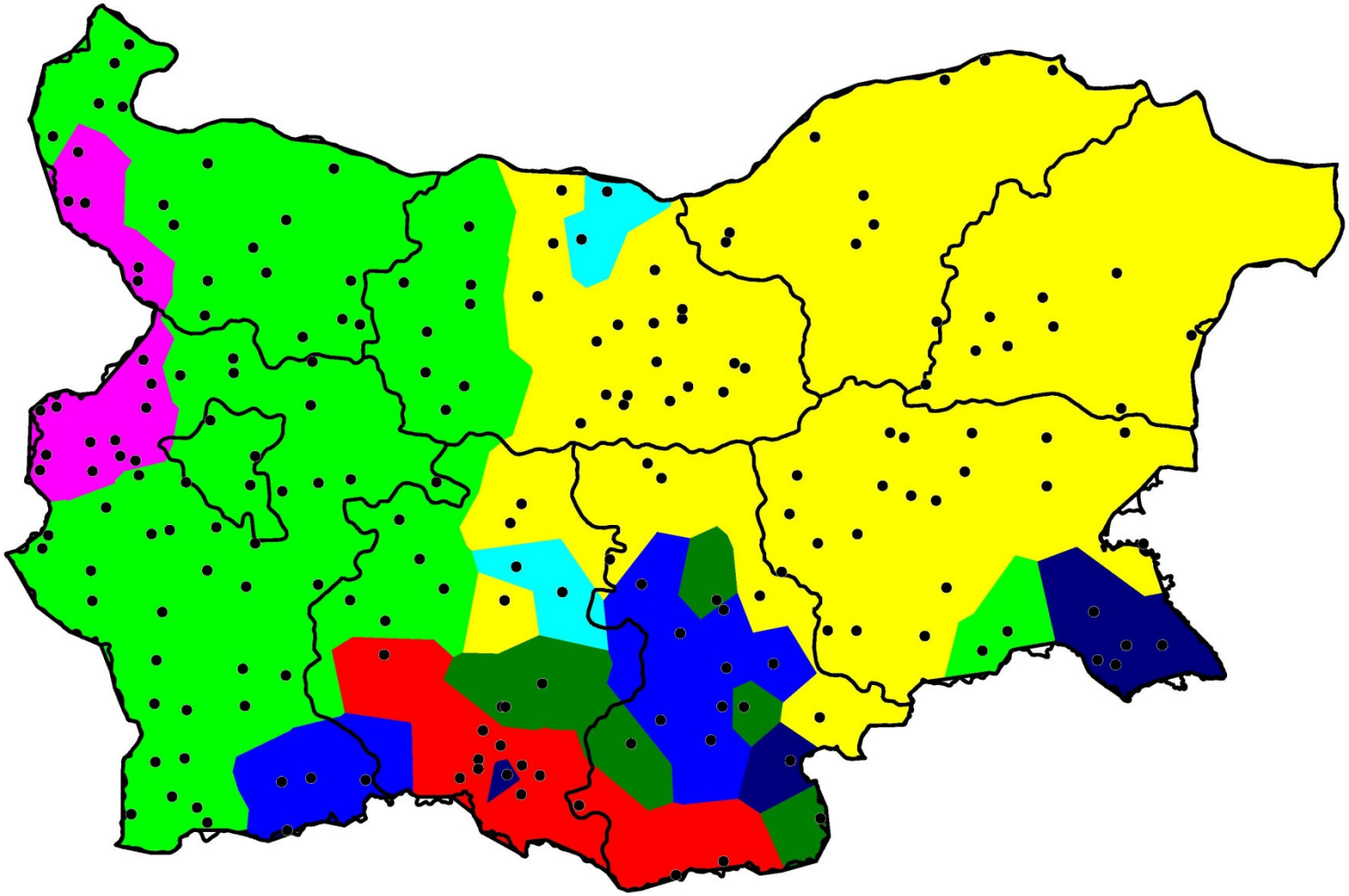
- $d_{k[ij]} = (1/2 \times d_{ki}) + (1/2 \times d_{kj})$

- Linkage within groups
- Number of elements not taken into account

WPGMA



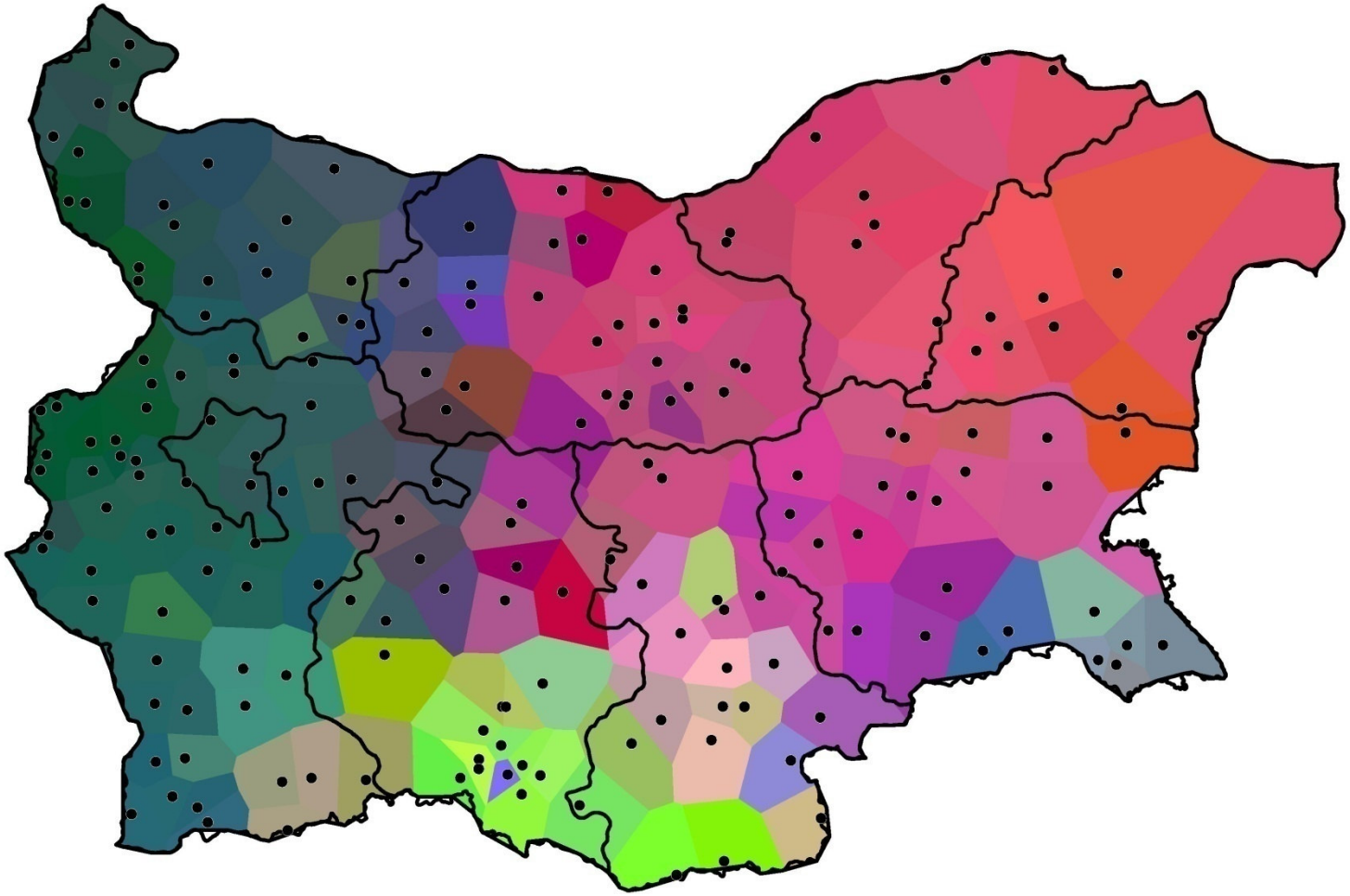
WPGMA



MultiDimensional Scaling

- What is MDS?
  - No clustering
  - A different way to visualize dialect areas
  - Allows insight in dialect-similarities
  - Less sensitive to changes in data
  - Dimensions

MultiDimensional Scaling



Conclusion

- Clustering shows similarity in data
- Useful for exploratory data analysis or data-grouping (dialect areas)
- Know your data!
  - Clustering always has output
  - Choose your method carefully

- Questions?